PSY 201: Statistics in Psychology

Lecture 01
Statistics are everywhere

Greg Francis

Purdue University

Fall 2023

MAKING JUDGMENTS

we have to make judgments all the time

- Do nicotine patches help people stop smoking?
- Is Pepsi better than Coke?
- How is alcohol consumption related to depression?
- Is this a good buy for a stereo?
- ...

A PROBLEM

people are not very good at answering these kinds of questions. we make *systematic* errors (take PSY 200, PSY 285 or PSY 318)
people in the "know" can take advantage of these tendencies

- politicians
- retailers
- drug companies
- "activists"

Let's look at an example

DECISION MAKING

Imagine you are getting a loan to purchase a car. You get three offers from different institutions. For each one you have to put some money down up front, and you don't have a lot of available cash. The loans also differ in interest rate. A higher rate means you will end up paying more for the loan. You estimate that each 0.1% increase in the interest rate is going to cost you about $200 over the life of the loan.

Bank	Money down	Interest rate
1st Federal	$2000	5.3%
United Savings	$2000	5.6%
National Federated	$2600	5.0%

DECISION MAKING

Imagine you are getting a loan to purchase a boat. You get three offers from different institutions. For each one you have to put some money down up front, and you don't have a lot of available cash. The loans also differ in interest rate. A higher rate means you will end up paying more for the loan. You estimate that each 0.1% increase in the interest rate is going to cost you about $200 over the life of the loan.

Bank	Money down	Interest rate
1st Federal	$4000	8.3%
United Savings	$4900	8.0%
National Federated	$4600	8.0%

AVOIDING COLDS

- Many people recommend the herb *echinacea* to reduce the severity of colds and/or to increase your immunity to getting a cold
- What should happen if echinacea *does* work?
- What should happen if echinacea *does not* work?
- Why is it popular?

SIGNIFICANCE

- people are easily influenced by the *context* in which they make decisions
- this is a problem, because context is easily (and subtly) manipulated
- it is important to *learn* how to make decisions properly
- STATISTICS
- it is not always easy...but it is worth it

COURSE GOALS

1. Descriptive statistics
   - How to describe data.
   - Using graphs.
   - How to summarize data.
2. Inferential statistics
   - Hypothesis testing.
   - Comparing descriptive statistics.
   - Designing good experiments.

WHY IS IT HARD?

- several reasons
  - Little differences in presentation can make a big difference in understanding.
  - It is hard to get good measurements.
  - It involves mathematics.
  - It goes against our intuitions (anecdotal evidence).
  - If you don't ask the right type of question it is worthless.
  - Sometimes the answer is "I don't know."

COURSE OUTLINE

- statistical terms
  - describing data
  - percentiles
  - normal distribution
  - correlation
  - **EXAM 1 (10%)**
- Significance tests
  - probability
  - signal detection theory
  - hypothesis testing
  - power
  - estimation
  - **EXAM 2 (10%)**
- various types of hypothesis testing
  - Proportions, correlations
  - Two sample means
  - Two sample proportions, correlations
  - **EXAM 3 (10%)**
- ANOVA
  - Multiple testing
  - Contrasts
  - Power
  - Dependent
- **FINAL (15%)** (cumulative)
  - Beware scheduling of the final exam!

TEXTBOOK

- On-line, free (to you). Set up instructions in the paper copy of the syllabus.
- Readings are assigned and monitored (10% of your class grade)
- Finishing a reading means that you answer the questions at the bottom of the page, or that you go through the entire demonstration/simulation
- Due dates and times are listed in the syllabus. The specific sections to read are listed on the Reading Assignments page on the textbook web site

HOMEWORK

- homework counts for 20% of your class grade
- finishing means that you get the correct answer (unlimited guesses)
- Due dates and times are listed in the syllabus. The specific questions are listed on the Homework Assignments page of the textbook web site

STATLAB

- On-line experiments where you generate your own data and then do a statistical analysis (15% of your class grade)
- You need to complete all the questions to get credit for a lab assignment
- Due dates are listed on the syllabus

ATTENDANCE

- Mandatory, we will check every class period (5% of your class grade)
- You are allowed 6 misses before you are penalized

PRACTICE EXAMS

- I have posted practice exams on the course web site. You need to complete the exam and submit it to the TA by the date/time indicated in the syllabus (5% of your class grade)
- Use the feedback from the TA to prepare for the real exam

COMPUTER SOFTWARE

- The textbook provides nice tools for calculating many things.
- Oftentimes the homework requires that you use those tools
- It is useful to have some skills with a spreadsheet to perform simple computations and to format data

GRADING

- straight scale
  - 98% –100% A+
  - 93% – 97% A
  - 90% – 92% A-
  - 88% – 89% B+
  - 83% – 87% B
  - 80% – 82% B-
  - 78% – 79% C+
  - 73% – 77% C
  - 70% – 72% C-
  - 68% – 69% D+
  - 63% – 67% D
  - 60% – 62% D-
  - 0% – 59% F

OFFICE HOURS

- Psychological Sciences Building
  Room 3186
  494-6934
- Monday, Wednesday, Friday
  2:00 - 3:00 pm
  or by appointment.
- email: `gfrancis@purdue.edu`

LECTURE NOTES

- reduced format of 6 slides to a page
- available on the class web page

  `http://www.psych.purdue.edu/~gfrancis/Classes/PSY201/indexF23.html`

TEACHING ASSISTANT

- Victoria Jakicic
- OFFICE: PSYCH 3188
- OFFICE HOURS: Tuesday and Thursday, 1:00–2:30 pm
- Email: vjakicic@purdue.edu

NEXT TIME

- variables:
  - independent
  - dependent
- measurement scales
  - nominal
  - ordinal
  - interval
  - ratio
- descriptive statistics

What is our national security threat?

PSY 201: Statistics in Psychology

Lecture 02
Measurement scales
Descriptive statistics
What is our national security threat?

Greg Francis

Purdue University

Fall 2023

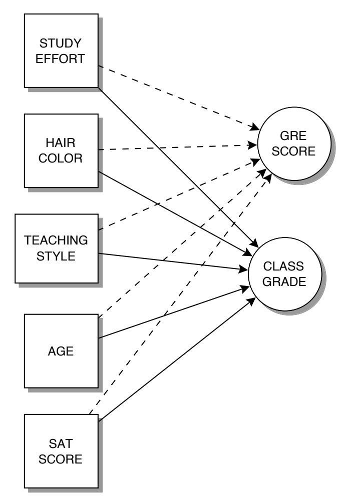
VARIABLES

factors that affect data e.g.
study performance of college students taking a statistics course
variables include

- teaching style
- age
- SAT scores
- class grade
- study effort
- hair color...

DEPENDENT VARIABLES

e.g.
class grade
GRE scores



DEPENDENT VARIABLES

- researchers are interested in how dependent variables change as other variables change
  - see how the dependent variables *depend* on other variables)
- other variables are called **independent** variables
  - researcher either keeps track of or controls the values of independent variables

INDEPENDENT VARIABLES

two types

1. researcher manipulates variable
   e.g. drug dosage, teaching style,...
2. variable classifies
   e.g. hair color, eye color, SAT scores,...

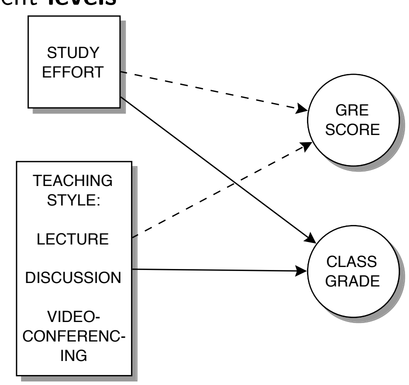
study wants to know how the dependent variable *changes* with changes in the independent variables

LEVELS OF VARIABLES

independent variables can have different **levels**

e.g.
three methods of teaching style

1. Lecture.
2. Discussion.
3. Videoconferencing



EXAMPLE

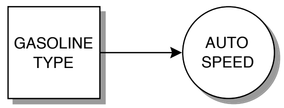
Does gasoline type affect car speed?

- take a car
  - fill it with different types of gasoline
  - measure top speed
- keep many things **constant**
  - same driver
  - same car
  - same course
  - same weather
  - ....

if you started changing these, they would become independent variables

VARIABLES

- independent variable: gasoline type
- levels of independent variable
  - Amoco
  - Sunoco
  - Crystal Flash
  - Marathon
  - ...
- dependent variable: auto speed



MEASUREMENT

- studies need to identify variables and measure them
- different variables have different scales of measurement
- four scales of measurement:
  least precise to most precise
  - nominal
  - ordinal
  - interval
  - ratio

NOMINAL SCALE

- classification of objects into categories
- e.g.
  - nationality
  - color of eyes
  - gender
  - names of objects
- no **order** to the categories!

NOMINAL SCALE

- two key properties
  - data categories are mutually exclusive.
  - data categories have no logical order.
- numbers can designate categories
  1. blue eyes
  2. brown eyes
  3. green eyes
- but the order of numbers does not imply order of categories, because there really is no order

ORDINAL SCALE

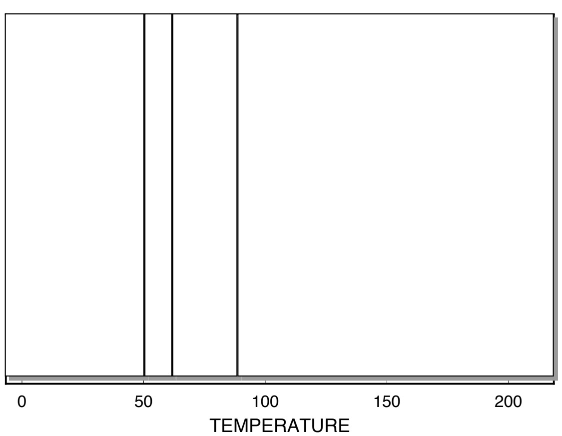
- ordered classification
- e.g.
  - grading system A,B,C,D,F
  - warmth: cold, cool, warm, hot
  - aggressive, timid
- **order** is important and means something

ORDINAL SCALE

- numbers can be used to designate categories
- e.g. warmth
  1. cold
  2. cool
  3. warm
  4. hot
- order of numbers agrees with order of categories

ORDINAL SCALE

- but size of number does not correspond to **amount** of relevant characteristic
- e.g., warm (3) does not necessarily have 2 more units of warmth than cold (1)



ORDINAL SCALE

characteristics

- data categories are mutually exclusive.
- data categories have some logical order.
- data categories are scaled according to the amount of the particular characteristic they posses.

USING SCALES

- One needs to pick items that have a "natural" scale to convey certain types of information
- Thus, for example, colors are typically at the nominal scale of measurement
- this makes them a poor choice for labeling of ordinal data because people do not automatically know what the different colors mean
- this was a problem for the National security warning system, which used colors to indicate different threat levels
- Which is more severe: green threat or blue threat?

MATCHING SCALES

- The problem was that the scales of threat (ordinal scale) and color (nominal scale) do not match. Thus, news reports of the threat level invariably do not list only the color but also the associated phrase with each report. The color scale was of no use at all (they were dropped in 2011).

INTERVAL SCALE

- equal unit scale
- e.g.
  - temperature (Fahrenheit or Celsius)
  - IQ scores (try to be)
  - most tests
- no beginning to scale
- zero point is just another category

INTERVAL SCALE

- numbers can be used to designate categories
  e.g.
  - 22° F $\rightarrow$ level of heat
  - 25° F $\rightarrow$ level of heat
  - 28° F $\rightarrow$ level of heat
- order of numbers agrees with order of categories
- number differences agree with characteristic differences (e.g., 3° F)

INTERVAL SCALE

- Intelligence quotient scores
  - 50 IQ
  - 100 IQ
  - 150 IQ
- an adult with a 50 IQ should have 50 fewer units of intelligence than a person with a 100 IQ
- a person with a 100 IQ should have 50 fewer units of intelligence than a person with a 150 IQ
- however, you cannot say that a genius (150 IQ) is 1.5 times as intelligent as an average (100 IQ)

INTERVAL SCALE

- zero point
- 0 temperature does not mean no heat (in F and C)
- 0 IQ does not mean no intelligence
- 50° F is *not* twice as hot as 25° F.
- an IQ of 100 is *not* twice as smart as an IQ of 50

WHY ZERO MATTERS

- I can create an equivalent interval scale that preserves all the differences
$$\text{NEW}_{\text{IQ}} = \text{OLD}_{\text{IQ}} + 20$$
- differences are still the same
  - $150 \rightarrow 170$
  - $100 \rightarrow 120$
  - $50 \rightarrow 70$
- but the ratios are all different 170 is not 1.5 times 120! Multiplication makes no sense!
- if zero meant absence of trait, I could not create an equivalent interval scale, zero would have to correspond to zero, and nothing else.

INTERVAL SCALE

characteristics

- data categories are mutually exclusive.
- data categories have some logical order.
- data categories are scaled according to the amount of the particular characteristic they posses.
- equal differences in the characteristic are represented by equal differences in the numbers.
- the value 0 is just another value on the scale.

RATIO SCALE

- what we normally think of as measurement
- e.g.
  - height
  - weight
  - energy
- zero point corresponds to the lack of a characteristic

RATIO SCALE

- numbers can be used to designate categories
  e.g.
  - 25 meters $\rightarrow$ distance
  - 5 meters $\rightarrow$ distance
  - 0 meters $\rightarrow$ no distance
- order of numbers agrees with order of categories
- number differences agree with characteristic differences

RATIO SCALE

- Kelvin temperature scale measures heat energy
- e.g.
  - 0° K $\rightarrow$ no heat energy
  - 25° K $\rightarrow$ heat energy
  - 50° K $\rightarrow$ heat energy

RATIO SCALE

- zero point
- 0 distance means no distance
- 0° K temperature means no heat
- 50 meters *is* twice as far as 25 meters
- 50° K *is* two times as much heat energy as 25° K.

RATIO SCALE

- data categories are mutually exclusive.
- data categories have some logical order.
- data categories are scaled according to the amount of the particular characteristic they posses.
- equal differences in the characteristic are represented by equal differences in the numbers.
- the value 0 reflects the absence of the characteristic.

MEASUREMENT SCALE

- how do you identify what scale is appropriate?
- measures at a "higher" scale can also be used at a lower scale, but not vice-versa
- the correct scale often depends on how you intend to *use* the data, and not so much on the intrinsic properties of the things you measure
- e.g. I can use person *names* as
  - nominal scale (code different people)
  - ordinal scale (alphabetize by name)

SCALES

- qualitative variables: generally discrete categories
  - nominal data
  - ordinal data
- quantitative variables: generally continuous
  - interval data
  - ratio data
- sometimes data looks like it is qualitative when it is actually quantitative (e. g., temperature readings do not usually use decimals, but they could)

POPULATION

- all members of a specified group
- e.g.,
  - all students in this class
  - all Purdue students
  - all patients with Alzheimer's disease
- measure of a population characteristic is called a **parameter**
- e.g.,
  - mean grade in class
  - highest grade in class
  - lowest grade in class

SAMPLE

- a subset of all members of a specified group, e.g.
  - all students in this class, relative to all Purdue students
  - all Purdue students, relative to all college students nationwide
  - all Alzheimer's patients, relative to all ill patients
- measures of a sample characteristic are called **statistics**, e.g.,
  - mean grade in class
  - highest grade in class
  - lowest grade in class
- we will use a statistic to **infer** properties of the corresponding population

CONCLUSIONS

- variables
  - dependent
  - independent
- measurement scales
- important issues for interpreting data
- important for applying statistical approaches

NEXT TIME

- working with data
- displaying data
- summarizing data

Why the space shuttle Challenger *blew up.*

PSY 201: Statistics in Psychology

Lecture 03

Plots

Why the space shuttle blew up.

Greg Francis

Purdue University

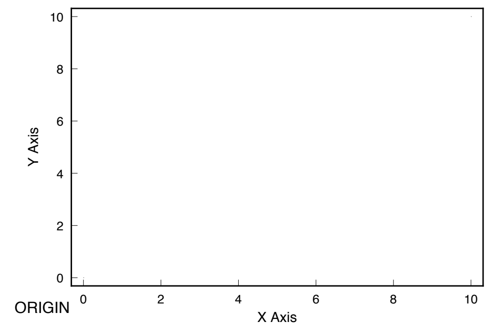
Fall 2023

DATA

GOAL:

- organize data in a way that helps us understand it
- often take advantage of visual interpretations
- particularly important for very large sets of data

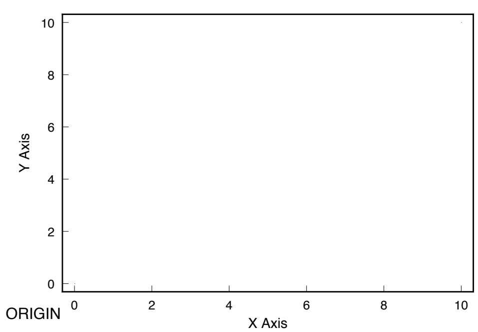
GRAPHS



- plot one variable against another

PLOTTING

- you make a graph to convey information
- place the dependent variable on the *y*-axis and the independent variable on the *x*-axis



- avoid everything else that might get in the way!

SPACE SHUTTLE

- January 28, 1986
- O-ring leaked
- the Challenger exploded 59 seconds after liftoff

SPACE SHUTTLE

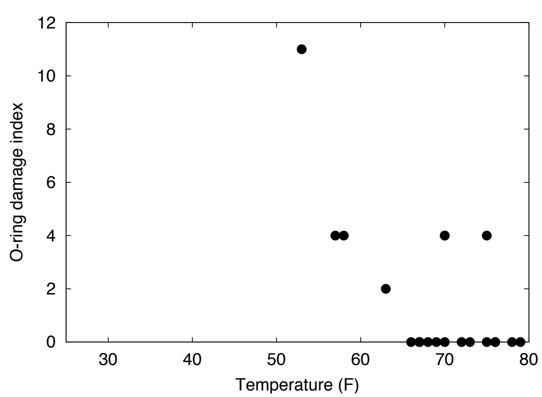
- January 28, 1986
- O-ring leaked
- the Challenger exploded 59 seconds after liftoff

SPACE SHUTTLE

- the night before engineers warned O-rings would leak in cold (29°) weather
- the engineers failed to make their case, and the shuttle blew up
- they failed to *present* their data in a way to convince others

THE DATA

- previous launches showed damage to the O-rings increased as temperature got colder



THE MISTAKES

- when trying to convince NASA scientists to cancel the liftoff engineers:
  - used tables (not bad by itself, but a graph is often more convincing)

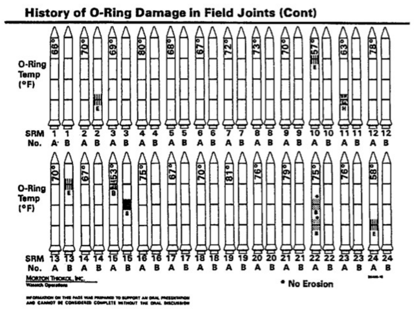
HISTORY OF O-RING TEMPERATURES
(DEGREES - F)

MOTOR	MBT	AMB	O-RING	WIND
DM-4	68	36	47	10 MPH
DM-2	76	45	52	10 MPH
QM-3	72.5	40	48	10 MPH
QM-4	76	48	51	10 MPH
SRM-15	52	64	53	10 MPH
SRM-22	77	78	75	10 MPH
SRM-25	55	26	29 27	10 MPH 25 MPH

  - distributed information across several tables

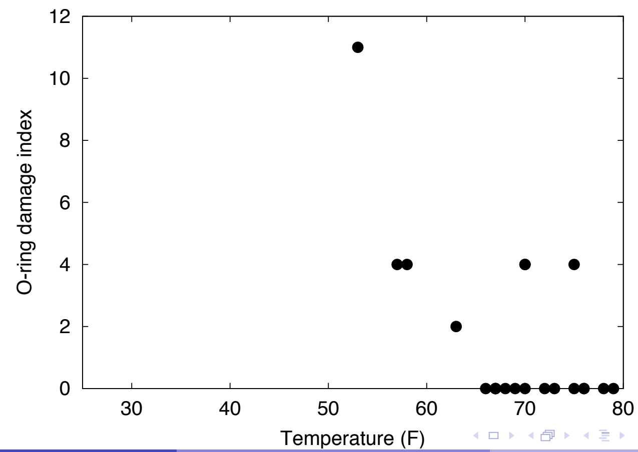
THE MISTAKES

- when trying to convince NASA scientists to cancel the liftoff engineers:
  - cluttered graphics with irrelevant information (motor type, date of launch,...)



THE MISTAKES

- when trying to convince NASA scientists to cancel the liftoff engineers:
  - failed to point out that all good launches were in warm temperatures
  - failed to point out that the forecasted temperature (29°) was *much* colder than for any other launch (good or bad)

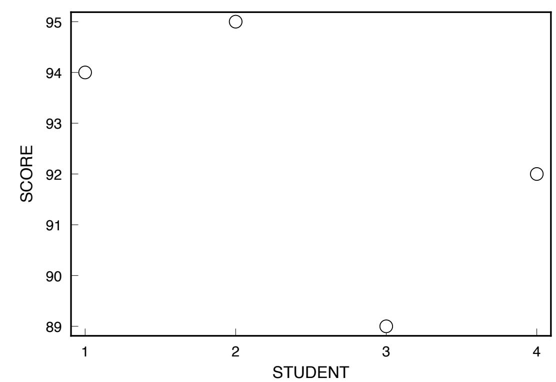


THE LESSON

- when trying to convince someone of something, you must present it properly
- avoid fancy graphics and 3D perspectives
- keep it simple
- present the right information
- we will go over some basics of graphing...

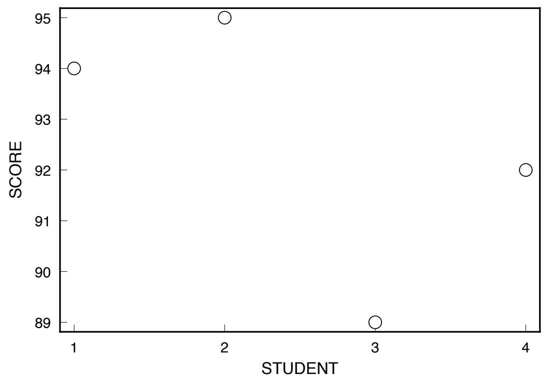
GRAPH

- using a small data set of four student's grades



GRAPH

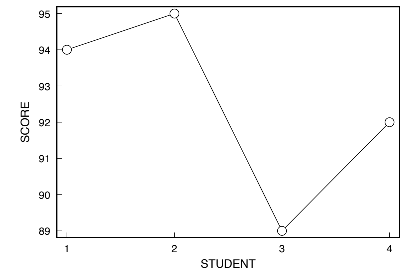
- using a small data set of four student's grades

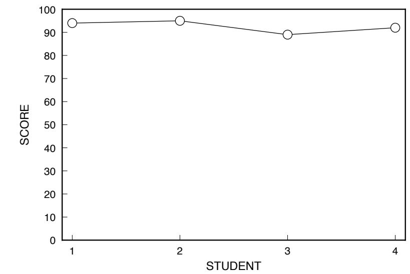


- what measurement scale is the student variable?
- what measurement scale is the score variable?

DATA CURVE

- it sometimes helps to connect the points
- How well did the third student do?
- changing the axis' scale makes the information look different, even though it isn't
- what matters is whether the graph conveys the intended information!





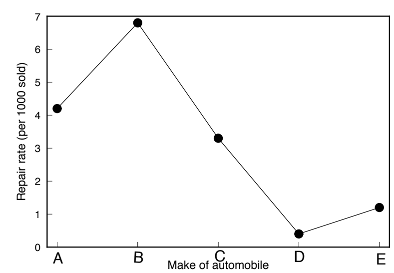
GRAPH TYPE

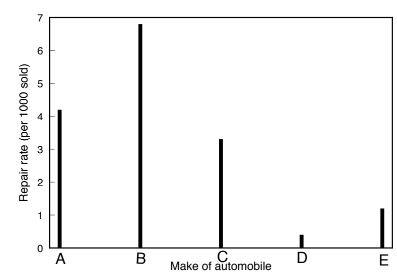
- type of data often determines what type of graph to draw
- previous graph plotted ratio (or interval) data against nominal data
- consider the following data

Make of Automobile	Repair Rate (per 1000 sold)
A	4.2
B	6.8
C	3.3
D	0.4
E	1.2

- the graph should **not** suggest continuity of automobile make

WHICH IS BETTER?



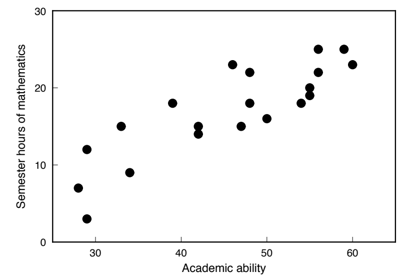


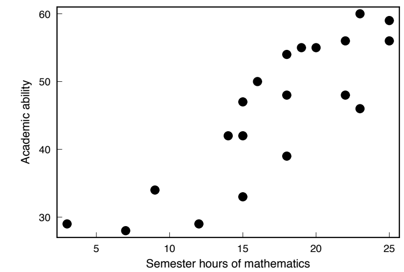
SCATTERGRAMS

- sometimes you want to look at co-occurrences of data

Student	Academic Ability Score	Hours of Mathematics
1	54	18
2	29	3
3	42	14
4	60	23
5	33	15
6	28	7
7	56	22
8	48	18
...	...	...

SCATTERGRAMS





GRAPHS

- Very useful for giving an overview of many types of data sets
- Useful for identifying trends in the data and relationships between variables
- Limited in that they depend on the viewer's interpretive abilities and sometimes graphs breakdown for really big or really small data sets
- We prefer more quantitative approaches

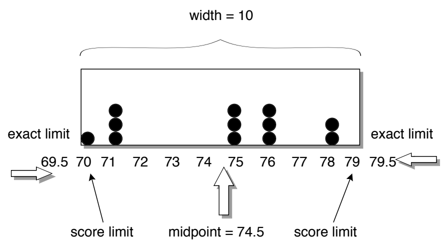
FREQUENCY

- for large data sets we cannot present all the scores
- we often look at the *number* or *frequency* of scores within certain limits
- we look at how scores are spread out across different values
- this reduces the number of *presented* scores and improves understanding

CLASS INTERVAL

Terminology

- width: exact upper limit - exact lower limit
- midpoint: value halfway between upper limit and lower limit
- exact limits: exact boundaries of interval
  - matter when we start to work with frequency distributions!
- score limits: highest and lowest possible scores that fall in the interval



FREQUENCIES

- compare a set of scores
  95, 22, 45, 45, 12, 79, 83, 46, 89, 96, 75, 33, 86, 57, 69, 94, 83, 75, 77, 88, 92, 85, 31, 69
- to frequencies

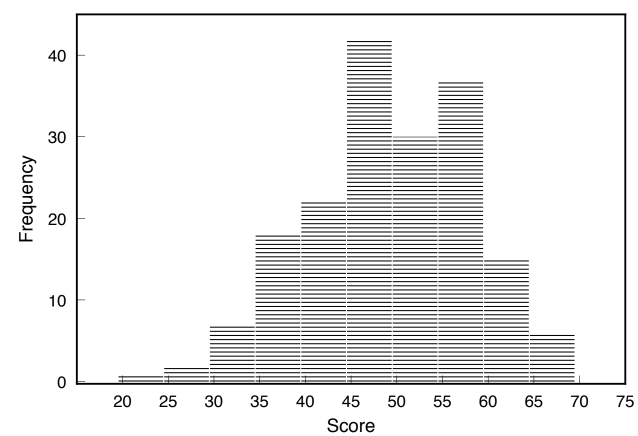
Class Interval	*f*
10–19	1
20–29	1
30–39	2
40–49	3
50–59	1
60–69	2
70–79	4
80–89	6
90–99	4

FREQUENCIES

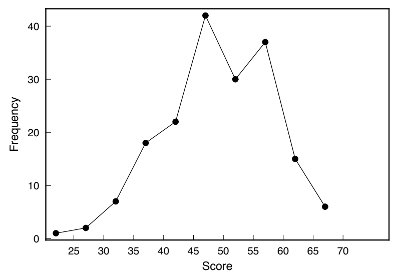
- ADVANTAGES
  - easier to see distribution of scores
  - easier to interpret data
- DISADVANTAGES
  - loss of information
  - individual scores are missing
  - midpoint score is often best guess
- often use frequency information to **supplement** other information (depends on your needs)

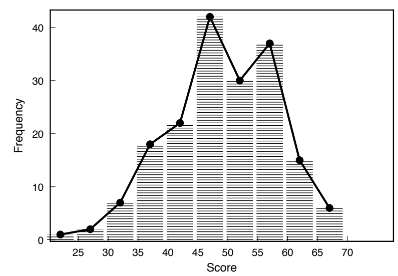
HISTOGRAMS

frequency versus score class interval



FREQUENCY POLYGON



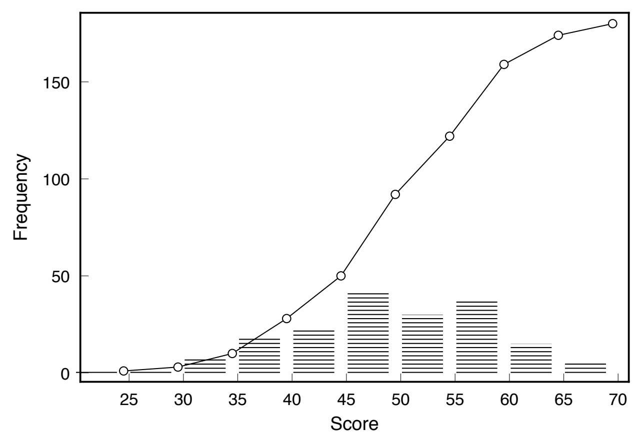


CUMULATIVE FREQUENCIES

- frequency distribution tells us how many scores in each class interval
- cumulative frequency distribution tells us how many scores in all class intervals below a specific score

Midpoint	f	cf
67	6	180
62	15	174
57	37	159
52	30	122
47	42	92
42	22	50
37	18	28
32	7	10
27	2	3
22	1	1

CUMULATIVE FREQUENCY DISTRIBUTION



Note: the point on the polygon has it's x-coordinate at the upper limit of the corresponding class interval

PERCENTAGES

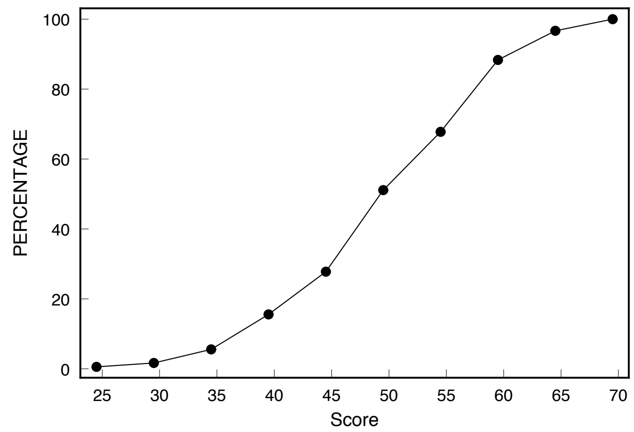
$$\% = \frac{\text{frequency}}{\text{total number of scores}}$$

$$c\% = \frac{\text{cumulative frequency}}{\text{total number of scores}}$$

Midpoint	f	cf	%	c%
67	6	180	3.33	100
62	15	174	8.33	96.67
57	37	159	20.56	88.34
52	30	122	16.67	67.78
47	42	92	23.33	51.11
42	22	50	12.22	27.78
37	18	28	10.00	15.56
32	7	10	3.89	5.56
27	2	3	1.11	1.67
22	1	1	0.56	0.56

OGIVE

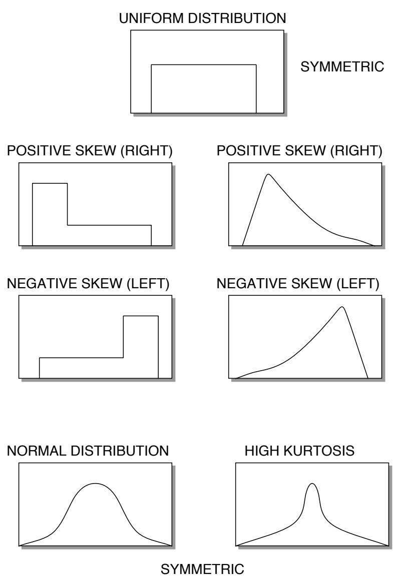
- plot cumulative frequency percentage against upper score class interval
- gives percentile points (next time)



FREQUENCY DISTRIBUTIONS

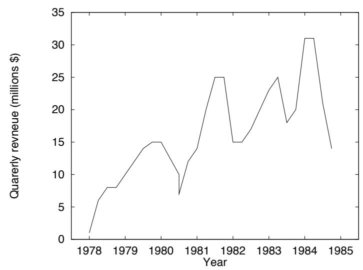
- useful to compare shapes
- any shape is possible
- some shapes are particularly important
  - uniform distribution
  - skewed distribution (long tail)
  - symmetric distribution
  - normal distribution
  - kurtosis (peakedness)

DISTRIBUTIONS



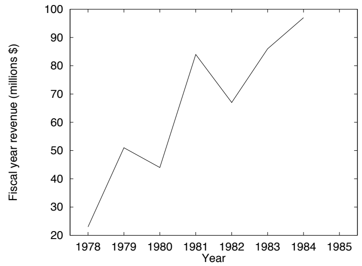
DISTRIBUTIONS

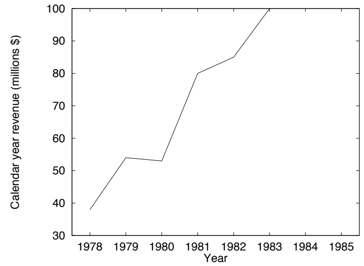
- with large data sets you *have* to group data together to make it manageable
- how you do it can sometimes have a profound effect on what people conclude
- consider revenue from a company: grouped by *quarterly* revenue



DISTRIBUTIONS

- now look at the data when grouped by fiscal or calendar year





DISTRIBUTIONS

- with computers people can now sift through huge amounts of data and present only those graphs that support what they want you to think
- a suspicious person might presume that the graphs you *do* see are the *best possible* for advancing the presenter's view
- the only way out of this is to either trust the presenter, or have access to the data and and knowledge to understand it

HONESTY

- so how you define class intervals can determine how you (or someone else) will interpret the data
- statistics don't lie (they are just numbers)
- but you could (and some people do) select certain statistics to make people believe one thing versus another
- the only thing you can do about this effect is to be aware that it exists
- you need to be aware of the limitations of the data and be on guard against things that might influence you

CONCLUSIONS

- graphing
- frequencies
- distributions
- remember: the goal is to correctly present information

NEXT TIME

- percentiles
- percentile ranks

How to score the SAT.

PSY 201: Statistics in Psychology

Lecture 04
Describing distributions
How to score the SAT.

Greg Francis

Purdue University

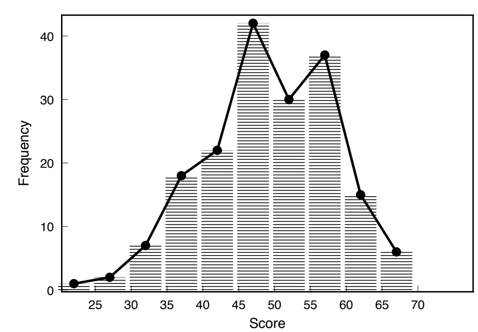
Fall 2023

DISTRIBUTIONS

- As we saw last time, a well-drawn graph conveys a lot of useful information...
- but a poorly drawn graph can mislead and confuse.
- We would like a **quantitative** method of describing distributions
- may not entirely avoid misinformation, but at least the limitations will be identifiable

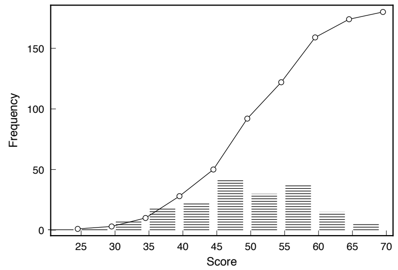
FREQUENCY DISTRIBUTIONS

- A data set of exam scores can be described in many ways
  - frequency versus score class interval



CUMULATIVE

- A data set of exam scores can be described in many ways
  - cumulative distributions



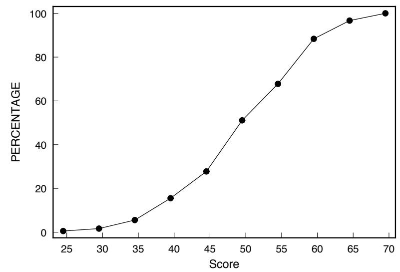


TABLE FORMAT

- A data set of exam scores can be described in many ways
  - frequency table

Exact Limits	Midpoint	f	cf	%	c%
64.5–69.5	67	6	180	3.33	100
59.5–64.5	62	15	174	8.33	96.67
54.5–59.5	57	37	159	20.56	88.34
49.5–54.5	52	30	122	16.67	67.78
44.5–49.5	47	42	92	23.33	51.11
39.5–44.5	42	22	50	12.22	27.78
34.5–39.5	37	18	28	10.00	15.56
29.5–34.5	32	7	10	3.89	5.56
24.5–29.5	27	2	3	1.11	1.67
19.5–24.5	22	1	1	0.56	0.56

DISTRIBUTION USES

- summarize data
- indicate most frequent data values
- indicate amount of variation across data values
- allows us to interpret a single score in the context of other scores
- we will explore quantitative methods to describe distributions

PERCENTILE

- point in a distribution at (or below) which a given percentage of scores is found
- written as

$$P_{\text{percentage}}$$

- 28th percentile is written as P_{28}
- 99th percentile is written as P_{99}
- ...

PERCENTILE

- what are the data values for the lowest 60% of the population?
- several steps
  1. Find out how many data values make up 60% of the population.
  2. Find the lowest class interval in the cumulative frequency distribution that includes at least that many data values.
  3. Estimate how far into the class interval you must go to reach exactly the percentile.
- works for any percentage!

CALCULATIONS

- find P_{60} using the above data set of scores
  (1) number of scores making up 60% of student scores is

  $$(180)(0.60) = 108$$

  In general, calculate

  $$(n)(p)$$

  where n is the size of the population (number of scores) and p is the percentage in decimal form

CALCULATIONS

(2) lowest class interval in the *cf* including 108 scores is with midpoint 52

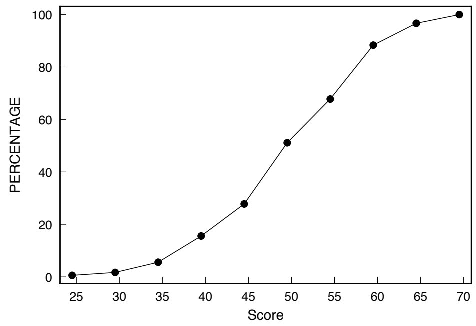
Exact Limits	Midpoint	f	cf	%	c%
64.5–69.5	67	6	180	3.33	100
59.5–64.5	62	15	174	8.33	96.67
54.5–59.5	57	37	159	20.56	88.34
49.5–54.5	**52**	**30**	**122**	**16.67**	**67.78**
44.5–49.5	47	42	92	23.33	51.11
39.5–44.5	42	22	50	12.22	27.78
34.5–39.5	37	18	28	10.00	15.56
29.5–34.5	32	7	10	3.89	5.56
24.5–29.5	27	2	3	1.11	1.67
19.5–24.5	22	1	1	0.56	0.56

CALCULATIONS

- so we know that the percentile is somewhere between 49.5 and 54.5. We want a more precise **estimate**
- we need to know
  - width of class interval (5)
  - frequency of scores in the class interval containing the percentile point (30)
  - exact lower limit of class interval containing the percentile point (49.5)
  - *cf* of scores **below** the class interval containing the percentile point (92)
  - remaining number of scores in class interval containing the percentile point (108 - 92 = 16)

CALCULATIONS

- estimate of percentile point
- go into the interval the remaining (unaccounted for) percentage



CALCULATIONS

$$P_X = ll + \left(\frac{np - cf}{f_i}\right)(w)$$

- ll = exact lower limit of the interval containing the percentile point
- n = total number of scores
- p = X/100, proportion corresponding to percentile (decimal form)
- cf = cumulative frequency of scores **below** the interval containing the percentile point
- f_i = frequency of scores **in** the interval containing the percentile point
- w = width of class interval

PERCENTILE RANK

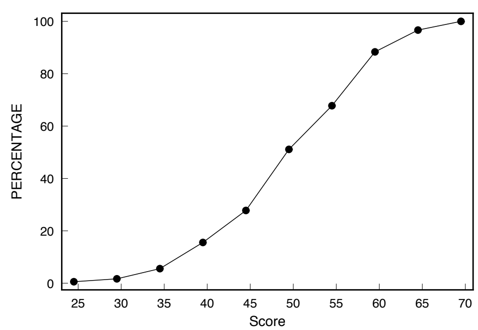
- given a particular data value, what percentage of data values are smaller?
- e.g. given a score on a test, what percentage of scores were lower?
- sort of the reverse of percentile
- for a data value of 39, we write the percentile rank as

$$PR_{39}$$

- (Used on achievement tests!)

OGIVE

- plot cumulative frequency percentage against score class interval (gives percentile rank)



CALCULATIONS

$$PR_X = \left\{ \frac{cf + (f_i)(X - ll)/w}{n} \right\} (100)$$

- X = score for which percentile rank is to be determined
- cf = cumulative frequency of scores **below** the interval containing the score X
- ll = exact lower limit of the interval containing X
- w = width of class interval containing X
- f_i = frequency of scores in the interval containing X
- n = total number of scores

CALCULATIONS

$$PR_X = \left\{ \frac{cf + (f_i)(X - ll)/w}{n} \right\} (100)$$

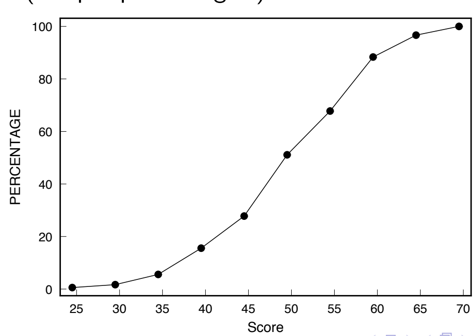
$$PR_{39} = \left\{ \frac{10 + (18)(39 - 34.5)/5}{180} \right\} (100)$$

$$PR_{39} = 14.556$$

Exact Limits	Midpoint	f	cf	%	c%
64.5–69.5	67	6	180	3.33	100
59.5–64.5	62	15	174	8.33	96.67
54.5–59.5	57	37	159	20.56	88.34
49.5–54.5	52	30	122	16.67	67.78
44.5–49.5	47	42	92	23.33	51.11
39.5–44.5	42	22	50	12.22	27.78
34.5–39.5	**37**	**18**	**28**	**10.00**	**15.56**
29.5–34.5	32	7	*10*	3.89	5.56
24.5–29.5	27	2	3	1.11	1.67
19.5–24.5	22	1	1	0.56	0.56

LIMITATIONS

- percentiles help *describe* a data value relative to its frequency distribution
- but they have some drawbacks
  - percentiles use an ordinal scale
  - equal differences in percentiles do not indicate equal differences in raw scores!
  - class intervals with higher frequency cover a broader range of percentiles (steeper part of ogive)



LIMITATIONS

- percentiles exaggerate differences in scores when lots of people have similar scores
- underestimate actual differences when lots of people have very different scores
- differences in percentiles should **not** be compared across different distributions!!!
  - only provide information on relative ranking of scores: ordinal scale!
  - cannot be meaningfully averaged, summed, multiplied,...
- fixing these problems requires additional terms for describing distributions (central tendency)

CONCLUSIONS

- percentiles
- percentile ranks

NEXT TIME

- central tendency
  - mode
  - median
  - mean

Does a company deserve a tax break?

PSY 201: Statistics in Psychology

Lecture 06

Variability

How to make IQ scores look good.

Greg Francis

Purdue University

Fall 2023

DESCRIPTION

- central tendency gives an indication of where most, many, or average, scores are
- also want some idea of how much variability exists in a distribution of scores
  - range
  - mean deviation
  - variance
  - standard deviation

RANGE

- Highest score - lowest score

Name	Sex	Score
Greg	Male	95
Ian	Male	89
Aimeé	Female	94
Jim	Male	92

- 95 - 89 = 6

PROBLEM

- range is **very** sensitive to "extreme" scores

Name	Sex	Score
Greg	Male	95
Ian	Male	89
Aimeé	Female	94
Jim	Male	92
Bob	Male	32

- 95 - 32 = 63
- one score makes a big difference!

MEAN DEVIATION

- we can decrease sensitivity to extreme scores by considering deviations from a measure of central tendency
- a deviation score is

$$x_i = X_i - \overline{X}$$

- we define the mean deviation as:

$$\text{MD} = \frac{\Sigma|X_i - \overline{X}|}{n} = \frac{\Sigma|x_i|}{n}$$

- where $|x_i|$ means: "absolute value of x_i"
- why do we take the absolute value instead of just summing deviations?

VARIANCE

- mean deviation turns out to be mathematically messy
- squaring also removes minus signs!
- sum of squares

$$SS = \Sigma(X_i - \overline{X})^2 = \Sigma(x_i)^2$$

- variance is the average sum of squares
- calculation depends on whether scores are from a population or a sample

POPULATION

- a population includes **all** members of a specified group
- variance is defined as:

$$\sigma^2 = \frac{SS}{N} = \frac{\Sigma(X_i - \mu)^2}{N} = \frac{\Sigma(x_i)^2}{N}$$

- where
  - μ is the mean of the population
  - N is the number of scores in the population

SAMPLE

- a sample includes a **subset** of scores from a population
- variance is defined as:

$$s^2 = \frac{SS}{n-1} = \frac{\Sigma(X_i - \overline{X})^2}{n-1} = \frac{\Sigma(x_i)^2}{n-1}$$

- where
  - $\overline{X}$ is the mean of the sample
  - n is the number of scores in the sample
- why the differences? Don't worry for now. Just know the calculations.

SAMPLE VARIANCE

- deviation formula:

$$s^2 = \frac{\Sigma(x_i)^2}{n-1}$$

- alternative (but equivalent) calculation is the raw score formula

$$s^2 = \frac{SS}{n-1} = \frac{\Sigma(X_i)^2 - [(\Sigma X_i)^2/n]}{n-1}$$

- use whichever formula is simpler!

EXAMPLE

Name	Sex	Score
Greg	Male	95
Ian	Male	89
Aimeé	Female	94
Jim	Male	92

- since we have the raw scores, we use the raw score formula (we assume a sample)

$$s^2 = \frac{SS}{n-1} = \frac{\Sigma(X_i)^2 - [(\Sigma X_i)^2/n]}{n-1}$$

$$\Sigma X_i^2 = (95)^2 + (89)^2 + (94)^2 + (92)^2 = 34246$$

$$(\Sigma X_i)^2/n = (95 + 89 + 94 + 92)^2/4 = \frac{(370)^2}{4} = \frac{136900}{4} = 34225$$

- so,

$$s^2 = \frac{34246 - 34225}{3} = \frac{21}{3} = 7$$

SUM OF SQUARES

- earlier we calculated the squared deviation from the mean

$$\sum x_i^2 = \sum (X_i - \overline{X})^2$$

$$= (95 - 92.5)^2 + (89 - 92.5)^2 + (94 - 92.5)^2 + (92 - 92.5)^2$$

$$= (2.5)^2 + (-3.5)^2 + (1.5)^2 + (-0.5)^2 = 0$$

$$= 6.25 + 12.25 + 2.25 + 0.25 = 21.0$$

- we can use that to calculate variance with the deviation score formula:

$$s^2 = \frac{\Sigma x_i^2}{n-1} = \frac{21}{3} = 7$$

- Same as before!
- Note! variance cannot be negative

STANDARD DEVIATION

- variance is in **squared** units of measurement
  - distance: squared meters
  - weight: squared kilograms
  - temperature: squared degrees
  - ...
- standard deviation is in the same units as the scores!
- square root of variance

STANDARD DEVIATION

- deviation score formula:

$$s = \sqrt{s^2} = \sqrt{\frac{SS}{(n-1)}} = \sqrt{\frac{\Sigma(x_i)^2}{(n-1)}}$$

- raw score formula:

$$s = \sqrt{s^2} = \sqrt{\frac{SS}{(n-1)}} = \sqrt{\frac{\Sigma(X_i)^2 - [(\Sigma X_i)^2/n]}{n-1}}$$

EXAMPLE

Name	Sex	Score
Greg	Male	95
Ian	Male	89
Aimeé	Female	94
Jim	Male	92

- since we have the raw scores, we use the raw score formula to calculate variance

$$s^2 = \frac{SS}{n-1} = \frac{\Sigma(X_i)^2 - [(\Sigma X_i)^2/n]}{n-1}$$

- we calculated earlier that the variance equals:

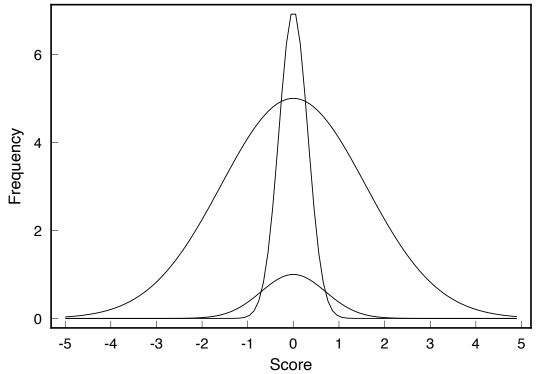
$$s^2 = \frac{34246 - 34225}{3} = \frac{21}{3} = 7$$

- and then the standard deviation equals:

$$s = \sqrt{s^2} = \sqrt{7} \approx 2.646$$

WHY BOTHER?

- the value of the standard deviation gives us an idea of how spread out scores are
- larger standard deviations indicate that scores are more spread out



WHY BOTHER?

- we will use standard deviation to let us estimate how different a score is relative to the central tendency of the distribution
- we can then compare (in a certain sense) **across** distributions!

STANDARD SCORE

- also called z-score

$$\text{Standard score} = \frac{\text{raw score} - \text{mean}}{\text{standard deviation}}$$

$$z = \frac{X - \overline{X}}{s}$$

- indicates the number of standard deviations a raw score is above or below the mean

EXAMPLE

- if
$$\overline{X} = 26$$
- and
$$s = 4$$
- and you have (among others) the scores $X_1 = 16$, $X_2 = 32$, $X_3 = 28$
- then

$$z_1 = \frac{X_1 - \overline{X}}{s} = \frac{16 - 26}{4} = -2.5$$

$$z_2 = \frac{X_2 - \overline{X}}{s} = \frac{32 - 26}{4} = 1.5$$

$$z_3 = \frac{X_3 - \overline{X}}{s} = \frac{28 - 26}{4} = 0.5$$

PROPERTIES

- when a raw score is **above** the mean, its z-score is positive
- when a raw score is **below** the mean, its z-score is negative
- when a raw score **equals** the mean, its z-score is zero
- absolute size of the z-score indicates how far from the mean a raw score is

UNITS

- z-scores work in units of standard deviation
- new numbers for same information!
- just like converting units for other familiar measures
  - length: feet into meters, miles into kilometers
  - weight: pounds into kilograms
  - temperature: fahrenheit into celsius
  - data: raw score units into standard deviation units
- trick!: standard deviation units depend on your particular set of data!

PROPERTIES

- z-scores are data
- we can find distributions, means, and standard deviations
- special properties of z-score distributions
  - The shape of the distribution of standard scores is identical to that of the original distribution of raw scores.
  - The mean of a distribution of z-scores will always equal 0.
  - The variance (and standard deviation) of a distribution of z-scores always equals 1.

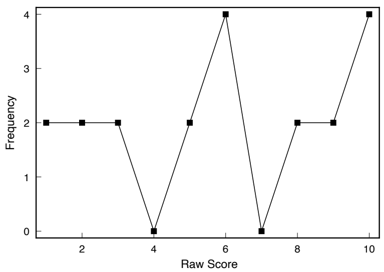
EXAMPLE

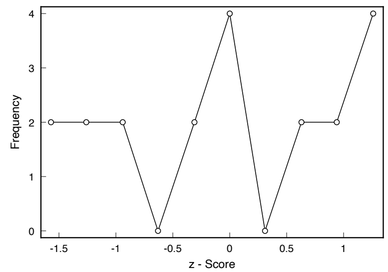
- A simple data set to play with
  - when a raw score is **above** the mean, its z-score is positive
  - when a raw score is **below** the mean, its z-score is negative
  - when a raw score **equals** the mean, its z-score is zero
  - absolute size of the z-score indicates how far from the mean a raw score is

Subject	Raw score	z-score
1	10	1.26
2	9	0.94
3	3	-0.94
4	10	1.26
5	9	0.94
6	2	-1.26
7	2	-1.26
8	10	1.26
9	5	-0.31
10	5	-0.31
11	1	-1.57
12	6	0.0
13	8	0.63
14	6	0.0
15	6	0.0
16	1	-1.57
17	3	-0.94
18	6	0.0
19	10	1.26
20	8	0.63
$n = 20$		
$\overline{X}$	6.0	0.0
s	3.18	1.0

EXAMPLE

- compare distributions of raw scores and z-scores
- shape is the same





USES

- suppose we want to compare the scores of a student in several classes
- we know the student's score, the mean score, the standard deviation, and the student's z-score

Subject	X	$\overline{X}$	s	z
Psychology	68	65	6	0.50
Mathematics	77	77	9	0.00
History	83	89	8	-0.75

- comparison of **raw scores** suggests that student did best in history, mathematics, then psychology
- comparison of **z-scores** suggests that student did best in psychology, mathematics, then history (relative to other students)

TRANSFORMED SCORES

- sometimes z-scores are unattractive
  - zero mean
  - negative values
- need to convert same information into a new distribution with a new mean and standard deviation

$$X' = (s')(z) + \overline{X}'$$

- where
  - X' = new or transformed score for a particular individual
  - s' = desired standard deviation of the distribution
  - z = standard score for a particular individual
  - $\overline{X}'$ = desired mean of the distribution

TRANSFORMED SCORES

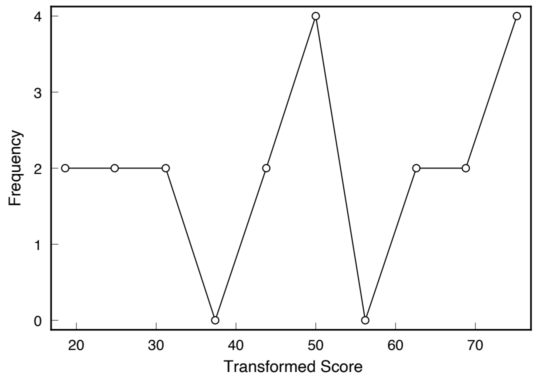
- GOAL: make data understandable; IQ scores, personality tests,...
- NOTE: you can change the mean and standard deviation all you want, but it does **not** change the information in the data
- shape remains the same!
  - conversion back to z-scores would produce the same z-scores!
  - a percentile maps to the corresponding transformed score

TRANSFORMED SCORES

- if we transform the scores from our earlier data set using

$$X' = 20X + 50$$

- we get



CONCLUSIONS

- variance
- standard deviation
- standard scores

NEXT TIME

- a very important distribution
- normal distribution

Describing everyone's height.

PSY 201: Statistics in Psychology

Lecture 07

Normal distribution

Describing everyone's height.

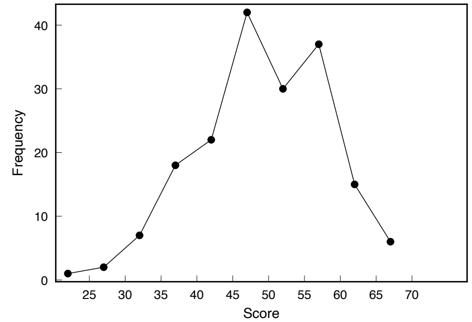
Greg Francis

Purdue University

Fall 2023

DISTRIBUTION

- frequency of scores plotted against score



- frequency $\rightarrow$ likelihood, probability

GOAL

- describe (summarize) distributions
  - shape: unimodal, bimodal, skew,...
  - central tendency: mode, median, mean
  - variation: range, variance, standard deviation
- summarizing forces you to lose information
- some **theoretical** distributions are special!
  - a few numbers completely specify the distribution

NORMAL DISTRIBUTION

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/2\sigma^2}$$

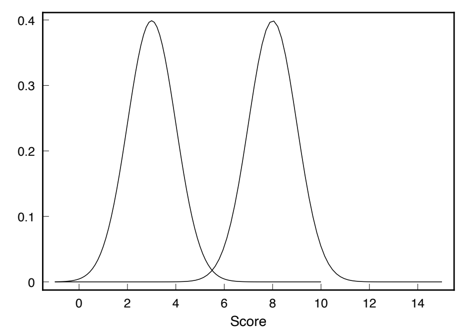
- Y height of the curve for any given value of X in the distribution of scores
- π mathematical value of the ratio of the circumference of a circle to its diameter. A constant (3.14159.....)
- e base of the system of natural logarithms. A constant (2.7183...)
- μ mean of the distribution of scores
- σ standard deviation of a distribution of scores

sometimes written as

$$Y = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-(X-\mu)^2/2\sigma^2\right]$$

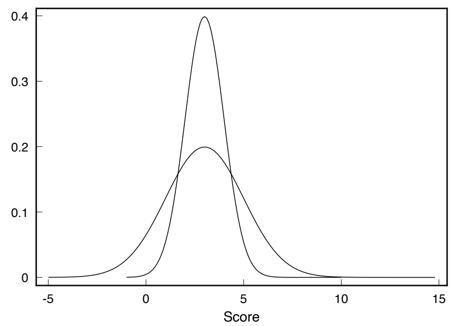
PARAMETERS

- a **family** of distributions
- member of the family is designated by the mean μ and standard deviation σ
- changing μ shifts the curve to the left or the right
  - shape remains the same



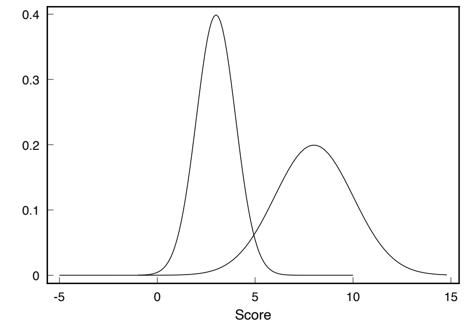
PARAMETERS

- changing σ changes the **spread** of the curve
- compare normal distributions for $\sigma = 1$ and $\sigma = 2$, both with $\mu = 3$



PARAMETERS

- changing μ and σ together produces predictable results



PROPERTIES

- all normal distributions have the following in common
  - Unimodal, symmetrical, bell shaped, maximum height at the mean.
  - A normal distribution is continuous. X must be a **continuous** variable, and there is a corresponding value of Y for each X value.
  - A normal distribution asymptotically approaches the X axis.

STANDARD NORMAL

- remember z-scores:
  - 0 mean
  - 1 standard deviation
- if the z-scores are normally distributed

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/2\sigma^2}$$

- becomes

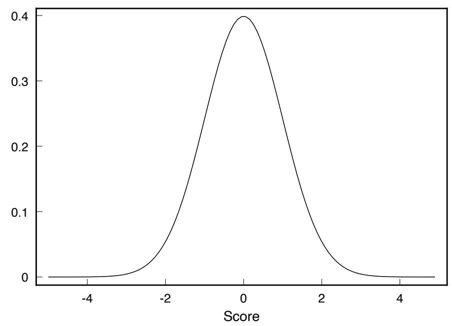
$$Y = \frac{1}{1\sqrt{2\pi}} e^{-(z-0)^2/2(1^2)}$$

- or

$$Y = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

STANDARD NORMAL

- looks like



SIGNIFICANCE

- It turns out that lots of frequency distributions can be described as a normal distribution
- for example, an estimate of height

SIGNIFICANCE

- It turns out that lots of frequency distributions can be described as a normal distribution
  - intelligence scores
  - weight
  - reaction times
  - judgment of distance
  - rating of personality
  - ...
- almost any situation where small independent components come together

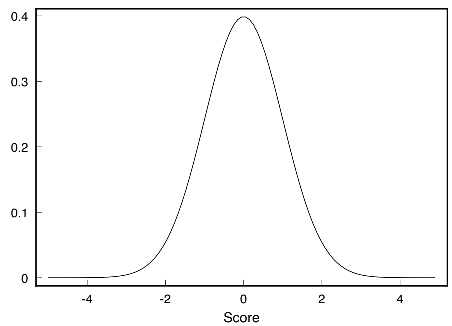
SIGNIFICANCE

- when the distribution is a normal distribution, we can describe the distribution by just specifying
  - Mean: $\overline{X}$
  - Standard deviation: s
  - Noting it is a normal distribution
- that's all we need!
- That's part of our goal: describe distributions

STANDARD NORMAL

- assume you have a standard normal distribution (don't worry about where it came from)

$$Y = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$



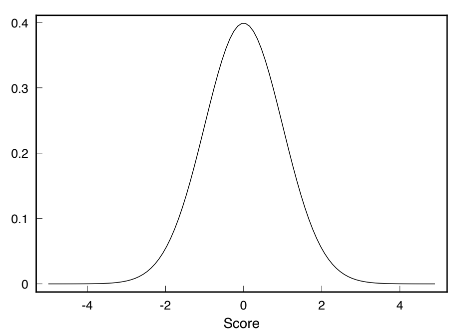
- if your distribution is normal, you can create a standard normal by converting to z-scores

USE

- same as all other distributions
  - identify key aspects of the data
  - percentiles
  - percentile rank
  - proportion of scores within a range
  - ...
- make it easier to interpret data significance!

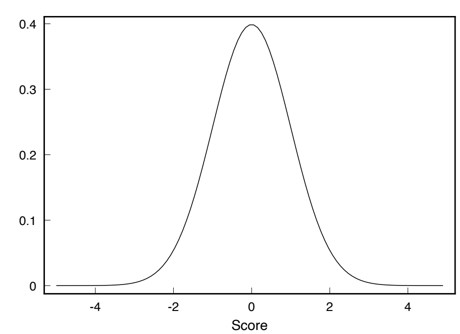
STANDARD NORMAL

- total area under the curve **always** equals 1.0
- area under the curve from the mean (0) to one tail equals 0.5



STANDARD NORMAL

- area under the curve one standard deviation away from the mean is approximately 0.3413
- area under the curve two standard deviations away from the mean is approximately 0.4772
- area under the curve three standard deviations away from the mean is approximately 0.4987



CONCLUSIONS

- normal distribution
  - equations
  - properties
  - standard normal equations

NEXT TIME

- area under the curve
- proportions
- percentiles
- percentile ranks

Business decisions.

PSY 201: Statistics in Psychology

Lecture 08
Normal distribution
Business decisions.

Greg Francis

Purdue University

Fall 2023

NORMAL DISTRIBUTIONS

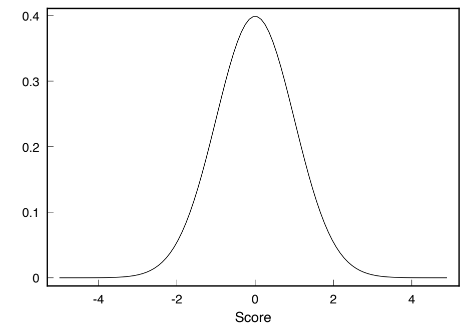
- when the distribution is a normal distribution, we can describe the distribution by just specifying
  - Mean: $\overline{X}$
  - Standard deviation: s
  - Noting it is a normal distribution
- that's all we need!
- That's part of our goal: describe distributions

USE

- same as all other distributions
  - identify key aspects of the data
  - percentiles
  - percentile rank
  - proportion of scores within a range
  - ...
- make it easier to interpret data significance!

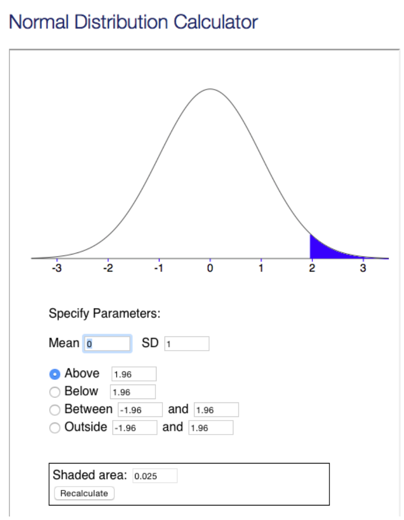
AREA UNDER CURVE

- proportional to the frequency of scores within the designated endpoints
- suppose you want to know the proportion of scores between the mean and another score (z-score)



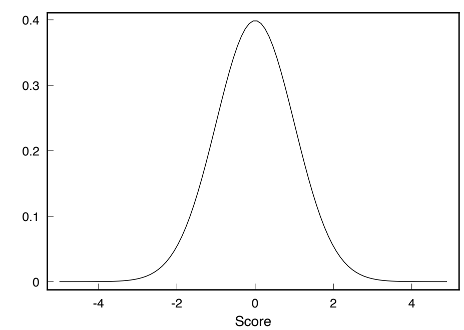
AREA UNDER CURVE

- solving for the area requires calculus and numerical analysis (ack!)
- fortunately, we can also use computers
- our text provides



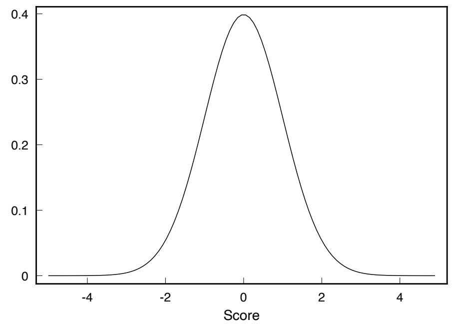
CALCULATOR

- how would you find the area between $z = -0.3$ and $z = 2.4$?



CALCULATOR

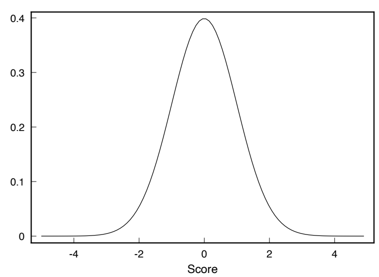
- how would you find the area below $z = 1.4$?



PROPORTIONS

- suppose you have 250 scores from a test that are normally distributed
- you want to know how many scores are **between** 1.0 standard deviation below the mean and 1.5 standard deviations above the mean
- two steps
  1. calculate the area under the standard normal between $z = -1.0$ and $z = 1.5$.
  2. convert the area under the curve to number of scores

PROPORTIONS



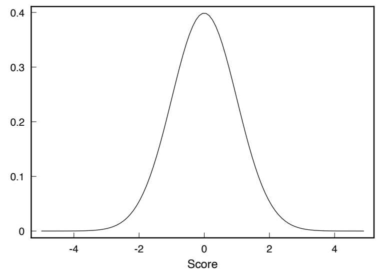
- We find that 77.45% of the scores lie between one standard deviation below the mean and 1.5 standard deviations above the mean
- so how many scores are in that range?
- multiply the total number of scores (250) with the percent in the range (decimal form)

$$(0.7745) \times (250) = 193.625 \approx 194$$

PROPORTIONS

- suppose you have 250 scores from a test that are normally distributed
- you want to know how many scores are **below** 0.5 standard deviations above the mean, and how many scores are **beyond** 2.5 standard deviations above the mean.
- two steps
  1. calculate the area under the standard normal below $z = 0.5$ and above $z = 2.5$.
  2. convert the area under the curve to number of scores

PROPORTIONS

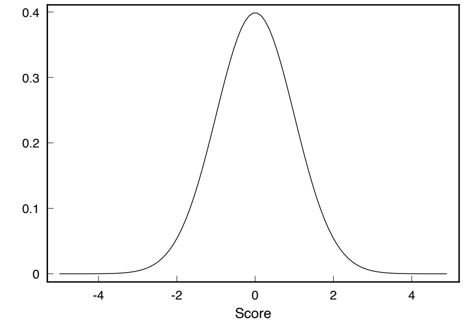


- We find that 69.77% of the scores lie below 0.5 standard deviation above the mean or beyond 2.5 standard deviations above the mean
- so how many scores are in that range?
- multiply the total number of scores (250) with the percent in the range (decimal form)

$$(0.6997) \times (250) = 174.925 \approx 175$$

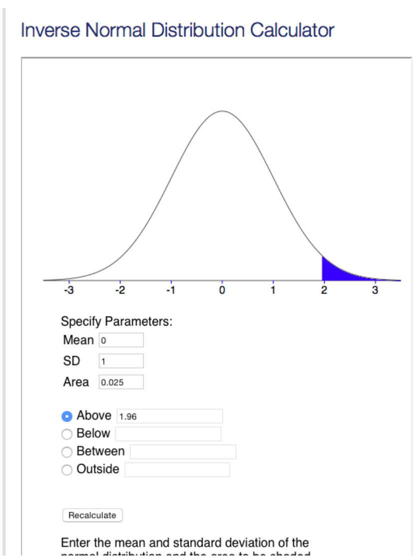
PERCENTILES

- Xth percentile is score for which X percent of scores fall at or below
- 50th percentile is the median (and the mean!)



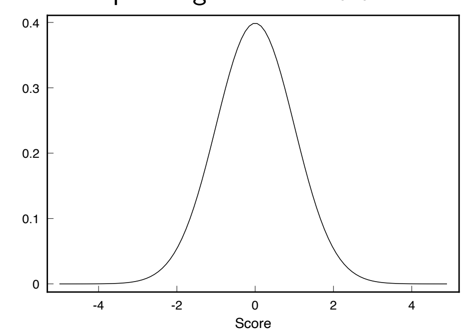
PERCENTILES

- The Inverse Normal Calculator gives the z-score that corresponds to different areas
- Click "Below" to make it fill in from the left side



EXAMPLE

- to find P_{75} for a standard normal curve, enter Area$= 0.75$
- and find that the corresponding z-score is 0.674



- what about P_{25}?

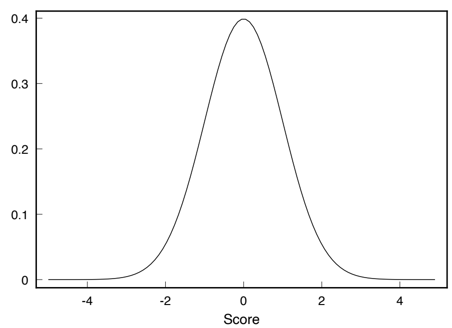
EXAMPLE

- Symmetry!

$$P_{25} = -P_{75}$$

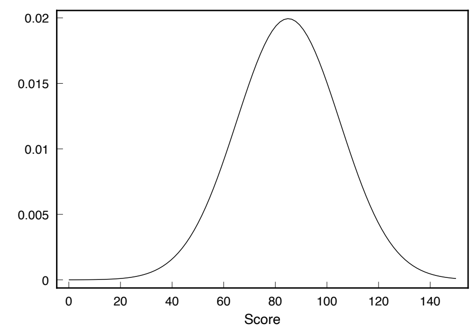
- in general for $X < 50$,

$$P_X = -P_{100-X}$$



CONVERSION

- suppose you have a **normal** distribution with a mean of 85 and a standard deviation of 20
- how would you find the 70th percentile?



z-scores

- Indirect way:
  1. Calculate percentile of z-score distribution.
  2. Convert z-score back to a raw score.
- from z-score we can calculate

$$X = (s)(z) + \overline{X}$$

- the online-app shows that for a standard normal, $P_{70} = 0.5244$, so

$$X = (20)(0.5244) + 85 = 95.49$$

- Or, just change the mean and the standard deviation of the normal distribution in the on-line app

BUSINESS DECISION

- suppose you are part of a company manufacturing what you think will be the "next big thing" in men's pants

BUSINESS DECISION

- You want to produce pants that will fit the center of the distribution of men's waist sizes
- To maximize profit, there is no need to make pants for men with really small or really large waists because there are so few such people
- According to the National Health and Nutrition Examination Survey the distribution of waist circumference is approximately normal with (in centimeters)

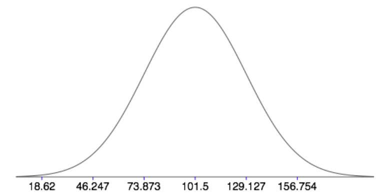
$$\mu = 101.5$$

- (around 40 inches)

$$\sigma = 27.6$$

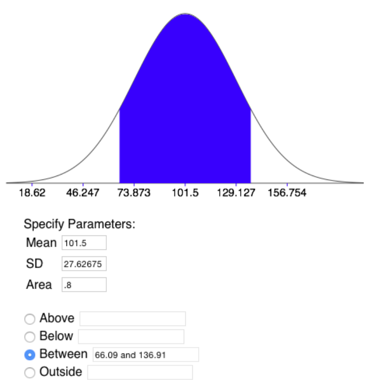
BUSINESS DECISION

- What size waists do you manufacture to cover the middle 80% of the distribution of waist sizes?



BUSINESS DECISION

- What size waists do you manufacture to cover the middle 80% of the distribution of waist sizes?



- (Obviously, there are more things to consider: costs, how many sizes, customer preferences,...)

BUSINESS DECISION

- You plan to set up a canoe business on the Wabash River. You want to purchase canoes that will be able to carry 90% of 3-person families. Canoes that carry more weight cost more, so you want canoes that hold the lower 90% of people (mother, father, child)
- Statistics (pounds)
  - Adult women:

$$\mu = 168.5, \ \sigma = 67.7$$

  - Adult men:

$$\mu = 195.7, \ \sigma = 68.0$$

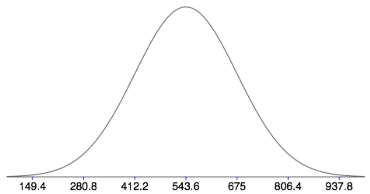
  - Children (18 year old):

$$\mu = 179.4, \ \sigma = 89.7$$

BUSINESS DECISION

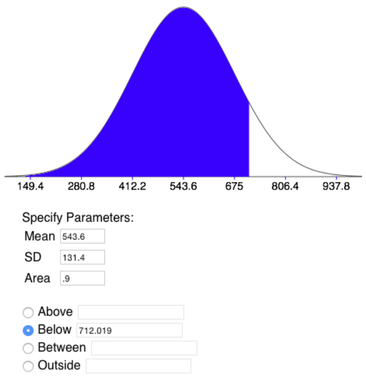
- For a *family* we add the means and the variances
- Family:

$$\mu = 168.5 + 195.7 + 179.4 = 543.6$$
$$\sigma^2 = (67.7)^2 + (68.0)^2 + (89.7)^2 = 17261$$
$$\sigma = 131.4$$



BUSINESS DECISION

- To be able to hold 90% of families, you need a canoe that holds weight of the 90th percentile



CONCLUSIONS

- normal distribution
- area under curve
- proportions
- percentiles

NEXT TIME

- percentile ranks
- examples

A statistical approach to assigning grades.

PSY 201: Statistics in Psychology

Lecture 09
Normal distribution
A statistical approach to assigning grades.

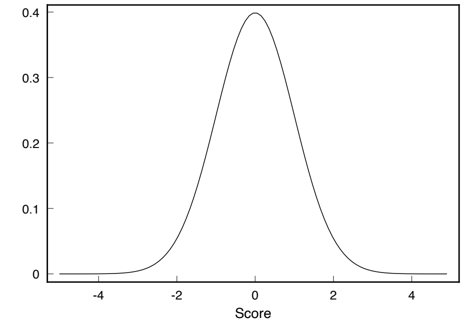
Greg Francis

Purdue University

Fall 2023

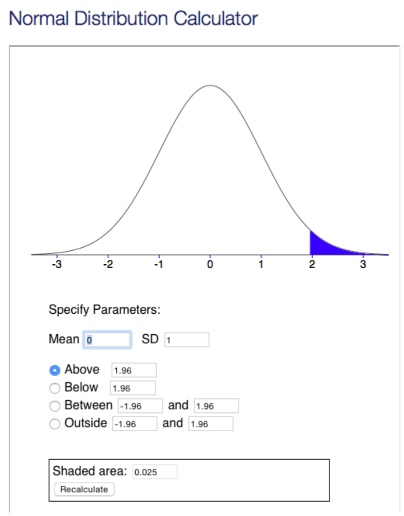
PERCENTILE RANKS

- area under the curve below a particular score (same as we did before)



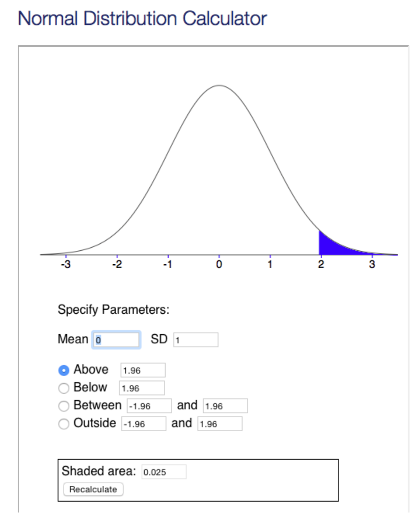
PERCENTILE RANKS

- suppose you have a **normal** distribution with a mean of 85 and a standard deviation of 20
- how would you find the percentile rank of raw score 65?



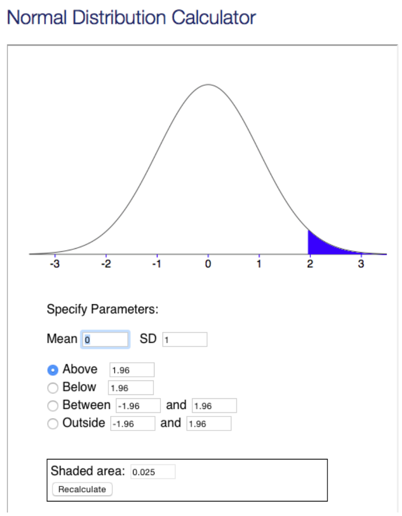
EXAMPLE

- A set of 200 scores is normally distributed with a mean of 60 and a standard deviation of 12.
- How many scores lie between the values of 48 and 80? 65 and 75? 34 and 52?
- Normal Distribution Calculator
- **proportion** of scores is area under the curve



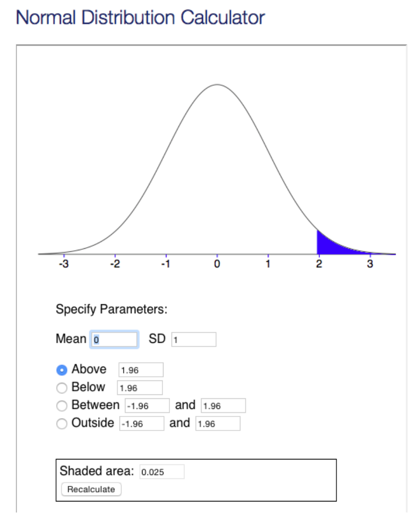
EXAMPLE

- the area is the **proportion** of scores
- to get the number of scores, we multiply the proportion times the total number of scores
  - number of scores between 48 and 80 is
    $200 \times 0.7938 = 158.76$
- We do the same thing for the other cases...



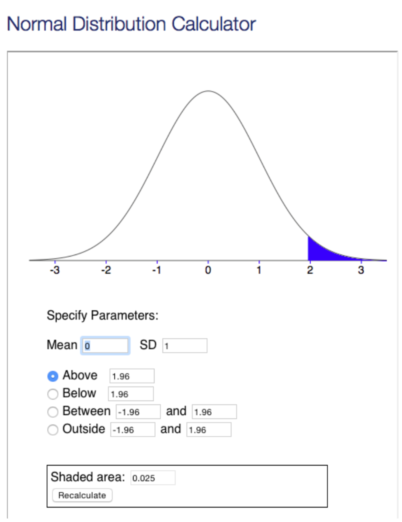
EXAMPLE

- To get the number of scores between 65 and 75 we calculate:
- Total area $= 0.2316$
  - Number of scores between 65 and 75 $=$
    $200 \times 0.2316 = 46.32$



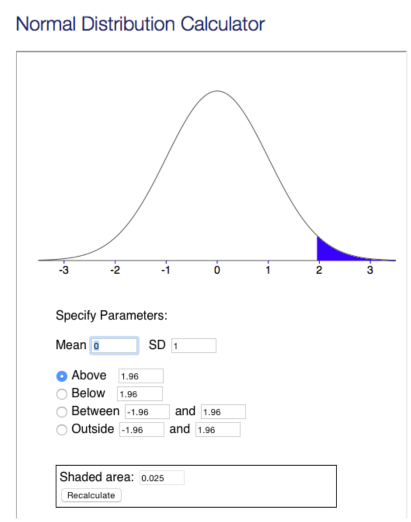
EXAMPLE

- to find scores between 34 and 52 we find that:
- area $= 0.2364$
  - Number of scores between 34 and 52 $=$
    $200 \times 0.2364 = 47.28$



EXAMPLE

- How many scores exceed the values of 80, 60, and 40?
- area under the normal curve greater than 80 is 0.0475
  - So the number of scores greater than 80 is:
    $200 \times 0.0475 = 9.5$
- Same approach for the other scores

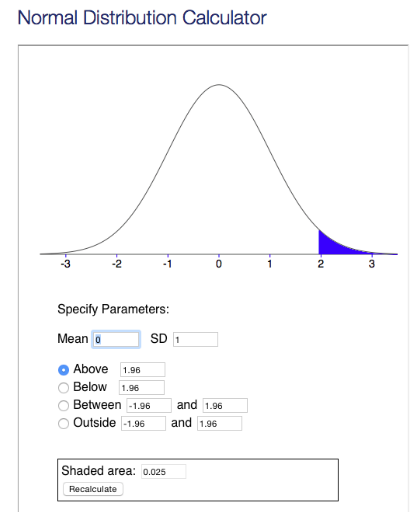


EXAMPLE

- area under the normal curve greater than 60 is 0.5
  - so the number of scores greater than 60 is $200 \times 0.5 = 100$.
- area under the normal curve greater than 40 is 0.9525
  - so the number of scores greater than 40 is $200 \times 0.9525 = 190.5$

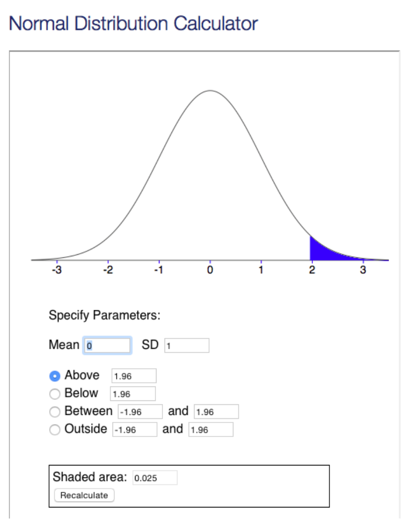
EXAMPLE

- How many scores are less than the values of 35, 50 and 75?
- area below 35 is 0.0188
  - Number of scores less than 35 is $200 \times 0.0188 = 3.76$



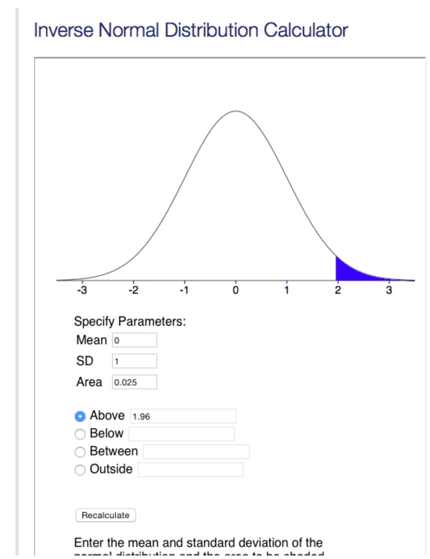
EXAMPLE

- area below 50 is 0.2033
  - Number of scores less than 50 is $200 \times 0.2033 = 40.66$
- area below 75 is $0.3944 + 0.5 = 0.8944$
  - Number of scores less than 75 is $200 \times 0.8944 = 178.88$



EXAMPLE

- Find P_{35}, P_{80}, PR_{55}, PR_{70}.
- For percentiles, we use the Inverse Normal Calculator



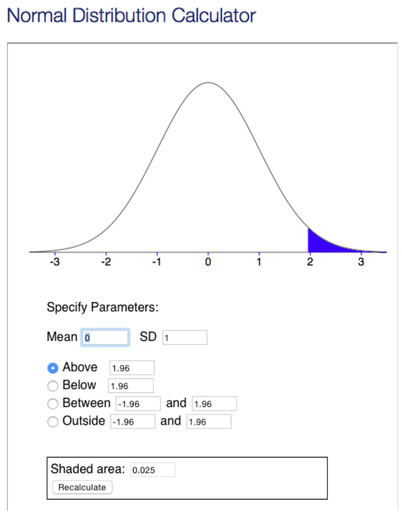
EXAMPLE

- For Percentile Ranks, use the Normal Distribution Calculator
- Find area under the normal for scores **less** than these scores

  area less than 55 $\rightarrow$ 0.3372

  area less than 70 $\rightarrow$ 0.7967

- in percentiles these mean:
  - $PR_{55} = 33.72$
  - $PR_{70} = 79.67$



ASSIGNING GRADES

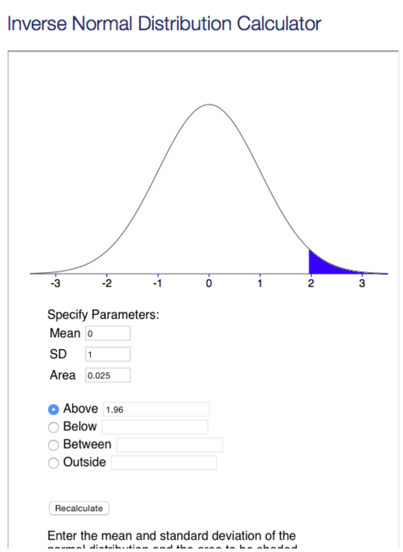
- A statistics instructor tells the class that grading will be based on the normal distribution. He plans to give 10 percent A's, 20 percent B's, 40 percent C's, 20 percent D's, and 10 percent F's.
- If the final examination scores have a mean of 75 and a standard deviation of 9.6, what is the range of scores for each grade?

A range

- To find the A range, we need to find what score corresponds to the top 10%.
- Use the Inverse Normal Calculator:

$$P_{90} = 87.30$$

- So the A range is any score greater than 87.30

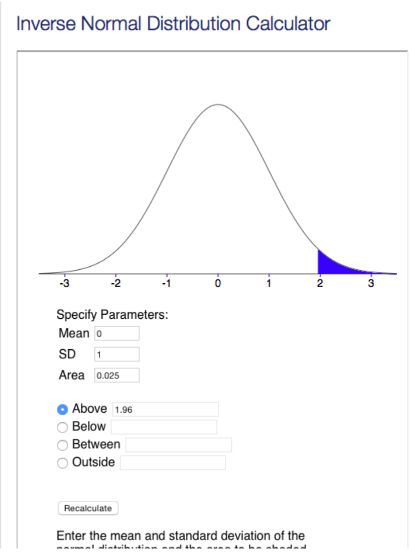


B range

- B range must include 20% of scores.
- Must be less than 87.30.
- We need to find P_{70}!! (lower limit)

$$P_{70} = 80.03$$

- So the B range is between 80.03 and 87.30

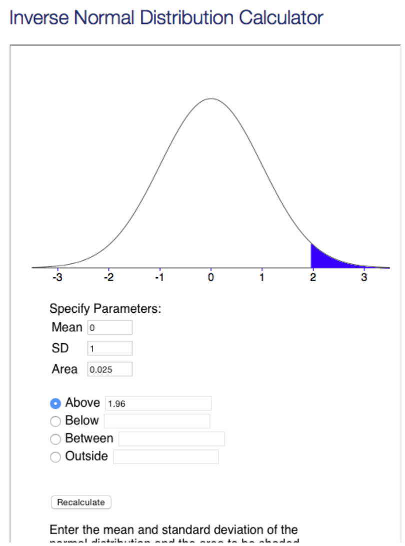


C range

- C range must include 40% of scores.
- Must be less than 80.03.
- We need to find P_{30}!! (lower limit)
- Use the Inverse Normal Calculator:

$$P_{30} = 69.97$$

- So the C range is between 69.97 and 80.03

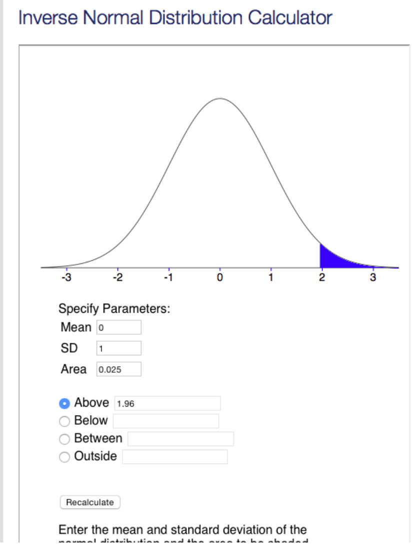


D range

- D range must include 20% of scores.
- Must be less than 69.97.
- We need to find P_{10}!! (lower limit)
- Use the Inverse Normal Calculator:

$$P_{10} = 62.70$$

- So the D range is between 62.70 and 69.97
- of course the F range is anything below 62.70



CONCLUSIONS

- the on-line calculator makes these problems fairly easy to compute
- it still takes effort to think about what you actually need
- looking at graphs helps a lot!

NEXT TIME

- correlation
- identifying relationships between data sets

How changes in one variable correspond to changes in another variable.

PSY 201: Statistics in Psychology

Lecture 10

Correlation

How changes in one variable correspond to change in another variable.

Greg Francis

Purdue University

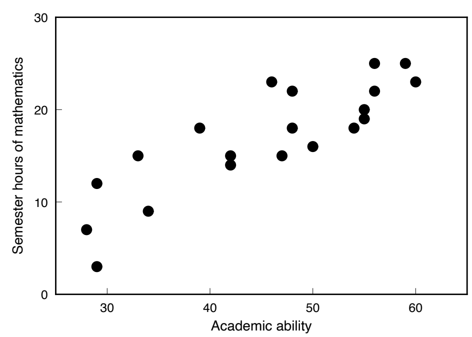
Fall 2023

CORRELATION

- two variables may be related
  - SAT scores, GPA
  - hours in therapy, self-esteem
  - grade on homeworks, grade on exams
  - number of risk factors, probability of getting AIDS
  - height, points in basketball
  - ...
- how do we show the relationship?
- scattergrams

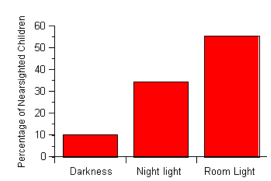
SCATTERGRAMS

- plot value of one variable against the value of the other variable



RELATIONSHIPS

- Identifying these types of relationships is one of the key issues in statistical analysis
- Consider a 1999 study that reported a relationship between the use of nightlights in a child's room and the tendency of the child to need glasses



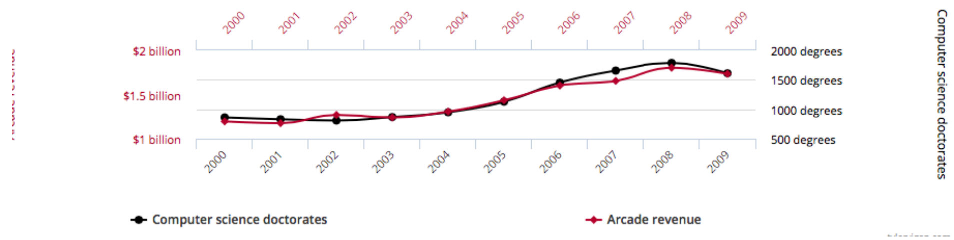
- My daughter slept with a nightlight. Was I a bad father?

COMPLICATIONS

- Clearly there is a relationship between using a nightlight and needing glasses
- However, it's not clear what the nature of the relationship involves
- It *could* be that the extra light somehow influences the child's eyes and causes the need for glasses
- Or it could be that needing glasses will somehow co-occur with the use of a nightlight (e.g., children who need glasses will want a night light, or their parents will want a nightlight)
- Finding a relationship is necessary for establishing causation, but it is not enough

SPURIOUS CORRELATION

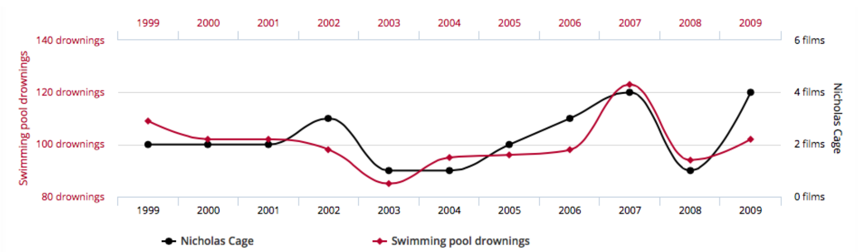
- Since so many variables get measured, it is easy to identify spurious correlations
- Sometimes there is an explanation for the relationship:



- (increased use of technology)

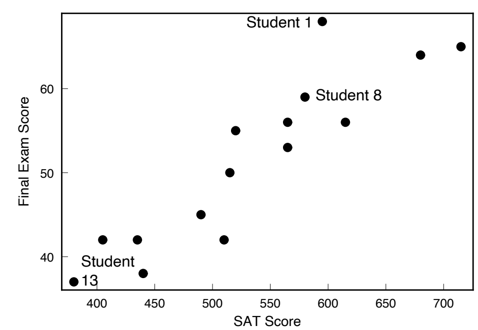
SPURIOUS CORRELATION

- Since so many variables get measured, it is easy to identify spurious correlations
- Sometimes there is no explanation for the relationship:



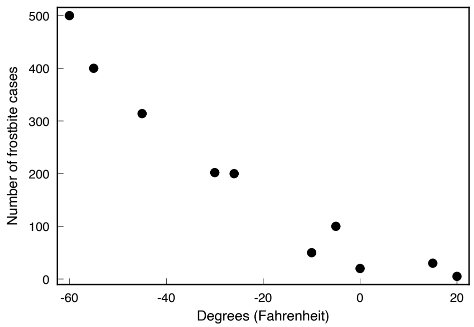
POSITIVE CORRELATION

- First, we need to understand how to quantify the existence of a relationship.
- Increases in the value of one variable tend to occur with increases in the value of the other variable
- SAT scores and exam scores



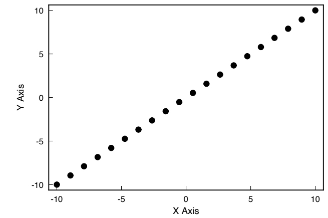
NEGATIVE CORRELATION

- Increases in the value of one variable tend to occur with **decreases** in the value of the other variable
- temperature and number of people with frostbite

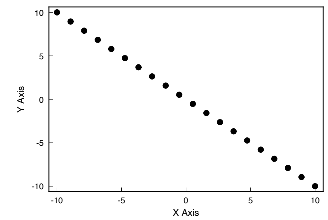


PERFECT CORRELATIONS

- perfect positive correlation

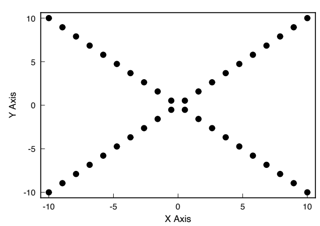


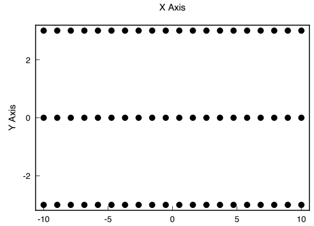
- perfect negative correlation



NO CORRELATION

- no correlation
- balance of larger and smaller values





CORRELATION COEFFICIENT

- quantitative measure of correlation
- bounded between

$$-1.0 \ \& \ +1.0$$

- correlation coefficient of -1.0 indicates perfect negative correlation
- correlation coefficient of +1.0 indicates perfect positive correlation
- correlation coefficient of 0.0 indicates **no** correlation
- values in between give ordinal measures of relationship

PEARSON r

- Pearson product-moment correlation coefficient
- one correlation coefficient for **quantitative** data (the most important one)

$$r = \frac{\text{degree to which } X \text{ and } Y \text{ vary together}}{\text{degree to which } X \text{ and } Y \text{ vary separately}}$$

- several formulas
  - z-scores
  - Deviation scores
  - Raw scores
  - Covariance
- all give the same result!

z SCORES

- Two steps
  - Convert raw scores into z scores
  - Find the mean of cross-products

$$r_{xy} = \frac{\Sigma z_x z_y}{n-1}$$

z SCORES

- what does this calculation do?
- suppose you have two distributions that have a positive correlation
- then a large value of X will be above $\overline{X}$ and have a positive z_x score
- and a corresponding Y will be above $\overline{Y}$ and have a positive z_y score
- Thus the cross-product

$$z_x z_y$$

- will be positive

PEARSON r

- also a small value of X will be below $\overline{X}$ and have a negative z_x score
- and the corresponding Y will be below $\overline{Y}$ and have a negative z_y score
- Thus

$$z_x z_y$$

- will again be positive
- to find the average, sum all the products (positive numbers) we divide by $n-1$

$$r_{xy} = \frac{\Sigma z_x z_y}{n-1}$$

- still a positive number!

PEARSON r

- exactly the opposite is true for negatively correlated distributions
- then a large value of X will be above $\overline{X}$ and have a positive z_x score
- and a corresponding Y will be **below** $\overline{Y}$ and have a **negative** z_y score
- Thus

$$z_x z_y$$

- will be negative

PEARSON r

- while a small value of X will be below $\overline{X}$ and have a negative z_x score
- and the corresponding Y will be **above** $\overline{Y}$ and have a **positive** z_y score
- Thus

$$z_x z_y$$

- will again be negative
- to find the average, sum all the products (negative numbers) we divide by $n-1$

$$r_{xy} = \frac{\Sigma z_x z_y}{n-1}$$

- still a negative number!

DEVIATION FORMULA

- it is awkward to convert to z scores
- we can get the same number with deviation scores

$$x = X - \overline{X}$$

$$y = Y - \overline{Y}$$

- deviation score formula

$$r_{xy} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

RAW SCORE FORMULA

- it is awkward to calculate deviation scores
- raw score formula

$$r_{xy} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{\left[n\Sigma X^2 - (\Sigma X)^2\right]\left[n\Sigma Y^2 - (\Sigma Y)^2\right]}}$$

COVARIANCE FORMULA

$$\text{covariance} = s_{xy} = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{n-1}$$

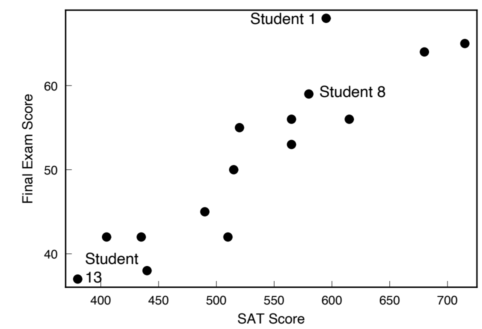
- average cross-product of deviation scores (similar to variance)
- Pearson r turns out to be:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- where s_x and s_y are the standard deviations of their respective distributions

EXAMPLE

X	Y
595	68
520	55
715	65
405	42
680	64
490	45
565	56
580	59
615	56
435	42
440	38
515	50
380	37
510	42
565	53



EXAMPLE

- standard score formula

$$r_{xy} = \frac{\Sigma z_x z_y}{n-1} = \frac{12.67}{14} = 0.905$$

X	Y	z_x	z_y	$z_x z_y$
595	68	0.63	1.64	1.03
520	55	-0.15	0.35	-0.05
715	65	1.88	1.34	2.52
405	42	-1.34	-0.94	1.26
680	64	1.51	1.24	1.87
490	45	-0.46	-0.64	0.29
565	56	0.32	0.45	0.14
580	59	0.48	0.74	0.36
615	56	0.84	0.45	0.38
435	42	-1.03	-0.94	0.97
440	38	-0.97	-1.33	1.29
515	50	-0.20	-0.15	0.03
380	37	-1.60	-1.43	2.29
510	42	-0.25	-0.94	0.24
565	53	0.32	0.15	0.05
$\sum X = 8010$	$\sum Y = 772$	$\sum z_x = 0.0$	$\sum z_y = 0.0$	$\sum z_x z_y = 12.67$

EXAMPLE

- deviation score formula

$$r_{xy} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{12332.00}{\sqrt{(130460.0)(1429.72)}} = 0.903$$

X	Y	x	y	xy
595	68	61.0	16.53	1008.33
520	55	-14.0	3.53	-49.42
715	65	181.0	13.53	2448.93
405	42	-129.0	-9.47	1221.63
680	64	146.0	12.53	1829.38
490	45	-44.0	-6.47	284.68
565	56	31.0	4.53	140.43
580	59	46.0	7.53	346.38
615	56	81.0	4.53	366.93
435	42	-99.0	-9.47	937.53
440	38	-94.0	-13.47	1266.18
515	50	-19.0	-1.47	27.93
380	37	-154.0	-14.47	2228.38
510	42	-24.0	-9.47	227.28
565	53	31.0	1.53	47.43
$\sum X = 8010$	$\sum Y = 772$	$\sum x = 0.0$	$\sum y = 0.0$	$\sum xy = 12332.00$

- $\Sigma x^2 = 130460.0$ and $\Sigma y^2 = 1429.72$

EXAMPLE

- raw score formula

$$r_{xy} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{\left[n\Sigma X^2 - (\Sigma X)^2\right]\left[n\Sigma Y^2 - (\Sigma Y)^2\right]}}$$

$$\frac{(15)(424580) - (8010)(772)}{\sqrt{[(15)(4407800) - (8010)^2][(15)(41162) - (772)^2]}} = 0.903$$

X	Y	XY
595	68	40460
520	55	28600
715	65	46475
405	42	17010
680	64	43520
490	45	22050
565	56	31640
580	59	34220
615	56	34440
435	42	18270
440	38	16720
515	50	25750
380	37	14060
510	42	21420
565	53	29945
$\sum X = 8010$	$\sum Y = 772$	$\sum XY = 424580$

- $\Sigma X^2 = 4407800$ and $\Sigma Y^2 = 41162$

EXAMPLE

- covariance formula

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{880.86}{(96.53)(10.11)} = 0.903$$

- where,

$$s_{xy} = \frac{\Sigma xy}{n-1} = \frac{12332}{14} = 880.86$$

$$s_x = \sqrt{\frac{\Sigma x^2}{n-1}} = \sqrt{\frac{130460}{14}} = 96.53$$

$$s_y = \sqrt{\frac{\Sigma y^2}{n-1}} = \sqrt{\frac{1429.72}{14}} = 10.11$$

CORRELATION

- r measures correlation between two variables
- **not** just any two variables
  - The two variables must be **paired observations**.
  - Variables must be quantitative (interval or ratio scale).

CONCLUSIONS

- correlation
- scattergrams
- Pearson r
- formulas

NEXT TIME

- factors affecting r
- interpreting r

Is there a link between IQ and problem solving ability?

PSY 201: Statistics in Psychology

Lecture 11

Correlation

Is there a relationship between IQ and problem solving ability?

Greg Francis

Purdue University

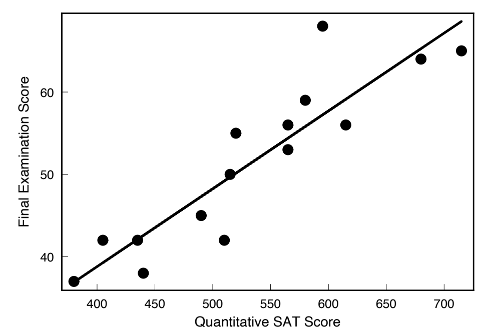
Fall 2023

CORRELATION

- suppose you get $r \approx 0$.
- Does that mean there is no correlation between the data sets?
- many aspects of the data may affect the value of r
  - Linearity of data.
  - Homogeneity of group.
  - Size of group.
  - Restricted range.

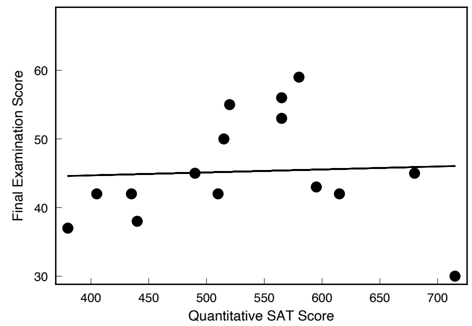
LINEARITY

- r is partly an index of how well a straight line fits the data set
- Here, $r = 0.903$



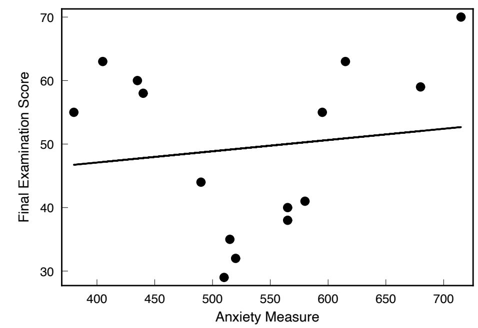
NONLINEARITY

- when data points don't fall along a single line (nonlinear data)
- Here, $r = 0.05$



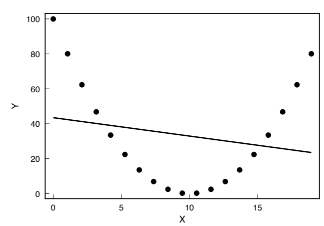
NONLINEARITY

- there are lots of types of nonlinearities
- curvilinear relationship
- Here, $r = 0.131$

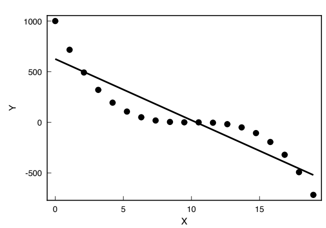


NONLINEARITY

- It can get complicated
- $r = -0.20$



- $r = -0.91$



BOTTOM LINE

- Pearson r is an index of a **linear** relationship between variables
- if another (nonlinear) relationship exists, r might not notice it
- Pearson r measures only simple relationships between variables
- if r is small, you might want to plot a scattergram to look at the data to notice if other relationships exist

HOMOGENEITY

- suppose you get $r \approx 0$, and you cannot detect any type of nonlinear relationship
- Does this mean there is no correlation between the variables?
- Not necessarily, it may be that the data does not have enough variation in it
- Correlation measures how variable X changes with variable Y
- if one doesn't change much, there won't be a strong correlation

HOMOGENEITY

- consider the covariance formula

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- where, covariance is

$$s_{xy} = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{n-1}$$

- if there is little change in Y from $\overline{Y}$, s_{xy} is going to be small because $+/-$ variations in $X - \overline{X}$ will be weighted by small values of $Y - \overline{Y}$
- similarly, s_y is going to be small, so we divide a small number by a small number

HOMOGENEITY

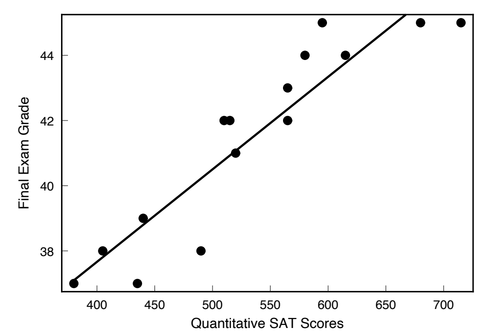
- intuitively

$$r = \frac{\text{degree to which } X \text{ and } Y \text{ vary together}}{\text{degree to which } X \text{ and } Y \text{ vary separately}}$$

- if one of those variables (or both) is not varying much at all, r will be small
- you need enough variability across both sets of scores to adequately measure correlation

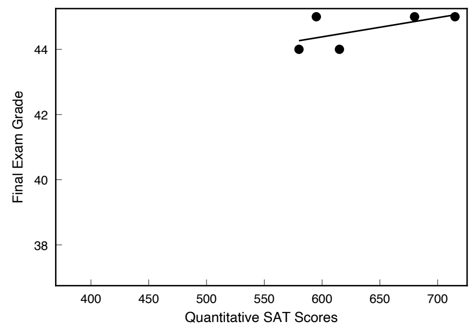
HOMOGENEITY

- the effects of homogeneity can be subtle
- relationship between SAT scores and Final exam grade
- $r = 0.92$



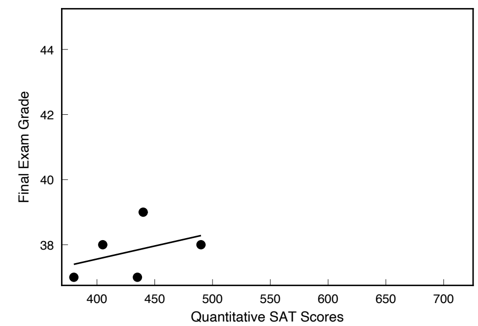
HOMOGENEITY

- suppose we looked at the relationship among only the best students
- (those with final exam scores above 44)
- $r = 0.62$



HOMOGENEITY

- or worst students
- (those with final exam scores below 40)
- $r = 0.62$



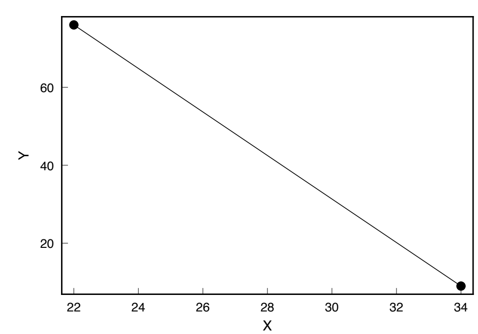
- correlation drops!

SIGNIFICANCE

- if you have $r \approx 0$, it may be because there is not enough variation in your data set
- e.g.
  - IQ and problem solving is probably unrelated among a group of geniuses
  - IQ and problem solving is probably unrelated among a group of idiots
  - IQ and problem solving is probably strongly related among a mix of geniuses, idiots, and normals

SIZE OF GROUP

- suppose you have only two data points
- you can always draw a straight line connecting them
- which implies perfect correlation
- $r = -1.0$



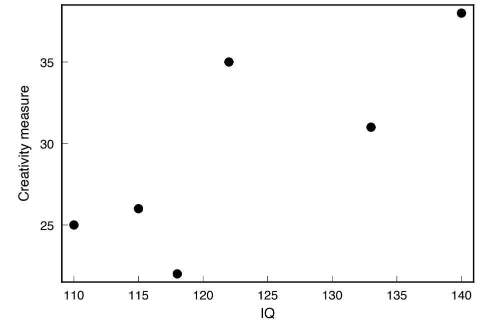
- (correlation doesn't tell us anything useful!)

SIZE OF GROUP

- if you have enough data points for correlation to be meaningful (> 2), and you have enough variation in the data, then
- size of group is not important in determining the **value** of r
- we will see later that it is important in determining the **accuracy** of the relationship (hypothesis testing)

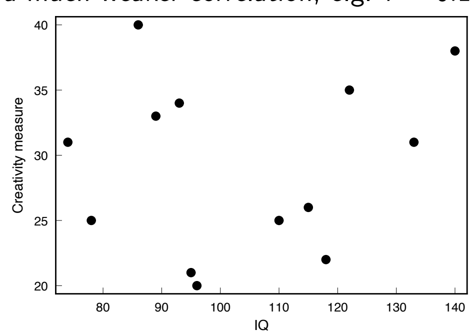
RESTRICTED RANGE

- if you sample data from a limited range you may not be able to trust the correlation values in general
- e.g., suppose you want to study relationships between IQ and creativity
- if you sample college students you will probably get IQ's between 110 and 140
- perhaps you find a strong correlation, e.g. $r = 0.78$



RESTRICTED RANGE

- if you sample from the general population (not just college students) you would get a larger range of IQs
- you may find a much weaker correlation, e.g. $r = 0.12$



RESTRICTED RANGE

- of course, it could be that you fail to find a large r over a restricted range, but a larger range finds a large r (this is slightly different from the issue of homogeneity)
- in general
- a correlation measure applies **only** to the range of values used to compute it
- you **cannot** extend the correlation value to other ranges

INTERPRETATION OF r

- if we calculate a value of r
- How do we know what it means?
- How do we compare r values for different data sets?
- Rule of thumb

$\|r\|$	Interpretation
0.9 to 1.0	Very high correlation
0.7 to 0.9	High correlation
0.5 to 0.7	Moderate correlation
0.3 to 0.5	Low positive correlation
0.0 to 0.3	Little if any correlation

SCALE OF r

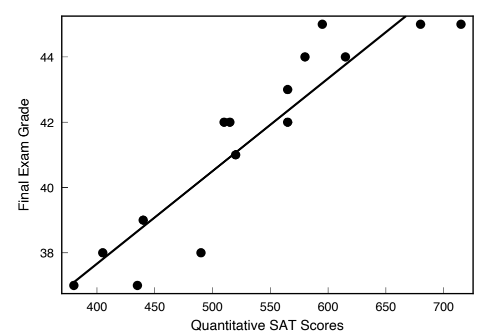
- values of r are **ordinal** measures of correlation
  - higher r values indicate larger correlation
  - equal spacings of r values may not indicate equal spacings of correlation
- thus, $r = 0.90$ is **not** twice as correlated as $r = 0.45$
- the difference in correlation between $r = 0.90$ and $r = 0.75$ is **not** the same as the difference in correlation between $r = 0.60$ and $r = 0.45$.

VARIANCE

- we can interpret r in terms of variance
- correlation coefficient indicates relationships between variables
- also indicates proportion of **individual differences** that can be associated with individual differences of another variable

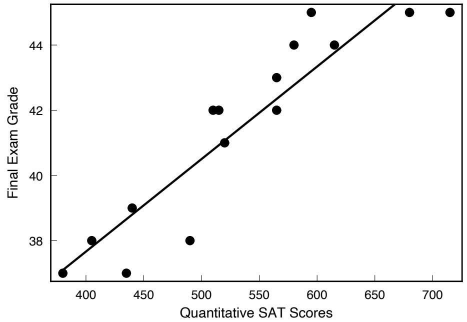
VARIANCE

- the idea is embedded in mathematical models
- assume you want to **predict** the final exam score when you know the SAT score
- line predicts score (could go in reverse too)



VARIATION

- deviation of a final exam score from the mean value can be due to deviation accounted for by SAT scores, or due to something else



VARIATION

- it turns out that
$$r^2 = \frac{s_a^2}{s_y^2}$$
- where:
  - s_y^2 = the total variance in y
  - s_a^2 = the variance in Y associated with variance in X
- thus, r^2 is the **proportion** of variance in Y accounted for with variance in X
- we are skipping the mathematical details (thank you!)
- called the coefficient of determination

CONCLUSIONS

- Pearson r
- size
- interpretation

NEXT TIME

- probability
- rules
- significance

Why casinos make money.

PSY 201: Statistics in Psychology

Lecture 12
Probability
Why casinos make money.

Greg Francis

Purdue University

Fall 2023

DESCRIPTIVE STATISTICS

- most of what we have discussed so far is called descriptive statistics
  - distributions
  - graphs
  - central tendency
  - variation
  - correlation
- **describe** sets of data

INFERENTIAL STATISTICS

- given a set of data from a sample
- we want to **infer** something about the entire population
  - mean
  - standard deviation
  - correlation
  - ...
- never with certainty, but with **probability**

PROBABILITY

- number between 0 and 1
- probability of event A is written as

$$P(A)$$

- if

$$P(A) = 1.0$$

- it indicates with certainty that event A will happen
- if

$$P(A) = 0$$

- it indicates with certainty that event A will **not** happen

PROBABILITY LAWS

- there are specific rules to probability
- we want to know the probability of many events, pairs of events, contingent events,...
- how to calculate probabilities depends upon
  - Complements
  - Mutually exclusive compound events
  - Nonmutually exclusive events
  - Statistically independent joint events
  - Statistically dependent joint events

SINGLE EVENTS

- precise definition requires high-level mathematics
- intuitive definition is that probability of a single event is the ratio of the number of possible outcomes that include the event to the total number of possible outcomes

$$P(\text{a die coming up 3}) = \frac{\text{Number of outcomes that include 3}}{\text{Total number of outcomes}}$$

$$P(\text{a die coming up 3}) = \frac{1}{6} \approx 0.167$$

1 2 3 4 5 6

COMPLEMENTS

- suppose we know the probability $P(A)$, where A is some event
- then if $\overline{A}$ represents "not A" (called the complement of A)

$$P(\overline{A}) = 1.0 - P(A)$$

- when $A =$ turning up a 3 on a die, $\overline{A}$ means turning up anything other than a 3
- since $P(A) = 0.167$
- $P(\overline{A}) = 1.0 - 0.167 = 0.833$

1 2 3 4 5 6

COMPOUND EVENTS

- sometimes we know the probability of two events A and B, and we want to know the probability of event A or B
- e.g.

$$P(\text{turning up a 3 or a 4 on a die})$$

- these are **mutually exclusive events**
- one or the other

MUTUALLY EXCLUSIVE

- for mutually exclusive compound events, calculating the probability of the compound is easy
- consider probability of rolling numbers on a die

$$P(\text{a 3 or a 4}) = P(3) + P(4)$$

$$P(\text{turning up a 3 or a 4 on a die}) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

1 2 3 4 5 6

- in general, if A and B are mutually exclusive

$$P(A \text{ or } B) = P(A) + P(B)$$

NONMUTUALLY EXCLUSIVE

- sometimes events are **not** mutually exclusive
- e.g.
  - A = turning up a number ≤ 3 on a die: $P(A) = \frac{1}{2}$
  - B = turning up an odd number on a die: $P(B) = \frac{1}{2}$
- what is $P(A \text{ or } B)$?

1 2 3 4 5 6

- cannot just add probabilities because numbers common to A and B get counted twice!

NONMUTUALLY EXCLUSIVE

- subtract out common probability

$$P(\text{number} \leq 3 \text{ or odd}) =$$

$$P(\leq 3) + P(\text{odd}) - P(\leq 3 \textbf{ and } \text{odd}) =$$

$$\frac{1}{2} + \frac{1}{2} - \frac{2}{6} = \frac{3}{6} + \frac{3}{6} - \frac{2}{6} = \frac{4}{6} = \frac{2}{3}$$

- in general

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

- when the events are mutually exclusive, $P(A \text{ and } B) = 0$, and we get the rule for mutually exclusive events

JOINT EVENTS

- if we know $P(A)$ and $P(B)$, what is $P(A \text{ and } B)$?
- both events must occur (simultaneously or successively)
- e.g.

$$P(\text{3 on a die and HEAD on a coin flip})$$

STATISTICAL INDEPENDENCE

- events are independent if the occurrence of one event does not affect the probability of the other event occurring
- e.g., rolling a 3 on a die has no effect on whether or not a coin will come up HEADS

$$P(\text{3 on die}) = \frac{1}{6}$$

$$P(\text{HEADS}) = \frac{1}{2}$$

- so

$$P(\text{3 and HEADS}) = P(3) \times P(\text{HEADS})$$

$$P(\text{3 and HEADS}) = \frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$$

MULTIPLICATION

- why multiply probabilities of joint events?
- probability is ratio of the number of outcomes including an event to the total number of possible outcomes
- for the joint event "3 on a die and HEADS", the possible outcomes are

1H, 2H, 3H, 4H, 5H, 6H
1T, 2T, 3T, 4T, 5T, 6T

- count up the possibilities!

SAMPLING WITH REPLACEMENT

- suppose we have 10 numbered balls in a jar
- the probability of drawing ball 3 is $\frac{1}{10}$
- if we put the ball back, the probability of drawing ball 3 again is $\frac{1}{10}$ (same for any ball)
- each event (drawing a ball) is independent from previous events
- in general for independent events A and B,

$$P(A \text{ and } B) = P(A) \times P(B)$$

SAMPLING WITHOUT REPLACEMENT

- many times the probability of an event **does** depend on other events
- e.g., suppose we have ten numbered balls in a jar
- the probability of drawing ball 3 is $\frac{1}{10}$
- suppose we draw ball 2; leaving nine balls in the jar
- the probability of drawing ball 3 is now $\frac{1}{9}$

CONDITIONAL PROBABILITIES

- we can describe the effect of other events by identifying conditional probabilities
- e.g.

$$P(\text{drawing ball 3 given that ball 2 was already drawn})$$

$$P(\text{ball 3}|\text{ball 2})$$

- in general the probability of event A, given event B is written as

$$P(A|B)$$

- no direct way of calculating from $P(A)$ or $P(B)$

NONINDEPENDENT EVENTS

- when

$$P(A) = P(A|B)$$

- we say events A and B are independent
- otherwise the events are nonindependent (dependent)

JOINT PROBABILITY

- if we know $P(A)$ and $P(B|A)$ then we can calculate the joint probability

$$P(A \text{ and } B) = P(A)P(B|A)$$

- if we know $P(B)$ and $P(A|B)$ then we can calculate the joint probability

$$P(A \text{ and } B) = P(B)P(A|B)$$

- same number!
- if events are independent, this rule is the same as before because

$$P(A|B) = P(A)$$

EXAMPLE

- what is the probability of drawing ball 2 and then ball 3 from a jar with ten numbered balls?
- we know that

$$P(\text{drawing ball 2 from the full jar}) = \frac{1}{10}$$

$$P(\text{drawing ball 3}|\text{ball 2 is drawn from the full jar}) = \frac{1}{9}$$

- so

$$P(\text{drawing ball 3 and drawing ball 2}) =$$

$$P(\text{drawing ball 2 from the full jar})\times$$

$$P(\text{drawing ball 3}|\text{ball 2 is drawn from the full jar}) =$$

$$\frac{1}{10} \times \frac{1}{9} = \frac{1}{90}$$

RANDOMNESS

- we assume coin flips, rolling dice, samples from jars are **random** events
- unpredictable for a specific instance
- predictable on average over lots of samples (likelihood of happening)
- randomness is sometimes a good thing

CONCLUSIONS

- probability
- mutually exclusive events
- compound events
- independence

NEXT TIME

- review for exam
- **SECTION EXAM 1**
- fun problems with probability

PSY 201: Statistics in Psychology

Lecture 13

Probability

Coincidences are rarely interesting.

Greg Francis

Purdue University

Fall 2023

PROBABILITY

number between 0 and 1
probability of event A is written as

$$P(A)$$

if

$$P(A) = 1.0$$

it indicates with certainty that event A will happen
if

$$P(A) = 0$$

it indicates with certainty that event A will **not** happen

EVERYDAY EVENTS

- people often have misconceptions about the way probabilities interact
- things that seem rare may not actually be
- interesting to analyze the probability of events that seem unusual
  - Julius Ceasar
  - Hitting streaks
  - Predictive dreams
  - Shared birthdays
  - Con games with cards

JULIUS CAESAR

- Some 2000 years ago (or so) Julius Caesar is said to have gasped "You too, Brutus? Then I die." as his friend stabbed him to death
- What are the chances that you just inhaled a molecule that came out of his mouth?
- Surprisingly good! Almost 0.99.
- Assumes
  - Caesar's dying breath contained about $A = 2.2 \times 10^{22}$ molecules
  - Those molecules are free and distributed around the globe evenly.
  - Your inward breath contained about $B = 2.2 \times 10^{22}$ molecules
  - The atmosphere contains about $N = 10^{44}$ molecules

JULIUS CAESAR

- If there are N molecules and Caesar exhaled A of them, then the probability that any given molecule you inhale is from Caesar is

$$P(\text{m from C}) = \frac{A}{N} = 2.2 \times 10^{-23}$$

- which is very small!
- So the probability that any given molecule you inhale is **not** from Caesar is the complement:

$$P(\text{m not from C}) = 1 - \frac{A}{N} = 1 - 2.2 \times 10^{-23}$$

JULIUS CAESAR

- So the probability of inhaling B molecules that are not from Caesar is

$$P(\text{breath not from C}) = \left(1 - \frac{A}{N}\right)^B \approx 0.01$$

- So the probability of your breath containing at least one molecule from Caesar is approximately 1-0.01 $=$ 0.99!

HITTING STREAKS

- Pete Rose set a National League record with 44 consecutive games with a safe hit
- this is impressive, but is it rare?
- Rose batted around 0.300 (had a safe hit 30% of the time)
- so, assuming 4 at bats per game, the probability of **not** getting a hit during a game is

$$P(\text{no hit}) = (1 - 0.3)^4 = 0.24$$

- So the probability of getting at least one hit is 1-0.24 $=$ 0.76.

HITTING STREAKS

- Still, the probability of getting hits in any given sequence of 44 games is

$$P(\text{44 streak}) = (0.76)^{44} = 0.000005699$$

- and the probability of not getting a streak is

$$P(\text{not 44 streak}) = 1 - (0.76)^{44} = 0.999994301$$

HITTING STREAKS

- But there are 162 games in a season, so there are 118 sets of 44 consecutive games.
- Thus, the probability of not getting a streak of hits in at least 44 consecutive games out of a 162 game season is:

$$P(\text{no streak}) = (0.999994)^{118} = 0.999327$$

- so the probability of a 44-game streak is

$$P(\text{streak}) = 1 - (0.999994)^{118} = 0.000672$$

- (includes the possibility of streaks of more than 44 games)
- Still very rare!

HITTING STREAKS

- But how many players have been in the Major Leagues at any given time? (say 30 that bat like Rose)
- the probability that **every** player will **not** get a streak of at least 44 games in a given year is

$$P(\text{no streak}) = (0.9993)^{30} = 0.9800$$

- So probability that at least one player gets such a streak is

$$1.0 - 0.980027651 = 0.019972349$$

- still small!

HITTING STREAKS

- And how many years has baseball been played? (say 100)
- the probability that **every** year everyone will **not** get a streak of at least 44 games in a given year is

$$P(\text{no streak}) = (0.9800)^{100} = 0.1329$$

- So probability that at least one player on some year gets such a streak is

$$1.0 - 0.132994269 = 0.867005731$$

- which is pretty good odds!
- Thus, we can expect that Rose's streak will be broken eventually (unless pitchers become much better)

PREDICTIVE DREAMS

- ever dream something and had it come true?
- Many people take this occurence as evidence of extrasensory perception and "other worlds". But it's actually not that uncommon from a probabilitistic point of view
- suppose that the probability that a night's dream matches some later event in life is 1 in 10000

$$P(\text{predictive dream}) = 0.0001$$

- Then the chance that a dream is non-predictive is

$$P(\text{non predictive dream}) = 1 - 0.0001 = 0.9999$$

- assume that dream predictiveness is independent

PREDICTIVE DREAMS

- With 365 days a year, the probability that all 365 nights have non-predictive dreams is

$$P(\text{non predictive}) = (0.9999)^{365} = 0.96415$$

- so the probability that an individual has a predictive dream during a year is

$$P(\text{predictive}) = 1.0 - 0.96415 = 0.03585$$

- or about 3.6% of people have a predictive dream during a year
- considering that there are billions of people, this corresponds to millions of dreams (and lots of people talk about them!)

PREDICTIVE DREAMS

- but what about for an individual?
- over a span of 20 years, the probability that **all** your dreams are non predictive is

$$P(\text{non predictive}) = (0.96415)^{20} = 0.481$$

- which means that the probability of having a predictive dream is

$$P(\text{predictive}) = 1.0 - 0.481 = 0.519$$

- better than 50% chance!
- It might be unusual to **not** have had a predictive dream!

SHARED BIRTHDAYS

- ever been amazed to find that a group of people has two members with a shared birthday?
- you shouldn't be; it is not much of a coincidence
- Consider that a year has 366 days (counting February 29)
- to be **certain** that a group of people has a common birthday you would need a group of size 367
- what if we were willing to be just 50% certain of a shared birthday? How big would the group need to be?
- the surprising answer is 23

SHARED BIRTHDAYS

- what is the probability that a group of 23 people have no shared birthdays?
- how many ways to have birthdates from 23 people?

$$(366)^{23} = 9.1214727 \times 10^{58}$$

- How many ways to have 23 birthdates with no shared birthdays?

$$366 \times 365 \times 364 \times \ldots \times 344 = 4.5030611 \times 10^{58}$$

SHARED BIRTHDAYS

- probability of no shared birthdays is the number of ways to have no shared birthdays divided by the number of ways to have birthdays

$$P(\text{no shared}) = \frac{4.5030 \times 10^{58}}{9.1214 \times 10^{58}} = 0.4936$$

- so the probability of at least one shared birthday is

$$P(\text{shared}) = 1.0 - 0.4936 = 0.5063$$

- just about 50%
- Test it!

CON GAMES

- Here is a game that is played on the streets of some cities
- A man has 3 cards
  - Card 1: Black on both sides.
  - Card 2: Red on both sides.
  - Card 3: Black on one side and red on the other.
- He drops the cards in a hat, turns around and asks you to pick a card. Then he asks you to show him only one side of the card.
- Suppose you show him a red side. Now the man knows that the card cannot be Card 1 (black on both sides) and the card in your hand must be either Card 2 or Card 3.
- He offers you a bet of even money that he can guess the card. Is this a fair bet?

CON GAMES

- It might seem that this is fair. After all, the card in your hand is either Card 2 or Card 3. He has a 50% chance of guessing correctly, right?
- No.

CON GAMES

- Given that you have shown him one red side he knows that what you have shown is either:
  - first side of card 2
  - second side of card 2
  - red side of card 3
- Thus, of the possibilities, two are consistent with his guess of Card 2, and only one is consistent with your option of Card 3. He wins two-thirds of the time.

CONCLUSIONS

- probability
- apply to lots of situations
- coincidences are not as interesting as you might expect

NEXT TIME

- Decision making from noisy data
- Signal detection

Is that your phone?

PSY 201: Statistics in Psychology

Lecture 14

Signal detection

Is that your phone?

Greg Francis

Purdue University

Fall 2023

DETECTION IN NOISE

- Suppose you have to determine if there is a line of dots in a random field of dots (on-line example)
- Your ability to do the task depends on
  - The number of dots in the field
  - The position of the dots in the field
  - How much effort you put in the task

DETECTION IN NOISE

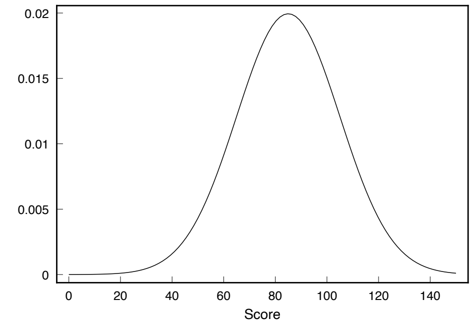
- Lots of tasks are essentially the same kind of situation
- what corresponds to noise in each situation?
  - Did you skip lunch at least one time last month?
  - Is that your phone ringing?
  - Does zinc shorten a cold?
  - Are men taller than women?

MEASUREMENT

- We suppose that there is some number that "measures" what you are interested in
  - Did you skip lunch at least one time last month?: strength of familiarity or memorability
  - Is that your phone ringing?: similarity to your ringtone?
  - Does zinc shorten a cold?: duration of a cold
  - Are men taller than women?: height

DISTRIBUTIONS

- Assume a normal distribution
- Mean is "noiseless" measurement
- Variation from mean is due to noise being added

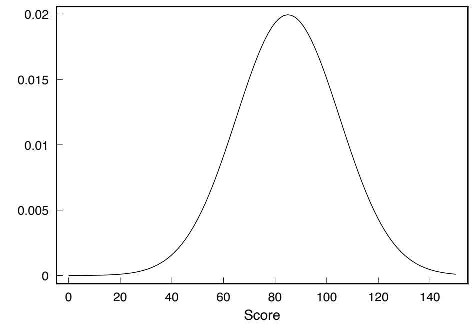


DISTRIBUTIONS

- There may be many sources of noise
  - Variation in the environment
  - Variation in your perceptual systems
  - Variation in your memory
- and many more!

DISTRIBUTIONS

- Suppose your measurement is drawn randomly from the distribution, then the area under the curve indicates the probability of getting a measurement over the specified region



DISTRIBUTIONS

- There are two **distributions** that you have to consider. One when the signal/effect is present and one when it is not:
  - Did you skip lunch at least one time last month?: strength of familiarity when you did skip lunch **and** strength of familiarity when you did not skip lunch
  - Is that your phone ringing?: similarity to your ringtone when it is your phone **and** similarity to your ringtone when it is not your phone
  - Does zinc shorten a cold?: duration of a cold when zinc works **and** duration of a cold when zinc does not work
  - Are men taller than women?: height difference when men are taller **and** height difference when men are the same height as women

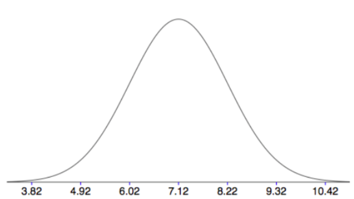
DISCRIMINATION

- To make a decision, you are trying to determine
  - whether your measurement was randomly drawn from a distribution where the signal/effect is present
  - whether your measurement was randomly drawn from a distribution where the signal/effect is not present

ZINC AND COLDS

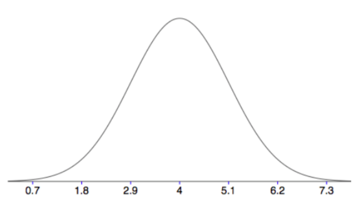
- Based on published research, if you do not take zinc tablets, the duration (in days) of a cold follows a normal distribution with

$$\mu = 7.12,\ \sigma = 1.1$$



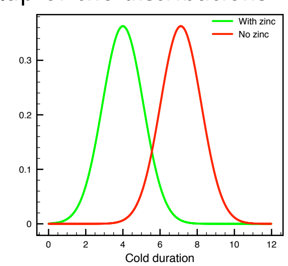
- If you take zinc tablets, the duration (in days) of a cold follows a normal distribution with

$$\mu = 4.00,\ \sigma = 1.1$$



ZINC AND COLDS

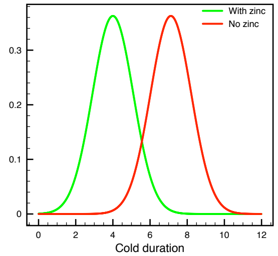
- Together, some overlap of the distributions



- Suppose you sample a person who has a cold and find the duration. Using just that information, you want to decide whether the person took zinc or not.
- Easy cases:
  - X=10
  - X=15
  - X=2
  - X=0.5

ZINC AND COLDS

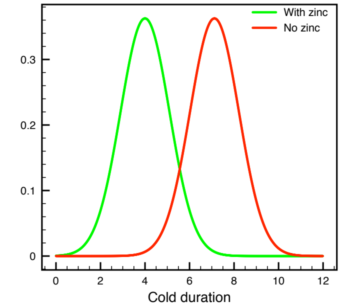
- Together, some overlap of the distributions



- Suppose you sample a person who has a cold and find the duration. Using just that information, you want to decide whether the person took zinc or not.
- Hard cases:
  - X=6
  - X=5

ZINC AND COLDS

- Together, some overlap of the distributions



- We want to quantify how *different* the distributions are
- How much they do **not** overlap
- Signal-to-noise ratio
- (it's a z-score!)

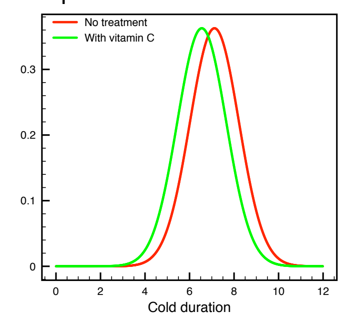
d-prime

- We take the mean of the "no zinc" distribution (noise alone) and compute distance to the mean of the "with zinc" distribution
- in standardized units

$$d' = \frac{\mu_{NZ} - \mu_{WZ}}{\sigma} = \frac{7.12 - 4.00}{1.1} = 2.02$$

VITAMIN C AND COLDS

- Together, lots of overlap of the distributions



- We take the mean of the "no treatment" distribution (noise alone) and compute distance to the mean of the "with vitamin C" distribution
- in standardized units

$$d' = \frac{\mu_{NT} - \mu_{WC}}{\sigma} = \frac{7.12 - 6.55}{1.1} = 0.52$$

DISCRIMINATION

- It is often easy to identify which distribution a measurement came from if d' is big
  - big difference in means, relative to the standard deviation
- It is often hard to identify which distribution a measurement came from if d' is small
  - small difference in means, relative to the standard deviation

DISCRIMINATION

- the same issues apply for lots of situations
- Suppose you are walking your dog who yelps in pain and runs to you
- You think he might have been bitten by a snake
- you have a “measure” of snake-bite evidence (bump on nose, paws are shaking,...)
- you want to determine whether your dog was bitten by a snake

DISCRIMINATION

- is your measurement a random sample from a distribution where your dog was bitten by a snake?
- or
- is your measurement a random sample from a distribution where your dog was not bitten by a snake?
- the separation of the distributions indicates whether the discrimination will be easy or hard
- actually describing the means and standard deviations of these distributions might be challenging!

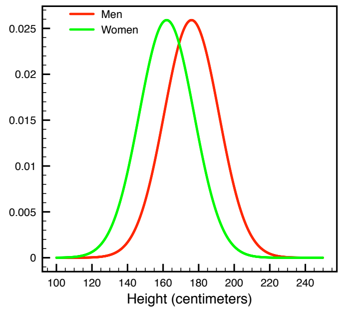
BAD NEWS

- For lots of situations, the d' value is quite small
- Within psychology, some rules of thumb are:
  - d'=0.2 is considered a “small” effect
  - d'=0.5 is considered a “medium” effect
  - d'=0.8 is considered a “large” effect

BAD NEWS

- For lots of situations, the d' value is quite small
- The difference of heights between men and women is roughly

$$d' = \frac{176 - 162}{15.4} = 0.90$$



CONCLUSIONS

- signal-to-noise ratio
- standard score
- d'
- Separation of distributions
- discrimination

NEXT TIME

- Making decisions
- Criterion

Making decisions.

PSY 201: Statistics in Psychology

Lecture 15
Signal detection
Making decisions.

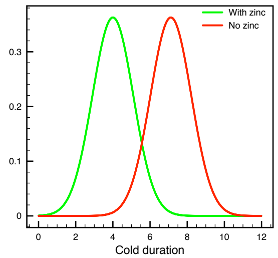
Greg Francis

Purdue University

Fall 2023

ZINC AND COLDS

- Distributions of cold duration when taking zinc or not taking zinc overlap somewhat



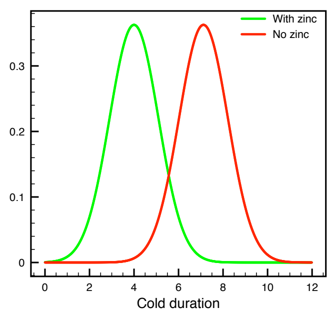
$$d' = \frac{\mu_{NZ} - \mu_{WZ}}{\sigma} = \frac{7.12 - 4.00}{1.1} = 2.02$$

ZINC AND COLDS

- Suppose you sample a person who has a cold and find the duration.
- Using just that information, you want to decide whether the person took zinc or not (e.g., you advised your friend to take the zinc, but he bought a generic version on the Internet and you suspect the tablets do not actually contain zinc).
  - If the cold duration is long, you conclude the tablets do not contain zinc
  - If the cold duration is short, you conclude the tablets do contain zinc

ZINC AND COLDS

- Distributions of cold duration when taking zinc or not taking zinc overlap somewhat



- We want to define a *decision criterion* to separate short and long cold durations
- Suppose we set our criterion to be

$$C = 4$$

DECISION OUTCOMES

	State of nature	
Decision made	Tablets contain zinc	Tablets do not contain zinc
Decide tablets contain zinc	Hit	False Alarm
Decide tablets do not contain zinc	Miss	Correct Rejection

- When making decisions in noise there is **always** the risk of making errors!
- We want to think about the probability of different outcomes

WITH ZINC

- Suppose the tablets really do contain zinc, then when you make a decision you either make:
  - Hit (if you decide the tablets contain zinc)
  - Miss (if you decide the tablets do not contain zinc)
- We know $\mu_{WZ} = 4$ and $\sigma = 1.1$. If we use a criterion of $C = 4$, how often do we make hits and misses?
- (Use the on-line calculator)
  - Hit: P(decide contains zinc — tablet contains zinc) = 0.5
  - Miss: P(decide no zinc — tablet contains zinc) = 0.5

NO ZINC

- Suppose the tablets really do not contain zinc, then when you make a decision you either make:
  - False Alarm (if you decide the tablets contain zinc)
  - Correct Rejection (if you decide the tablets do not contain zinc)
- We know $\mu_{NZ} = 7.12$ and $\sigma = 1.1$. If we use a criterion of $C = 4$, how often do we make false alarms and correct rejections?
- (Use the on-line calculator)
  - False Alarm: P(decide contains zinc — tablet has no zinc) = 0.0023
  - Correct Rejection: P(decide no zinc — tablet has no zinc) = 0.9977

DECISION OUTCOMES

$$P(\text{correct decision}) = P(\text{decide contains zinc}|\text{tablet contains zinc}) \times P(\text{tablet contains zinc}) +$$
$$P(\text{decide no zinc}|\text{tablet has no zinc}) \times P(\text{tablet has no zinc})$$

- If it is equally likely that the tablets contain zinc or do not contain zinc, then the probability that you make a correct decision is:

$$0.5 \times 0.5 + 0.9977 \times 0.5 = 0.74885$$

DIFFERENT CRITERION

- Suppose the tablets really do contain zinc; we know $\mu_{WZ} = 4$ and $\sigma = 1.1$. If we use a criterion of $C = 5$, how often do we make hits and misses?
  - Hit: $P(\text{decide contains zinc}|\text{tablet contains zinc}) = 0.8183$
  - Miss: $P(\text{decide no zinc}|\text{tablet contains zinc}) = 0.1817$
- Suppose the tablets really do not contain zinc; we know $\mu_{NZ} = 7.12$ and $\sigma = 1.1$. If we use a criterion of $C = 5$, how often do we make false alarms and correct rejections?
  - False Alarm: P(decide contains zinc — tablet has no zinc) = 0.027
  - Correct Rejection: P(decide no zinc — tablet has no zinc) = 0.973

DECISION OUTCOMES

$$P(\text{correct decision}) = P(\text{decide contains zinc}|\text{tablet contains zinc}) \times P(\text{tablet contains zinc}) +$$
$$P(\text{decide no zinc}|\text{tablet has no zinc}) \times P(\text{tablet has no zinc})$$

- If it is equally likely that the tablets contain zinc or do not contain zinc, then the probability that you make a correct decision is:

$$0.8183 \times 0.5 + 0.973 \times 0.5 = 0.89565$$

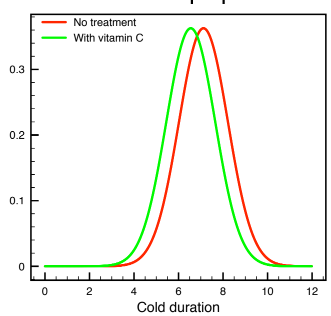
- Using $C = 5$ produces better outcomes (more likely to make the right decision) than using $C = 4$.
- What would be the **optimal** criterion?

TRADE OFFS

- Setting the decision criterion always involves trade offs. In our situation of cold durations and zinc in tablets:
  - Increasing $C \rightarrow$ more hits, more false alarms
  - Deceasing $C \rightarrow$ more misses, more correct rejections
- You generally cannot avoid some errors when making decisions under noisy situations

OVERLAP

- For vitamin C, the durations overlap quite a bit



- We take the mean of the "no treatment" distribution (noise alone) and compute distance to the mean of the "with vitamin C" distribution
- in standardized units

$$d' = \frac{\mu_{NT} - \mu_{WC}}{\sigma} = \frac{7.12 - 6.55}{1.1} = 0.52$$

OVERLAP

- Suppose the tablets really do contain vitamin C; we know $\mu_{WC} = 6.55$ and $\sigma = 1.1$. If we use a criterion of $C = 5$, how often do we make hits and misses?
  - Hit: $P(\text{decide contains vitamin C}|\text{tablet contains vitamin C}) = 0.0794$
  - Miss: $P(\text{decide no vitamin C}|\text{tablet contains vitamin C}) = 0.9206$
- Suppose the tablets really do not contain vitamin C; we know $\mu_{NT} = 7.12$ and $\sigma = 1.1$. If we use a criterion of $C = 5$, how often do we make false alarms and correct rejections?
  - False Alarm: $P(\text{decide contains vitamin C}|\text{tablet has no vitamin C}) = 0.027$
  - Correct Rejection: $P(\text{decide no vitamin C}|\text{tablet has no vitamin C}) = 0.973$

DECISION OUTCOMES

$$P(\text{correct decision}) =$$
$$P(\text{decide contains vitamin C}|\text{tablet contains vitamin C}) \times P(\text{tablet contains vitamin C}) +$$
$$P(\text{decide no vitamin C}|\text{tablet has no vitamin C}) \times P(\text{tablet has no vitamin C})$$

- If it is equally likely that the tablets contain vitamin C or do not contain vitamin C, then the probability that you make a correct decision is:
$$0.0794 \times 0.5 + 0.973 \times 0.5 = 0.5262$$
- Not much better than a random guess!

OVERLAP

- Suppose the tablets really do contain vitamin C; we know $\mu_{WC} = 6.55$ and $\sigma = 1.1$. If we use a criterion of $C = 6.835$ (optimal), how often do we make hits and misses?
  - Hit: $P(\text{decide contains vitamin C}|\text{tablet contains vitamin C}) = 0.6022$
  - Miss: $P(\text{decide no vitamin C}|\text{tablet contains vitamin C}) = 0.39778$
- Suppose the tablets really do not contain vitamin C; we know $\mu_{NT} = 7.12$ and $\sigma = 1.1$. If we use a criterion of $C = 6.835$, how often do we make false alarms and correct rejections?
  - False Alarm: $P(\text{decide contains vitamin C}|\text{tablet has no vitamin C}) = 0.39778$
  - Correct Rejection: $P(\text{decide no vitamin C}|\text{tablet has no vitamin C}) = 0.6022$

DECISION OUTCOMES

$$P(\text{correct decision}) =$$
$$P(\text{decide contains vitamin C}|\text{tablet contains vitamin C}) \times P(\text{tablet contains vitamin C}) +$$
$$P(\text{decide no vitamin C}|\text{tablet has no vitamin C}) \times P(\text{tablet has no vitamin C})$$

- If it is equally likely that the tablets contain vitamin C or do not contain vitamin C, then the probability that you make a correct decision is:
$$0.6022 \times 0.5 + 0.6022 \times 0.5 = 0.6022$$
- Not great, but you cannot do better!

CONCLUSIONS

- signal-to-noise ratio
- decision criterion
- decision outcomes
- performance
- trade-offs

NEXT TIME

- Underlying distributions

Can you read my mind?

PSY 201: Statistics in Psychology

Lecture 16
Underlying distributions
Can you read my mind?

Greg Francis

Purdue University

Fall 2023

DISTRIBUTION

- representation of all possible outcomes
- area under the curve represents relative frequency of events
- completely describes an aspect of a situation relative to a particular variable
- often theoretical curves (but not always)

DICE ROLES

Die 1	Die 2	Sum	Difference
1	1	2	0
1	2	3	1
1	3	4	2
1	4	5	3
1	5	6	4
1	6	7	5
2	1	3	1
2	2	4	0
2	3	5	1
2	4	6	2
2	5	7	3
2	6	8	4
3	1	4	2
3	2	5	1
3	3	6	0
3	4	7	1
3	5	8	2
3	6	9	3
4	1	5	3
4	2	6	2
4	3	7	1
4	4	8	0
4	5	9	1
4	6	10	2
5	1	6	4
5	2	7	3
5	3	8	2
5	4	9	1
5	5	10	0
5	6	11	1
6	1	7	5
6	2	8	4
6	3	9	3
6	4	10	2
6	5	11	1
6	6	12	0

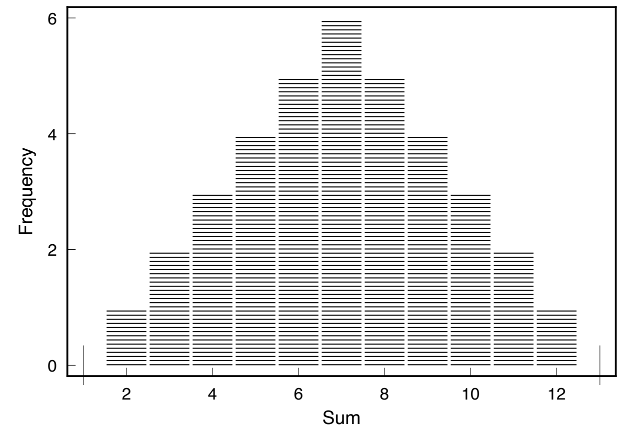
DISTRIBUTION

- we can identify the underlying distribution of the sum of dice variable

Sum	f
1	0
2	1
3	2
4	3
5	4
6	5
7	6
8	5
9	4
10	3
11	2
12	1
13	0

DISTRIBUTION

- same type of stuff we did earlier



- frequency of every possible outcome of the variable Sum

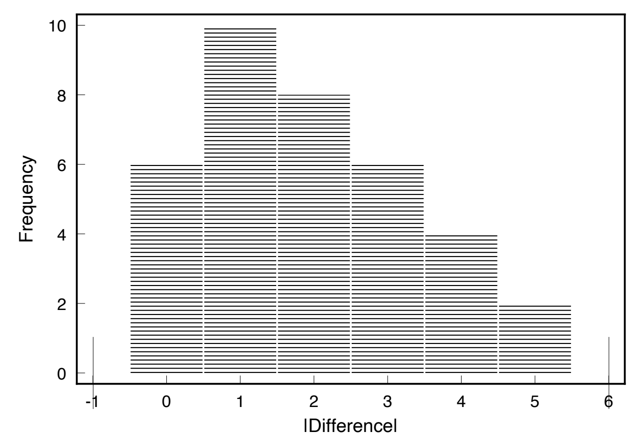
VARIABLE

- a distribution is specific to a variable (x-coordinate)
- suppose instead of the sum of dice roles, we look at the distribution of the absolute value of the difference of dice roles

$\|$Difference$\|$	f
0	6
1	10
2	8
3	6
4	4
5	2
6	0

DISTRIBUTION

- the underlying distribution is different because we are considering a different variable



USE

- once we have the underlying distribution we can calculate probabilities

$$P(A) = \frac{\text{Number of outcomes that include } A}{\text{Total number of possible outcomes}}$$

- you better believe a casino cares about this!
- so does the government
- in practice statisticians generally work with theoretical distributions

BINOMIAL DISTRIBUTION

- suppose you have a situation where there are only two possible outcomes from an action
- e.g., flip a coin: H or T
- each flip is **independent** of the other flips
- how many H's do you get if you flip the coin over and over (or flip many identical coins at once)?

BINOMIAL DISTRIBUTION

- suppose you flip the coin twice
- the possible outcomes are

First coin	Second Coin	Number H
H	H	2
H	T	1
T	H	1
T	T	0

- can produce a frequency distribution table

Number H's	f
0	1
1	2
2	1

BINOMIAL

- from Webster
  1. a mathematical expression consisting of two terms connected by a plus sign or minus sign
  2. a biological species name consisting of two terms

$$(H+T)$$

- to find out how many H's and how many T's, for two coin flips, square the binomial

$$(H+T)^2 = H^2 + 2HT + T^2$$

- or

$$(H+T)^2 = HH + 2HT + TT$$

- coefficient in front identifies how many of each combination

BINOMIAL DISTRIBUTION

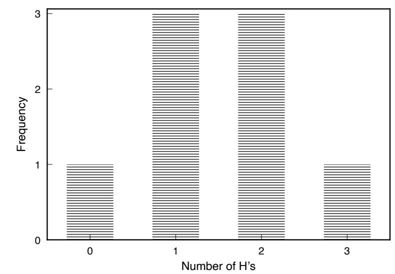
- suppose you flip the coin thrice
- the possible outcomes are

First coin	Second Coin	Third Coin	Number H
H	H	H	3
H	T	H	2
T	H	H	2
T	T	H	1
H	H	T	2
H	T	T	1
T	H	T	1
T	T	T	0

BINOMIAL DISTRIBUTION

- can produce a frequency distribution table

Number H's	f
0	1
1	3
2	3
3	1



BINOMIAL

- for three flips, cube the binomial

$$(H+T)^3 = H^3 + 3H^2T + 3HT^2 + T^3$$

- or

$$(H+T)^3 = HHH + 3HHT + 3HTT + TTT$$

- coefficient of each term indicates number of occurrences!
- this approach works in general (combinatorics)

BINOMIAL DISTRIBUTION

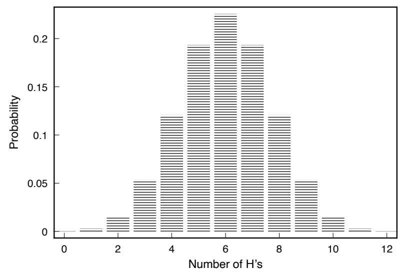
- in turns out that for m flips of the coin the **probability** of getting x number of H's is

$$P(x \text{ number of H's}) = \frac{m!}{x!(m-x)!}(0.5)^m$$

- where $x! = (x)(x-1)(x-2)...(2)(1)$ is "x-factorial" (don't worry about it)
- works for probabilities other than 0.5 too (slightly more complicated)
- Your textbook provides an on-line calculator for any probability

BINOMIAL AND NORMAL

- for $m = 12$



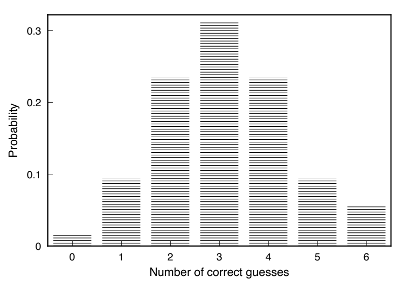
- looks a lot like a normal distribution
- for $m > 20$, the difference is very small

USE

- suppose you have a friend who always drinks Sprite, claiming it is better than 7-Up
- you test your friend's ability to distinguish between Sprite and 7-Up
- Your friend sips two glasses of soda, one containing Sprite and the other 7-Up. Your friend must decide which is the Sprite. You do this 6 times. (Glasses are identical, randomized for tasting first,...)
- Your friend identifies the glass containing Sprite every time. Now you need to decide if your friend really knows his stuff or is just lucky.

USE

- You need to know the probability of **guessing** 6 correct identifications out of 6 trials
- binomial distribution gives us exactly what we want to know



USE

- the probability of guessing correctly 6 out of 6 times is very small (0.0156)
- most likely your friend can tell Sprite from 7-Up
- we will be using distributions like this a lot!
- compare performance to guessing performance
- performance we get from experimentation
- guessing performance we get through tables and calculations (can be complicated)

MIND READING?

- Suppose I flip a coin and look at the upward side.
- Can people read my mind?
- Suppose we took 10 people and asked them to guess which side I saw.
- Some will guess correctly, just by luck.
- How often must people guess correctly before we decide they can read my mind?

MIND READING?

- Each person guessing has a 1 in 2 chance of being correct. So if each person was guessing, how many would we expect to guess correctly?
- What is the probability for each number of guessing correctly?: its a binomial distribution
- can produce a table of probabilities
- It would be surprising (rare) if people were correct 8, 9, or 10 times out of 10.

Number correct	p
0	0.0010
1	0.0098
2	0.0439
3	0.1172
4	0.2051
5	0.2461
6	0.2051
7	0.1172
8	0.0439
9	0.0098
10	0.0010

MIND READING?

- Let's see if people can read my mind:
- measure ability to read my mind
- get number correct
- see if it is "rare enough" for us to conclude they can read minds
  - By using the on-line Binomial distribution calculator

CONCLUSIONS

- underlying distributions
- binomial distribution
- started hypothesis testing

NEXT TIME

- sampling distribution of the mean
- properties of sampling distributions

Marvel at my predictive powers!

PSY 201: Statistics in Psychology

Lecture 17
Sampling distribution of the mean
Marvel at my predictive powers!

Greg Francis

Purdue University

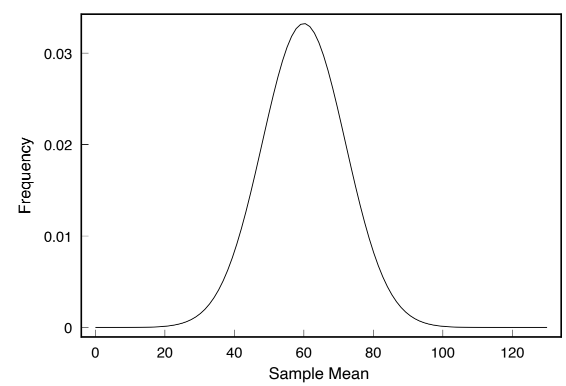
Fall 2023

SAMPLING

- suppose we have a **population** with a mean μ and a standard deviation σ
- suppose we take a **sample** from the population and calculate a sample mean $\overline{X}_1$
- suppose we take a **different** sample from the population and calculate a sample mean $\overline{X}_2$
- suppose we take a **different** sample from the population and calculate a sample mean $\overline{X}_3$

DISTRIBUTION

- the different $\overline{X}_i$ sample means that are calculated will be related to each other because they all come from the same population, which has a population mean of μ
- we can consider a **distribution** of the sample means (same idea as distribution of sum of dice roles)



DISTRIBUTION

- this distribution involves frequencies of **means** rather than frequencies of **scores**
- for most of inferential statistics we do **not** deal with the frequency distribution of scores
- A sampling distribution is the underlying distribution of values of the statistic under consideration, from all possible samples of a given size.
- currently, the statistic is the sample mean $\overline{X}$

SAMPLING DISTRIBUTION

- how do we get the sampling distribution?
- e.g., suppose you have a population of 5 people with math scores
  - and you take sample sizes of 3
- you must consider every possible group of 3 people from the population
  - turns out there are 10 such groups
- NOTE: the number of samples is greater than the size of the population!

CENTRAL LIMIT THEOREM

- fortunately, there are theorems that tell us what the distribution will look like
- as the sample size (n) increases, the sampling distribution of the mean for simple random samples of n cases, taken from a population with a mean equal to μ and a finite variance equal to σ^2, approximates a normal distribution
- another theorem based on unbiased estimation tells us that the mean of the **sampling distribution** is μ

STANDARD ERROR

- theorems on unbiased estimates also give us the sampling distribution variance and standard deviation
- denote the sampling distribution variance as

$$\sigma^2_{\overline{X}}$$

- it turns out that

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$$

- where
  - σ^2 = variance in the population
  - n = size of sample

STANDARD ERROR

- of course the standard deviation of the sampling distribution is the square root of the variance

$$\sigma_{\overline{X}} = \sqrt{\sigma^2_{\overline{X}}}$$

- or

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

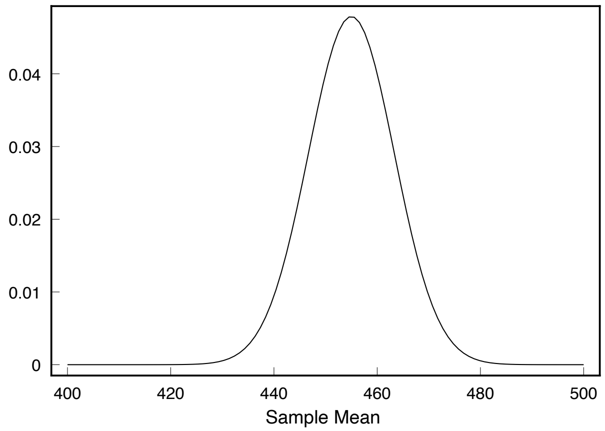
- also called the **standard error of the mean**

WHY BOTHER?

- suppose you know that for a population, $\mu = 455$ and $\sigma = 100$ (an example involving SAT scores)
- then we know the following about a **sampling distribution** involving samples sizes of 144 students
  - The distribution is normal.
  - The mean of the distribution is 455.
  - The standard error of the mean is $100/\sqrt{144} = 8.33$.

WHY BOTHER?

- this is something we can work with!



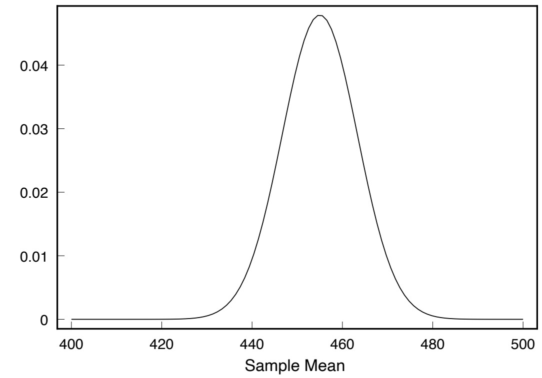
- calculate percentages, proportions, percentile ranks

PROBABILITY

- we can answer questions like
- what is the probability of randomly selecting a sample with a mean $\overline{X}$ such that

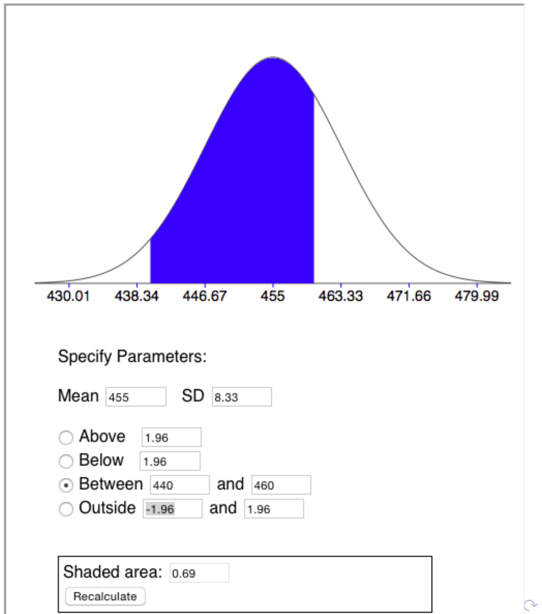
$$440 < \overline{X} < 460 \text{ ?}$$

- area under the curve



PROBABILITY

- everything is just like before
- area under the curve
- We use the normal distribution calculator with Mean=455 and SD=8.33



SAMPLING DISTRIBUTION

- the sampling distribution has two critical properties
  - As sample size (n) increases, the sampling distribution of the mean becomes more like the normal distribution in shape, even when the population distribution is not normal.
  - As the sample size (n) increases, the variability of the sampling distribution of the mean decreases (the standard error decreases).

SHAPE

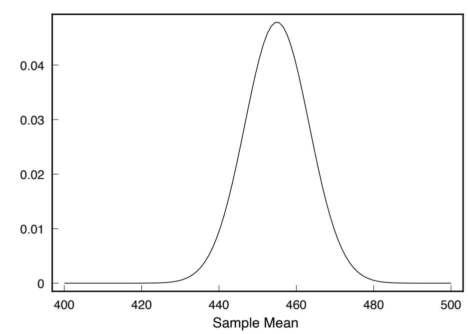
- with large sample sizes, all sampling distributions of the mean look like normal distributions
- means the conclusions we draw from sampling distributions are not dependent on the shape of the population distribution!
- a remarkable result that is due to the central limit theorem

VARIABILITY

- from our calculation of standard error:

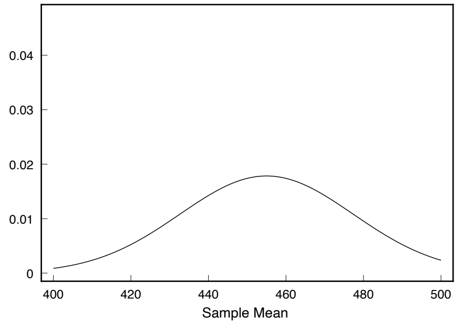
$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

- we see that increasing n makes for smaller values of $\sigma_{\overline{X}}$
- e.g. for $n = 144$ in our previous example $\sigma_{\overline{X}} = 8.33$



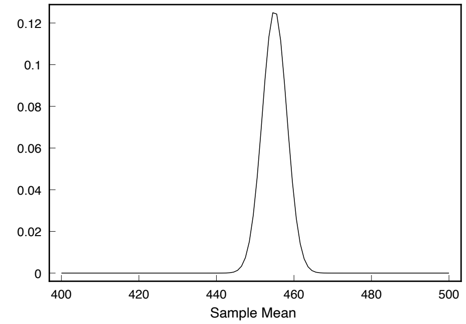
VARIABILITY

- but if $n = 20$,
$$\sigma_{\overline{X}} = \frac{100}{\sqrt{20}} = 22.36$$
- compare to the 8.33 with $n = 144$



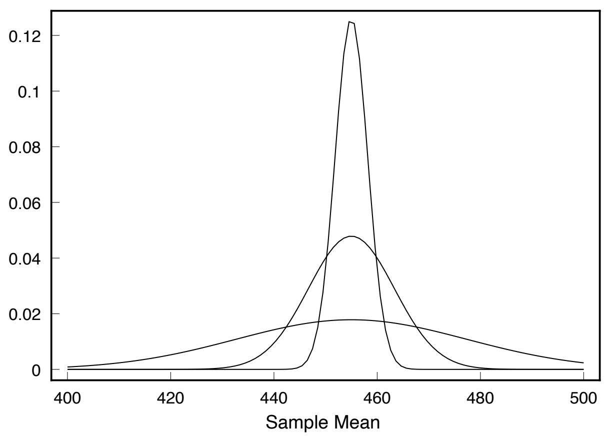
VARIABILITY

- OR if $n = 1000$,
$$\sigma_{\overline{X}} = \frac{100}{\sqrt{1000}} = 3.16$$
- compare to the 8.33 with $n = 144$



VARIABILITY

- increasing the sample size decreases the variability of sample means
- makes sense if you think about it



SAMPLING

- to use the sampling distribution like we want to, we must have **random** samples
- without random sampling, our calculations about probability of sample means are not valid (this will get more important later)
- lots of methods of sampling that emphasize different aspects of the data

WHY STATISTICS WORKS

- we have two ways of finding the sampling distribution of the mean
  - gather lots of samples, calculate means and standard deviations (virtually impossible)
  - estimate the mean and standard deviation of the population, use central limit theorem (relatively easy)
- the central limit theorem allows us to do inferential statistics, without it, much of this course would not exist (actually there is one other way to do statistics...)

EXAMPLE

- let's create a sampling distribution
- two things
  - Write down the height of your father (in inches) on the papers going around the room.
  - Sample the height measure of 10 people close to you.

EXAMPLE

- I'll sit down and calculate the **population** mean (μ) and standard deviation (σ)
- you calculate the **sample** mean ($\overline{X}$) for the 10 scores you have

$$\overline{X} = \frac{\Sigma X_i}{10}$$

EXAMPLE

- OK, I get

$$\mu = \frac{\Sigma X_i}{N} =$$

$$\sigma = \sqrt{\frac{\Sigma(X_i)^2 - [(\Sigma X_i)^2 / N]}{N}} =$$

EXAMPLE

- with this information, I can **predict** the frequency of sample means each of you calculated
- I predict that most of you calculated sample means close to

$$\overline{X} = \mu =$$

- moreover, I predict that the distribution of sample means is **normal**
- lets plot the sample means you calculated

EXAMPLE

- Let's calculate the standard deviation of the sampling distribution of the mean heights as

$$\sigma_{\overline{X}} = \sqrt{\frac{\Sigma(\overline{X})^2 - [(\Sigma\overline{X})^2 / N]}{N}} =$$

- I predict that it will be very close to

$$\frac{\sigma}{\sqrt{10}} =$$

CONCLUSIONS

- sampling distribution of the mean looks like a normal distribution
- methods of calculating mean and standard deviation **if** μ and σ are known
- samples must be randomly selected

NEXT TIME

- hypothesis testing
- using the sampling distribution (in what looks to be reverse!)
- null hypothesis

Why I don't use herbal medicines.

PSY 201: Statistics in Psychology

Lecture 18

Hypothesis testing of the mean

Why I don't use herbal medicines.

Greg Francis

Purdue University

Fall 2023

SUPPOSE

- we think the mean value of a population of SAT scores is $\mu = 455$
- we can take a sample of the population and calculate the sample mean of SAT scores $\overline{X} = 535$
- we can make some statement about how rare it is to get a result like $\overline{X} = 535$ (what we did last time)
- **and** if such a result is very rare
- we can make a statement about how unreasonable it is that our original thought is true!

HYPOTHESIS TESTING

- in hypothesis testing we consider how reasonable a hypothesis is, given the data that we have
- if the hypothesis is reasonable (consistent with the data), we assume it could be true
- if the hypothesis is unreasonable (inconsistent with the data), we assume it is false
- deciding on what hypotheses to test is critically important!

HYPOTHESIS TESTING

- four steps:
  1. State the hypothesis and criterion.
  2. Compute the test statistic.
  3. Compute the p value.
  4. Make a decision.

HYPOTHESIS

- conjecture about one or more population parameters
- e.g.
  - $\mu = 455$
  - $\mu_1 = \mu_2$
  - $\sigma = 3.5$
  - $r = 0.76$
  - ...
- in inferential statistics we always test the **null hypothesis**: H_0

NULL HYPOTHESIS

- H_0 is the assumption of no relationship, or no difference. e.g.
  - H_0: no relationship between variables
  - H_0: no difference between treatment groups
- We want the H_0 to be *specific* so that we can define a sampling distribution
- the alternative hypothesis, H_a is the other possibility. e.g.
  - H_0: $\mu = 455$
  - H_a: $\mu \neq 455$
- does not say what μ is, but says what it is not!

NULL HYPOTHESIS

- what's wrong with herbal medicines?
- nothing necessarily, but I don't know that they are any good (and they may be bad)
- lots of reports that they help people (but how can they be sure)
- need to start by assuming that a medicine does nothing, and **prove** that the assumption is false!
- anecdotal reports are just about worthless

NULL HYPOTHESIS

- often times (almost always) the goal of statistical research is to reject the null hypothesis, so that the only alternative is to accept H_a
- similar to an indirect proof. e.g.
  - show that the angles of a triangle sum to 180° by assuming that they do not and then finding a contradiction
- why this approach?
  - it is much easier to show that something is false (H_0) than to show that something is true (H_a)
- understanding of relationship between variables or differences between groups often requires many experiments!

STATE THE HYPOTHESIS

- before doing anything else, we need to make certain that we understand the tested hypothesis
- for the SAT example

$$H_0 : \mu = 455$$
$$H_a : \mu \neq 455$$

- sometimes this is the most difficult step in designing an experiment
- to start, we will worry only about hypotheses about the population mean, μ

SIGNAL DETECTION

- The task is almost the same as deciding whether a measurement came from a noise-alone (null hypothesis) distribution or a signal-and-noise (alternative hypothesis) distribution
- How well you can do is determined by the signal-to-nose ratio (d'), but that value is typically unknown
- we set a criterion using only the null hypothesis (noise-alone distribution)

CRITERION

- we will examine the data to see if we should reject H_0
- we will do that by comparing the sample mean, $\overline{X}$, to the hypothesized value of the population mean, μ
- the bottom-line is whether $\overline{X}$ is sufficiently different from μ to reject H_0
- but we have to consider four things to quantify the term *sufficiently different*
  - standard scores
  - errors in hypothesis testing
  - level of significance
  - region of rejection

STANDARD SCORES

- we previously used standard scores to indicate how much a given score deviates from a distribution mean
- We do the same kind of thing here, but we want to know how a sample mean, $\overline{X}$ deviates from what the sampling distribution would be if the null hypothesis is true
- We give the standard score a special term:

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}}$$

- We compute everything else using the sampling distribution of this t value: the t distribution, which is similar to a normal distribution with fatter tails and requires degrees of freedom:

$$df = n - 1$$

DECISIONS

- after deciding to reject or not reject H_0 there are four possible situations
  - A true null hypothesis is rejected. (False alarm)
  - ** A true null hypothesis is not rejected. (Correct rejection)
  - A false null hypothesis is not rejected. (Miss)
  - ** A false null hypothesis is rejected. (Hit)
- errors are unavoidable
- we want to minimize the probability of making errors, given the particular data set we have

ERRORS

- two types of errors:
  - **Type I error**: when we reject a true null hypothesis (false alarm).
  - **Type II error**: when we do not reject a false null hypothesis (miss).

	State of nature	
Decision made	H_0 true	H_0 false
Reject H_0	Type I error	Correct decision
Do not reject H_0	Correct decision	Type II error

- generally, decreasing the probability of making one type of error increases the probability of making the other type of error

ERRORS

- suppose you have a new, untested, and expensive treatment for cancer
- you run a test to judge whether the drug is better than existing drugs
- if you reject H_0, indicating that the drug **is** more effective, when in fact it is not, people will spend a lot of money for no reason (Type I error)
- if you fail to reject H_0, indicating that the drug is not effective, when in fact it is, people will not use the drug (Type II error)
- scientific research tends to focus on avoiding Type I errors

SIGNIFICANCE LEVEL

- alpha (α) level
- indicates probability of Type I error
- frequently we choose $\alpha = 0.05$ or $\alpha = 0.01$
- that is, the corresponding decision to reject H_0 may produce a Type I error 5% or 1% of the time
- a statement about how much error we will accept
- usually chosen **before** the data is gathered
  depends upon use of the analysis

REGION OF REJECTION

- α is a probability
- it identifies how much risk of Type I error we are willing to take (rejecting H_0 when it is true)
- consider our example of SAT scores

$$H_0 : \mu = 455$$

- suppose we also know the sample standard deviation

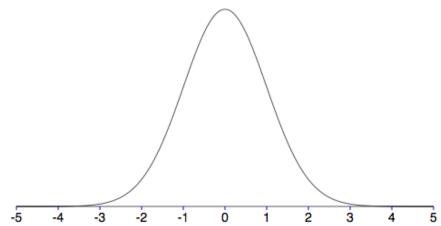
$$s = 100$$

- and our sample size is $n = 144$

REGION OF REJECTION

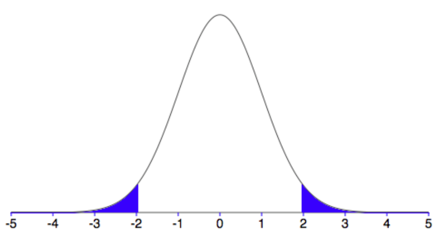
- we know that the sampling distribution of t is:
  - A t distribution with $df = n - 1 = 143$.
  - Has a mean of $\mu = 0$, if H_0 is true
  - Has a standard error of the mean

$$s_{\overline{X}} = \frac{s}{\sqrt{n}} = \frac{100}{\sqrt{144}} = 8.33$$



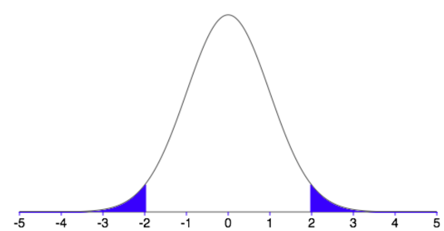
REGION OF REJECTION

- area under the curve represents the probability of getting the corresponding t values, if the H_0 is true
- the extreme tails of the sampling distribution correspond to what should be very rare t values, and thus very rare sample means



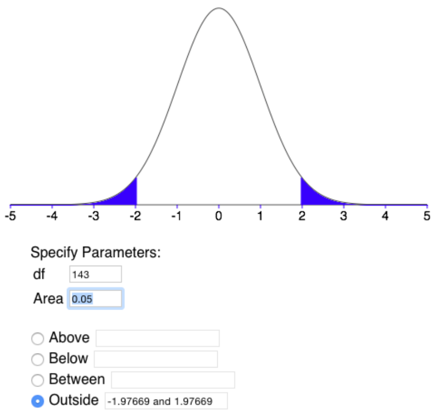
REGION OF REJECTION

- we shade in the extreme α percentage of the sampling distribution
- called the region of rejection
- if our data produces a t value in the region of rejection, we reject H_0 because it is unlikely that we would get such a value if the H_0 were true.



REGION OF REJECTION

- values of sample means at the beginning of the region of rejection
- NOTE: α is split up in each tail
- called a two-tailed or non-directional test



TEST STATISTIC

- if the t-score is beyond ± 1.977, it is very unlikely to have occurred if the H_0 is true.
- we have the following data:
  - $\mu = 455$, H_0
  - $n = 144$, sample size
  - $\overline{X} = 535$, observed value for sample statistic
  - $s = 100$, value of the standard deviation of the population
  - $s_{\overline{X}} = 8.33$, standard error (calculated earlier)
- from this we can calculate the t-score

TEST STATISTIC

- we want to know how different $\overline{X}$ is from the hypothesized μ in terms of standard error units

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}}$$

$$t = \frac{535 - 455}{8.33} = 9.60$$

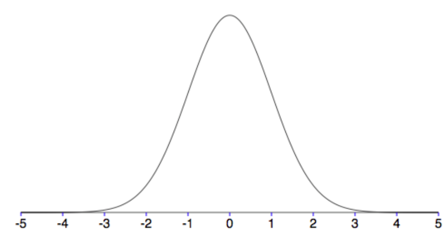
- the standard score is the **test statistic** for testing H_0 about a population mean

DECIDING ABOUT H_0

- compare the test statistic to the critical value

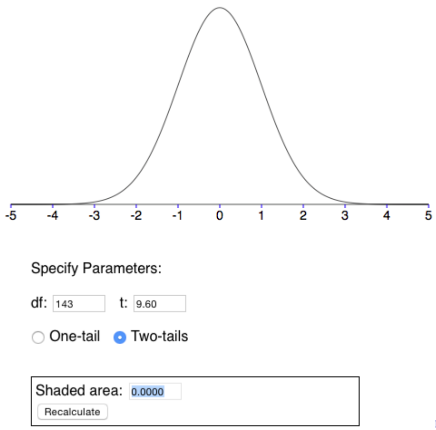
$$t = 9.60 > 1.977 = t_{cv}$$

- indicates that the sample mean $\overline{X}$ is extremely rare, given the assumed population mean μ, by chance (random sampling)



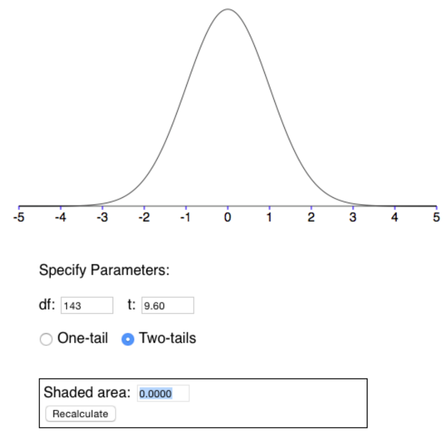
p-VALUE

- another way to do it (advocated by your text) is to use the t-value to compute the probability of getting a t-value more extreme than what you found
- p-value
- t distribution calculator



p-VALUE

- We find $p \approx 0$
- Since the probability is small ($< .05$), then we conclude that the H_0 is probably not true



DECISIONS

- since the p value is smaller than the α we set, we reject

$$H_0 : \mu = 455$$

- in favor of the alternative hypothesis

$$H_a : \mu \neq 455$$

- but there is still a chance that H_0 is true!

CONCLUSIONS

- null hypothesis
- rejecting H_0
- Type I error
- Type II error

NEXT TIME

- Test statistic
- Deciding about H_0

Why clinical studies use thousands of subjects.

PSY 201: Statistics in Psychology

Lecture 19

Hypothesis testing of the mean

Why clinical studies use thousands of subjects.

Greg Francis

Purdue University

Fall 2023

SUPPOSE

- we think the mean value of a population of SAT scores is $\mu = 455$
- we can take a sample of $n = 144$ from the population and calculate the sample mean of SAT scores $\overline{X} = 535$ with sample standard deviation $s = 100$

HYPOTHESIS TESTING

- four steps
  1. State the hypothesis and criterion.
  2. Compute the test statistic.
  3. Compute the p value
  4. Make a decision.

RECAP OF LAST TIME

- (1) State the hypotheses and set the criterion

$$H_0 : \mu = 455$$
$$H_a : \mu \neq 455$$

- $\alpha = 0.05$

RECAP OF LAST TIME

- (2) Compute the test statistic

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}}$$
$$t = \frac{535 - 455}{8.33} = 9.60$$

- (3) Compute the p-value (using the t-distribution calculator with $df = n - 1$):

$$p \approx 0$$

- (4) Make a decision: $p < \alpha$, so reject H_0
  - the found sample mean would be a very rare event if H_0 were true

DIFFERENT MEAN

- suppose we had the same situation as before, but we had instead found

$$\overline{X} = 465$$

- (1) State the hypotheses and set the criterion (unchanged!)

$$H_0 : \mu = 455$$
$$H_a : \mu \neq 455$$

- $\alpha = 0.05$
- (2) Compute the test statistic

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}}$$
$$t = \frac{465 - 455}{8.33} = 1.20$$

DIFFERENT MEAN

- (3) Compute the p-value (using the t-distribution calculator with $df = n - 1$):

$$p = 0.2301$$

- (4) Make a decision: $p > \alpha$, so do **not** reject H_0
- the found sample mean would not be very rare if H_0 were true
  - if the null hypothesis is true, then the probability that $|\overline{X}| \geq 465$ would be found by random sampling is greater than .05

SAMPLE SIZE

- suppose we had the same situation as before, but we had instead found

$$\overline{X} = 465$$

- with a sample size of $n = 500$
- (1) State the hypotheses and set the criterion

$$H_0 : \mu = 455$$
$$H_a : \mu \neq 455$$

- $\alpha = 0.05$

SAMPLE SIZE

- (2) Compute the test statistic

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}}$$

- we need to recompute $s_{\overline{X}}$

$$s_{\overline{X}} = \frac{s}{\sqrt{n}} = \frac{100}{\sqrt{500}} = 4.47$$

$$t = \frac{465 - 455}{4.47} = 2.24$$

SAMPLE SIZE

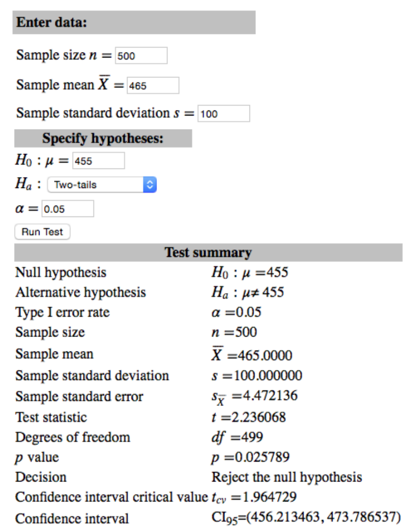
- (3) Compute the p-value (using the t-distribution calculator with $df = n - 1 = 499$):

$$p = 0.0251$$

- (4) Make a decision: $p < \alpha$, so do reject H_0
  - the found sample mean would be a rare event if H_0 were true. The probability that $|\overline{X}| \geq 465$ would be found by random sampling is less than .05

CALCULATOR

- you need to understand the math and calculations, but generally you should not **do** it



Enter data:

Sample size $n =$ 500

Sample mean $\overline{X} =$ 465

Sample standard deviation $s =$ 100

Specify hypotheses:

$H_0 : \mu =$ 455

$H_a :$ Two-tails

$\alpha =$ 0.05

Run Test

Test summary	
Null hypothesis	$H_0 : \mu = 455$
Alternative hypothesis	$H_a : \mu \neq 455$
Type I error rate	$\alpha = 0.05$
Sample size	$n = 500$
Sample mean	$\overline{X} = 465.0000$
Sample standard deviation	$s = 100.000000$
Sample standard error	$s_{\overline{X}} = 4.472136$
Test statistic	$t = 2.236068$
Degrees of freedom	$df = 499$
p value	$p = 0.025789$
Decision	Reject the null hypothesis
Confidence interval critical value	$t_{cv} = 1.964729$
Confidence interval	$CI_{95} = (456.213463, 473.786537)$

CLINICAL TRIALS

- often hear about medical studies that track thousands of patients
- why do they need so many people?
- a larger sample makes for less variation in the sampling distribution of the mean

$$s_{\overline{X}} = \frac{s}{\sqrt{n}}$$

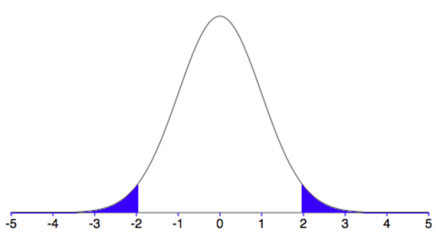
- thus, if the null hypothesis really is false, you are more likely to reject it with a larger sample
- if the null hypothesis is really true, you are not more likely to reject it (no extra mistakes with a larger sample size!)

COMMENTS

- several things are worth noting
  - The α probability is about the *process* of making decisions. It controls Type I error rates, but for any given decision you do not know if you made an error or not.
  - Even when we reject H_0, there is always a chance that it is true.
  - Even when we do not reject H_0, there is always a chance that it is false.
  - The statement $p < 0.05$ is about the **statistic** given the hypothesis, not about the hypothesis. We never conclude that H_0 is false with probability 0.95.
  - Technically, we have done all of this before.
  - These techniques are quantifiable.
  - No inclusion of knowledge about the direction of difference.

DIRECTIONAL HYPOTHESIS

- we choose a significance level, α
- indicates probability of Type I error
- earlier, we split this error across the two tails of the sampling distribution

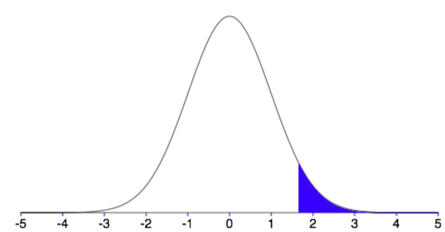


DIRECTIONAL HYPOTHESIS

- suppose we know (or strongly suspect) that if the sample mean $\overline{X}$ is different from the population mean μ, it will be **greater**
- then we don't need to worry about the left-side tail

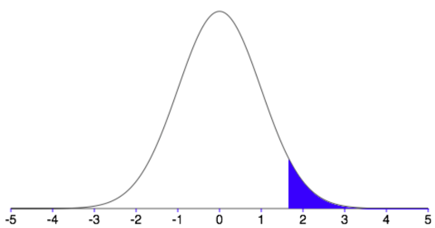
$$H_0 : \mu = 455$$

$$H_a : \mu > 455$$



REGION OF REJECTION

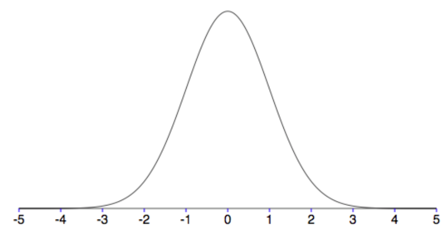
- if we only have to worry about one tail, the region of rejection (in that tail) is larger!
- with $df = 143$, last 5% starts with a t-score of 1.656
- we can reject H_0 when the difference between $\overline{X}$ and μ is smaller!



EXAMPLE

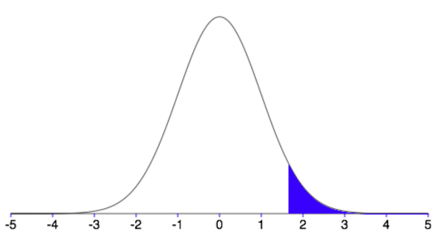
- we know that the sampling distribution of t is:
  - A t distribution with $df = 143$.
  - Has a mean of $\mu = 0$.
  - Has a standard error of the mean

$$s_{\overline{X}} = \frac{s}{\sqrt{n}} = \frac{100}{\sqrt{144}} = 8.33$$



REGION OF REJECTION

- area under the curve represents the probability of getting the corresponding t values, given that H_0 is true
- the extreme right tail of the sampling distribution corresponds to what should be very rare t values
- critical t-score value is 1.656



TEST STATISTICS

- we compute test statistic

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}}$$

$$t = \frac{535 - 455}{8.33} = 9.60$$

- greater than critical value

$$9.60 > 1.656$$

- reject H_0
- The same decision is found by computing the p-value

$$p \approx 0 < \alpha = 0.05$$

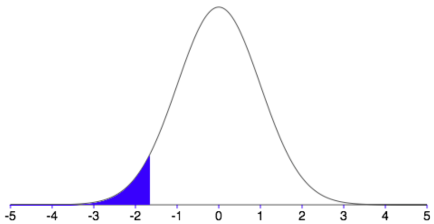
EXAMPLE

- suppose everything was the same, except we had hypotheses:

$$H_0 : \mu = 455$$

$$H_a : \mu < 455$$

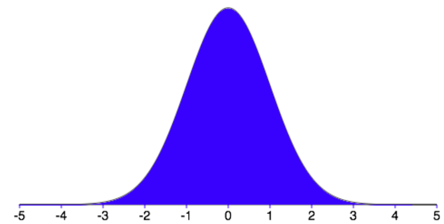
- then we would shift the region of rejection to the **left** tail



EXAMPLE

- the critical t-score value becomes -1.656
- with our sample mean of $\overline{X} = 535$, and $z = 9.60$,
- we **cannot** reject H_0

$$p \approx 1 > \alpha = 0.05$$



CALCULATOR

- you need to understand the math and calculations, but generally you should not **do** it

Enter data:

Sample size n = 144

Sample mean $\overline{X}$ = 535

Sample standard deviation s = 100

Specify hypotheses:

$H_0 : \mu$ = 455

H_a : Negative one-tail

α = 0.05

Run Test

Test summary	
Null hypothesis	$H_0 : \mu = 455$
Alternative hypothesis	$H_a : \mu < 455$
Type I error rate	$\alpha = 0.05$
Sample size	$n = 144$
Sample mean	$\overline{X} = 535.0000$
Sample standard deviation	$s = 100.000000$
Sample standard error	$s_{\overline{X}} = 8.333333$
Test statistic	$t = 9.600000$
Degrees of freedom	$df = 143$
p value	$p = 1.000000$
Decision	Do not the reject null hypothesis
Confidence interval critical value	$t_{cv} = 1.976692$
Confidence interval	$CI_{95} = (518.527565, 551.472435)$

CONCLUSIONS

- hypothesis testing
- sample size
- directional test

NEXT TIME

- Designing experiments
- Power
- Selecting sample size

Plan ahead!

PSY 201: Statistics in Psychology

Lecture 20
Power
Plan ahead!

Greg Francis

Purdue University

Fall 2023

BREAKFAST

- Consider an example from the text:
- A runner typically does not have breakfast before going on a 5K run. She wonders if eating breakfast before the run would influence her running time.
- She wants to design an experiment to test whether breakfast influences her running time. She knows that without breakfast her mean running time (in minutes) is

$$H_0 : \mu = 22.5$$

- she will test against

$$H_a : \mu \neq 22.5$$

- with $\alpha = 0.05$
- She plans to try running with breakfast and measure her running time, but she needs to know how many how many days she should try this. (What is an appropriate sample size?)

SAMPLE SIZE

- We know that sample size matters for hypothesis testing.
- Standard error gets smaller

$$s_{\overline{X}} = \frac{s}{\sqrt{n}}$$

- and $df = n - 1$ gets bigger which makes for smaller tails in the t distribution
- More data is better, but it comes with a cost.
  - Experiment takes longer.
  - It may be trouble to wake up earlier to have breakfast

EASY APPROACH

- Suppose the runner decides to try running for a week and then run a hypothesis test. She runs every day, so the sample size will be $n = 7$.
- Is this a good strategy? Is the experiment likely to work, even if breakfast does change mean running time?
- There is no way of knowing whether $n = 7$ is a large enough sample; it depends on how much change breakfast causes and how much variability there is in the data

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}}$$

- Of course, before running the experiment, we do not know $\overline{X}$ or s. However, perhaps we can estimate these values or use something meaningful.

ESTIMATE s

- The runner keeps track of her past running times (that's how she knows the mean is $\mu = 22.5$ minutes)
- The same data allows her to compute the standard deviation of her past running times. Let us suppose it is $\sigma = 2.2$ minutes. It seems reasonable to suppose that over the week when she eats breakfast the standard deviation of running times will be about the same. Thus, the standard error of her mean running time for the week with breakfast will be similar to:

$$s_{\overline{X}} = \frac{s}{\sqrt{n}} = \frac{2.2}{\sqrt{7}} = 0.8315$$

- Of course, it will vary from sample to sample.

ESTIMATE $\overline{X}$

- Past data cannot tell us how much breakfast should change running times, but the runner might have some idea of how much *matters* to her
- To motivate her to wake up earlier and have breakfast before running, eating breakfast needs to shorten her mean running time by at least 2 minutes. Thus, she hopes that when eating breakfast her running time measures are from a distribution with $\mu = 20.5$ minutes.
- We set a specific alternative hypothesis:

$$H_a : \mu = 20.5$$

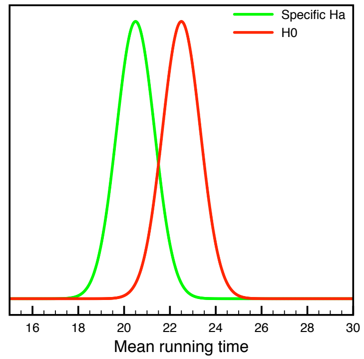
- If she runs the experiment, she will typically get $\overline{X}$ close to 20.5 minutes, but it will vary from sample to sample

SIGNAL DETECTION

- With all this information, we have something similar to signal detection
- Noise-alone distribution is the H_0, breakfast does not affect running times
- Signal-and-noise distribution is the specific H_a, where breakfast reduces running times by 2 minutes
- $\sigma = 2.2$ minutes, so with $n = 7$, $s_{\overline{X}} = 0.8315$
- The hypothesis test procedure establishes criterion t values, if we get data with t bigger than those criterion values, we will reject H_0
- What is the probability we will reject H_0 if the specific H_a is true? This is the "hit"' rate.

SIGNAL DETECTION

- Graphically, if H_0 is true, then we will get sample means from the red distribution
- If the specific H_a is true, then we will get sample means from the green distribution



HYPOTHESIS TEST

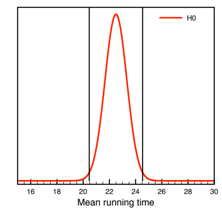
- When doing the hypothesis test, the runner will compute a t-value and see if it is more extreme than

$$t_{cv} = \pm 2.44691$$

- In terms of mean running times, this corresponds to:

$$\overline{X}_{lower} = \mu - s_{\overline{X}} t_{cv} = 22.5 - (0.8315)(2.44691) = 20.465$$

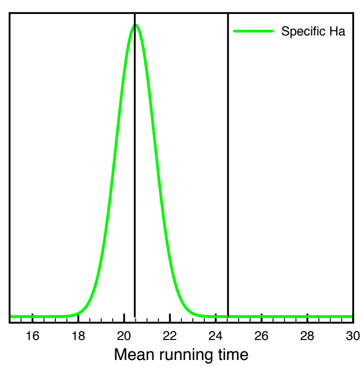
$$\overline{X}_{upper} = \mu + s_{\overline{X}} t_{cv} = 22.5 + (0.8315)(2.44691) = 24.535$$



- Note the region of rejection!

POWER

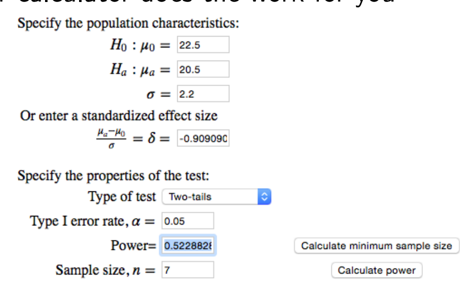
- If the specific H_a is true, then the probability of rejecting the null is the area under the distribution for the specific H_a over the region of rejection



- Looks pretty close to 0.5

CALCULATOR

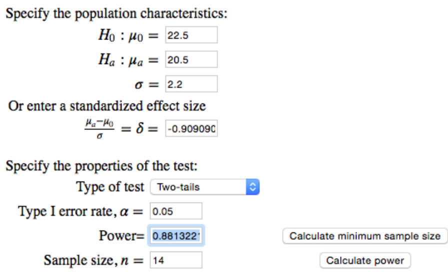
- Online power calculator does the work for you



- Even if the effect exists, the probability that your hypothesis test will show the effect is only 0.52.
- Maybe it is not worth doing the experiment.

INCREASE SAMPLE SIZE

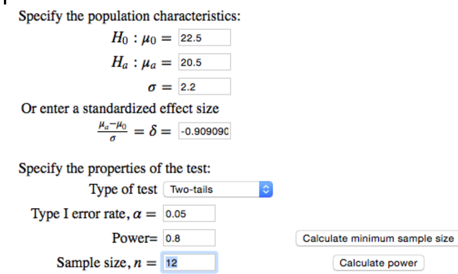
- Or, the runner may decide it is worth trying to run with breakfast for two weeks. Then $n = 14$.
- We use the on-line power calculator to find the probability such an experiment would reject the H_0 if there is an effect.



- Now the probability of finding an effect is 0.88!

FINDING SAMPLE SIZE

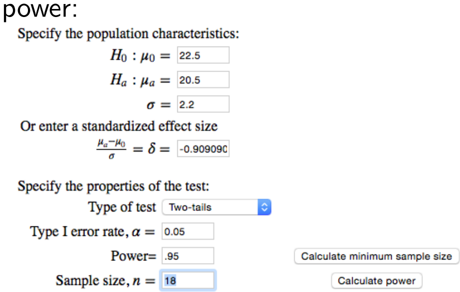
- Perhaps it makes more sense to identify the smallest sample that will give you a desired power.
- We enter the desired power and click on the *Calculate minimum sample size* button.
- To get 80% power:



- the runner needs to gather data over 12 runs

FINDING SAMPLE SIZE

- If you want more power, you have to pay for it with a larger sample size.
- To get 95% power:



- the runner needs to gather data over 18 runs

POWER ESTIMATES

- Note, the power probabilities depend on the effect being as large as what is given. The runner used a "smallest meaningful" effect size for her, $\mu_a = 20.5$, a 2 minute reduction compared to the null hypothesis, $\mu_0 = 22.5$
- The true effect may be larger or smaller than this difference. If the true effect is larger, the experiment will be even more likely to reject H_0 than the estimated power. The experiment may use more resources than is necessary, but it will still work.
- If the true effect is smaller, the experiment will be less likely to reject H_0 than the estimated power. The experiment may not work, but the runner hardly cares because the effect is not big enough for her, anyhow.

SIZE OF EFFECT

- Power increases as the difference between μ_0 and μ_a increases.
  - Bigger signal is easier to detect.
- Suppose the runner used a "smallest meaningful" effect of 3 minutes, so $\mu_a = 19.5$ compared to the null hypothesis, $\mu_0 = 22.5$

Specify the population characteristics:
$H_0 : \mu_0 =$ 22.5
$H_a : \mu_a =$ 19.5
$\sigma =$ 2.2
Or enter a standardized effect size
$\frac{\mu_a - \mu_0}{\sigma} = \delta =$ -1.363636

Specify the properties of the test:
Type of test Two-tails
Type I error rate, $\alpha =$ 0.05
Power= .95 Calculate minimum sample size
Sample size, $n =$ 10 Calculate power

- the runner needs to gather data over 10 runs to have 95% power

DIRECTIONAL HYPOTHESIS

- One-tailed tests are more powerful than two-tailed tests, provided the effect is in the correct tail
- Using $\mu_a = 20.5$ compared to $\mu_0 = 22.5$, the runner might use a one-tailed test when analyzing the data. We noted earlier that $n = 18$ gave us 95% power for a two-tailed test. If she plans to gather that much data but use a (Negative) one-tail test, the power is larger

Specify the population characteristics:
$H_0 : \mu_0 =$ 22.5
$H_a : \mu_a =$ 20.5
$\sigma =$ 2.2
Or enter a standardized effect size
$\frac{\mu_a - \mu_0}{\sigma} = \delta =$ -0.909090

Specify the properties of the test:
Type of test Negative one-tail
Type I error rate, $\alpha =$ 0.05
Power= 0.9799340 Calculate minimum sample size
Sample size, $n =$ 18 Calculate power

Enter the population characteristics by entering either the mean and standard deviation of each population or the standardized effect size. Select the type of test and the Type I error rate. Enter either a desired power value or a sample size and click the corresponding button to either *Calculate minimum sample size* or *Calculate power* .

DIRECTIONAL HYPOTHESIS

- One-tailed tests are more powerful than two-tailed tests, provided the effect is in the correct tail
- Using $\mu_a = 20.5$ compared to $\mu_0 = 22.5$, the runner might use a one-tailed test when analyzing the data. We noted earlier that $n = 18$ gave us 95% power for a two-tailed test. If she plans to use a (Negative) one-tail test, a smaller sample can be used to get the same power:

Specify the population characteristics:
$H_0 : \mu_0 =$ 22.5
$H_a : \mu_a =$ 20.5
$\sigma =$ 2.2
Or enter a standardized effect size
$\frac{\mu_a - \mu_0}{\sigma} = \delta =$ -0.909090

Specify the properties of the test:
Type of test Negative one-tail
Type I error rate, $\alpha =$ 0.05
Power= .95 Calculate minimum sample size
Sample size, $n =$ 15 Calculate power

Enter the population characteristics by entering either the mean and standard deviation of each population or the standardized effect size. Select the type of test and the Type I error rate. Enter either a desired power value or a sample size and click the corresponding button to either *Calculate minimum sample size* or *Calculate power* .

α CRITERION

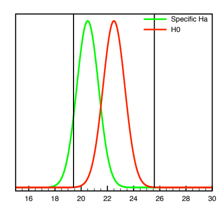
- Suppose the runner plans to use a criterion of $\alpha = 0.01$. Then, when doing a two-tailed hypothesis test with $n = 7$, the runner will compute a t-value and see if it is more extreme than

$$t_{cv} = \pm 3.70743$$

- In terms of mean running times, this corresponds to:

$$\overline{X}_{lower} = \mu - s_{\overline{X}} t_{cv} = 22.5 - (0.8315)(3.70743) = 19.417$$

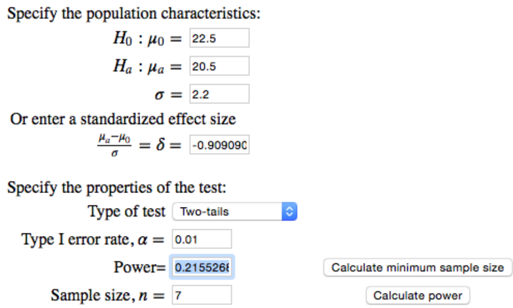
$$\overline{X}_{upper} = \mu + s_{\overline{X}} t_{cv} = 22.5 + (0.8315)(3.70743) = 25.583$$



- Power will be smaller!

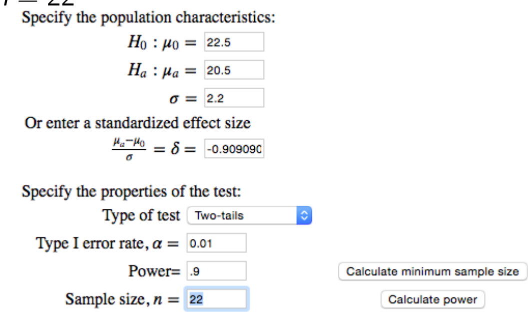
α CRITERION

- Using the power calculator, we find that with $\alpha = 0.01$, $n = 7$, $\mu_0 = 22.5$, $\mu_a = 20.5$, and $\sigma = 2.2$ for a two-tailed test, power of 0.216.



α CRITERION

- Using the power calculator, we find that with $\alpha = 0.01$, $\mu_0 = 22.5$, $\mu_a = 20.5$, and $\sigma = 2.2$ for a two-tailed test, to get a power of 0.9, we need $n = 22$



TRADE OFFS

- Experimental design always involves trade offs
- You want studies with large power (probability of rejecting the null hypothesis)
- You can only estimate power by hypothesizing how big the effect is, and estimating the variability of your data
- Bigger samples provide more power (but cost resources: time and money)
- Reducing the Type I error rate (α) also decreases power
  - Signal Detection Theory
  - Type I error corresponds to false alarms
  - Power corresponds to hits

CONCLUSIONS

- power
- experimental design
- sample size
- you should do it before gathering data

NEXT TIME

- Estimating means
- Confidence intervals

How tall is the room?

PSY 201: Statistics in Psychology

Lecture 21

Estimation of population mean

How tall is the room?

Greg Francis

Purdue University

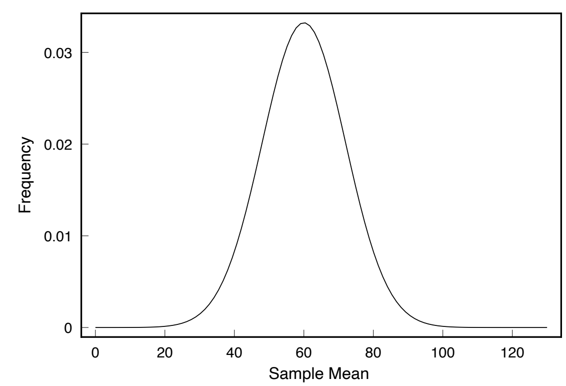
Fall 2023

LAST TIME

- we know how to check if a sample mean, $\overline{X}$, is statistically significantly different from a hypothesized population mean, μ.
- but sometimes we have no idea what μ is!
- we would like to be able to **estimate** μ using the sample data we have
- statistical estimation

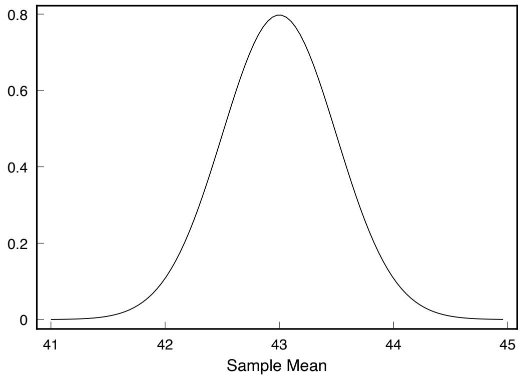
POINT ESTIMATION

- single value that represents the best estimate of a population value
- when we want to estimate μ, the best point estimate is the sample mean $\overline{X}$
- but the estimate depends on which sample we select!



INTERVAL ESTIMATION

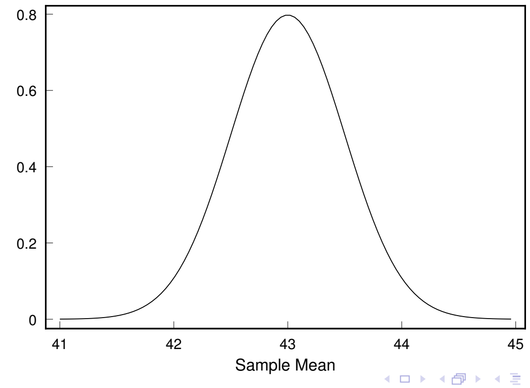
- we get a better idea of the value of μ by considering a **range** of values that are likely to contain μ
- we will show how to build up **confidence intervals** using the properties of the sampling distribution of the mean



σ KNOWN

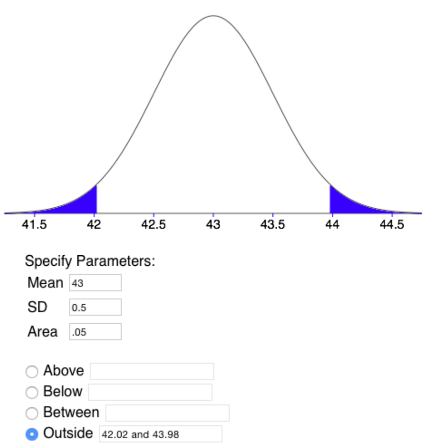
- to demonstrate our technique, suppose we have a population of scores with $\mu = 43$, $\sigma = 10$
- from the population we get the sampling distribution for samples of size $n = 400$ with

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = 0.50$$



INTERVAL ESTIMATION

- with the sampling distribution we can calculate (using the online Inverse Normal Distribution Calculator) that 95% of all sample means will lie between 42.02 and 43.98
- but since we do not really know the value of μ, we must **estimate** it



CONFIDENCE INTERVALS

- construct an interval around the observed statistic, $\overline{X}$

$$\text{CI} = \text{statistic} \pm (\text{critical value})\ (\text{standard error of the statistic})$$

$$\text{CI} = \overline{X} \pm (t_{cv})(s_{\overline{X}})$$

- where
  - $\overline{X}$ is the sample mean
  - t_{cv} is the critical value using the appropriate t distribution for the desired level of confidence
  - $s_{\overline{X}}$ is the estimated standard error of the mean

$$s_{\overline{X}} = \frac{s}{\sqrt{n}}$$

LEVEL OF CONFIDENCE

- degree of confidence that computed interval contains μ
- usually complement of level of significance, α
- level of confidence is $(1-\alpha)$
- calculating the critical value t_{cv} is the same!
- e.g., for $\alpha = 0.05$, $(1-\alpha) = 0.95$, and

$$t_{cv} = 1.9659$$

- (using the Inverse t Distribution calculator with df=399)

CONFIDENCE INTERVAL

- suppose we calculate $\overline{X} = 44.6$
- the confidence interval is then

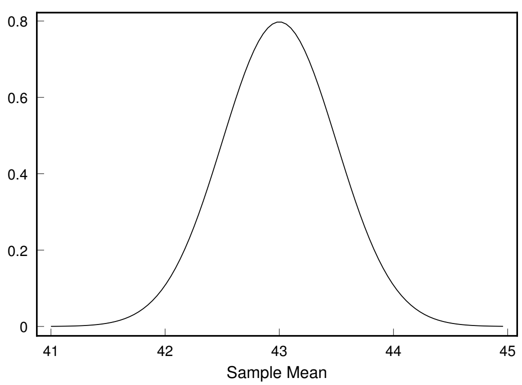
$$CI = \overline{X} \pm (t_{cv})(s_{\overline{X}})$$

$$CI_{95} = 44.6 \pm (1.9659)(0.50)$$

$$CI_{95} = (43.62, 45.58)$$

CONFIDENCE INTERVAL

- this means we are 95% confident that the interval (43.62, 45.58) contains the unknown value μ
  - Our procedure for producing the interval contains μ 95% of the time
- note: if $\mu = 43$ like was said originally, we are **wrong**!
  - CI does not contain μ (no way to avoid error completely)!



EXAMPLE

- Guess the height of this room in feet, and write down your guess on a piece of paper
- Now go around the room and get 10 guesses from other random people
- Then, tell me your guess
- Calculate the mean and standard deviation for your sample (use the on-line calculator for a one-sample t test)

$$\overline{X} = \frac{\sum X_i}{n}$$

$$s = \sqrt{\frac{\sum_i X_i^2 - [(\sum_i X_i)^2/n]}{n-1}}$$

- I'll calculate the *population* mean for the class
- each of you will calculate a confidence interval, for your sample, with $\alpha = 0.05$

CONFIDENCE INTERVAL

$$\text{CI} = \overline{X} \pm (t_{cv})(s_{\overline{X}})$$

- Calculate standard error of the mean

$$s_{\overline{X}} = \frac{s}{\sqrt{n}} = \frac{s}{\sqrt{10}} =$$

- we have

$$\text{d.f.} = n - 1 = 10 - 1 = 9$$

- so from the Inverse t Distribution Calculator, we find that

$$t_{cv} = 2.262$$

CONFIDENCE INTERVALS

- thus

$$\text{CI}_{95} = \overline{X} \pm (t_{cv})(s_{\overline{X}})$$

$$\text{CI}_{95} = \overline{X} \pm (2.262)(s_{\overline{X}})$$

$$\text{CI}_{95} = (\quad , \quad)$$

WHAT DOES THIS MEAN?

- we conclude with 95% confidence that your interval contains μ
- this is a probabilistic statement about the **interval**
- μ is a **parameter**, a fixed number

$$\mu =$$

- different samples produce different confidence intervals, but 95% of the time the interval would contain μ
- check

CONFIDENCE

- we **never** say that a specific confidence interval contains μ with probability 0.95
- either the interval contains μ or it does not
- we **can** say that the procedure of producing CI's produce intervals that contain μ with probability 0.95
- we do talk about the **confidence** that an interval includes μ
- we would say that the confidence interval contains μ with confidence of 0.95
- the confidence is in the **procedure** of calculating CIs

CONCLUSIONS

- estimation
- confidence intervals
- t distribution
- interpretation

NEXT TIME

- more on estimation
- relationship between confidence intervals and hypothesis testing
- statistical precision

Less than 5% of published psychological research should be wrong (and why that probably isn't true).

PSY 201: Statistics in Psychology

Lecture 22

Estimation of population mean

Less than 5% of published psychological research should be wrong (and why that probably isn't true).

Greg Francis

Purdue University

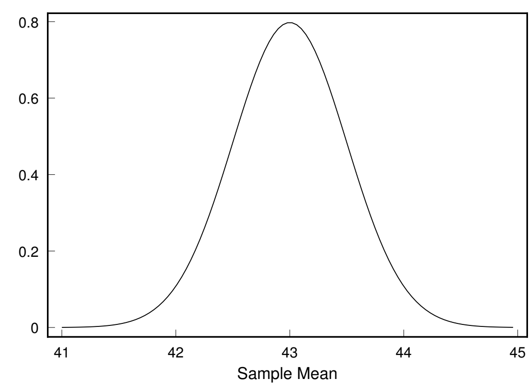
Fall 2023

LAST TIME

- construct an interval around an observed statistic, $\overline{X}$

$$\text{CI} = \text{statistic} \pm (\text{critical value})\ (\text{standard error of the statistic})$$

$$\text{CI} = \overline{X} \pm (t_{cv})(s_{\overline{X}})$$



CONFIDENCE

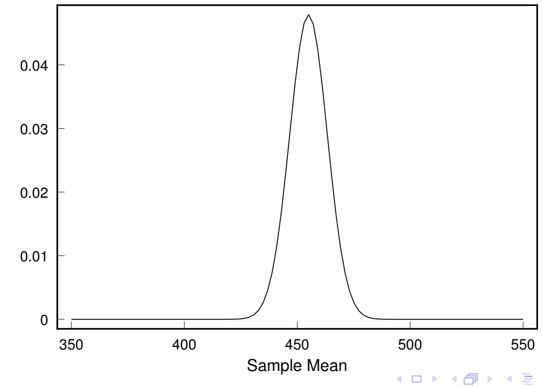
- we **never** say that a specific 95% confidence interval contains μ with probability 0.95
- either the interval contains μ or it does not
- we **can** say that the procedure of producing CI's produce intervals that contain μ with probability 0.95
- we do talk about the **confidence** that an interval includes μ
- we would say that the confidence interval contains μ with confidence of 0.95
- the confidence is in the **procedure** of calculating CIs

HYPOTHESIS TESTING

- remember SAT data:

$$H_0 : \mu = 455$$
$$H_a : \mu \neq 455$$

- calculated sampling distribution
  - for $\alpha = 0.05$, $\overline{X} = 535$, $s_{\overline{X}} = 8.33$
  - $t = 9.6$, $p \approx 0$, we rejected H_0



CONFIDENCE INTERVAL

- given our data, we could also compute confidence intervals around $\overline{X} = 535$
- $t_{cv} = \pm 1.96$

$$\text{CI} = \overline{X} \pm (t_{cv})(s_{\overline{X}})$$
$$\text{CI}_{95} = 535 \pm (1.96)(8.33)$$
$$\text{CI}_{95} = (518.67, 551.33)$$

COMPARISON

- note: the rejected $H_0 : \mu = 455$ is consistent with the CI
- 455 is **not** in the 95% confidence interval (518.67, 551.33)
- CI contains only tenable values of μ, given the sampled data

CI AND HYPOTHESIS TESTS

- CIs ask: which values of μ would it be reasonable for me to get the value of $\overline{X}$ that I found?
- Hypothesis tests ask: is the value of $\overline{X}$ I found consistent with a hypothesized value of μ?
- "reasonable" and "consistent" are defined relative to Type I error (α), and confidence (1-α)

HYPOTHESIS TESTING

- constructing a CI is like testing a large number of non-directional hypotheses simultaneously:

$$H_0 : \mu = 435$$

$$H_0 : \mu = 22$$

$$H_0 : \mu = 522$$

$$H_0 : \mu = 549$$

$$H_0 : \mu = 563$$

- anything in the CI (518.67, 551.33) would be not be rejected, anything not in the CI would be rejected

EXAMPLE

- On the papers going around the room, write down the number of math-based courses you have taken at college (include physics, engineering, and computer science, if it was largely math-based)
- Now go around the room and sample this information from 6 other people
- Calculate the mean and standard deviation for your sample (use the on-line calculator of the textbook)

$$\overline{X} = \frac{\sum X_i}{n}$$

$$s = \sqrt{\frac{\sum_i X_i^2 - [(\sum_i X_i)^2/n]}{n-1}}$$

HYPOTHESIS TEST

- (1) State the hypothesis and set the criterion:

$$H_0 : \mu = 3$$

$$H_a : \mu \neq 3$$

- $\alpha = 0.05$
- (2) Compute test statistic:
  - Calculate standard error of the mean

$$s_{\overline{X}} = \frac{s}{\sqrt{n}} = \frac{s}{\sqrt{6}} =$$

  - Compute the t-value

$$t = \frac{\overline{X} - 3}{s_{\overline{X}}} =$$

- (3) Compute the p-value:
  - use the t Distribution calculator with $df = n - 1 = 5$ to compute

$$p =$$

- (4) Make your decision:

CONFIDENCE INTERVAL

$$\text{CI} = \overline{X} \pm (t_{cv})(s_{\overline{X}})$$

- Calculate standard error of the mean

$$s_{\overline{X}} = \frac{s}{\sqrt{n}} = \frac{s}{\sqrt{6}} =$$

- with $n = 6$, $df = 5$, so the Inverse t Distribution calculator gives $t_{cv} = 2.571$
- plug everything into your CI formula

COMPARISON

- Who rejected H_0?
- Who have the value 3 outside their CI?
- Should be similar!
- Now repeat everything for $H_0 : \mu = 4$, using the textbook's online calculator
- notice what is required in the new calculations!

STATISTICAL PRECISION

- consider the equation for confidence intervals

$$\text{CI} = \overline{X} \pm (t_{cv})(s_{\overline{X}})$$

- where
  - $\overline{X}$ is the sample mean
  - t_{cv} is the critical value using the appropriate t distribution for the desired level of confidence
  - $s_{\overline{X}}$ is the estimated standard error of the mean

$$s_{\overline{X}} = \frac{s}{\sqrt{n}}$$

- smaller t_{cv} or $s_{\overline{X}}$ produce narrower widths

STATISTICAL PRECISION

- since

$$s_{\overline{X}} = \frac{s}{\sqrt{n}}$$

- increasing the sample size n produces narrow widths of CI
- narrower widths imply greater precision about where μ is located
- increasing n also modifies t_{cv} by changing degrees of freedom

$$\text{df} = n - 1$$

- larger df leads to smaller t_{cv}
  (see the Inverse t Calculator)

STATISTICAL PRECISION

- we can also change t_{cv} by changing the level of confidence
- larger level of confidence, implies smaller α, which implies larger t_{cv}, which implies larger width of CI
- makes sense, we become more confident the interval includes μ by broadening the interval
- of course, then we are less sure about the value μ

PUBLISHED DATA

- most researchers in the behavioral sciences use $\alpha = 0.05$
- this means that they make a Type I error only 5% of the time (or less)
- no way to completely avoid making mistakes
- this makes it quite likely that **some** of the data in published journals is wrong
- it is important in science to double (and triple) check everything
- if a bit of data is tremendously important, better replicate the experimental finding

PUBLISHING CHALLENGES

- Researchers often use statistical significance as a way of identifying what findings should be published
- If only findings with $p < .05$ are published, then journals can be filled with findings where H_0 is actually true
- even if H_0 is true, around 5% of samples will produce a significant p value
- If non-significant findings are not published, then it becomes hard to interpret the findings that actually are published (publication bias)

SAMPLING CHALLENGES

- Suppose you run a study with $n = 20$ subjects and get $p = .07$. This does not meet the $\alpha = 0.05$ criterion.
- It is tempting to add an additional 10 subjects (for a total of $n = 30$) and do the analysis again
- This is a problem because you have given yourself an extra chance to get a significant outcome. Your Type I error is bigger than the $\alpha = 0.05$ that you intended.
- Cannot add subjects to an experiment and re-analyze. Nor can you stop data collection when you get a significant result (data peeking, optional stopping).
- The sampling distribution is only valid for a **fixed** sample size. In the above cases, the sample size is not fixed.
- To avoid these problems, you have to plan your experiment carefully in advance (power).

PRECISION FOCUS

- One way to avoid these issues is to run your study to focus on measuring things "well enough".
- You might want to keep gathering data until the width of a 95% confidence interval is "small enough"
- Then you could test the H_0
- Of course, you have to come up with some definition of small enough

CONCLUSIONS

- estimation
- confidence intervals
- relationship with hypothesis testing
- statistical precision
- challenges

NEXT TIME

- more hypothesis testing
- tests for a proportion

Can you read my mind: Part II?

PSY 201: Statistics in Psychology

Lecture 23

Hypothesis tests for a proportion

Can you read my mind? Part II

Greg Francis

Purdue University

Fall 2023

HYPOTHESIS TESTING

- four steps
  1. State the hypothesis and the criterion
  2. Compute the test statistic.
  3. Compute the p-value.
  4. Make a decision

HYPOTHESIS TESTING

- we need to know the properties of the sampling distribution
- for the mean, the central limit theorem tells us that the sampling distribution is normal, and specifies the mean and standard deviation (standard error)
- area under the curve of the sampling distribution gives probability of getting that sampled value, or values more extreme (p-value)
- for other types of statistics, the sampling distribution is different
  - area under the curve of sampling distribution still gives probability of getting that sampled value, or values more extreme
- proportion

HYPOTHESIS TESTING

- the approach is still basically the same
- we compute

$$\text{Test statistic} = \frac{\text{statistic - parameter}}{\text{standard error of the statistic}}$$

- and use it to compute a p-value, which we compare to α

PROPORTION

- many times we want to know what proportion (P) of a population has a certain trait
  - Own a phone.
  - Are a democrat.
  - Are a republican.
  - Own a computer.
  - ...
- dichotomous population (have trait or do not)
- percentages

PROPORTION

- we can take a random sample and calculate a sample proportion p
- we can test hypotheses about the population parameter P
  e.g.

$$H_0 : P = 0.5$$

$$H_a : P \neq 0.5$$

- the sampling distribution of p is the binomial distribution
- for large samples it is very close to the normal distribution

STANDARD ERROR

- an estimate of the standard error of the sampling distribution (standard error of the sample proportion) is:

$$s_p = \sqrt{\frac{PQ}{n}}$$

  - P = population proportion possessing characteristic
  - $Q = 1 - P$ = population proportion not possessing characteristic
  - n = sample size
- now we can apply the techniques of hypothesis testing!

PEPSI CHALLENGE

- several years ago Pepsi sponsored the **Pepsi Challenge** where you sampled Coke and Pepsi and decided which tasted better
- after testing hundreds of people, they found that more than half the Coke drinkers preferred Pepsi (63%)
- how would we test to see if the proportion of people who preferred Pepsi over Coke was a significant proportion (different from chance)?

HYPOTHESIS

- Step 1. State the hypothesis and criterion. By chance we would expect the proportion of people that preferred Pepsi would be 50%

$$H_0 : P = 0.5$$

$$H_a : P \neq 0.5$$

- Let's set our level of significance at $\alpha = 0.05$, two-tailed test

CRITERION

- Step 2. Compute the test statistic. Suppose the sample proportion is

$$p = \frac{189}{300} = 0.63$$

- Let's suppose $n = 300$ people were tested, and so the standard error of the sample proportion is:

$$s_p = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{(0.5)(0.5)}{300}} = 0.02886$$

TEST STATISTIC

- the test statistic is:

$$z = \frac{p - P}{s_p} = \frac{0.63 - 0.5}{0.02886} = 4.50$$

- Step 3. Compute the p-value. We use the Normal Distribution Calculator to compute

$$p \approx 0$$

- Step 4. Make a decision. Since $p < \alpha = 0.05$, we can reject H_0!
  - If $P = 0.5$, the probability of getting $p = 0.63$, or an even bigger difference from $P = 0.5$, from a random sample of 300 people is less than 0.05.
  - The observed difference is a significant difference.

CONFIDENCE INTERVALS

- Let's construct a confidence interval with level of confidence $1 - \alpha = 0.95$
- The critical value z_{cv} is found from the Inverse Normal Distribution Calculator

$$z_{cv} = 1.96$$

- so

$$CI_{95} = p \pm (1.96)(s_p)$$

- For the confidence interval, we recompute the standard error by using the estimate from the sample

$$s_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.63)(0.37)}{300}} = 0.0279$$

$$CI_{95} = 0.63 \pm (1.96)(0.0279)$$

$$CI_{95} = (0.57, 0.68)$$

- which does not include the chance level $P = 0.5$

MIND READING

- I am going to pick one of the following words as a "special" word
- You try to read my mind as to which one is "special"
- write it down on a sheet of paper. I'll write down my chosen word on a sheet of paper
  - COMPUTER
  - STEREO
  - BICYCLE
  - STAPLER
  - BOOKCASE
  - DESK

MIND READING

- Now, I tell you my special word, and we find out how many of you were correct. We are measuring p, the sample proportion
- we can test whether you can read my mind
- (1) State the hypothesis and the criterion
  - the null hypothesis is that you cannot read my mind, so we say that

$$H_0 : P = \frac{1}{6} = 0.167$$

$$H_a : P \neq 0.167$$

  - where 0.167 is what you would get just by guessing
- $\alpha = 0.10$

MIND READING

- (2) Compute the test statistic

$$s_p = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{(0.167)(0.833)}{n}} = \sqrt{\frac{0.1391}{n}} =$$

$$z = \frac{p - P}{s_p} =$$

- (3) Which we plug in to the Normal Distribution Calculator to find the p-value
- (4) Make a decision
- We can do it all with the One Sample Proportion Test Calculator in the textbook

POWER

- How would we design a good experiment to test Mind Reading abilities?
- How big a sample do we need to have a 90% chance of rejecting the H_0?
- Conceptually, this is the same issue as estimating power or sample size for a hypothesis test of means
- We just need to use the sampling distribution for a sample proportion instead of the sampling distribution for a sample mean

POWER

- We have to set the specific proportion for the alternative hypothesis
- Suppose we plan to test

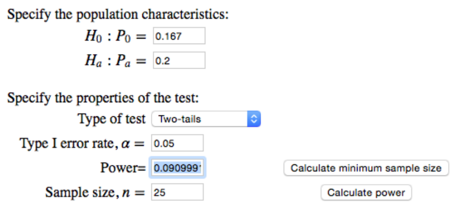
$$H_0 : P = 0.167, H_a : P \neq 0.167$$

- and we set the specific alternative as

$$H_a : P_a = 0.2$$

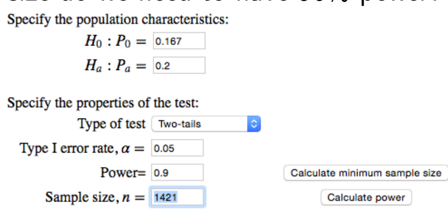
- What is the probability that a random sample of $n = 25$ will reject the H_0?
- The on-line calculator does all the work!

POWER


Specify the population characteristics:
$H_0 : P_0 =$ 0.167
$H_a : P_a =$ 0.2

Specify the properties of the test:
Type of test: Two-tails
Type I error rate, $\alpha =$ 0.05
Power= 0.090999
Sample size, $n =$ 25
Calculate minimum sample size
Calculate power

- Less than 10% chance of rejecting the null hypothesis
- What sample size do we need to have 90% power?


Specify the population characteristics:
$H_0 : P_0 =$ 0.167
$H_a : P_a =$ 0.2

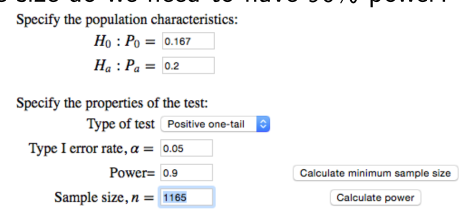
Specify the properties of the test:
Type of test: Two-tails
Type I error rate, $\alpha =$ 0.05
Power= 0.9
Sample size, $n =$ 1421
Calculate minimum sample size
Calculate power

POWER

- Suppose we plan to test

$$H_0 : P = 0.167, H_a : P > 0.167$$

- What sample size do we need to have 90% power?


Specify the population characteristics:
$H_0 : P_0 =$ 0.167
$H_a : P_a =$ 0.2

Specify the properties of the test:
Type of test: Positive one-tail
Type I error rate, $\alpha =$ 0.05
Power= 0.9
Sample size, $n =$ 1165
Calculate minimum sample size
Calculate power

POWER

- Let's use the proportion we found for the class as the specific alternative value
  - Power?
  - Sample size for 90% power?

CONCLUSIONS

- testing significance of proportions
- confidence intervals for proportions
- power for tests of proportions

NEXT TIME

- hypothesis testing of correlations
- Fisher z transform
- another t test
- confidence interval
- power

Is there a correlation between homework and exam scores?

PSY 201: Statistics in Psychology

Lecture 24

Hypothesis testing for correlations

Is there a correlation between homework and exam grades?

Greg Francis

Purdue University

Fall 2019

HYPOTHESIS TESTING

- four steps
  1. State the hypothesis and the criterion
  2. Compute the test statistic.
  3. Compute the p-value.
  4. Make a decision

HYPOTHESIS TESTING

- we need to know the properties of the sampling distribution
- for the mean, the central limit theorem tells us that the sampling distribution is normal, and specifies the mean and standard deviation (standard error)
- area under the curve of the sampling distribution gives probability of getting that sampled value, or values more extreme (p-value)
- for other types of statistics, the sampling distribution is different
  - area under the curve of sampling distribution still gives probability of getting that sampled value, or values more extreme
- correlation coefficient

HYPOTHESIS TESTING

- the approach is still basically the same
- we compute

$$\text{Test statistic} = \frac{\text{statistic - parameter}}{\text{standard error of the statistic}}$$

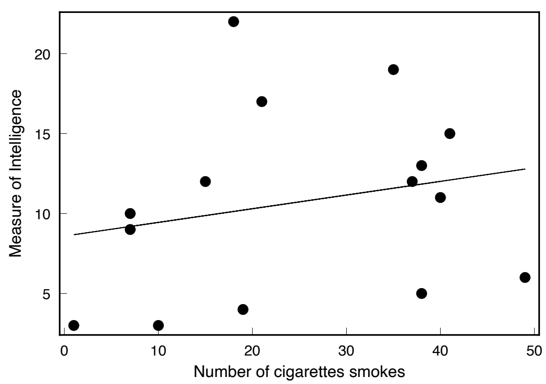
- and use it to compute a p-value, which we compare to α

CORRELATION COEFFICIENT

- from a population with scores X and Y, we can calculate a correlation coefficient
- let ρ be the correlation coefficient **parameter** of the population
- let r be the correlation coefficient **statistic** from a random sample of the population

SAMPLING

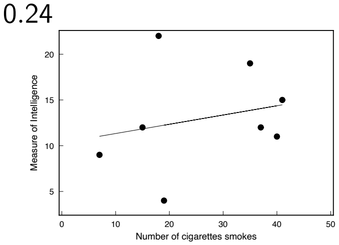
- Suppose $\rho = 0.22$



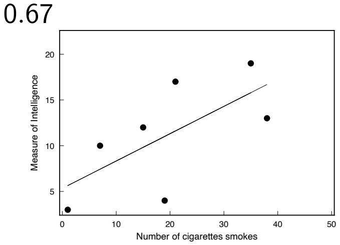
- depending on which points we sample, the computed r will take different values

RANDOM SAMPLING

- $r = 0.24$

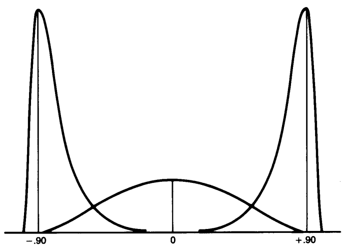


- $r = 0.67$



SAMPLING DISTRIBUTION

- frequency of different r values, given a population parameter ρ
- **not** usually a normal distribution!
- often skewed to the left or the right
- cannot find area under curve!

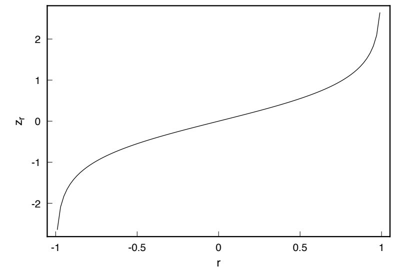


FISHER z TRANSFORM

- formula for creating new statistic

$$z_r = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right)$$

- where $\log_e$ is the "natural logarithm" function
  - also sometimes designated as ln



- textbook provides a r to z' calculator (reversible!)

FISHER z TRANSFORM

- for large samples, the sampling distribution of z_r is normally distributed
  - regardless of the value of ρ
- with a mean

$$z_\rho = \frac{1}{2}\log_e\left(\frac{1+\rho}{1-\rho}\right)$$

- and with standard error (standard deviation of the sampling distribution)

$$s_{z_r} = \sqrt{\frac{1}{n-3}}$$

- where n is the sample size

FISHER z TRANSFORM

- means we can use all our knowledge about hypothesis testing with normal distributions for the transformed scores!
- online calculator converts r to z_r (it calls it z')
- e.g.

$$r = -0.90 \rightarrow z_r = -1.472$$

$$r = 0 \rightarrow z_r = 0$$

$$r = 0.45 \rightarrow z_r = 0.485$$

- we can convert back the other way from $z_r \rightarrow r$ too!

HYPOTHESIS TESTING

- Suppose we study a population of data that we think has a correlation of 0.65. We want to test the hypothesis with a sample size of $n = 30$.
- e.g. family income and attitudes about democratic childrearing
- Step 1. State the hypothesis and criterion

$$H_0 : \rho = 0.65$$

$$H_a : \rho \neq 0.65$$

- two-tailed test

$$\alpha = 0.05$$

HYPOTHESIS TESTING

- Step 2. Compute the test statistics
- suppose from our sampled data we get

$$r = 0.61$$

- we need to convert it to a z_r score

$$r = 0.61 \rightarrow z_r = 0.709$$

- and calculate standard error

$$s_{z_r} = \sqrt{\frac{1}{n-3}} = \sqrt{\frac{1}{27}} = 0.192$$

HYPOTHESIS TESTING

- now we calculate the test statistic

$$\text{Test statistic} = \frac{\text{statistic - parameter}}{\text{standard error of the statistic}}$$

$$z = \frac{z_r - z_\rho}{s_{z_r}} = \frac{0.709 - 0.775}{0.192} = -0.344$$

- Step 3. Compute the p-value. From the Normal Distribution calculator, we compute

$$p = 0.7346$$

HYPOTHESIS TESTING

- Step 4. Make a decision.

$$p = 0.7346 > 0.05 = \alpha$$

- H_0 is **not** rejected at the 0.05 significance level
  - The probability of getting $r = 0.61$ (or a value further away from 0) with a random sample, if $\rho = 0.65$, is greater than 0.05.
  - The observed difference is not a significant difference.

A SPECIAL CASE

- hypothesis testing of correlation coefficients can always use Fisher's z transform

$$H_0 : \rho = a$$
$$H_a : \rho \neq a$$

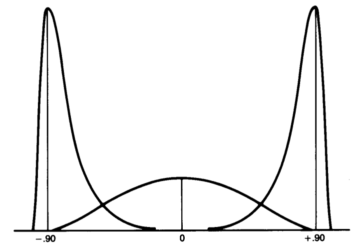
- special case $a = 0$:

$$H_0 : \rho = 0$$
$$H_a : \rho \neq 0$$

- Is there a significant correlation coefficient?
- Is there a linear relationship between two variables?

SAMPLING DISTRIBUTION

- while we needed Fisher's z transformation to convert the sampling distribution into a normal distribution
- it is not necessary for testing $\rho = 0$



SAMPLING DISTRIBUTION

- for $\rho = 0$ the sampling distribution of the test statistic is a t distribution with df $= n - 2$
  - two sets of scores, minus 1 from each set
- no need to convert with Fisher z transform
- we follow the same procedures as before
  1. State the hypothesis. $H_0 : \rho = 0$ and set the criterion
  2. Compute the test statistic.
  3. Compute the p-value
  4. Make a decision.

HYPOTHESIS TESTING

- everything is the same, except the test statistic calculation is a bit different
- it turns out that an estimate of the standard error is:

$$s_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

- so that the test statistic is:

$$t = \frac{r - \rho}{s_r} = r\sqrt{\frac{n - 2}{1 - r^2}}$$

- we use this with a t distribution to compute a p-value

EXAMPLE

- $n = 32$ scores calculated to get $r = -0.375$
  1. State the hypothesis. $H_0 : \rho = 0$, $H_a : \rho \neq 0$, $\alpha = 0.05$
  2. Compute the test statistic

$$t = r\sqrt{\frac{n-2}{1-r^2}} = (-0.375)\sqrt{\frac{30}{0.859}} = -2.216$$

  3. Compute the p value using the t Distribution calculator with df$=n-2=30$

$$p = 0.0344$$

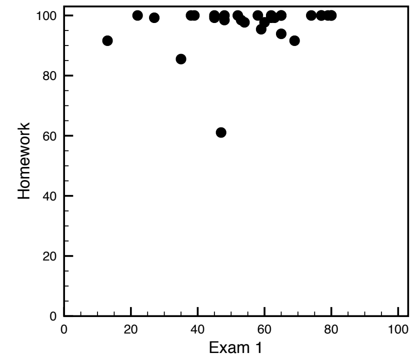
  4. Interpret the results: $p = 0.0344 < 0.05 = \alpha$; reject H_0

EXAMPLE

- I took the percentage of the first five homework grades and correlated it with the first exam scores

$$\rho = 0.2123$$

- Is this a significant correlation?



CAREFULL!

- If I treat the class as a *population*, the correlation simply is what it is. Significance is not an issue!
- If I treat the class as a *sample* of students who do homework and take exams in statistics, then I can ask about statistical significance

CAREFULL!

- is $r = 0.2123$ significantly different from 0? I have $n = 30$ scores
- Compute the test statistics.

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 1.1496$$

- use the t Distribution calculator with df$=n-2=28$

$$p = 0.769$$

- Interpret the results: $p = 0.26 > 0.05 = \alpha$, do not reject H_0

READING?

- For Homework and Reading, $r = 0.8964$. I have $n = 30$ scores
- Compute the test statistics.

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 10.70$$

- use the t Distribution calculator with df$=n-2=28$

$$p \approx 0$$

- Interpret the results: $p \approx 0 < 0.05 = \alpha$, reject H_0

CAREFUL!

- When we conclude a test is statistically significant, we base that on the observation that observed data (or more extreme) would be rare if the H_0 were true
- But if we make multiple tests from a single sample, our calculations of probability may be invalid.
- We performed two hypothesis tests from one sample of students.
- Each test has a chance of producing a significant results, even if H_0 is true
- It is not appropriate to just run various tests with one data set, if all you are doing is looking for significant results (fishing)
- You have to do a different type of statistical analysis

CONFIDENCE INTERVAL

- Always use the Fisher z transform
- Build interval as a Fisher z score and then convert to correlation (r value)

$$CI = z_r \pm z_{cv} s_{z_r}$$

- the correlation between homework and reading scores:

$$CI_{95} = 1.453 \pm (1.96)(0.192) = (1.076, 1.831)$$

- when we convert to r values:

$$(0.792, 0.950)$$

POWER

- How would we design a good experiment to test a correlation?
- How big a sample do we need to have a 90% chance of rejecting the H_0?
- Conceptually, this is the same issue as estimating power or sample size for a hypothesis test of means
- We just need to use the sampling distribution for the Fisher z transform of the sample correlation instead of the sampling distribution for a sample mean

POWER

- We have to specify the specific correlation for the alternative hypothesis
- Suppose we plan to test

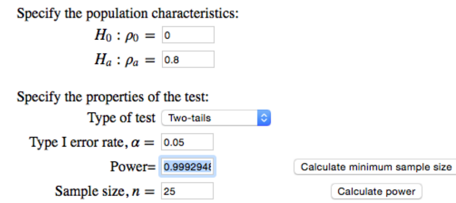
$$H_0 : \rho = 0, H_a : \rho \neq 0$$

- and we set the specific alternative as

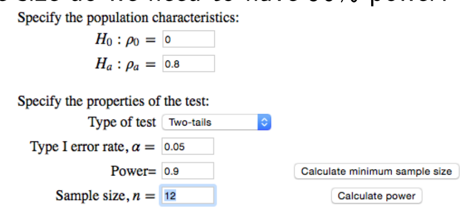
$$H_a : \rho_a = 0.8$$

- What is the probability that a random sample of $n = 25$ will reject the H_0?
- The on-line calculator does all the work!

POWER



- Higher than 99.9% chance of rejecting the null hypothesis
- What sample size do we need to have 90% power?



- However, whether these calculations make sense depends on whether $\rho = 0.8$ in reality.

CONCLUSIONS

- correlation coefficient
- Fisher z transform
- testing significance of correlation
- confidence interval
- power

NEXT TIME

- hypothesis testing of two means
- homogeneity of variance
- confidence interval
- robustness and assumptions

Check yourself before you wreck yourself.

PSY 201: Statistics in Psychology

Lecture 25

Hypothesis testing for two means

Check yourself before you wreck yourself.

Greg Francis

Purdue University

Fall 2023

HYPOTHESIS TESTING

$$H_0 : \mu = a$$

$$H_a : \mu \neq a$$

$$H_0 : \rho = a$$

$$H_a : \rho \neq a$$

- always compare **one-sample** to a hypothesized population parameter
- sometimes we want to compare two (or more) population parameters

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

TWO-SAMPLE CASE FOR THE MEAN

- useful when you want to compare means of two groups
  - different teaching methods
  - survival with and without drug
  - depression with and without treatment
  - height of males and females
- the null hypothesis is that there is no difference between the means

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

- or another way to say the same thing

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 \neq 0$$

DIFFERENCE OF MEANS

- since we want to compare the difference of two population means
- our statistic should be the difference of two sample means

$$\overline{X}_1 - \overline{X}_2$$

- and we will compare that statistic to the hypothesized value of the parameter

$$H_0 : \mu_1 - \mu_2 = 0$$

- if the statistic is much different from the hypothesized parameter, we will reject H_0
- same approach as before, different sampling distribution

INDEPENDENCE

- drawing a sample with a particular value of $\overline{X}_1$ should not affect the probability of drawing a sample with any other particular value of $\overline{X}_2$
- remember statistical independence

$$P(X \text{ and } Y) = P(X) \times P(Y)$$

- same idea here

INDEPENDENCE

- in practice this means we need to be careful about how we sample
- if comparing treatments, randomly divide a random sample into an **experimental group** and a **control group**
- Thus, even if you hope your new treatment will save lives, you have to have one group of patients without the treatment (maybe even a "sham" treatment).
  - It seems cruel, but you cannot assume the treatment works, you have to *demonstrate* it.
- take random samples from each population (no overlap, so no risk of dependence)
- avoid situations like repeating subjects:
  - e.g., comparing depression for the same subjects before and after treatment
  - there are ways to test this situation, but not with these techniques

HOMOGENEITY OF VARIANCE

- to carry out hypothesis testing we need to calculate standard error
- to get standard error we need to estimate (or know) the standard deviation
- since we sample two groups, we need a **pooled estimate** of σ^2
- to get a pooled estimate we need to be certain that

$$\sigma_1^2 = \sigma_2^2$$

- note this is a statement about the **populations**
  we would not expect the sample variances to be identical

HYPOTHESIS TESTING

- we want to compare population means from two populations
- we have
  - $H_0 : \mu_1 = \mu_2$
  - $\sigma_1^2 = \sigma_2^2 = \sigma^2$
  - Independent samples of size n_1 and n_2
- although we draw two random samples (one from each population), we are only interested in one statistic

$$\overline{X}_1 - \overline{X}_2$$

- but we need to know the sampling distribution for this statistic

SAMPLING DISTRIBUTION OF DIFFERENCES

- it turns out that the sampling distribution is familiar
  - Shape: As sample sizes get large, distribution becomes normal.
  - Central tendency: The mean of the sampling distribution equals $\mu_1 - \mu_2$.
  - Variability: The standard deviation of the sampling distribution (standard error of the difference between means) is

$$\sigma_{\overline{X}_1-\overline{X}_2} = \sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

- We have to estimate σ from our data
- our estimate is called the **pooled estimate** because we use scores from both samples

FORMULAS

- deviation formula

$$s^2 = \frac{\Sigma \left(X_{i1} - \overline{X}_1 \right)^2 + \Sigma \left(X_{i2} - \overline{X}_2 \right)^2}{n_1 + n_2 - 2}$$

- deviations relative to the sample mean of each sample!
- raw score form:

$$s^2 = \frac{\left[\Sigma X_{i1}^2 - \left(\Sigma X_{i1} \right)^2 / n_1 \right] + \left[\Sigma X_{i2}^2 - \left(\Sigma X_{i2} \right)^2 / n_2 \right]}{n_1 + n_2 - 2}$$

  - X_{i1} refers to the ith score from sample 1
  - X_{i2} refers to the ith score from sample 2
  - n_1 refers to the number of scores in sample 1
  - n_2 refers to the number of scores in sample 2

FORMULAS

- variances

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- where
  - s_1^2 is the variance among scores in sample 1
  - s_2^2 is the variance among scores in sample 2

STANDARD ERROR

- we use the pooled s to calculate an estimate of standard error for the sampling distribution of differences

$$s_{\overline{X}_1-\overline{X}_2} = \sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

- this gives us an estimate of the standard deviation of the sampling distribution of the difference of sample means
- we need to know one more thing

DEGREES OF FREEDOM

- we have two samples with (possibly) different numbers of scores
- the degrees of freedom in sample 1

$$\text{df} = n_1 - 1$$

- from sample 2

$$\text{df} = n_2 - 1$$

- added together gives the result (depends on independence!)

$$\text{df} = n_1 + n_2 - 2$$

- (same as in denominator of standard deviation estimate)

HYPOTHESIS TESTING

- now we have everything we need to apply the techniques of hypothesis testing
  1. State the hypothesis and the criterion.
  2. Compute the test statistic.
  3. Compute the p-value.
  4. Make a decision.

EXAMPLE

- A neurosurgeon believes that lesions in a particular area of the brain, called the thalamus, will decrease pain perception. If so, this could be important in the treatment of terminal illness accompanied by intense pain. As a first attempt to test this hypothesis, he conducts an experiment in which 16 rats are randomly divided into two groups of 8 each. Animals in the experimental group receive a small lesion in the part of the thalamus thought to be involved in pain perception. Animals in the control group receive a comparable lesion in a brain area believed to be unrelated to pain. Two weeks after surgery each animal is given a brief electrical shock to the paws. The shock is administered with a very low intensity level and increased until the animal first flinches. In this manner, the pain threshold to electric shock is determined for each rat. The following data are obtained. Each score represents the current level (milliamperes) at which flinching is first observed. The higher the current level, the higher is the pain threshold.

HYPOTHESIS

- Step 1. State the hypotheses and the criterion.
- Directional hypothesis because we expect the lesion will increase the threshold.

$$H_0 : \mu_1 = \mu_2 \text{ or } \mu_1 - \mu_2 = 0$$

- (lesion makes no difference)

$$H_a : \mu_1 < \mu_2 \text{ or } \mu_1 - \mu_2 < 0$$

- (lesion increases pain threshold, less sensitivity)
- we will set $\alpha = 0.05$ for a one-tailed test
- We expect a negative t value (see H_a)

DATA

- now we consider the data from the experiment. The researcher gets the following

Control Group (False lesion) X_1	Experimental Group (Thalamic lesion) X_2
0.8	1.9
0.7	1.8
1.2	1.6
0.5	1.2
0.4	1.0
0.9	0.9
1.4	1.7
1.1	0.7

COMPUTING TEST STATISTIC

- Step 2. we have $n_1 = 8$, $n_2 = 8$
- from the data we calculate

$$\overline{X}_1 = 0.875$$

$$\overline{X}_2 = 1.3625$$

$$\overline{X}_1 - \overline{X}_2 = -0.4875$$

$$s^2 = 0.403$$

- (using any formula you want), so that the estimate of standard error is

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{0.403\left(\frac{1}{8} + \frac{1}{8}\right)} = 0.2015$$

COMPUTING THE TEST STATISTIC

- $$\text{Test statistic} = \frac{\text{Statistic} - \text{Parameter}}{\text{Standard Error of the Statistic}}$$

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_{\overline{X}_1 - \overline{X}_2}}$$

$$t = \frac{(0.875 - 1.3625) - 0}{0.2015} = -2.419$$

- Step 3. Compute the p-value.
  - we need to calculate the degrees of freedom

$$\text{df} = n_1 + n_2 - 2 = 16 - 2 = 14$$

- We use the t Distribution Calculator to compute

$$p = 0.015$$

INTERPRET RESULTS

- Step 4. Make a decision.
  - our interpretation of the test is that the difference between the calculated sample means, or a even bigger difference, would have occurred by chance less than 5% of the time if the null hypothesis were true
  - in practice, this means that the study supports the theory that lesions to the thalamus decrease pain perception
  - significant result
  - This means you have support for the idea that the surgery **did** affect pain perception

CONFIDENCE INTERVAL

- Basic formula for all confidence intervals:

$$CI = \text{statistic} \pm (\text{critical value})(\text{standard error})$$

- for a difference of sample means

$$CI = (\overline{X}_1 - \overline{X}_2) \pm t_{cv} s_{\overline{X}_1 - \overline{X}_2}$$

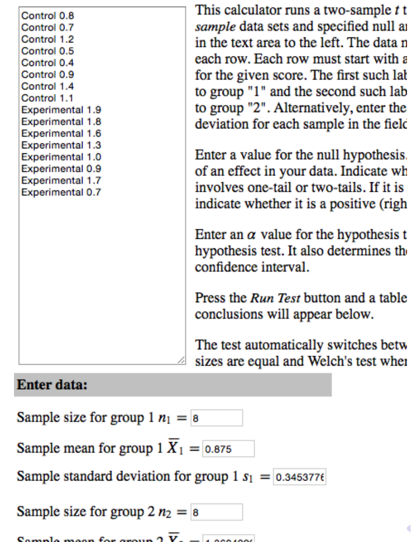
- We already have most of the terms (we get t_{cv} from the Inverse t-distribution calculator, so

$$CI_{95} = (0.875 - 1.3625) \pm (2.1448)(0.2015)$$

$$CI_{95} = (-0.9197, -0.0553)$$

ONLINE CALCULATOR

- The calculations are not complicated, but it is usually better to use a computer. You have to properly format the data.



ONLINE CALCULATOR

- You need to understand how to pull out the information you want

Test summary	
Type of test	Standard
Null hypothesis	$H_0 : \mu_1 - \mu_2 = 0$
Alternative hypothesis	$H_a : \mu_1 - \mu_2 < 0$
Type I error rate	$\alpha = 0.05$
Label for group 1	Control
Sample size 1	$n_1 = 8$
Sample mean 1	$\overline{X}_1 = 0.8750$
Sample standard deviation 1	$s_1 = 0.345378$
Label for group 2	Experimental
Sample size 2	$n_2 = 8$
Sample mean 2	$\overline{X}_2 = 1.3625$
Sample standard deviation 2	$s_2 = 0.453360$
Pooled standard deviation	$s = 0.403002$
Sample standard error	$s_{\overline{X}_1 - \overline{X}_2} = 0.201501$
Test statistic	$t = -2.419342$
Degrees of freedom	$df = 14$
p value	$p = 0.014873$
Decision	Reject the null hypothesis
Confidence interval critical value	$t_{cv} = 2.144787$
Confidence interval	$CI_{95} = (-0.919677, -0.055323)$

ASSUMPTIONS

- The t-test that we use for hypothesis tests of means is based on three key assumptions
  - The *population* distributions are normally distributed. Matters for small sample sizes.
  - Independent scores. For a two-sample t-test, the scores are uncorrelated between populations. (We deal with this case soon.)
  - Homogeneity of variance. For a two-sample t-test, the populations have the same variance (or standard deviation).
- If these assumptions do not hold, then the t-distribution that we calculate is not an accurate description of the sampling distribution.

ROBUSTNESS?

- Deviation from normal distributions for the populations does not matter very much, especially for large samples. If we run many tests, we see the Type I error rate pretty close to what is intended by setting α (e.g., $\alpha = 0.05$)
- Show in Robustness Simulation Demonstration in the textbook (12.4)
- This is true for varying sample sizes

HOMOGENEITY OF VARIANCE

- to carry out hypothesis testing we need to calculate standard error
- to get standard error we need to estimate (or know) the standard deviation of the population
- since we sample two groups, we used a **pooled estimate** of σ^2
- to get a pooled estimate we need to be certain that

$$\sigma_1^2 = \sigma_2^2$$

- we need consider what happens when homogeneity does not hold

ROBUSTNESS?

- For a two-sample t-test, if $n_1 = n_2$, then having $\sigma_1^2 \neq \sigma_2^2$ does not matter very much.
- If we run many tests, we see the Type I error rate pretty close to what is intended by setting α (e.g., $\alpha = 0.05$), especially for larger sample sizes
- Show in Robustness Simulation Demonstration in the textbook (12.4)
- Shape of the population distributions does not matter very much.

ROBUSTNESS?

- For a two-sample t-test, if $n_1 \neq n_2$, then having $\sigma_1^2 \neq \sigma_2^2$ matters a lot.
- If we run many tests, we see the Type I error rate is *much different* than what is intended by setting α (e.g., $\alpha = 0.05$)
- Type I error rate is around 37% if big σ^2 is paired with small n
- Type I error rate is around 0.2% if big σ^2 is paired with big n
- Show in Robustness Simulation Demonstration in the textbook (12.4)
- Shape of the population distributions does not matter very much.

HOMOGENEITY OF VARIANCE

- Our concern is about population variances (σ_1^2 and σ_2^2) not about sample variances (s_1^2 and s_2^2)
- It is possible to do a hypothesis test for variances

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_a : \sigma_1^2 \neq \sigma_2^2$$

- Note, it would be nice if we did **not** reject H_0, because then we could use our original method
- if we reject H_0, we must make some adjustments to hypothesis testing for the means

HOMOGENEITY OF VARIANCE

- We are not actually going to do the hypothesis test for homogeneity of variance
- It is messy and (a bit) confusing
- Just remember:
  - If the sample sizes are equal, then you are fine with the standard method.
  - If the sample sizes are unequal, then you might want to worry about homogeneity of variance. If $s_1^2 \approx s_2^2$, then you are probably also fine
- If you think you do not have homogeneity of variance, then you can run a revised version of the test (next time). Some people (including your textbook) recommend this as the default approach.

CONCLUSIONS

- comparing two means
- independent samples
- more flexible than one-sample case
- many more experiments can be tested
- same basic technique

NEXT TIME

- Welch's test
- Power

Planning a replication study

PSY 201: Statistics in Psychology

Lecture 26
Hypothesis testing for two means
Planning a replication study.

Greg Francis

Purdue University

Fall 2023

TESTING MEANS

- we want to test

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 \neq 0$$

- but the techniques of last time require $\sigma_1^2 = \sigma_2^2$
- pooled estimate of variance

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

- and use the t-distribution

REVISED TESTING MEANS

- when

$$\sigma_1^2 \neq \sigma_2^2$$

- we must make two changes
  - different estimate of standard error of the difference $s_{\overline{X}_1 - \overline{X}_2}$
  - adjustment of degrees of freedom
- still use the t distribution

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$$

- or

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{s_{\overline{X}_1}^2 + s_{\overline{X}_2}^2}$$

DEGREES OF FREEDOM

- when $\sigma_1^2 \neq \sigma_2^2$ we calculate df as:

$$df = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$$

- or

$$df = \frac{\left(s_{\overline{X}_1}^2 + s_{\overline{X}_2}^2\right)^2}{\left(s_{\overline{X}_1}^2\right)^2/(n_1 - 1) + \left(s_{\overline{X}_2}^2\right)^2/(n_2 - 1)}$$

- looks (and is) messy
- just a matter of plugging in numbers carefully
- still use the t-test as before!
- We call it Welch's test

EXAMPLE

- A researcher wants to know if single or married parents are more satisfied with their status. She randomly samples 61 single and 161 married parents. Each parent rates her/his marital status satisfaction, with higher scores indicating greater satisfaction. The researcher wants to know if there is a difference between the population means of single versus married parents.
- data summary

Variable	n	$\overline{X}$	s	$s_{\overline{X}}$
Group 1	61	2.6557	0.602	0.077
Group 2	161	2.7516	0.461	0.036

HYPOTHESES

$$H_0 : \mu_1 - \mu_2 = 0$$

- indicating there is no difference in satisfaction between the two groups

$$H_a : \mu_1 - \mu_2 \neq 0$$

- indicating there is a difference in satisfaction between the two groups
- not an ordered hypothesis because we do not know who might be more satisfied
- level of significance is set at $\alpha = 0.05$

WORRY ABOUT HOMOGENEITY

- We do not know the true values of σ_1 and σ_2, but we notice that $n_1 < n_2$ and that $s_1 > s_2$.
- This makes us worry that maybe our Type I error rate will be off (and maybe too big), so we use Welch's t-test
- The online calculator in the textbook uses Welch's test unless $n_1 = n_2$

TEST STATISTIC

- pooled standard error estimate

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$$

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\left(\frac{(0.602)^2}{61} + \frac{(0.461)^2}{161}\right)}$$

$$s_{\overline{X}_1 - \overline{X}_2} = 0.075683$$

TEST STATISTIC

- the formula for the test statistic is

$$\text{Test statistic} = \frac{\text{Statistic} - \text{Parameter}}{\text{Standard Error}}$$

- or

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_{\overline{X}_1 - \overline{X}_2}}$$

- or

$$t = \frac{(2.6557 - 2.75162) - (0)}{0.075683} = -1.267$$

TEST STATISTIC

- adjusted degrees of freedom

$$df = \frac{\left(s^2_{\overline{X}_1} + s^2_{\overline{X}_2}\right)^2}{\left(s^2_{\overline{X}_1}\right)^2/(n_1 - 1) + \left(s^2_{\overline{X}_2}\right)^2/(n_2 - 1)}$$

$$df = \frac{\left((0.077)^2 + (0.036)^2\right)^2}{\left((0.077)^2\right)^2/(61 - 1) + \left((0.036)^2\right)^2/(161 - 1)}$$

$$df = 87.995$$

p VALUE

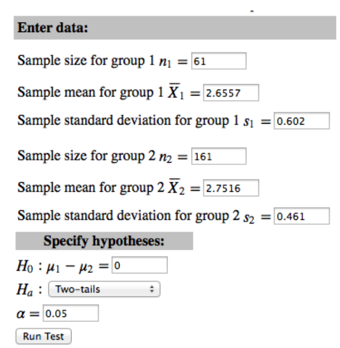
- From the t-distribution calculator, we find (for a two-tailed test with $df = 87.995$) that

$$p = 0.208455 > \alpha = 0.05$$

- we do not reject H_0
  - there is no evidence that satisfaction with marital status differs for married versus single parents
  - the probability that the observed (or more extreme) difference in means would occur by chance if $\mu_1 - \mu_2 = 0$ is greater than 0.05

ONLINE CALCULATOR

- As always, it is best to use a computer. We can enter the summary statistics.



ONLINE CALCULATOR

- You need to understand how to pull out the information you want

Test summary	
Type of test	Welch's Test
Null hypothesis	$H_0 : \mu_1 - \mu_2 = 0$
Alternative hypothesis	$H_a : \mu_1 - \mu_2 \neq 0$
Type I error rate	$\alpha = 0.05$
Label for group 1	Group 1
Sample size 1	$n_1 = 61$
Sample mean 1	$\overline{X}_1 = 2.6557$
Sample standard deviation 1	$s_1 = 0.602000$
Label for group 2	Group 2
Sample size 2	$n_2 = 161$
Sample mean 2	$\overline{X}_2 = 2.7516$
Sample standard deviation 2	$s_2 = 0.461000$
Pooled standard deviation	$s =$NA
Sample standard error	$s_{\overline{X}_1-\overline{X}_2} = 0.075683$
Test statistic	$t = -1.267121$
Degrees of freedom	$df = 87.99504946388605$
p value	$p = 0.208455$
Decision	Do not the reject null hypothesis
Confidence interval critical value	$t_{cv} = 1.987291$
Confidence interval	CI_{95}=(-0.246305, 0.054505)

CONFIDENCE INTERVAL

- Basic formula for all confidence intervals:

$$CI = \text{statistic} \pm (\text{critical value})(\text{standard error})$$

- for a difference of sample means

$$CI = (\overline{X}_1 - \overline{X}_2) \pm t_{cv} s_{\overline{X}_1 - \overline{X}_2}$$

- We already have most of the terms (we get t_{cv} from the Inverse t-distribution calculator, so

$$CI_{95} = (2.6557 - 2.7516) \pm (1.987291)(0.075683)$$

$$CI_{95} = (-0.246305, 0.054505)$$

POWER

- Power is treated much the same as for the one-sample case
- We just have to keep track of whether we are using the standard t-test or Welch's test
- The on-line calculator of our textbook does this for you automatically
- Power is very important when designing an experiment

REPLICATION

- An important characteristic of science is *replication*
- Show that the same methods and measures produce the same results
- "Hard" sciences are very good at this (e.g., physics, chemistry)
- Sciences that depend on statistics face challenges
- We *always* face a risk of making a Type I or a Type II error
- Thus, successful replication is not expected even for real effects
- You can mitigate these problems by designing good replication studies that use the same methods, but have high power

INTERESTING STUDY

- Consider a study on how nonconformity can induce higher status in certain environments
- Participants were 52 shop assistants working in downtown Milan, Italy boutiques (Armani, Burberry, Christian Dior, La Perla, Les Copains, and Valentino)
- Two groups of 26 each read a vignette:
- Imagine that a woman is entering a luxury boutique in downtown Milan during summer. She looks approximately 35 years old.
- Nonconforming condition (Group 1): She is wearing plastic flip-flops and she has a Swatch on her wrist.
- Conforming condition (Group 2): She is wearing sandals with heels and she has a Rolex on her wrist.
- Rate the status of the woman on a scale of 1–7 (bigger means higher status)

INTERESTING STUDY

- The results are:
- Nonconforming

$$\overline{X}_1 = 4.8$$

- Conforming

$$\overline{X}_2 = 4.2$$

$$t(50) = 2.1$$

$$p = 0.0408$$

REPLICATION

- You want to repeat the study, but it is not easy to get shop assistants from high end stores (you might have to go to Chicago for your subjects)
- The online power calculator requires you to enter estimates of:

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_0 : \mu_{a1} - \mu_{a2} \approx \overline{X}_1 - \overline{X}_2 = 0.6$$

$$\sigma_1, \sigma_2$$

REPLICATION

- for the standard deviations, we use some algebra. We know that for the reported t-test:

$$2.1 = t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} = \frac{0.6}{s_{\overline{X}_1 - \overline{X}_2}}$$

- so

$$s_{\overline{X}_1 - \overline{X}_2} = \frac{0.6}{2.1} = 0.28571$$

- We can assume the standard t-test was used, so $\sigma_1 = \sigma_2$. Thus

$$s_{\overline{X}_1 - \overline{X}_2} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = s\sqrt{\frac{1}{26} + \frac{1}{26}} = s(0.27735)$$

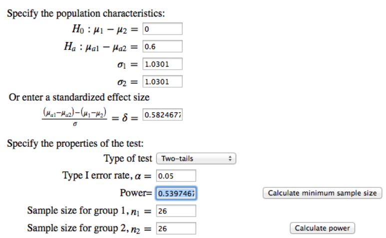
- so

$$s = \frac{0.28571}{0.27735} = 1.030157$$

- which we can use for both σ_1 and σ_2

REPLICATION

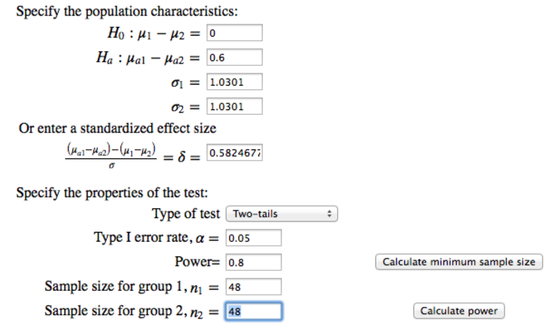
- Oftentimes researchers just use the same sample size as a previous study. After all, that study worked, so it must be an appropriate sample size, right?
- No, if we use $n_1 = n_2 = 26$, the on-line power calculator gives power=0.5397



- this should make sense because the $p = 0.04$ in the original study is just below the $\alpha = 0.05$ criterion
- if we take a different random sample, we will get a different p value, almost half the time it will be bigger than α

REPLICATION

- Suppose you want 80% power
- The calculator tells you that you need $n_1 = n_2 = 48$ participants. Nearly twice as big as the original study!



- if you do the replication correctly, you typically run a *better* study than the original
- That is common in science, where new experiments are better than old experiments

EFFECT SIZE

- The power calculator computes a term called δ. This is an estimate of d' between the null and specific alternative distributions. Bigger values of δ mean it is easier to notice a difference. It can be computed from the means and standard deviation estimates that you provide to the power calculator.
- An estimate, d, can also be computed from the t value and sample sizes

$$d = t\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (2.1)\sqrt{\frac{1}{26} + \frac{1}{26}} = 0.5824$$

- Often called Cohen's d

EFFECT SIZE

- You might worry that the effect size of the original study is an overestimate
- After all, if the researchers had not found a significant difference, they might not have published their paper (publication bias)
- A conservative approach is to divide the estimated effect size half, and do the power calculation from that new effect size.
- Thus, we can directly enter:

$$\delta = \frac{d}{2} = \frac{0.5824}{2} = 0.2912$$

- The power calculator now tells us that to have 80% power, we need $n_1 = n_2 = 187$ subjects
- This could be a very difficult experiment to run

CONCLUSIONS

- Welch's test
- Power
- Replication

NEXT TIME

- hypothesis testing for dependent samples
- sampling distribution
- standard error

Within is better than between.

PSY 201: Statistics in Psychology

Lecture 27

Hypothesis testing for dependent sample means

Within is better than between.

Greg Francis

Purdue University

Fall 2023

DEPENDENT SAMPLES

- two samples of data are dependent when each score in one sample is paired with a specific score in the other sample
- e.g
  - testing the same set of subjects before and after treatment
  - matched subjects in two groups (match along IQ before treatment and test mathematics)

CORRELATED DATA

- when samples are dependent, the scores across samples may be correlated
- suggests that you can (partly) predict one from other
- removes some of the randomness from the samples
- generally a good thing (more control over variables)
- but requires slightly different analysis

VARIABLE

- the variable of interest for dependent groups is the difference scores

$$d_i = X_{1i} - X_{2i}$$

- where
  - the ith scores in each group are matched (same subject)
  - X_{1i} is the ith score in the first group
  - X_{2i} is the ith score in the second group
- note that for dependent groups $n_1 = n_2 = n$, so we can calculate n difference scores

HYPOTHESIS

- we can calculate the mean of the difference scores for the sample

$$\overline{d} = \frac{\Sigma d_i}{n}$$

- which is the same as

$$\overline{d} = \overline{X}_1 - \overline{X}_2$$

- we would like to know if the mean of difference scores for the **population** $(\mu_1 - \mu_2)$ is different from zero

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 \neq 0$$

- This is actually the same as a one-sample t test!

STANDARD ERROR

- we estimate the standard error with

$$s_{\overline{d}} = \frac{s_d}{\sqrt{n}}$$

- where s_d is the standard deviation of the difference scores

$$s_d = \sqrt{\frac{\Sigma\left(d_i - \overline{d}\right)^2}{n-1}}$$

- or in raw score form

$$s_d = \sqrt{\frac{\Sigma d_i^2 - \left(\Sigma d_i\right)^2/n}{n-1}}$$

TEST STATISTIC

- the test statistic is the same in form as all those before

$$\text{Test statistic} = \frac{\text{Statistic} - \text{Parameter}}{\text{Standard Error}}$$

- for our specific situation it is

$$t = \frac{\overline{d} - (\mu_1 - \mu_2)}{s_{\overline{d}}}$$

- which is used to compute a p-value in a t-distribution with $n-1$ degrees of freedom

EXAMPLE

- do your thoughts control your autonomic processes?
- relax and take your pulse for 30 seconds
  - write the number down (X_1)
- picture yourself running and then take your pulse for 30 seconds
  - write the number down (X_2)
- we want to know if the mean difference across the two measurements (samples) is different from zero

EXAMPLE

- the measurements are dependent because if you tend to have a high pulse rate, it will be high for both measurements
- but we are interested in the **difference**, so the overall rate is unimportant
- calculate the difference of your pulse rates

$$d_i = X_{1i} - X_{2i}$$

HYPOTHESIS

- (1) our null hypothesis is that there is no effect of imagination on pulse rate

$$H_0 : \mu_1 - \mu_2 = 0$$

- the alternative hypothesis is that there is an effect

$$H_a : \mu_1 - \mu_2 < 0$$

- note, this is a directional hypothesis because we suspect that thinking about exercise should **increase** heart rate
- we will use a level of significance of $\alpha = 0.05$

DATA

- take a sample of pulse rate differences from ten people
- with your sampled data calculate the sample mean

$$\overline{d} = \frac{\sum d_i}{10}$$

- and the sample standard deviation (you can use the on-line calculator)

$$s_d = \sqrt{\frac{\Sigma d_i^2 - (\Sigma d_i)^2 / 10}{9}}$$

- and the estimate of the standard error

$$s_{\overline{d}} = \frac{s_d}{\sqrt{10}}$$

TEST STATISTIC

- (2) now calculate the test statistic as

$$t = \frac{\overline{d} - (\mu_1 - \mu_2)}{s_{\overline{d}}}$$

- since we assume $\mu_1 - \mu_2 = 0$ this is

$$t = \frac{\overline{d}}{s_{\overline{d}}}$$

p VALUE

- (3) You can compute the corresponding p-value with the t-Distribution Calculator for a one-tailed test
- with your sample of 10 people you have

$$df = n - 1 = 9$$

- if

$$p < \alpha = 0.05$$

- you reject H_0

DECISION

- (4) if you reject H_0 that means there is evidence that imagination of exercise **does** affect heart rate
- if you do not reject H_0 that means there is no evidence that imagination of exercise affects heart rate
- if you reject, that means that if $\mu_1 - \mu_2 = 0$, then the probability of the observed (or a more extreme) sample mean $\overline{d}$ value is less than 0.05

SIGNIFICANCE VS. IMPORTANCE

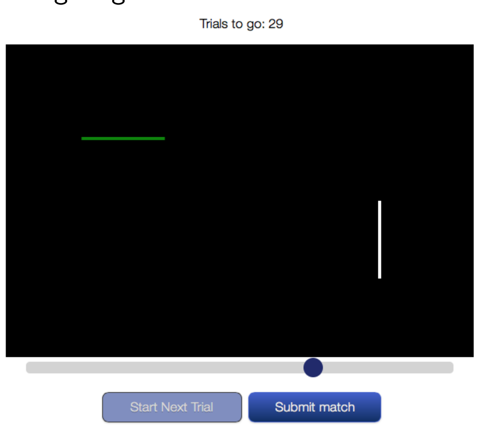
- if you failed to reject H_0, it may have been because you had too small a sample, n,
- or may have been because there was no real difference
- in principle, it is hard to believe that imagined running has **no effect at all** on pulse rate
  - Surely the brain uses energy differently during imagined running compared to not
- the effect might be very small, so small that our experiment cannot find it

SIGNIFICANCE VS. IMPORTANCE

- on the other hand
- probably everyone had a sample difference $\overline{d}$ that was non-zero
- but some people probably did not reject H_0
- we cannot just look at numbers like $\overline{d}$ and take them at face value
- the statistical procedures keep us from rushing to conclusions that are unwarranted

POWER

- The computation of power is very similar (actually, identical) to the one-sample t-test situation
- Consider the STATLAB Horizontal-Vertical illusion. Across the class, we have data from 26 students that reports the mean (for each student) matching length for a horizontal and a vertical line.



POWER

- Suppose you want to test for a difference between matching lengths for a horizontal and vertical line.

$$H_0 : \mu_1 - \mu_2 = 0$$

- How many subjects should you use to have 90% power?
- We can use the STATLAB data to estimate power and then compute an appropriate sample size
- From the STATLAB data we find:

$$\overline{X}_1 - \overline{X}_2 = 99.3213 - 105.6313 = -6.31$$

$$s_d = 4.655993$$

ONLINE CALCULATOR

Specify the population characteristics:

$H_0 : \mu_1 - \mu_2 =$ 0

$H_a : \mu_{a1} - \mu_{a2} =$ -6.31

Enter the standard deviation of each population and the correlation between scores...

$\sigma_1 =$ NA

$\sigma_2 =$ NA

Population correlation: $\rho =$ NA

...or enter the standard deviation of difference scores:

$\sigma_d =$ 4.6559

Or enter a standardized effect size:

$\frac{(\mu_{a1}-\mu_{a2})-(\mu_1-\mu_2)}{\sigma_d} = \delta =$ -1.355269

Specify the properties of the test:

Type of test: Two-tails

Type I error rate, $\alpha =$ 0.05

Power= 0.9 — Calculate minimum sample size

Sample size (pairs of scores) $n =$ 8 — Calculate power

ONLINE CALCULATOR

- You get exactly the same numbers using the one-sample calculator:

Specify the population characteristics:

$H_0 : \mu_0 =$ 0

$H_a : \mu_a =$ -6.31

$\sigma =$ 4.6560

Or enter a standardized effect size

$\frac{\mu_a-\mu_0}{\sigma} = \delta =$ -1.35524C

Specify the properties of the test:

Type of test: Two-tails

Type I error rate, $\alpha =$ 0.05

Power= 0.9 — Calculate minimum sample size

Sample size, $n =$ 8 — Calculate power

ONLINE CALCULATOR

- Instead of using s_d (the standard deviation of the differences), you could use the standard deviation of each group and the correlation between scores:

Specify the population characteristics:

$H_0 : \mu_1 - \mu_2 =$ 0

$H_a : \mu_{a1} - \mu_{a2} =$ -6.31

Enter the standard deviation of each population and the correlation between scores...

$\sigma_1 =$ 3.5502

$\sigma_2 =$ 6.4328

Population correlation: $\rho =$ 0.7073

...or enter the standard deviation of difference scores:

$\sigma_d =$ 4.656027(

Or enter a standardized effect size:

$\frac{(\mu_{a1}-\mu_{a2})-(\mu_1-\mu_2)}{\sigma_d} = \delta =$ -1.355232

Specify the properties of the test:

Type of test: Two-tails

Type I error rate, $\alpha =$ 0.05

Power= 0.9 — Calculate minimum sample size

Sample size (pairs of scores) $n =$ 8 — Calculate power

INDEPENDENT TEST

- Suppose you wanted to do the experiment with different subjects assigned to different line orientations. How many subjects do you need?
- Independent Means Power Calculator

Specify the population characteristics:

$H_0 : \mu_1 - \mu_2 =$ 0

$H_a : \mu_{a1} - \mu_{a2} =$ -6.31

$\sigma_1 =$ 3.5502

$\sigma_2 =$ 6.4328

Or enter a standardized effect size

$\frac{(\mu_{a1}-\mu_{a2})-(\mu_1-\mu_2)}{\sigma} = \delta =$ NA

Specify the properties of the test:

Type of test: Two-tails

Type I error rate, $\alpha =$ 0.05

Power= 0.9 — Calculate minimum sample size

Sample size for group 1, $n_1 =$ 16

Sample size for group 2, $n_2 =$ 16 — Calculate power

- 4 times as many subjects for an independent means experiment!

WITHIN vs. BETWEEN

- A within subjects (dependent test) design is usually more powerful than a between subjects (independent test)
- This is because you are able to remove one source of variability from the standard error
  - The variability in overall score values
- Standard error reflects variability between conditions
- A between subjects calculation of variability includes variability between conditions and variability across subjects (more variability!)

CONCLUSIONS

- dependent samples
- very important for lots of interesting tests
- more powerful than independent tests

NEXT TIME

- two-sample case for independent proportions
- hypothesis testing
- confidence interval
- power

What is a "margin of error"?

PSY 201: Statistics in Psychology

Lecture 28

Hypothesis testing for independent proportions

What is a "margin of error"?

Greg Francis

Purdue University

Fall 2023

PROPORTIONS

- we want to test hypotheses about proportions of populations

$$H_0 : P_1 = P_2$$

$$H_a : P_1 \neq P_2$$

- or, the same thing is

$$H_0 : P_1 - P_2 = 0$$

$$H_a : P_1 - P_2 \neq 0$$

- (or we could use directional hypotheses)

SAMPLING DISTRIBUTION

- the statistic is the difference between two **independent** sample proportions

$$p_1 - p_2$$

- need to know sampling distribution and standard error
- turns out that for large sample sizes (and some constraints on the proportions), the sampling distribution is approximately normal with a mean equal to the difference of population proportions

$$P_1 - P_2$$

STANDARD ERROR

- the standard error of the difference between independent proportions is

$$s_{p_1-p_2} = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

  - where p is the average proportion across the groups

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

  - or an equivalent formula is

$$p = \frac{f_1 + f_2}{n_1 + n_2}$$

  - with f_1 and f_2 being the **frequencies** of occurrences in each sample, respectively.
  - Also

$$q = 1 - p$$

INDEPENDENT SAMPLES

- we now have everything we need to carry out hypothesis testing of proportions for the two-sample case when the samples are independent
- That is we can calculate the test statistic

$$z = \frac{(p_1 - p_2) - (P_1 - P_2)}{s_{p_1-p_2}}$$

- when $H_0 : P_1 - P_2 = 0$ this is simply

$$z = \frac{(p_1 - p_2)}{s_{p_1-p_2}}$$

- and look up a p-value with the normal distribution calculator

CONFIDENCE INTERVALS

- we can create confidence intervals too
- the general formula is

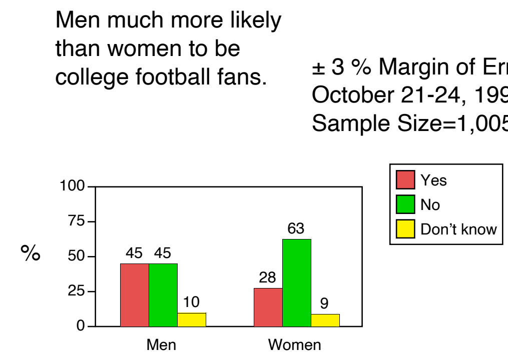
$$CI = \text{statistic} \pm (\text{critical value }) \times (\text{standard error})$$

- for the difference of proportions it becomes

$$CI = (p_1 - p_2) \pm (z_{cv})(s_{p_1 - p_2})$$

AN EXAMPLE

- A Gallup poll sampled 1005 adults and asked, "Are you a fan of college football, or not?"



FOOTBALL

- Is the sex difference a real difference between populations?
- (1) State the hypothesis.

$$H_0 : P_1 - P_2 = 0$$

$$H_a : P_1 - P_2 \neq 0$$

- Set the criterion
  - we'll use $\alpha = 0.01$

FOOTBALL

- (2) Find the test statistic
- We are interested in the proportion of people who answer "yes" to the question.
- We really need to know n_1 and n_2, but we do not have that information. We'll assume $n_1 = 502$ and $n_2 = 503$, for males and females, respectively.
- We need p, the average proportion across the groups,

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{(502)(0.45) + (503)(0.28)}{1005} = 0.364$$

- and q

$$q = 1 - p = 1 - 0.364 = 0.636$$

FOOTBALL

- we use p and q to get standard error

$$s_{p_1-p_2} = \sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$= \sqrt{(0.364)(0.636)\left(\frac{1}{502}+\frac{1}{503}\right)} = 0.03035$$

- and can now compute the test statistic

$$z = \frac{(p_1-p_2)-(P_1-P_2)}{s_{p_1-p_2}}$$

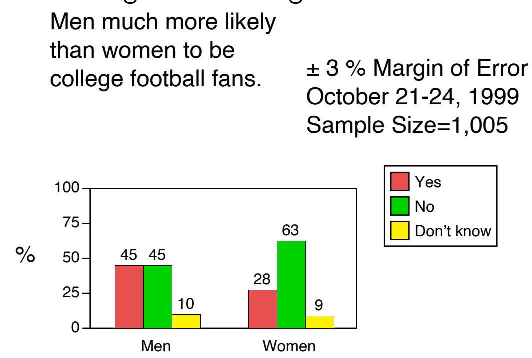
$$z = \frac{(0.45-0.28)-0}{0.03035} = 5.6013$$

- (3) Find the p-value from the normal distribution calculator

$$p \approx 0$$

FOOTBALL

- (4) Make a decision
- Reject H_0. The samples suggest that men and women have different proportions of being fans of college football



ONLINE CALCULATOR

- Small rounding differences

Enter data:

Sample size for group 1: $n_1 =$ 502

Number of scores with trait for group 1: $f_1 =$ 226

Sample size for group 2: $n_2 =$ 503

Number of scores with trait for group 2: $f_2 =$ 141

Specify hypotheses:

$H_0 : P_1 - P_2 =$ 0

$H_a :$ Two-tails

$\alpha =$ 0.01

Run Test

Test summary	
Null hypothesis	$H_0 : P_1 - P_2 = 0$
Alternative hypothesis	$H_a : P_1 - P_2 \neq 0$
Type I error rate	$\alpha = 0.01$
Sample size for group 1	$n_1 = 502$
Sample size for group 2	$n_2 = 503$
Sample proportion for group 1	$p_1 = 0.4502$
Sample proportion for group 2	$p_2 = 0.2803$
Pooled proportion	$p = 0.3652$
Standard error	$s_{p_1-p_2} = 0.030376$
Test statistic	$z = 5.592688$
p value	$p = 0.000000$
Decision	Reject the null hypothesis
Confidence interval critical value	$z_{cv} = 2.576236$
Confidence interval	$CI_{99} = (0.091626, 0.248136)$

FOOTBALL

- note, we take a sample of 1005 people, and we draw conclusions about *everyone* in the US
- that is remarkable, and it works because we know the properties of the sampling distribution
- of course, our conclusion could be wrong. There is a chance, $\alpha < 0.01$, that even if H_0 were true that we would get a difference of sample proportions like this.

FOOTBALL

- what's that margin of error about?
- You see in lots of polls that there is a "margin of error of $\pm 3\%$" (or $\pm 5\%$,...)
- It's the range of a confidence interval

$$CI = (p_1 - p_2) \pm (z_{cv})(s_{p_1 - p_2})$$

$$CI_{99} = (0.45 - 0.28) \pm (2.576)(0.03035)$$

$$CI_{99} = (0.45 - 0.28) \pm 0.07818$$

- going 0.08 (or 8%) above and below the difference of the sample proportions
- where does 3% come from?

FOOTBALL

- Try building a CI_{95}

$$CI = (p_1 - p_2) \pm (z_{cv})(s_{p_1 - p_2})$$

$$CI_{95} = (0.45 - 0.28) \pm (1.96)(0.03035)$$

$$CI_{90} = (0.45 - 0.28) \pm 0.059486$$

- going 0.06 (or 6%) above and below the difference of the sample proportions
- where does 3% come from?

MARGIN OF ERROR

- In this particular case, the Gallup organization seems to be reporting the $\pm$ range of a 95% confidence interval for the proportion of the entire set of data, under the worst case (when s_p is as big as it possibly could be)

$$CI = p \pm (z_{cv})(s_p)$$

$$s_p = \sqrt{\frac{p(1-p)}{n}}$$

- is biggest for $p = 0.5$
- with $n = 1005$

$$s_p = \sqrt{\frac{0.5(0.5)}{1005}} = 0.01577$$

- for a 95% CI,

$$z_{cv} = 1.96$$

- so

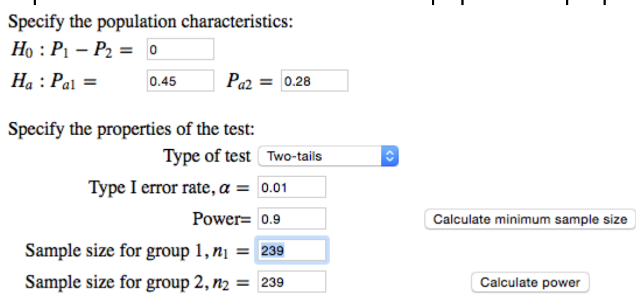
$$CI_{95} = p \pm (1.96)(0.01577) = p \pm 0.03091$$

MARGIN OF ERROR

- Journalists often refer to a margin of error, but they often do not really explain how it is calculated
- It *does* give an indication of how much variability is in the data
- If described properly, it could give us some information about the confidence of the estimate. But the confidence is almost never given
- I can make a CI (and the margin of error) big or small by varying my desired confidence.
- Moreover, remember, margin of error is not absolute, but only with confidence! (Sometimes the CI is wrong, and the "margin of error" is much bigger than the CI indicates.)

POWER

- This study is rather old (1999). You might decide to check whether there are similar results today. To design a new experiment, you can use the previous data as estimates of the population proportions.



POWER

- You have to think carefully about what you are going to compare. For example, rather than repeat the same experiment for people in 2018, you might be interested in checking whether the proportion of women who like college football has changed since 1999:

$$H_0 : P_1 - P_2 = 0$$

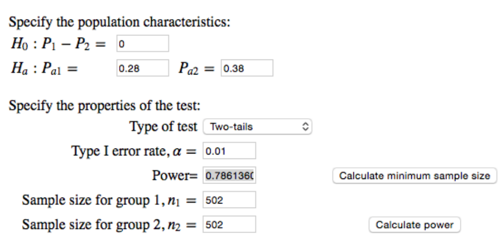
- That would be a different experiment, and thus it requires a different power analysis.

POWER

- For example, maybe an advertiser is willing to reconsider placing ads targeted to women during college football games, provided the proportion has increased by at least 0.1 since 1999.
- Then, our specific values for the alternative hypothesis are:

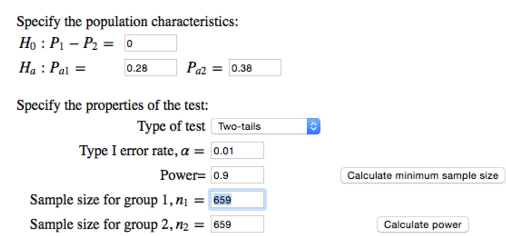
$$H_a : P_{a1} = 0.28, P_{a2} = 0.38$$

- If we use the same sample sizes as the 1999 study ($n_1 = 502$, $n_2 = 502$)



POWER

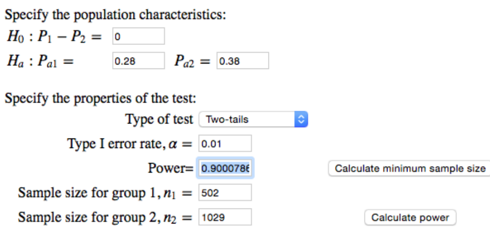
- To have 90% power for this study, we can find the minimum sample size



- but this does not help us much, as the sample from 1999 is fixed at $n_1 = 502$.

POWER

- We can fix n_1=502 and increase the n_2 value until we get a power of 0.9



- It takes a quite large sample to detect a 0.1 difference in proportions!

CONCLUSIONS

- two-sample case
- independent proportions
- confidence interval
- power

NEXT TIME

- two-sample case
- dependent proportions
- confidence interval
- power (tricky)

What do people think about death?

PSY 201: Statistics in Psychology

Lecture 29

Hypothesis testing for dependent proportions

What do people think about death?

Greg Francis

Purdue University

Fall 2023

DEPENDENT SAMPLES

- when the samples are not independent, hypothesis testing of proportions becomes a bit more complicated
- samples are dependent when each score in one sample is paired with a score in the other sample
- just like dependent samples for the mean, the problem is that the samples are not independent (not truly random) and we need to take that into account
  - This can be a good thing from a statistical point of view
  - We can remove some variability

EXAMPLE

- Testing the difference of proportions of individuals who pass each of two similar items on a test. (e.g. comparing pass/fail for two sets of students who get better than 600 SAT)
- Test the difference in proportions of individuals who support something before and after discussion.
- Comparing proportions of husbands and wives on an issue.

HYPOTHESES

- for dependent samples we set our hypotheses as

$$H_0 : P_1 - P_2 = 0$$

$$H_a : P_1 - P_2 \neq 0$$

SAMPLING DISTRIBUTION

- we need to know the sampling distribution and the standard error
- but first we need to design a contingency table that shows disagreement or dissimilarity in responses

		Group 2		
		NO	YES	
Group 1	YES	A	B	$A+B$
	NO	C	D	$C+D$
		$A+C$	$B+D$	$A+B+C+D=n$

- A is the number of scores that are "no" in group 2 and "yes" in group 1
- B is the number of scores that are "yes" in group 2 and "yes" in group 1
- C is the number of scores that are "'no" in group 2 and "no" in group 1
- D is the number of scores that are "yes" in group 2 and "no" in group 1

CONTINGENCY TABLE

- we then convert these to proportions

	Group 2		
	NO	YES	
Group 1 YES	a	b	$a+b$
NO	c	d	$c+d$
	$a+c$	$b+d$	$a+b+c+d=1.0$

- $a = A/n$ is the proportion of scores that are "no" in group 2 and "yes" in group 1
- $b = B/n$ is the proportion of scores that are "yes" in group 2 and "yes" in group 1
- $c = C/n$ is the proportion of scores that are "'no" in group 2 and "no" in group 1
- $d = D/n$ is the proportion of scores that are "yes" in group 2 and "no" in group 1

PROPORTIONS

- from the contingency table we can get the proportions of scores with the trait we are interested in
- this is what we need for our statistic

$$p_1 = b + d = \frac{B+D}{n}$$

$$p_2 = a + b = \frac{A+B}{n}$$

- but we need the contingency table for other reasons!

CONTINGENCY TABLES

- the sampling distribution is approximately normal with a mean of $P_1 - P_2$ if

$$A + D > 10$$

or

$$B + C > 10$$

- if not, do not use this test
- moreover, our estimate of standard error of the difference between dependent proportions is

$$s_{p_1-p_2} = \sqrt{\frac{a+d}{n}} = \sqrt{\frac{p_d}{n}} = \sqrt{\frac{(A+D)/n}{n}}$$

- which we get from the contingency table

HYPOTHESIS TESTING

- so to actually carry out the test, we compute the test statistic

$$z = \frac{(p_1 - p_2) - (P_1 - P_2)}{s_{p_1-p_2}}$$

- or, since our null hypothesis is that $(P_1 - P_2) = 0$

$$z = \frac{(p_1 - p_2)}{s_{p_1-p_2}}$$

- and then look up a p-value from the normal distribution calculator

EXAMPLE

- I took the Exam 1 grades and the Homework grades and for each student computed:
  - Trait 1: Grade on Exam 1 is ≥ 70 (C range)
  - Trait 2: Grade on Homework is ≥ 70 (C range)

Exam 1	Homework Grade
0	1
1	1
0	1
1	1
1	1
1	1
0	1
1	1
1	1
0	1
0	1
0	1
1	1
0	1
0	1
1	1
1	1
1	1
0	1
1	0
1	1
0	0
0	1
0	1
0	1
0	1
1	0
1	0

EXAMPLE

- We can test for a difference in proportions using the on-line calculator:

		Has Trait 2?	
		No	Yes
Has Trait 1?	Yes	$A =$ 3	$B =$ 11
	No	$C =$ 1	$D =$ 13

Specify hypotheses:

$H_0 : P_1 - P_2 =$ 0

$H_a :$ Two-tails

$\alpha =$ 0.05

Run Test

Test summary	
Null hypothesis	$H_0 : P_1 - P_2 = 0$
Alternative hypothesis	$H_a : P_1 - P_2 \neq 0$
Type I error rate	$\alpha = 0.05$
Sample size	$n = 28$
Sample proportion for group 1	$p_1 = 0.5000$
Sample proportion for group 2	$p_2 = 0.8571$
Disagreement proportion	$p_d = 0.5714$
Standard error	$s_{p_1 - p_2} = 0.142857$
Test statistic	$z = -2.500000$
p value	$p = 0.012419$
Decision	Reject the null hypothesis
Confidence interval critical value	$z_{cv} = 1.960395$
Confidence interval	$CI_{95} = (-0.637199, -0.077086)$

ANOTHER EXAMPLE

- Suppose we want to know if there is a difference in the proportion of students that oppose the death penalty and the proportion of students that support gun control.
- Raise your hand if (feel free to lie if you do not want others to know your true opinions)
  - A: You support gun control, but do not oppose the death penalty.
  - B: You support gun control and oppose the death penalty.
  - C: You do not support gun control and do not oppose the death penalty.
  - D: You do not support gun control, but do oppose the death penalty.

CONTINGENCY TABLE

	OPPOSE DEATH PENALTY		
SUPPORT	NO	YES	
GUN CONTROL	YES NO		

- I want to test

$$H_0 : P_1 - P_2 = 0$$
$$H_a : P_1 - P_2 \neq 0$$

- In words: the proportion of people supporting gun control is the same as the proportion of people who oppose the death penalty (individuals are always pro-life or pro-death)
- I will use $\alpha = 0.05$
- Note: Group 1 is the set of responses to the question about opposition to the death penalty
- Group 2 is the set of responses to the question about support of gun control

CRITERION

- I need to check if I can use the normal approximation to the sampling distribution
- check if

$$A + D > 10$$

or

$$B + C > 10$$

- if not, do not use this test
- We use the on-line calculator

INTERPRETATION

- If we reject H_0, that indicates the probability of getting the observed difference of proportions, or bigger difference, when the population parameters were equal is less than 0.05. We interpret that as meaning the population parameters are different.
- If we fail to reject H_0, that indicates the probability of getting the observed difference of proportions, or bigger, when the population parameters were equal is greater than 0.05. We do not have strong enough evidence to conclude that the population parameters are different.

CONFIDENCE INTERVAL

- easy to create confidence intervals too
- the general formula is

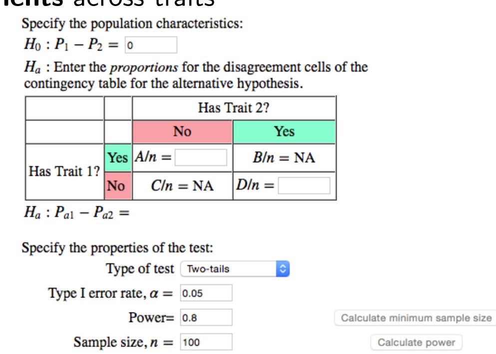
$$CI = \text{statistic} \pm (\text{critical value }) \times (\text{standard error})$$

- for the difference of dependent proportions it becomes

$$CI = (p_1 - p_2) \pm (z_{cv})(s_{p_1 - p_2})$$

POWER

- Computing power (and estimating minimum sample sizes) feels a bit awkward for dependent proportions
- Although you are testing the differences in proportions that *have* traits, the needed information is the proportions about **disagreements** across traits



- Let's look at an example

RETRIEVAL PRACTICE

- Studies have found that a good way to improve memory is to actively retrieve information from memory
  - practice test instead of study
- A common study goes like: Each subject reads two paragraphs about different topics (e.g., photosynthesis or leukemia). After reading:
  - For one paragraph the subject takes a practice test that requires them to recall information from the paragraph (e.g., "How does photosynthesis directly benefit our environment?")
  - For the other paragraph, the subject reads the paragraph a second time.
- A week later, subjects are tested on both paragraphs (new questions)

RETRIEVAL PRACTICE

- Typical results for correctly answering the final questions are something like:
- Retrieval practice: $p_1 = 0.44$
- Study: $p_2 = 0.28$
- From the raw data (you typically cannot get this information from what is published in scientific papers)

		Has Trait 2?	
		No	Yes
Has Trait 1?	Yes	$A =$ 6	$B =$ 10
	No	$C =$ 20	$D =$ 0

RETRIEVAL PRACTICE

- Suppose you want to explore retrieval practice in a new setting (statistics-related questions)
- The null hypothesis is no difference in proportions for retrieval versus study conditions

$$H_0 : P_1 - P_2 = 0$$

- the alternative hypothesis is that there is a some difference

$$H_a : P_1 - P_2 \neq 0$$

- You need a specific alternative hypothesis, and using the data from the previous study is a good starting point

$$H_a : P_{1a} - P_{a2} = 0.44 - 0.28 = 0.16$$

- but you specify it by identifying the disagreements in responses

RETRIEVAL PRACTICE

- We need

$$a = \frac{A}{n} = \frac{6}{36} = 0.1667$$

$$d = \frac{D}{n} = \frac{0}{36} = 0$$

- Suppose you want 0.9 for power

Specify the population characteristics:

$H_0 : P_1 - P_2 =$ 0

H_a : Enter the *proportions* for the disagreement cells of the contingency table for the alternative hypothesis.

		Has Trait 2?	
		No	Yes
Has Trait 1?	Yes	$A/n =$ 0.1667	$B/n =$ NA
	No	$C/n =$ NA	$D/n =$ 0

$H_a : P_{a1} - P_{a2} =$ 0.1667 - 0. = 0.1667

Specify the properties of the test:

Type of test: Two-tails

Type I error rate, $\alpha =$ 0.05

Power= 0.9 — Calculate minimum sample size

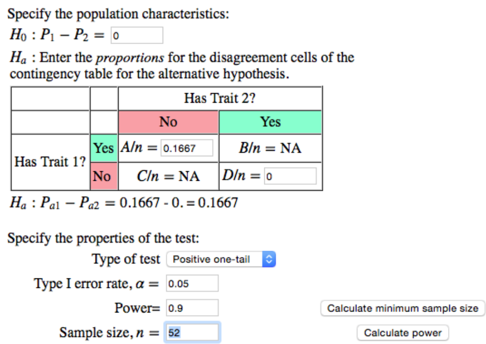
Sample size, $n =$ 64 — Calculate power

RETRIEVAL PRACTICE

- You might argue that a one-tailed test

$$H_0 : P_1 - P_2 > 0$$

- is appropriate because you know retrieval practice helps in most settings
- Then, to have 0.9 power:



CONCLUSIONS

- two-sample case
- dependent proportions
- confidence interval
- power

NEXT TIME

- Comparing two sample correlations
- Power

How careful are students?

PSY 201: Statistics in Psychology

Lecture 30

Hypothesis testing for two correlations

How careful are students?

Greg Francis

Purdue University

Fall 2023

CORRELATIONS

- you have two independent populations (tests with dependent samples are more complicated)
- each with two scores for which you can calculate a correlation coefficient. e.g.
  - male population of students
  - female population of students
- might want to compare correlations between verbal SAT and math SAT

HYPOTHESIS

- the null hypothesis to test is

$$H_0 : \rho_1 = \rho_2$$

- where ρ_1 is the correlation coefficient of scores in population 1
- and ρ_2 is the correlation coefficient of scores in population 2
- and the alternative hypothesis is

$$H_a : \rho_1 \neq \rho_2$$

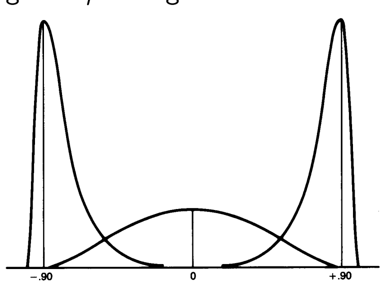
- or the same thing is:

$$H_0 : \rho_1 - \rho_2 = 0$$

$$H_a : \rho_1 - \rho_2 \neq 0$$

SAMPLING DISTRIBUTION

- to test H_0 we need to know the sampling distribution of the difference of correlation coefficients
- unfortunately, just like for the one-sample case, the sampling distribution changes as ρ changes



FISHER z TRANSFORMATION

$$z_r = \frac{1}{2} \log_e \left(\frac{1+r}{1-r} \right)$$

- which can be found by using the calculator in the textbook
- the sampling distribution of difference of z_r values is approximately normal and has a mean of

$$z_{\rho_1} - z_{\rho_2}$$

- which is zero, if H_0 is true

SAMPLING DISTRIBUTION

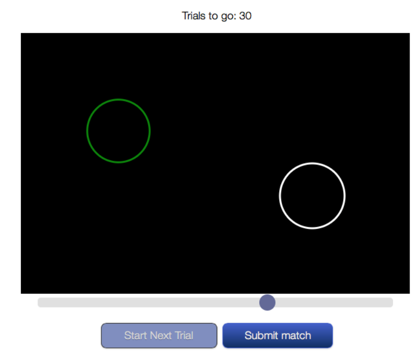
- so we know that the sampling distribution of the $z_{r_1} - z_{r_2}$ values is normally distributed and (if H_0 is true) has a mean of zero.
- all we need to know is the standard error of the difference between independent transformed correlation coefficients

$$s_{z_{r_1}-z_{r_2}} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$

- (note, you need more than three scores in each group)
- We just apply the same hypothesis testing approach as for other cases!

EXAMPLE

- In some tasks, correlation can be a measure of consistency or carefulness
- For example, in the STATLAB Weber's Law experiment, subjects adjust the size of a circle so it matches a target. There were two different target sizes, and we expect these matches to be correlated across subjects.



EXAMPLE

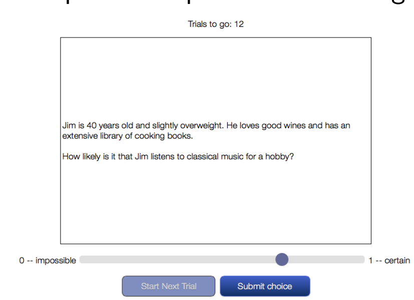
- Using the STATLAB data, we find that for the $n_1 = 30$ subjects who completed the Weber's Law lab, the correlation in matching sizes for the 10 pixel and 50 pixel targets is

$$r_1 = 0.1516$$

- In part this correlation reflects carefulness by the subject. If subjects are careless in their judgments, then they essentially add noise to their matching circles, and this will reduce the correlation

EXAMPLE

- In some tasks, correlation can be a measure of consistency or carefulness
- For example, in the STATLAB Typical Reasoning experiment, subjects rate the likelihood of certain characteristics of a described person. The descriptions were set up in a systematic way, so that some descriptions were expected to produce high ratings and other descriptions were expected to produce low ratings.



EXAMPLE

- Using the STATLAB data, we find that for the $n_2 = 29$ subjects who completed the Typical Reasoning lab, the correlation in likelihood ratings for the *Low typicality and two activities* and the *High typicality and two activities* is

$$r_2 = 0.1836$$

- In part this correlation reflects carefulness by the subject. If subjects are careless in their judgments, then they essentially add noise to their ratings, and this will reduce the correlation

EXAMPLE

- Are the correlations similar across the two tasks? They might seem like very different tasks, but to some extent, the correlations measure "effort" or "consistency" by the subjects.
- The overall strength of the correlation is less interesting than the similarity of the correlations. Some tasks may involve rather a lot of variability, so the correlation cannot be very large. Still, we can compare across tasks.
- Just looking at the correlations would not let us draw strong conclusions, but it could be part of a bigger argument.

EXAMPLE

- We use the on-line calculator to do the computations:

Sample correlation for group 1, $r_1 =$ 0.1516

Sample size for group 1, $n_1 =$ 30

Sample correlation for group 2, $r_2 =$ 0.1836

Sample size for group 2, $n_2 =$ 29

Specify hypotheses:

$H_0 : \rho_1 - \rho_2 =$ 0

$H_a :$ Two-tails

$\alpha =$ 0.05

Run Test

Test summary	
Null hypothesis	$H_0 : \rho_1 - \rho_2 = 0$
Difference of null Fisher z transforms	$z_{\rho_1} - z_{\rho_2} = 0.0000$
Alternative hypothesis	$H_a : \rho_1 - \rho_2 \neq 0$
Type I error rate	$\alpha = 0.05$
Label for group 1	
Sample size for group 1	$n_1 = 30$
Sample correlation for group 1	$r_1 = 0.1516$
Fisher z transform of r_1	$z_{r_1} = 0.1528$
Label for group 2	
Sample size for group 2	$n_2 = 29$
Sample correlation for group 2	$r_2 = 0.1836$
Fisher z transform of r_2	$z_{r_2} = 0.1857$
Sample standard error	$s_{z_r} = 0.274770$
Test statistic	$z = -0.119839$
p value	$p = 0.904611$
Decision	Do not the reject null hypothesis

MISSING?

- It is possible to compute a confidence interval for a difference of independent correlations.
- It is also possible to compute a hypothesis test for a difference of dependent correlations.
- However, these methods require some new ideas that we do not have time to go into (lots of special cases).

POWER

- Computing power for a test of independent correlations is conceptually similar to other power calculations
- however, the use of the Fisher z transform of correlations makes it difficult to have good intuition into how sample size relates to power
- Since we take the Fisher z transform of each correlation and then take the difference, a fixed difference of correlations does not necessarily produce a fixed difference of Fisher z transform values.

POWER

- For example,

$$r_1 - r_2 = 0.3 - 0.2 = 0.1$$

- corresponds to

$$z_{r_1} - z_{r_2} = 0.203 - 0.100 = 0.103$$

- but

$$r_1 - r_2 = 0.5 - 0.4 = 0.1$$

- corresponds to a larger value:

$$z_{r_1} - z_{r_2} = 0.549 - 0.424 = 0.125$$

POWER

- these differences mean that testing for a specific alternative

$$H_a : \rho_1 - \rho_2 = 0.5 - 0.4$$

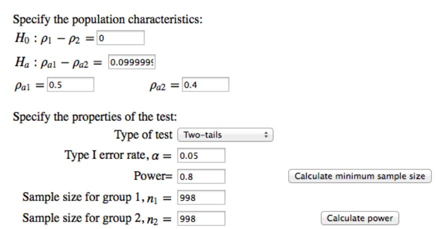
- is easier than testing for

$$H_a : \rho_1 - \rho_2 = 0.2 - 0.1$$

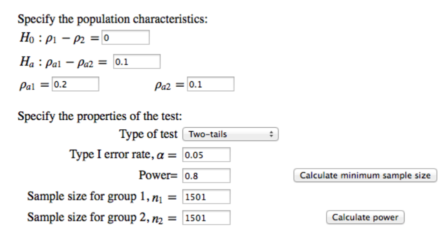
- Suppose you wanted 80% power for each test

POWER

$$H_a : \rho_1 - \rho_2 = 0.5 - 0.4$$



$$H_a : \rho_1 - \rho_2 = 0.2 - 0.1$$



- a difference of 1006 subjects across the two groups!

POWER

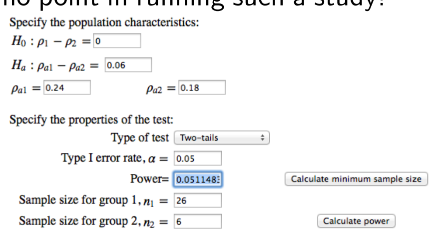
- Let's look at a specific example
- Height is correlated with economic success (income, wealth). Taller people are more successful
- Combining data across multiple studies suggests a difference in the correlation for men compared to women
  - Men: $r_1 = 0.24$
  - Women: $r_2 = 0.18$

POWER

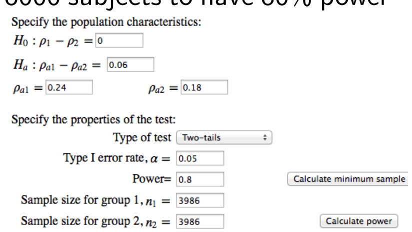
- Suppose you want to test this difference in correlations for Purdue engineering technology graduates. You will look at starting salaries and height.
- Engineering Technology degrees are given by Purdue Polytechnic, and each year it typically distributes BS degrees to 18 women and 78 men.
- You think you can get starting salary and height data for a third of the graduates. If the Purdue graduates are similar to the general population, what is the power of your study?

POWER

- there's no point in running such a study!



- You need a total of nearly 8000 subjects to have 80% power



- You simply cannot test for these kinds of small differences in correlations for relatively small populations (like Purdue graduates).

CONCLUSIONS

- two-sample case for correlations
- power
- trust the numbers!

NEXT TIME

- More than two comparisons
- Multiple testing

Error is sneaky.

PSY 201: Statistics in Psychology

Lecture 31

Multiple testing

Error is sneaky.

Greg Francis

Purdue University

Fall 2023

HYPOTHESIS TESTING

- we know how to test the difference of two means

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 \neq 0$$

- by using the t distribution and estimates of standard error
- what if you have more populations and what to know if they are all equal?

MULTIPLE t- TESTS

- if we have $K = 5$ population means, we might want to compare each mean to all the others
- requires

$$c = \frac{K(K-1)}{2} = 10$$

- different t-tests
- suppose each test is with $\alpha = 0.05$
- What is the Type I error?

MULTIPLE t- TESTS

- we have a risk of making a type I error for *each* t test
- since we have $c = 10$ different t-tests, with $\alpha = 0.05$, the Type I error rate becomes approximately

$$1 - (1 - \alpha)^c = 0.40$$

- bigger risk of error than you might expect!
- to be sure we do not make *any* Type I errors, we would need to set α much smaller to insure that Type I error rate is below 0.05!

ADJUST α

- To a first approximation, to make sure the Type I error rate for $c = 10$ tests is less than 0.05, you could set the α criterion for each t-test to be

$$\alpha = \frac{0.05}{c} = \frac{0.05}{10} = 0.005$$

- Then, the probability of any given test producing a Type I error is 0.005, and the probability than any of the 10 tests produces a Type I error is 0.05
- This is called the Bonferroni correction
- But decision making always involves trade offs.

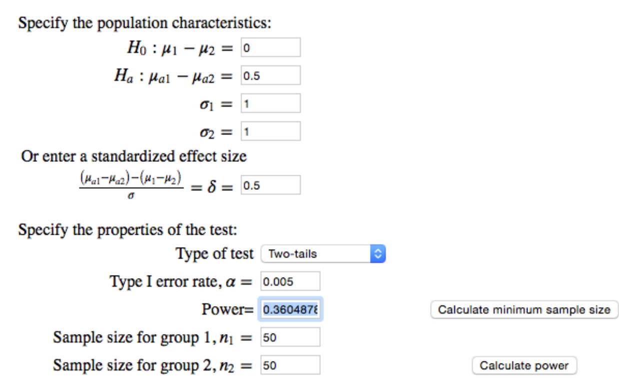
ADJUST α

- What kind of power do we have?
- Suppose $\sigma = 1$ and we take samples of size $n = 50$ for each condition
- If you use $\alpha = 0.005$, and one of the means, $\mu_1 = 0.5$, is *truly* different from the other four means, $\mu_2 = \mu_3 = \mu_4 = \mu_5 = 0$. What is the probability you will reject H_0?
- For the six tests that do not involve μ_1, the probability of any of them producing a Type I error is

$$1 - (1 - 0.005)^6 = 0.029$$

ADJUST α

- For the four tests involving μ_1, we can estimate power of each test, by using the on-line power calculator:



POWER

- We have four chances for one of the tests involving μ_1 to be significant, so the probably of at least one being significant is

$$1 - (1 - 0.360)^4 = 0.832$$

- On the other hand, the probably that each of those four experiments will reject H_0 is

$$(0.360)^4 = 0.0168$$

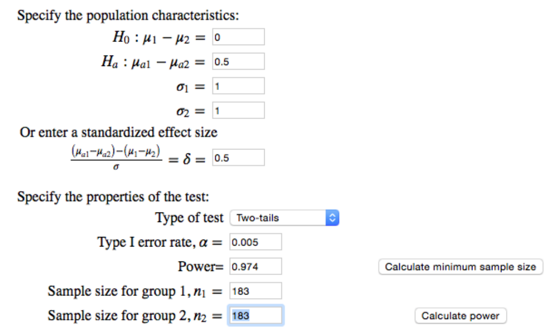
- So, you are almost surely going to draw *some* wrong conclusions

POWER

- If you want to have a 0.9 probability that all four tests involving μ_1 reject H_0, each test needs a power of

$$(0.9)^{1/4} = 0.974$$

- We can identify the required sample size for *each* condition



- this is an approximation because the tests are not independent

POWER

- Trying to control the error probabilities becomes complicated when you have multiple comparisons
- The probability of making at least one Type I error increases (power for detecting *something* increases)
- The probability of making at least one Type II error increases (power for the full set of differences decreases)
- This will always be true, but there are steps we can take to partially deal with the problem

DEMONSTRATION

- Open your five packages and count the number of *green* M&M's in each package. You have five numbers, $n = 5$, that make a sample
- within each sample, compute:

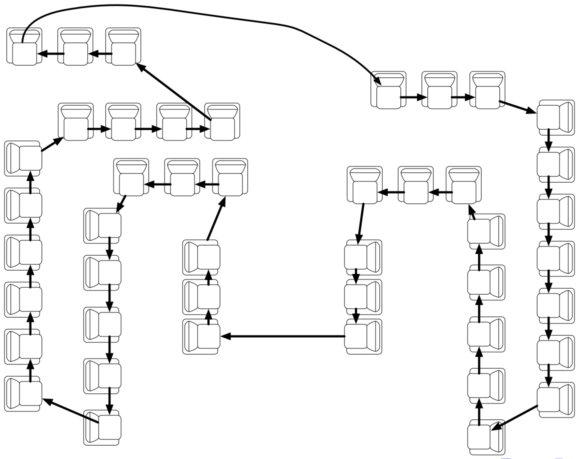
$$\overline{X}_k = \frac{\sum_i X_{ki}}{5}$$

$$s_k = \sqrt{\frac{\sum_i X_{ki}^2 - [(\sum_i X_{ki})^2/5]}{4}}$$

- Use the on-line calculator for *Descriptive Statistics*, if you want (use your phone). No need to log in.

DEMONSTRATION

- now, share your $\overline{X}_j$ and s_j with your "neighbor" by following the arrows below
- get $\overline{X}_j$ and s_j from your "neighbor"



DEMONSTRATION

- Run a hypothesis test to compare your mean to the mean of your neighbor
  - we'll assume homogeneity of variance
- (1) State the hypothesis:

$$H_0 : \mu_k = \mu_j$$

$$H_a : \mu_k \neq \mu_j$$

- and set the criterion
  $\alpha = 0.05$

DEMONSTRATION

- (2) Compute test statistic:

$$s^2 = \frac{(n_k - 1)s_k^2 + (n_j - 1)s_j^2}{n_k + n_j - 2} = \frac{(4)s_k^2 + (4)s_j^2}{8} =$$

$$s_{\overline{X}_k - \overline{X}_j} = \sqrt{s^2\left(\frac{1}{n_k} + \frac{1}{n_j}\right)} = \sqrt{s^2\left(\frac{1}{5} + \frac{1}{5}\right)} =$$

$$t = \frac{(\overline{X}_k - \overline{X}_j) - (0)}{s_{\overline{X}_k - \overline{X}_j}} =$$

$$df = n_k + n_j - 2 = 8$$

- (3) Compute the p-value using the t-distribution calculator
  - Instead we will identify the t_{cv} that corresponds to $p = 0.05$. It is $t_{cv} = 2.306$
- (4) Make a decision:

$$t = \quad < ? > \quad = \pm 2.306$$

DEMONSTRATION

- We know from the outset that H_0 is actually true here.

$$\mu_k = \mu_j$$

  - because all the samples are actually from the very same population (M&M packages from the same factory have a fixed ratio of colors)
- Still, just due to sampling errors, we expect to have some tests reject H_0. The probability of at least one is around:

$$1 - (1 - \alpha)^c =$$

- where c is the number tests (number of students in the class)

WHAT DO WE MAKE OF THIS?

- Not only is it a pain to compute multiple comparisons of means
- but it tends to lead to more Type I error than α indicates
- we could decrease α to a smaller value so that the overall Type I error is how we want it
- which will decrease power

CONCLUSIONS

- testing multiple means
- loss of control of Type I error

NEXT TIME

- there is a better method
- ANOVA
- two measures of variance

Measure twice, cut once.

PSY 201: Statistics in Psychology

Lecture 32

Analysis of Variance

Measure twice, cut once.

Greg Francis

Purdue University

Fall 2023

ANOVA VARIABLES

- independent variables: variable that forms groupings
- one-way ANOVA: one independent variable
- levels: number of groups, number of populations
- e.g. Method of teaching is an independent variable
- you may teach in 17 different ways (levels) and have 17 different sample groups with sample means

$$\overline{X}_1, \overline{X}_2, ...\overline{X}_{16}, \overline{X}_{17},$$

- so that for your hypothesis test you would want to test whether all the population means of the different levels are the same

ANOVA VARIABLES

- we need additional subscripts to keep track of variables

$$X_{ik}$$

- is the score for the ith subject in the kth level (group)

$$n_k$$

- is the number of scores in the kth level

$$\sum_i X_{ik}$$

- is the sum of scores in the kth level

$$\sum_k \sum_i^{n_k} X_{ik}$$

- is the sum of all scores

HYPOTHESES

- for one-way ANOVA the hypotheses are

$$H_0 : \mu_1 = \mu_2 = ... = \mu_K$$

$$H_a : \mu_i \neq \mu_k \text{ for some } i,\ k$$

- the null hypothesis is that all population means are the same
- the alternative hypothesis is that at least one mean is different from another

INTUITION

- the basic approach of ANOVA is to make two calculations of variance
  1. We can calculate variance of each group separately and combine them to estimate the variance of all scores. (within variance, s_W^2)
  2. We can also calculate the variance among all the group means, relative to a grand mean. (between variance, s_B^2)
- these estimates will be the same **if** H_0 is true!
- these estimates will be different **if** H_0 is not true!

INTUITION

- we compare the estimates using the F ratio

$$F = \frac{s_B^2}{s_W^2}$$

- f $F \approx 1$, do not reject H_0
- if $F > 1$, reject H_0
- how big depends on the sample sizes, significance, ...

SCORES

- what contributes to a particular score?
- assume a linear model

$$X_{ik} = \mu + \alpha_k + e_{ik}$$

  - X_{ik} is the ith score in the kth group
  - μ is the grand mean for the population, across all groups
  - $\alpha_k = \mu_k - \mu$ is the effect of belonging to group k
  - e_{ik} is random error associated with the score
- e_{ik} changes because of random sampling (normally distributed, mean of zero, σ^2)

SUM OF SQUARES

- we want to estimate σ^2 (variance of population if H_0 is true)
- need sum of squares

$$\Sigma_k \Sigma_i (X_{ik} - \overline{X})^2$$

- consider one score

$$(X_{ik} - \overline{X}) = (X_{ik} - \overline{X}_k) + (\overline{X}_k - \overline{X})$$

- so

$$(X_{ik} - \overline{X})^2 = [(X_{ik} - \overline{X}_k) + (\overline{X}_k - \overline{X})]^2$$

- or

$$(X_{ik} - \overline{X})^2 = (X_{ik} - \overline{X}_k)^2 + 2(\overline{X}_k - \overline{X})(X_{ik} - \overline{X}_k) + (\overline{X}_k - \overline{X})^2$$

SUM OF SQUARES

- if we sum across all subjects in category k

$$\sum_i^{n_k} (X_{ik} - \overline{X})^2 = \sum_i^{n_k} (X_{ik} - \overline{X}_k)^2 + 2(\overline{X}_k - \overline{X}) \sum_i^{n_k} (X_{ik} - \overline{X}_k) + \sum_i^{n_k} (\overline{X}_k - \overline{X})^2$$

- since deviations from a mean equal zero, this reduces to

$$\sum_i^{n_k} (X_{ik} - \overline{X})^2 = \sum_i^{n_k} (X_{ik} - \overline{X}_k)^2 + \sum_i^{n_k} (\overline{X}_k - \overline{X})^2$$

- in addition,

$$\sum_i^{n_k} (\overline{X}_k - \overline{X})^2 = n_k (\overline{X}_k - \overline{X})^2$$

- so we get

$$\sum_i^{n_k} (X_{ik} - \overline{X})^2 = \sum_i^{n_k} (X_{ik} - \overline{X}_k)^2 + n_k (\overline{X}_k - \overline{X})^2$$

SUM OF SQUARES

- now, we sum across the k groups to get the total sum of squares

$$\sum_k \sum_i (X_{ik} - \overline{X})^2 = \sum_k \left(\sum_i^{n_k} (X_{ik} - \overline{X}_k)^2 + n_k(\overline{X}_k - \overline{X})^2 \right)$$

- which becomes

$$\sum_k \sum_i (X_{ik} - \overline{X})^2 = \sum_k \sum_i^{n_k} (X_{ik} - \overline{X}_k)^2 + \sum_k n_k(\overline{X}_k - \overline{X})^2$$

- or

$$SS_T = SS_W + SS_B$$

- where
  - SS_T is the total sum of squares.
  - SS_W is the within sum of squares. Deviation of scores from the group mean.
  - SS_B is the between sum of squares. Deviation of group means from the grand mean.

WITHIN DEVIATIONS

$$SS_W = \sum_k \sum_i^{n_k} (X_{ik} - \overline{X}_k)^2$$

- what causes this to be greater than zero?
- since

$$X_{ik} = \mu + \alpha_k + e_{ik}$$

- $\mu + \alpha_k$ is fixed as i varies
- thus, deviations from $\overline{X}_k$ must be due to the e_{ik} term (random error)

ESTIMATE OF σ^2

- within each group, deviations from the mean are due to the error terms e_{ik}, so

$$s_k^2 = \frac{\sum_i (X_{ik} - \overline{X}_k)^2}{n_k - 1} \to \sigma^2$$

- to get a better estimate, pool across all groups (just like for two-sample t-test)

$$\frac{SS_W}{N-K} = MS_W \to \sigma^2$$

  - here MS_W stands for mean squares within
  - $N - K$ is the degrees of freedom

BETWEEN DEVIATIONS

$$SS_B = \sum_k n_k(\overline{X}_k - \overline{X})^2$$

- what causes this to be greater than zero?
- since

$$X_{ik} = \mu + \alpha_k + e_{ik}$$

- the mean of group k is

$$\overline{X}_k = \frac{\sum_i X_{ik}}{n_k} = \mu + \alpha_k + \frac{\sum_i e_{ik}}{n_k}$$

- as k changes, μ stays the same
- so any deviations from $\overline{X}$ are due to changes in α_k (changes between groups) or to changes in $\frac{\sum_i e_{ik}}{n_k}$ (random error)

ESTIMATE OF σ^2

- **if** H_0 is true, then all $\alpha_k = 0$ and any deviations must be due only to the random error terms ($\sum_i e_{ik}/n_k$)
- so we can again estimate σ^2 as

$$MS_B = \frac{SS_B}{K-1} = \frac{\sum_k n_k(\overline{X}_k - \overline{X})^2}{K-1} \to \sigma^2$$

  - here $K-1$ is degrees of freedom
- on the other hand, **if** H_0 is not true, then MS_B includes deviations due to α_k, so

$$MS_B > \sigma^2$$

F statistic

- so, we do not know what σ^2 is, but we have two estimates
  - MS_W: always estimates σ^2
  - MS_B: estimates σ^2 if H_0 is true. Larger than σ^2 if H_0 is false.
- compare the estimates by computing

$$F = \frac{MS_B}{MS_W}$$

- if H_0 is true, should get $F = 1$, if H_0 is not true, should get $F > 1$

F critical

- as always for inferential statistics, we need to know if F is significantly greater than 1.0
- depends on two degrees of freedom
- df numerator $= K - 1$
- df denominator $= N - K$
- look up p-value using the online F-distribution calculator

TESTING

- 4 STEPS
  1. State the hypothesis and set the criterion: $H_0 : \mu_1 = \mu_2 = ... = \mu_K$, $H_a : \mu_i \neq \mu_j$ for some i, j.
  2. Compute the test statistic $F = MS_B/MS_W$.
  3. Compute the p-value. Need to find the degrees of freedom.
  4. Make a decision.

EXAMPLE

- A college professor wants to determine the best way to present an important lecture topic to his class.
- He decides to do an experiment to evaluate three options. He solicits 27 volunteers from his class and randomly assigns 9 to each of three conditions.
- In condition 1, he lectures to the students.
- In condition 2, he lectures plus assigns supplementary reading.
- In condition 3, the students see a film on the topic plus receive the same supplementary reading as the students in condition 2.
- The students are subsequently tested on the material. The following scores (percentage correct) were obtained.

EXAMPLE

Lecture Condition 1	Lecture + Reading Condition 2	Film + Reading Condition 3
92	86	81
86	93	80
87	97	72
76	81	82
80	94	83
87	89	89
92	98	76
83	90	88
84	91	83

- No one does the calculations by hand. Always use a computer.

(1) HYPOTHESES

- for one-way ANOVA the hypotheses are

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_a : \mu_i \neq \mu_k \text{ for some } i, k$$

- Set $\alpha = 0.05$

(2) TEST STATISTIC

- Use the on-line calculator
- We have to format the data properly for the calculator
- One score to each line
- Indicate the level (no spaces) and then the score

Lecture	92
Lecture	86
...	
LectureReading	86
LectureReading	93
...	
FilmReading	81
FilmReading	80

- Order does not matter

(2) TEST STATISTIC

- Data could look like this when pasted into the calculator

```
Lecture    92
Lecture    86
Lecture    87
Lecture    76
Lecture    80
Lecture    87
Lecture    92
Lecture    83
Lecture    84
LectureReading    86
LectureReading    93
LectureReading    97
LectureReading    81
LectureReading    94
LectureReading    89
LectureReading    98
LectureReading    90
LectureReading    91
FilmReading    81
FilmReading    80
```


(2) TEST STATISTIC

- We read out the results of the analysis in the ANOVA summary table

Source	df	SS	MS	F	p-value
Between	2	408.0741	204.0370	7.2894	0.00336
Within	24	671.7778	27.9907		
Total	26	1079.8519			

- lots of information

(2) TEST STATISTIC

Source	df	SS	MS	F	p-value
Between	2	408.0741	204.0370	7.2894	0.00336
Within	24	671.7778	27.9907		
Total	26	1079.8519			

- We can double check things

$$F = \frac{MS_B}{MS_W} = \frac{204.0370}{27.9907} = 7.2894$$

$$MS_B = \frac{SS_B}{K-1} = \frac{408.0741}{3-1} = 204.0370$$

$$MS_W = \frac{SS_W}{N-K} = \frac{671.7778}{27-3} = 27.9907$$

(3) p VALUE

Source	df	SS	MS	F	p-value
Between	2	408.0741	204.0370	7.2894	0.00336
Within	24	671.7778	27.9907		
Total	26	1079.8519			

- between degrees of freedom (numerator)

$$df = K - 1 = 3 - 1 = 2$$

- within degrees of freedom (denominator)

$$df = N - K = 27 - 3 = 24$$

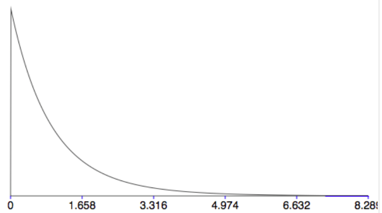
- Total degrees of freedom

$$df = N - 1 = 27 - 1 = 26$$

(3) p VALUE

Source	df	SS	MS	F	p-value
Between	2	408.0741	204.0370	7.2894	0.00336
Within	24	671.7778	27.9907		
Total	26	1079.8519			

- Check the p-value using the F distribution calculator



df numerator =	2
df denominator =	24
F =	7.2894
p =	0.00336

- Note, we just compute p from one tail, but this is equivalent to a two-tailed t-test.

(4) DECISION

- since

$$p = 0.00336 < .05 = \alpha$$

- we reject H_0. The methods of presentation are not equally effective.
- Note, does not tell us which pair of means are different!
- Look at means

Condition	Mean	Standard deviation	Sample size
Lecture	85.22222222222223	5.214829282387329	9
LectureReading	91	5.338539126015656	9
FilmReading	81.55555555555556	5.317685377847901	9

GENERALITY

- The great thing about ANOVA is that these basic steps stay the same even if you have many more means to be compared
- I happen to have data from 8 different classes that all completed an experiment where subjects responded as quickly as possible whether a set of letters formed a word or not
- The summary is the same format as above

GENERALITY

Source	df	SS	MS	F	p-value
Between	7	2324584.6485	332083.5212	6.6500	0.00000
Within	407	20324589.8142	49937.5671		
Total	414	22649174.4627			

Condition	Mean	Standard deviation	Sample size
Francis200F15	788.3333333333335	244.2585052255086	81
Francis200S16	756.0007352941174	204.17983832898088	68
Francis200F16	750.0464601769914	218.19667178177372	113
Francis200F17	756.6531914893621	214.33283856802967	94
FUSfall2018	766.1649999999998	172.00442964925605	30
Psy200Spring15	1167.3535714285715	360.9423454196428	14
FS16PSY200	776.26	224.8173218909571	10
PSY2008HKIED	849.6600000000002	191.92566073873397	5

- it would be the same format with 8000 classes!

CONCLUSIONS

- testing multiple means
- two estimates of population variance
- one estimate always estimates variance
- other estimate is true only if H_0 is true
- lets us test H_0

NEXT TIME

- interpreting ANOVA
- contrasts
- more multiple testing

Some thing versus which thing.

PSY 201: Statistics in Psychology

Lecture 33

Analysis of Variance

Some thing versus which thing.

Greg Francis

Purdue University

Fall 2023

ANOVA

- Test statistic:
$$F = \frac{MS_B}{MS_W}$$
$$F = \frac{\text{Estimated variability from noise and mean differences}}{\text{Estimated variability from noise}}$$
- if H_0 is true, and F is sufficiently larger than 1, then a rare event has happened. Since rare events are rare, when $F >> 1$ we suppose that H_0 is not true
- Rareness is established by the p value, which is gotten from an F distribution with $K-1$ df in the numerator and $N-K$ df in the denominator

HYPOTHESES

- The null is an *omnibus* hypothesis. It supposes no difference between any population means

$$H_0 : \mu_i = \mu_j \; \forall \; i, j$$

- the alternative is the complement

$$H_a : \mu_i \neq \mu_j \text{ for some } i, j$$

- Note, there is no *one-tailed* version of ANOVA

INTERPRETING

- I happen to have data from 8 different classes that all completed an experiment where subjects responded as quickly as possible whether a set of letters formed a word or not

Source	df	SS	MS	F	p-value
Between	7	2324584.6485	332083.5212	6.6500	0.00000
Within	407	20324589.8142	49937.5671		
Total	414	22649174.4627			

Condition	Mean	Standard deviation	Sample size
Francis200F15	788.3333333333335	244.2585052255086	81
Francis200S16	756.0007352941174	204.17983832898088	68
Francis200F16	750.0464601769914	218.19667178177372	113
Francis200F17	756.6531914893621	214.33283856802967	94
FUSfall2018	766.1649999999998	172.00442964925605	30
Psy200Spring15	1167.3535714285715	360.9423454196428	14
FS16PSY200	776.26	224.8173218909571	10
PSY2008HKIED	849.6600000000002	191.92566073873397	5

- The conclusion is that *at least one* population mean seems to be different from the other population means. **Something** is different
- The ANOVA does not tell you **which** mean is different from the others; or if more than one mean is different from others.

INTERPRETING

- It might be tempting to just look at the data and "wing it"
- For example, looking at the means, it seems that class *Psy200Spring15* has a much larger mean than any other class

Source	df	SS	MS	F	p-value
Between	7	2324584.6485	332083.5212	6.6500	0.00000
Within	407	20324589.8142	49937.5671		
Total	414	22649174.4627			

Condition	Mean	Standard deviation	Sample size
Francis200F15	788.3333333333335	244.2585052255086	81
Francis200S16	756.0007352941174	204.17983832898088	68
Francis200F16	750.0464601769914	218.19667178177372	113
Francis200F17	756.6531914893621	214.33283856802967	94
FUSfall2018	766.1649999999998	172.00442964925605	30
Psy200Spring15	1167.3535714285715	360.9423454196428	14
FS16PSY200	776.26	224.8173218909571	10
PSY2008HKIED	849.6600000000002	191.92566073873397	5

- But that class also has a small number of students ($n = 14$), and a large standard deviation ($s = 360.9$), so we would expect quite a bit of variability in the mean value. Maybe this big mean is not so rare, given the variability due to random sampling

INTERPRETING

- More than one mean might differ from other means
- Even if the mean for *Psy200Spring15* is different from the others, might other means also be different?

Source	df	SS	MS	F	p-value
Between	7	2324584.6485	332083.5212	6.6500	0.00000
Within	407	20324589.8142	49937.5671		
Total	414	22649174.4627			

Condition	Mean	Standard deviation	Sample size
Francis200F15	788.3333333333335	244.2585052255086	81
Francis200S16	756.0007352941174	204.17983832898088	68
Francis200F16	750.0464601769914	218.19667178177372	113
Francis200F17	756.6531914893621	214.33283856802967	94
FUSfall2018	766.1649999999998	172.00442964925605	30
Psy200Spring15	1167.3535714285715	360.9423454196428	14
FS16PSY200	776.26	224.8173218909571	10
PSY2008HKIED	849.6600000000002	191.92566073873397	5

- We would really like to know which means seem to be different from which other means

TYPE I ERROR

- Multiple testing problem
- To motivate ANOVA, we mentioned that it is problematic to just test all pairwise comparisons of group means. With 8 means, there would be 28 tests. So the Type I error rate would be around

$$1-(1-\alpha)^{28} = (1-0.95^{28}) = 0.76$$

- Instead of just testing all possible comparisons, suppose we first require that the ANOVA produces a significant result. If H_0 is true, the ANOVA should only conclude that some difference exists with a probability of 0.05 (or whatever you choose as α)

TYPE I ERROR

- Thus, we can control the overall Type I error rate by insisting that our data produce a significant ANOVA *before* we start testing different means
- We want to check that something is different before we check which means are different!
- If we now test *Psy200Spring15* against each of the other seven means, the Type I error rate can be no bigger than what it was for the ANOVA
- In fact, it has to be a bit smaller than the α used for the ANOVA because we have to satisfy two criteria
- If H_0 is true, 95% of the time, we never compare the means to each other

t tests

- One approach is to just run t tests (Welch's test) to compare different means
- For example, we can test *Psy200Spring15* against *Francis200F15*

Test summary	
Type of test	Welch's Test
Null hypothesis	$H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis	$H_a: \mu_1 - \mu_2 \neq 0$
Type I error rate	$\alpha = 0.05$
Label for group 1	Group 1
Sample size 1	$n_1 = 14$
Sample mean 1	$\overline{X}_1 = 1167.3536$
Sample standard deviation 1	$s_1 = 360.942345$
Label for group 2	Group 2
Sample size 2	$n_2 = 81$
Sample mean 2	$\overline{X}_2 = 788.3333$
Sample standard deviation 2	$s_2 = 244.258505$
Pooled standard deviation	s =NA
Sample standard error	$s_{\overline{X}_1-\overline{X}_2} = 76.322416$
Test statistic	$t = 4.966041$
Degrees of freedom	$df = 15.124024621983446$
p value	$p = 0.000165$
Decision	Reject the null hypothesis
Confidence interval critical value	$t_{cv} = 2.129928$
Confidence interval	CI_{95}=(216.458972, 541.581502)

t tests

- One approach is to just run t tests (Welch's test) to compare different means
- For example, we can test *Psy200Spring15* against *PSY2008HKIED*

Test summary	
Type of test	Welch's Test
Null hypothesis	$H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis	$H_a: \mu_1 - \mu_2 \neq 0$
Type I error rate	$\alpha = 0.05$
Label for group 1	Group 1
Sample size 1	$n_1 = 14$
Sample mean 1	$\overline{X}_1 = 1167.3536$
Sample standard deviation 1	$s_1 = 360.942345$
Label for group 2	Group 2
Sample size 2	$n_2 = 5$
Sample mean 2	$\overline{X}_2 = 849.6600$
Sample standard deviation 2	$s_2 = 191.925607$
Pooled standard deviation	s =NA
Sample standard error	$s_{\overline{X}_1-\overline{X}_2} = 171.445901$
Test statistic	$t = 1.853025$
Degrees of freedom	$df = 13.741233049477954$
p value	$p = 0.085474$
Decision	Do not the reject null hypothesis
Confidence interval critical value	$t_{cv} = 2.148582$
Confidence interval	CI_{95}=(-50.672018, 686.059158)

CONTRASTS

- There is a better (and more general approach)
- ANOVA assumes/requires homogeneity of variance

$$\sigma_i^2 = \sigma_j^2 \ \forall \ i, j$$

- For the t-test we pooled variances/standard deviations to get a better estimate of σ
- With more populations, we can pool all of the sample variances and thereby get a still better estimate
- Thus, even when we compare *Psy200Spring15* against *Francis200F15*, we can use the data from the other samples to get a better estimate of σ

POOLED ESTIMATE

- Fortunately, the pooled estimate of variance is easy to find
- We computed it in the ANOVA, it is MS_W

Source	df	SS	MS	F	p-value
Between	7	2324584.6485	332083.5212	6.6500	0.00000
Within	407	20324589.8142	49937.5671		
Total	414	22649174.4627			

Condition	Mean	Standard deviation	Sample size
Francis200F15	788.3333333333335	244.2585052255086	81
Francis200S16	756.0007352941174	204.17983832898088	68
Francis200F16	750.0464601769914	218.19667178177372	113
Francis200F17	756.6531914893621	214.33283856802967	94
FUSfall2018	766.1649999999998	172.00442964925605	30
Psy200Spring15	1167.3535714285715	360.9423454196428	14
FS16PSY200	776.26	224.8173218909571	10
PSY2008HKIED	849.6600000000002	191.92566073873397	5

- Thus, the standard error that we use for the t test is

$$s_{\overline{X}_1-\overline{X}_2} = \sqrt{MS_W\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

CONTRASTS

- For example, we can test *Psy200Spring15* against *Francis200F15*

$$s_{\overline{X}_1-\overline{X}_2} = \sqrt{MS_W\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \sqrt{(49937.5671)\left(\frac{1}{14}+\frac{1}{81}\right)} = 64.67984425$$

- Compare to the traditional t test, where

$$s_{\overline{X}_1-\overline{X}_2} = 76.322416$$

- So with the pooled variance, we get

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1-\overline{X}_2}} = \frac{379.02}{64.6798} = 5.8599$$

- Compare to $t = 4.966$ for the traditional t test
- The degrees of freedom is based on how many scores contribute to the variance calculation, so we get

$$df = N - K = 415 - 8 = 407$$

- compare to $df = n_1 + n_2 - 2 = 14 + 81 - 2 = 93$, for traditional t test (smaller with Welch's test)

CONTRASTS

- For example, we can test *Psy200Spring15* against *PSY2008HKIED*

$$s_{\overline{X}_1-\overline{X}_2} = \sqrt{MS_W\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \sqrt{(49937.5671)\left(\frac{1}{14}+\frac{1}{5}\right)} = 116.423$$

- Compare to the traditional t test, where

$$s_{\overline{X}_1-\overline{X}_2} = 171.446$$

- So with the pooled variance, we get

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1-\overline{X}_2}} = \frac{317.69}{116.423} = 2.7288$$

- The degrees of freedom is based on how many scores contribute to the variance calculation, so we get

$$df = N - K = 415 - 8 = 407$$

- so $p = 0.0066$
- Compare to $t = 1.853$ for the traditional t test
- compare to $df = n_1 + n_2 - 2 = 14 + 5 - 2 = 17$, and $p = 0.08$

BETTER IS BETTER

- With a contrast, we get a better estimate of $s_{\overline{X}_1 - \overline{X}_2}$, which sometimes means we can reject H_0. Not always, though.
- It is possible for a standard t test to reject H_0, but the corresponding contrast test does not reject H_0 (because the sample s^2 is smaller than MS_W)
- We do not have any cases like that in our current data set
- Generally speaking, using MS_W is better than using the pooled s^2 because more data contributes to the estimate

OTHER CONTRASTS

- Comparing two means is actually a special case of using contrasts
- We can also compare various *combinations* of means

Source	df	SS	MS	F	p-value
Between	7	2324584.6485	332083.5212	6.6500	0.00000
Within	407	20324589.8142	49937.5671		
Total	414	22649174.4627			

Condition	Mean	Standard deviation	Sample size
Francis200F15	788.3333333333335	244.2585052255086	81
Francis200S16	756.0007352941174	204.17983832898088	68
Francis200F16	750.0464601769914	218.19667178177372	113
Francis200F17	756.6531914893621	214.33283856802967	94
FUSfall2018	766.1649999999998	172.00442964925605	30
Psy200Spring15	1167.3535714285715	360.9423454196428	14
FS16PSY200	776.26	224.8173218909571	10
PSY2008HKIED	849.6600000000002	191.92566073873397	5

- For example, we might wonder if the mean for classes taught by Dr. Francis differs from the mean for classes not taught by Dr. Francis

OTHER CONTRASTS

- We set up *contrast weights*, c_i, for each class' mean
- Our null hypothesis will be

$$H_0 : \sum_{i=1}^{K} (c_i \mu_i) = 0$$

- and we require that the contrast weights sum to 0:

$$\sum_{i=1}^{K} c_i = 0$$

- Our alternative hypothesis is

$$H_a : \sum_{i=1}^{K} (c_i \mu_i) \neq 0$$

- (one-tailed tests are also possible)

TEST STATISTIC

- We compute the weighted sum of means

$$L = \sum_{i=1}^{K} (c_i \overline{X}_i)$$

- which has a standard error of:

$$s_L = \sqrt{MS_W \sum_{i=1}^{K} \frac{c_i^2}{n_i}}$$

- and our test statistic is

$$t = \frac{L}{s_L}$$

- which follows a t distribution with

$$df = N - K$$

  - where N is the sum of sample sizes across all groups and K is the number of groups

ONLINE CALCULATOR

- To compare the mean of the four classes taught by Dr. Francis to the mean of the other four classes, we use contrast weights of ± 1

Contrast test

To set up a contrast, enter a weight value for each population mean such that the weights sum to 0.

Specify hypotheses:

H_0: 1 $\mu_{Francis200F15}$ + 1 $\mu_{Francis200S16}$ + 1 $\mu_{Francis200F16}$ + 1 $\mu_{Francis200F17}$ + -1 $\mu_{FUSfall2018}$ + -1 $\mu_{Psy200Spring15}$ + -1 $\mu_{FS16PSY200}$ + -1 $\mu_{PSY2008HKIED}$ = 0

H_a: Two-tails

α 0.05

Run Contrast

Contrast test summary	
Null hypothesis	$H_0: (1)\mu_{Francis200F15} + (1)\mu_{Francis200S16} + (1)\mu_{Francis200F16} + (1)\mu_{Francis200F17} + (-1)\mu_{FUSfall2018} + (-1)\mu_{Psy200Spring15} + (-1)\mu_{FS16PSY200} + (-1)\mu_{PSY2008HKIED} = 0$
Alternative hypothesis	$H_a: (1)\mu_{Francis200F15} + (1)\mu_{Francis200S16} + (1)\mu_{Francis200F16} + (1)\mu_{Francis200F17} + (-1)\mu_{FUSfall2018} + (-1)\mu_{Psy200Spring15} + (-1)\mu_{FS16PSY200} + (-1)\mu_{PSY2008HKIED} \neq 0$
Type I error rate	α=0.05
Weighted sum of sample means	L= -508.4048511347672
Standard error	s_L =150.1229164077
Test statistic	t=-3.386590557260787
Degrees of freedom	df=407
p value	p=0.00077652497924241
Decision	Reject the null hypothesis

ONLINE CALCULATOR

- Other sets of contrast weights compare other combinations. For example, to contrast the mean of the non-US based class, *PSY2008HKIED*, against all the other classes, we could use:

To set up a contrast, enter a weight value for each population mean such that the weights sum to 0.

Specify hypotheses:

H_0: 1 $\mu_{Francis200F15}$ + 1 $\mu_{Francis200S16}$ + 1 $\mu_{Francis200F16}$ + 1 $\mu_{Francis200F17}$ + 1 $\mu_{FUSfall2018}$ + 1 $\mu_{Psy200Spring15}$ + 1 $\mu_{FS16PSY200}$ + -7 $\mu_{PSY2008HKIED}$ = 0

H_a: Two-tails

α 0.05

Run Contrast

Contrast test summary	
Null hypothesis	$H_0: (1)\mu_{Francis200F15} + (1)\mu_{Francis200S16} + (1)\mu_{Francis200F16} + (1)\mu_{Francis200F17} + (1)\mu_{FUSfall2018} + (1)\mu_{Psy200Spring15} + (1)\mu_{FS16PSY200} + (-7)\mu_{PSY2008HKIED} = 0$
Alternative hypothesis	$H_a: (1)\mu_{Francis200F15} + (1)\mu_{Francis200S16} + (1)\mu_{Francis200F16} + (1)\mu_{Francis200F17} + (1)\mu_{FUSfall2018} + (1)\mu_{Psy200Spring15} + (1)\mu_{FS16PSY200} + (-7)\mu_{PSY2008HKIED} \neq 0$
Type I error rate	α=0.05
Weighted sum of sample means	L= -186.80770827762535
Standard error	s_L =708.4755001381367
Test statistic	t=-0.2636756080361312
Degrees of freedom	df=407
p value	p=0.7921633394031113
Decision	Do not the reject null hypothesis

- You do not have to use integer values for the c_i terms, but it helps to avoid rounding issues.

ONLINE CALCULATOR

- It can be appropriate to set some weights equal to 0. For example, if you want to compare the mean from two classes in 2015 against the mean from three classes in 2016, you can set weights as:

To set up a contrast, enter a weight value for each population mean such that the weights sum to 0.

Specify hypotheses:

H_0: -3 $\mu_{Francis200F15}$ + 2 $\mu_{Francis200S16}$ + 2 $\mu_{Francis200F16}$ + 0 $\mu_{Francis200F17}$ + 0 $\mu_{FUSfall2018}$ + -3 $\mu_{Psy200Spring15}$ + 2 $\mu_{FS16PSY200}$ + 0 $\mu_{PSY2008HKIED}$ = 0

H_a: Two-tails

α 0.05

Run Contrast

Contrast test summary	
Null hypothesis	$H_0: (-3)\mu_{Francis200F15} + (2)\mu_{Francis200S16} + (2)\mu_{Francis200F16} + (0)\mu_{Francis200F17} + (0)\mu_{FUSfall2018} + (-3)\mu_{Psy200Spring15} + (2)\mu_{FS16PSY200} + (0)\mu_{PSY2008HKIED} = 0$
Alternative hypothesis	$H_a: (-3)\mu_{Francis200F15} + (2)\mu_{Francis200S16} + (2)\mu_{Francis200F16} + (0)\mu_{Francis200F17} + (0)\mu_{FUSfall2018} + (-3)\mu_{Psy200Spring15} + (2)\mu_{FS16PSY200} + (0)\mu_{PSY2008HKIED} \neq 0$
Type I error rate	α=0.05
Weighted sum of sample means	L= -1302.4463233434972
Standard error	s_L =249.66291788092158
Test statistic	t=-5.216819279364137
Degrees of freedom	df=407
p value	p=2.905150702225967e-7
Decision	Reject the null hypothesis

SPECIAL CASE

- Comparing two means is just a special case where the contrast weights for those means are set to ± 1 and the other weights are set to 0:

To set up a contrast, enter a weight value for each population mean such that the weights sum to 0.

Specify hypotheses:

H_0: -1 $\mu_{Francis200F15}$ + 0 $\mu_{Francis200S16}$ + 0 $\mu_{Francis200F16}$ + 0 $\mu_{Francis200F17}$ + 0 $\mu_{FUSfall2018}$ + 1 $\mu_{Psy200Spring15}$ + 0 $\mu_{FS16PSY200}$ + 0 $\mu_{PSY2008HKIED}$ = 0

H_a: Two-tails

α 0.05

Run Contrast

Contrast test summary	
Null hypothesis	$H_0: (-1)\mu_{Francis200F15} + (0)\mu_{Francis200S16} + (0)\mu_{Francis200F16} + (0)\mu_{Francis200F17} + (0)\mu_{FUSfall2018} + (1)\mu_{Psy200Spring15} + (0)\mu_{FS16PSY200} + (0)\mu_{PSY2008HKIED} = 0$
Alternative hypothesis	$H_a: (-1)\mu_{Francis200F15} + (0)\mu_{Francis200S16} + (0)\mu_{Francis200F16} + (0)\mu_{Francis200F17} + (0)\mu_{FUSfall2018} + (1)\mu_{Psy200Spring15} + (0)\mu_{FS16PSY200} + (0)\mu_{PSY2008HKIED} \neq 0$
Type I error rate	α=0.05
Weighted sum of sample means	L= 379.020238095238
Standard error	s_L =64.67984426050968
Test statistic	t=5.859943579466054
Degrees of freedom	df=407
p value	p=9.569574910273104e-9
Decision	Reject the null hypothesis

- This gives the same result as we computed previously

MULTIPLE TESTING

- There are an *enormous* number of different contrasts that you could create
- If you require a significant ANOVA before running any contrasts, then you can control the Type I error rate to be no higher than α
- However, we have a new kind of "conditional" Type I error
- Given that the ANOVA indicates there is some difference in means, what means (or combinations of means) differ? For some contrasts the H_0 is true, but, just due to random sampling, they indicate that there is a difference

MULTIPLE TESTING

- Thus, we have a new multiple testing problem for identifying the differences; even though we only get to that situation with probability α if the ANOVA omnibus H_0 is true
- Worse, it could be that $\mu_7 \neq \mu_8$, so you reject the ANOVA H_0
- but then you run contrasts for other means where $\mu_i = \mu_j$
- Generally, it is not a good idea to try all possible contrasts. Contrasts (and hypothesis testing in general) make the most sense when you have some specific plans to compare combinations of means

CONCLUSIONS

- interpreting an ANOVA
- identifying differences
- contrast tests

NEXT TIME

- power for ANOVA
- power for contrasts

Keep it simple!

PSY 201: Statistics in Psychology

Lecture 34

Power for Analysis of Variance

Keep it simple!

Greg Francis

Purdue University

Fall 2023

HYPOTHESES

- The null for an ANOVA is an *omnibus* hypothesis. It suppose no difference between any population means

$$H_0 : \mu_i = \mu_j \ \forall \ i, j$$

- the alternative is the complement

$$H_a : \mu_i \neq \mu_j \text{ for some } i, j$$

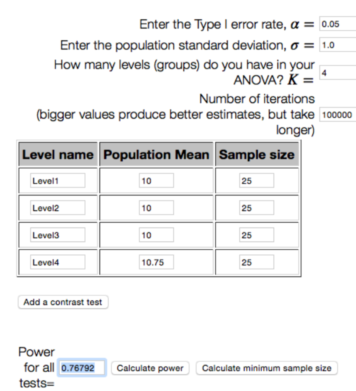
- To compute power, we have to provide the standard deviation, α, n's, and specific values for the means

POWER CALCULATOR

- For other power calculators, it was kind of easy to identify how power is affected by the specific alternative:
- bigger differences (between population means, proportions, or correlations) leads to more power
- That is also true for ANOVA, but it can be more complicated because there are multiple means

POWER CALCULATOR

- Consider a situation with $K = 4$ means (one different from the others):
- We estimate the power to be 0.76792



Enter the Type I error rate, α = 0.05

Enter the population standard deviation, σ = 1.0

How many levels (groups) do you have in your ANOVA? K = 4

Number of iterations (bigger values produce better estimates, but take longer) 100000

Level name	Population Mean	Sample size
Level1	10	25
Level2	10	25
Level3	10	25
Level4	10.75	25

Add a contrast test

Power for all tests= 0.76792 Calculate power Calculate minimum sample size

POWER CALCULATOR

- Consider a situation with $K = 8$ means (one different from the others):
- We estimate the power to be 0.70688.
- Power is affected by the ratio of the variability between group means and the variability within each group ($\sigma = 1$). If just one mean is different from the others, this ratio decreases as K gets bigger

Enter the Type I error rate, $\alpha =$ 0.05
Enter the population standard deviation, $\sigma =$ 1.0
How many levels (groups) do you have in your ANOVA? $K =$ 8
Number of iterations (bigger values produce better estimates, but take longer) 100000

Level name	Population Mean	Sample size
Level1	10	25
Level2	10	25
Level3	10	25
Level4	10.75	25
Level5	10	25
Level6	10	25
Level7	10	25
Level8	10	25

Add a contrast test

Power for all tests= 0.70688 Calculate power Calculate minimum sample size

POWER CALCULATOR

- Consider a situation with $K = 8$ means (four different from the others):
- We estimate the power to be 0.98324

Enter the Type I error rate, $\alpha =$ 0.05
Enter the population standard deviation, $\sigma =$ 1.0
How many levels (groups) do you have in your ANOVA? $K =$ 8
Number of iterations (bigger values produce better estimates, but take longer) 100000

Level name	Population Mean	Sample size
Level1	10	25
Level2	10	25
Level3	10	25
Level4	10	25
Level5	10.75	25
Level6	10.75	25
Level7	10.75	25
Level8	10.75	25

Add a contrast test

Power for all tests= 0.98324 Calculate power Calculate minimum sample size

POWER CALCULATOR

- Consider a situation with $K = 4$ means (two different from the others):
- We estimate the power to be 0.88392
- Thus, it is *not* just that power decreases as K increases. It depends on the values of the means

Enter the Type I error rate, $\alpha =$ 0.05
Enter the population standard deviation, $\sigma =$ 1.0
How many levels (groups) do you have in your ANOVA? $K =$ 4
Number of iterations (bigger values produce better estimates, but take longer) 100000

Level name	Population Mean	Sample size
Level1	10	25
Level2	10	25
Level3	10.75	25
Level4	10.75	25

Add a contrast test

Power for all tests= 0.88392 Calculate power Calculate minimum sample size

POWER CALCULATOR

- Consider a situation with $K = 4$ means (every mean is different from the others):
- We estimate the power to be 0.62368
- The biggest and smallest means differ by 0.75, just like previous cases, but that alone does not determine power

Enter the Type I error rate, $\alpha =$ 0.05
Enter the population standard deviation, $\sigma =$ 1.0
How many levels (groups) do you have in your ANOVA? $K =$ 4
Number of iterations (bigger values produce better estimates, but take longer) 100000

Level name	Population Mean	Sample size
Level1	10	25
Level2	10.25	25
Level3	10.5	25
Level4	10.75	25

Add a contrast test

Power for all tests= 0.62368 Calculate power Calculate minimum sample size

TRUST THE MATH

- With sufficient experience, you can learn to recognize what types of situations produce large (or small) power
- Until you get that experience, rely on the calculator (even after you get the experience you need the calculator to do the actual computations)
- It is still the case that larger samples lead to higher power.

EXAMPLE

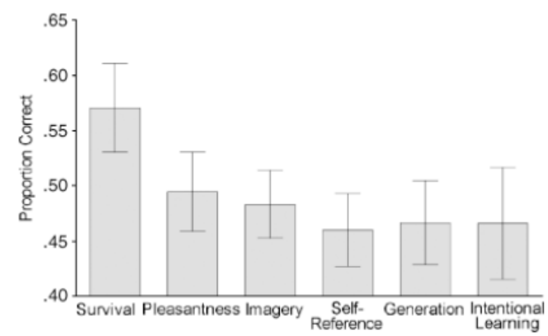
- There are lots of mnemonic tricks to try to improve your memory. They really do work!
- To compare these tricks we can use a standard memory test (Nairne, Pandeirada & Thompson, 2008):
- A subject is shown a word and asked to do some kind of task. This is repeated for 30 words.
- At the end of the experiment, the subject is asked to recall as many words as possible. Usually, this is a surprise memory task.
- For each subject, we compute the proportion of recalled words.
- We are interested in the mean value of the proportion across subjects.
- We can compare how well different tasks influence memory.

TASKS

- *Pleasantness*: Rate the pleasantness of the word on a scale from 1 to 5.
- *Imagery*: Rate how easy it is to form a mental image of the word on a scale from 1 to 5.
- *Self-reference*: Rate how easily the word brings to mind an important personal experience on a scale from 1 to 5.
- *Generation*: Words are partially scrambled; unscramble and then rate the pleasantness of the word on a scale from 1 to 5. (e.g., "iktten")
- *Survival*: Rate the relevance of the word for survival if you are stranded in the grasslands of a foreign land, on a scale from 1 to 5.
- *Intentional learning*: Try to remember the words for a future memory test.
- Different subjects are assigned to different conditions

ORIGINAL RESULTS

- Nairne, Pandeirada & Thompson (2008) found a big advantage for survival processing compared to the other methods. $n_i = 50$ for each group



- $F_{5,294} = 4.41$, $p = 0.00178$, $MS_W = 0.019$

NEW METHOD

- Suppose that you want to further explore these kinds of memory tricks. You think that the survival processing method does well because it gets subjects to be really engaged in thinking about the word. You come up with a new method
- *Vacation*: Rate the relevance of the word for enjoyment while on vacation at a fancy resort, on a scale from 1 to 5.
- You expect that the vacation task will do about the same as the survival task
- You worry that other details of the experiment may change the overall level of performance for all tasks, so you decide to repeat the full study, with the addition of your new, Vacation, task. So there will be seven groups.
- How do you plan an appropriate sample size?

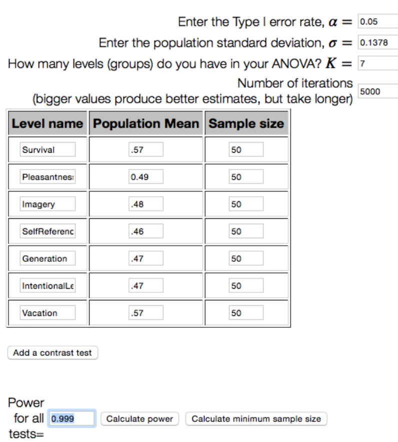
SPECIFIC MEANS

- As the values for the specific means, we can use the sample means found in the original study
- We get them from the figure
- For the Vacation task, we expect performance to be the same as the Survival task
- For the standard deviation, we can use the square root of MS_W

$$\sigma = \sqrt{MS_W} = \sqrt{0.019} = 0.1378$$

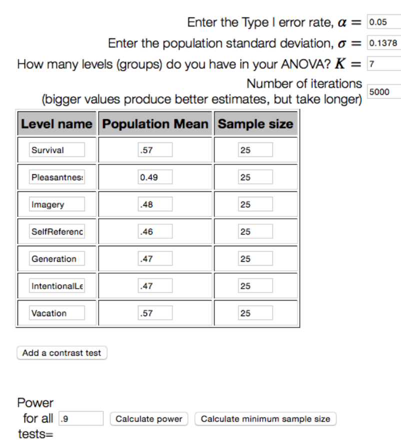
POWER FOR ANOVA

- Power is quite high (0.999) if we use $n_i = 50$, as in the original study



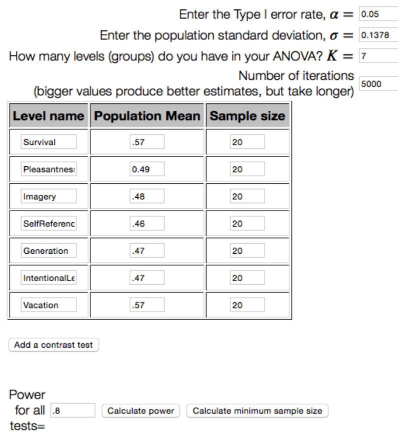
POWER FOR ANOVA

- If we accept power of 0.9, $n = 25$ subjects in each sample is sufficient



POWER FOR ANOVA

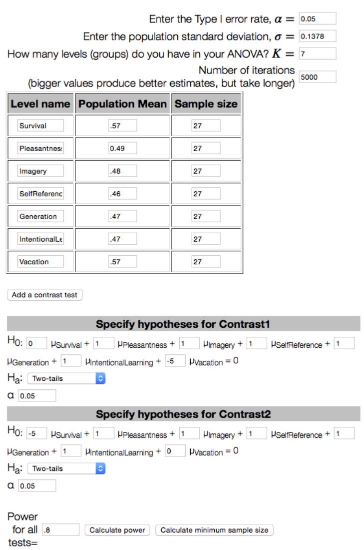
- If we accept power of 0.8, $n = 20$ subjects in each sample is sufficient



CONTRASTS

- However, just a significant ANOVA is not enough for what we are studying
- We want to show that the Vacation task is better than most of the other tasks (not including the Survival task)
- We also want to show that the Survival task is better than most of the other tasks (not including the Vacation task)
- For our hypothesis test, we will set up two contrasts to test Vacation and Survival against the other tasks:
- We need to include those contrasts in the power analysis (more subjects)

CONTRASTS



NULL

- You might also want to demonstrate that memory performance is the same for the Survival and Vacation tasks (after all, your idea is that both tasks are engaging, so they should have similar performance)
- Unfortunately, hypothesis testing cannot show that two groups have equal means (that would be *proving* the null hypothesis)
- Thus, we cannot set a sample size so that we are sure the Survival and Vacation tasks are equally effective for improving memory

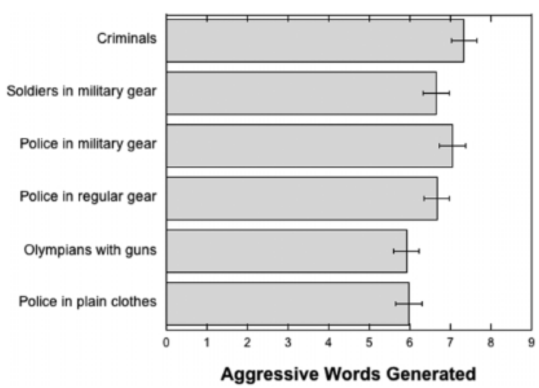
ANOTHER EXAMPLE

- Bushman (2018) investigated the "weapons effect": the mere presence of weapons can increase aggression
- Subjects were assigned to view a set of images of one type:
- Criminals, Soldiers, Police in military gear, Police in regular gear, Olympians with guns, Police in plain clothes
- Afterwards, complete a word fragment task:
  - C H O __ E
  - K I __ __
  - M U __ __ E R
  - C __ T
- each fragment can be completed to form an aggressive or non-aggressive word
- Count how many aggressive words are formed: measure of aggressive thoughts

EXAMPLE IMAGES

DATA

- Roughly $n = 100$ for each image set



MANY TESTS

- Conclusions are based on many contrasts
  - Significant ANOVA (some difference across image types)
  - Contrast between people with guns vs. plainclothes police (no guns): Weapon is important
  - Contrast between Olympians vs. Others: Person must intend to hurt others
  - Contrast between people with guns vs. Olympians: Weapon must be to hurt people
- Conclusion: only guns intended to shoot human targets prime aggressive thoughts

REPLICATION STUDY

- Suppose you want to replicate this study. To estimate power you use the means and standard deviation of the original finding. You want to see what happens if you use a similar sample size as the original study, $n = 100$, for each sample
- We enter the information in the ANOVA Power calculator

Enter the Type I error rate, $\alpha =$ 0.05
Enter the population standard deviation, $\sigma =$ 3.4
How many levels (groups) do you have in your ANOVA? $K =$ 6
Number of iterations (bigger values produce better estimates, but take longer) 5000

Level name	Population Mean	Sample size
Criminal	7.1	100
Soldiers	6.65	100
PoliceMilitar	6.8	100
PoliceRegula	6.7	100
Olympians	5.9	100
PolicePlainC	5.95	100

REPLICATION STUDY

- We set up each of the contrast tests in the ANOVA Power calculator:

Specify hypotheses for Contrast1
H_0: 1 $\mu_{Criminal}$ + 1 $\mu_{Soldiers}$ + 1 $\mu_{PoliceMilitaryGear}$ + 1 $\mu_{PoliceRegularGear}$ + 0 $\mu_{Olympians}$ + -4 $\mu_{PolicePlainClothes}$ = 0
H_a: Two-tails
α 0.05

Specify hypotheses for Contrast2
H_0: 1 $\mu_{Criminal}$ + 1 $\mu_{Soldiers}$ + 1 $\mu_{PoliceMilitaryGear}$ + 1 $\mu_{PoliceRegularGear}$ + -5 $\mu_{Olympians}$ + 1 $\mu_{PolicePlainClothes}$ = 0
H_a: Two-tails
α 0.05

Specify hypotheses for Contrast3
H_0: 1 $\mu_{Criminal}$ + 1 $\mu_{Soldiers}$ + 1 $\mu_{PoliceMilitaryGear}$ + 1 $\mu_{PoliceRegularGear}$ + -4 $\mu_{Olympians}$ + 0 $\mu_{PolicePlainClothes}$ = 0
H_a: Two-tails
α 0.05

REPLICATION STUDY

- When we hit the "Calculate power" button, we get:

Power for all tests= 0.3936 [Calculate power] [Calculate minimum sample size]

Test	Estimated Power
ANOVA	0.6828
Contrast1	0.625
Contrast2	0.6456
Contrast3	0.666

- Each test has around a 65% chance of rejecting its H_0, but the probability of **all** tests rejecting the H_0 for one set of samples is only around 40%.

ADJUSTING α

- Bushman (2018) was concerned about multiple tests increasing Type I error, so he set $\alpha = 0.025$ for the second contrast

Specify hypotheses for Contrast1
H_0: 1 $\mu_{Criminal}$ + 1 $\mu_{Soldiers}$ + 1 $\mu_{PoliceMilitaryGear}$ + 1 $\mu_{PoliceRegularGear}$ + 0 $\mu_{Olympians}$ + -4 $\mu_{PolicePlainClothes}$ = 0
H_a: Two-tails
α 0.05

Specify hypotheses for Contrast2
H_0: 1 $\mu_{Criminal}$ + 1 $\mu_{Soldiers}$ + 1 $\mu_{PoliceMilitaryGear}$ + 1 $\mu_{PoliceRegularGear}$ + -5 $\mu_{Olympians}$ + 1 $\mu_{PolicePlainClothes}$ = 0
H_a: Two-tails
α 0.025

Specify hypotheses for Contrast3
H_0: 1 $\mu_{Criminal}$ + 1 $\mu_{Soldiers}$ + 1 $\mu_{PoliceMilitaryGear}$ + 1 $\mu_{PoliceRegularGear}$ + -4 $\mu_{Olympians}$ + 0 $\mu_{PolicePlainClothes}$ = 0
H_a: Two-tails
α 0.05

ADJUSTING α

- When we hit the "Calculate power" button, we get:

Power for all tests= 0.3304 Calculate power Calculate minimum sample size

Test	Estimated Power
ANOVA	0.6776
Contrast1	0.6196
Contrast2	0.5598
Contrast3	0.6724

- The power of the second contrast drops a bit. The other power estimates change, but that is just a side effect of the calculations. We could increase the number of iterations to avoid these changes.
- The power for all tests drops from 40% to around 33%

SAMPLE SIZE

- How big a sample size do we need to have 80% power?
- $n = 223$, which means a total of $6 \times 223 = 1338$ subjects
- The power values would be distributed across the tests as:

Power for all tests= 0.8072 Calculate power Calculate minimum sample size

Test	Estimated Power
ANOVA	0.9712
Contrast1	0.9266
Contrast2	0.8752
Contrast3	0.95

SIMPLE IS BETTER

- If your conclusion depends on many hypothesis tests producing significant results, you should design your study to take into account all of those tests
- Adding tests always lowers power
- Complicated experiments require much larger samples than simple experiments
- Lots of studies that are published are woefully underpowered because they do not consider these details of experimental design

CONCLUSIONS

- power for ANOVA
- power for contrasts
- simple is better

NEXT TIME

- Dependent ANOVA
- Contrasts

Ignoring (some) variability.

PSY 201: Statistics in Psychology

Lecture 35
Analysis of Variance
Ignoring (some) variability.

Greg Francis

Purdue University

Fall 2023

ANOVA TESTING

- 4 STEPS
  1. State the hypothesis. : $H_0 : \mu_1 = \mu_2 = ... = \mu_K$, $H_a : \mu_i \neq \mu_j$ for some i, j.
  2. Set the criterion: α
  3. Compute the test statistic: $F = MS_B/MS_W$, degrees of freedom, and p-value
  4. Interpret results.

ASSUMPTIONS

- to use ANOVA for independent means validly, the data must meet some restrictions
  - The observations are **random** and **independent** samples from the populations.
  - The distributions of the populations from which samples are selected are **normal**.
  - The variances of the distributions in the populations are equal. **Homogeneity of variance**.

ASSUMPTIONS

- it turns out that
- independence of samples is critical
- violations of normality have small effects on Type I error rates
- violations of homogeneity of variance have a big effect if the population sizes are different
  - similar to the standard t test
- means that ANOVA is robust as long as the sample sizes are the same across populations

t tests

- if we have only two groups

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

- we can use either ANOVA or the (standard) t-test discussed previously
- they give identical results!

t tests

- it turns out that the F distribution for $K-1, N-K$ $(1, N-2)$ degrees of freedom is simply the t distribution for $N-2$ df, squared.

$$t^2 = F$$

- so using either technique produces the same results (reject or not reject)

EXAMPLE

- A sociologist wants to determine whether sorority or dormitory women date more often. He randomly samples 12 women who live in sororities and 12 women who live in dormitories and determines the number of dates they each have during the ensuing month. The following are the results.

Sorority Women, X_1	Dormitory Women, X_2
8	9
5	7
6	3
4	4
12	4
7	8
9	7
10	5
5	8
3	6
7	3
5	5
$\overline{X}_1$=6.750	$\overline{X}_2$ =5.750

t TEST

- test with $\alpha = 0.05$, two-tailed

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 \neq 0$$

- we have equal numbers of subjects, so we do not need to worry about homogeneity of variance
- from data we calculate the pooled estimate of population variance

$$s^2 = 5.570$$

t TEST

- so standard error is

$$s_{\overline{X}_1-\overline{X}_2} = \sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$s_{\overline{X}_1-\overline{X}_2} = 0.963$$

- and

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_{\overline{X}_1-\overline{X}_2}}$$

$$t = \frac{1.0}{0.963} = 1.038$$

$$df = n_1 + n_2 - 2 = 12 + 12 - 2 = 22$$

- From the t-distribution calculator, we find

$$p = 0.3105 > 0.05 = \alpha$$

- so do not reject H_0
- no evidence for a difference in number of dates

ANOVA

- The same hypotheses work for an ANOVA

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

- we can calculate

$$SS_B = 6.00$$

$$SS_W = 122.500$$

ANOVA

$$MS_B = \frac{SS_B}{K-1} = \frac{6.00}{1} = 6.00$$

$$MS_W = \frac{SS_W}{N-K} = \frac{122.500}{22} = 5.568$$

$$F = \frac{MS_B}{MS_W} = \frac{6.00}{5.568} = 1.078$$

- we have 1 df in the numerator and 22 df in the denominator, and we use the F-distribution calculator to find

$$p = 0.31042 > 0.05 = \alpha$$

- we do not reject H_0
- note:

$$F = 1.078 \approx 1.077 = (1.038)^2 = t^2$$

DEPENDENT MEASURES

- one way ANOVA deals with independent samples
- we want to consider a situation where all samples are "connected"
- e.g., tracking health patterns for a common set of patients across years; grades for a common set of students throughout school
- Often called a *within subjects* ANOVA or a *repeated measures* ANOVA
- there can be other kinds of dependencies
- e.g., IQ of first-born, second-born, and third-born siblings

SUM OF SQUARES

- scores for an "individual" are dependent
- scores for different "individuals" are independent

$$SS_T = SS_I + SS_O + SS_{Res}$$

- where
  - SS_T is the total sum of square
  - SS_I is the variation among individuals
  - SS_O is the variation among test occasions
  - SS_{Res} is any other type of variation

INDIVIDUALS

- the combined variation among individuals is

$$SS_I = \sum_i K \left(\overline{X}_i - \overline{X}\right)^2$$

- where

$$\overline{X}_i = \frac{\sum_k X_{ik}}{K}$$

- is the average for the ith individual across all observations
- SS_I deviation of individual means from overall mean
- does not correspond to SS_W or SS_B in the normal ANOVA
- we want to *ignore* this variability

OBSERVATIONS

- the combined variation across observations is

$$SS_O = \sum_k n \left(\overline{X}_k - \overline{X}\right)^2$$

- where

$$\overline{X}_k = \frac{\sum_i X_{ik}}{n}$$

- is the average for the kth observation across all subjects
- SS_O deviation of observation mean from overall mean
- similar to SS_B in the independent ANOVA

RESIDUAL

- we need a term that corresponds to SS_W
- we can directly calculate the total sum of squares

$$SS_T = \sum_k \sum_i \left(X_{ik} - \overline{X}\right)^2$$

- if there is variation beyond SS_I and SS_O, we can calculate it as

$$SS_{Res} = SS_T - SS_I - SS_O$$

- this is similar to SS_W
  - factors out variation due to individuals and variation due to observations

VARIANCE ESTIMATES

- $SS_{Res} > 0$ due to random sampling (choice of individuals)

$$MS_{Res} = \frac{SS_{Res}}{(K-1)(n-1)}$$

- estimates the variance of the population distribution
- the degrees of freedom associated with this estimate is

$$(K-1)(n-1)$$

VARIANCE ESTIMATES

- SS_O can vary due to random sampling, or due to differences across observations
- if H_0 is true

$$H_0 : \mu_1 = \mu_2 = ... = \mu_K$$

- then there are no population differences across observations, so all variation must be due to random sampling. So,

$$MS_O = \frac{SS_O}{K-1}$$

- estimates the variance of the population distribution **if H_0 is true**
  - otherwise it overestimates it

F RATIO

- as before we compare these estimates with the F statistic

$$F = \frac{MS_O}{MS_{Res}}$$

- if H_0 is true

$$F \approx 1.0$$

- if H_0 is not true

$$F > 1.0$$

- look up p value using $(K-1)$ and $(K-1)(n-1)$ degrees of freedom
- everything else is the same as before

EXAMPLE

- A school principal traces reading comprehension scores on a standardized test for a random sample of dyslexic students across three years. The data are given below. Complete the ANOVA using $\alpha = 0.05$.

Student	Third Grade	Fourth Grade	Fifth Grade
1	2.8	3.2	4.5
2	2.6	4.0	5.1
3	3.1	4.3	5.0
4	3.8	4.9	5.7
5	2.5	3.1	4.4
6	2.4	3.1	3.9
7	3.2	3.8	4.3
8	3.0	3.6	4.4

(1) HYPOTHESIS

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_a : \mu_i \neq \mu_k \text{for some } i \text{ and } k$$

- use $\alpha = 0.05$

(2) TEST STATISTIC

- It turns out that

$$SS_T = 18.66$$

and

$$SS_I = 5.67$$

and

$$SS_O = 12.0858$$

- so any remaining variation is residual

$$SS_{Res} = SS_T - SS_I - SS_O$$

$$SS_{Res} = 18.66 - 5.67 - 12.09 = 0.9075$$

- this cannot be negative!

(2) TEST STATISTIC

- now calculate

$$MS_O = \frac{SS_O}{K-1} = \frac{12.0858}{2} = 6.0429$$

- and

$$MS_{Res} = \frac{SS_{Res}}{(K-1)(n-1)} = \frac{0.9075}{14} = 0.0648$$

- and get the F statistic

$$F = \frac{MS_O}{MS_{Res}} = \frac{6.0429}{0.0648} = 93.22$$

(3) P-VALUE

- for the numerator (observation sum of squares) we have

$$df = K - 1 = 3 - 1 = 2$$

- for the denominator (residual sum of squares) we have

$$df = (K - 1)(n - 1) = (3 - 1)(8 - 1) = 14$$

- so from the F-distribution calculator, we find the $F = 93.22$ corresponds to

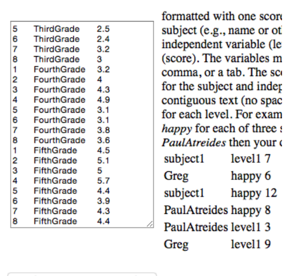
$$p \approx 0.000 < 0.05 = \alpha$$

(4) DECISION

- we reject H_0.
- there is evidence that the reading scores for these subjects are different across the years

CALCULATORS

- No one does these computations by hand. Computer programs do it for you. Your text provides a Dependent ANOVA One-Way calculator.
- You have to format the data correctly



Source	df	SS	MS	F	p-value
Individuals	7	5.6662	0.8095		
Occasions	2	12.0858	6.0429	93.2241	0.00000
Residual	14	0.9075	0.0648		
Total	23	18.6596			

CALCULATORS

- Extra information is important for interpreting the results
- means, correlations
  - Not always reported, but should be

Summary table

Condition	Mean	Standard deviation	Sample size
ThirdGrade	2.9250	0.4559	8
FourthGrade	3.7500	0.6392	8
FifthGrade	4.6625	0.5680	8

Correlation table

	ThirdGrade	FourthGrade	FifthGrade
ThirdGrade	1.0000	0.8383	0.6826
FourthGrade	0.8383	1.0000	0.8912
FifthGrade	0.6826	0.8912	1.0000

CONTRASTS

- We set up *contrast weights*, c_i, for each sample
- Our null hypothesis will be

$$H_0 : \sum_{i=1}^{K} (c_i \mu_i) = 0$$

- and we require that the contrast weights sum to 0:

$$\sum_{i=1}^{K} c_i = 0$$

- Our alternative hypothesis is

$$H_a : \sum_{i=1}^{K} (c_i \mu_i) \neq 0$$

- (one-tailed tests are also possible)

TEST STATISTIC

- We compute the weighted sum of means

$$L = \sum_{i=1}^{K} (c_i \overline{X}_i)$$

- which has a standard error of:

$$s_L = \sqrt{MS_{\text{Res}} \sum_{i=1}^{K} \frac{c_i^2}{n}}$$

- and our test statistic is

$$t = \frac{L}{s_L}$$

- which follows a t distribution with

$$df = (K-1)(n-1)$$

- where N is the sum of sample sizes across all groups and K is the number of groups

CALCULATORS

- A one-tailed contrast to compare scores in Third Grade against scores in Fourth Grade

Specify hypotheses:

H$_0$: -1 $\mu_{\text{ThirdGrade}}$ + 1 $\mu_{\text{FourthGrade}}$ + 0 $\mu_{\text{FifthGrade}}$ = 0

H$_a$: Positive one-tail

α 0.05

Run Contrast

Contrast test summary	
Null hypothesis	H$_0$: (-1)$\mu_{\text{ThirdGrade}}$ + (1)$\mu_{\text{FourthGrade}}$ + (0)$\mu_{\text{FifthGrade}}$= 0
Alternative hypothesis	H$_a$: (-1)$\mu_{\text{ThirdGrade}}$ + (1)$\mu_{\text{FourthGrade}}$ + (0)$\mu_{\text{FifthGrade}}$ > 0
Type I error rate	α=0.05
Weighted sum of sample means	L= 0.8250
Standard error	s_L =0.1273
Test statistic	t=6.4807
Degrees of freedom	df=14
p value	p=0.00001
Decision	Reject the null hypothesis

CALCULATORS

- A one-tailed contrast to compare scores in Fourth Grade against scores in Fifth Grade

Specify hypotheses:

H$_0$: 0 $\mu_{\text{ThirdGrade}}$ + -1 $\mu_{\text{FourthGrade}}$ + 1 $\mu_{\text{FifthGrade}}$ = 0

H$_a$: Positive one-tail

α 0.05

Run Contrast

Contrast test summary	
Null hypothesis	H$_0$: (0)$\mu_{\text{ThirdGrade}}$ + (-1)$\mu_{\text{FourthGrade}}$ + (1)$\mu_{\text{FifthGrade}}$= 0
Alternative hypothesis	H$_a$: (0)$\mu_{\text{ThirdGrade}}$ + (-1)$\mu_{\text{FourthGrade}}$ + (1)$\mu_{\text{FifthGrade}}$ > 0
Type I error rate	α=0.05
Weighted sum of sample means	L= 0.9125
Standard error	s_L =0.1273
Test statistic	t=7.1681
Degrees of freedom	df=14
p value	p=0.00000
Decision	Reject the null hypothesis

ASSUMPTIONS

- ANOVA for dependent measures depends on four assumptions
  - The sample was randomly selected for a population.
  - The dependent variable (e.g., reading scores) is normally distributed in the population.
    - deviations tend to not cause serious problems
  - The population variances for the test occasions are equal. (homogeneity of variance)
    - Can be compensated for sometimes
  - The population correlation coefficients between pairs of test occasion scores are equal.
    - Can be compensated for sometimes

CONCLUSIONS

- assumptions of one-way independent ANOVA
- ANOVA for dependent measures
- contrasts for dependent ANOVA
- assumptions of dependent ANOVA

NEXT TIME

- power for dependent ANOVA

Leverage relationships.

PSY 201: Statistics in Psychology

Lecture 36

Power for Dependent ANOVA

Leverage relationships.

Greg Francis

Purdue University

Fall 2023

HYPOTHESES

- The null for a dependent ANOVA is an *omnibus* hypothesis. It supposes no difference between any population means

$$H_0 : \mu_i = \mu_j \; \forall \; i, j$$

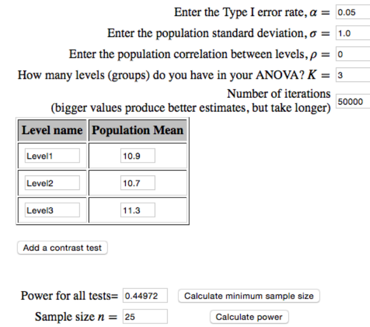
- the alternative is the complement

$$H_a : \mu_i \neq \mu_j \;\; \text{for some } i, j$$

- To compute power, we have to provide the standard deviation, α, n, specific values for the means, and the correlation (ρ) between the different measures

POWER CALCULATOR

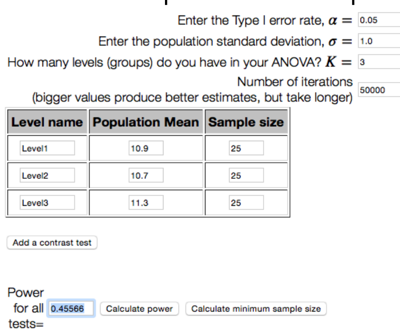
- Consider a situation with $K = 3$ dependent means (all different from each other), $n = 25$, and $\rho = 0$:



- We estimate the power to be 0.45

INDEPENDENT CALCULATOR

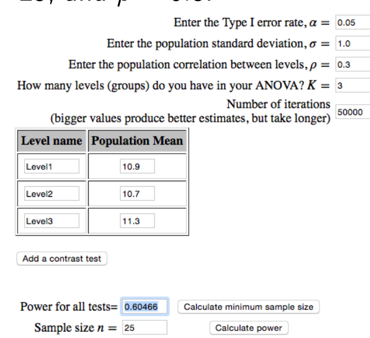
- Since we have $\rho = 0$, the means are independent. Thus, we get nearly the same result with the independent means power calculator



- We estimate the power to be 0.46
- when $\rho = 0$, the independent and dependent ANOVA are almost the same test

DEPENDENT POWER CALCULATOR

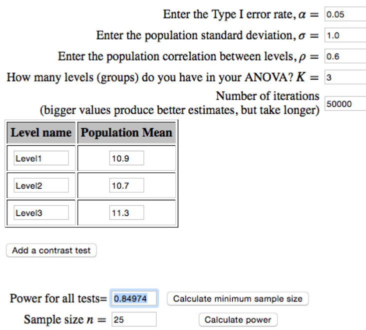
- Increasing the correlation increases the power
- Consider a situation with $K = 3$ dependent means (all different from each other), $n = 25$, and $\rho = 0.3$:



- We estimate the power to be 0.6

DEPENDENT POWER CALCULATOR

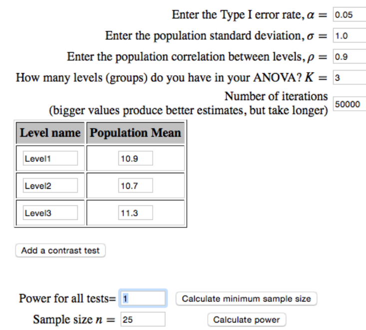
- Consider a situation with $K = 3$ dependent means (all different from each other), $n = 25$, and $\rho = 0.6$:



- We estimate the power to be 0.85

DEPENDENT POWER CALCULATOR

- Consider a situation with $K = 3$ dependent means (all different from each other), $n = 25$, and $\rho = 0.9$:



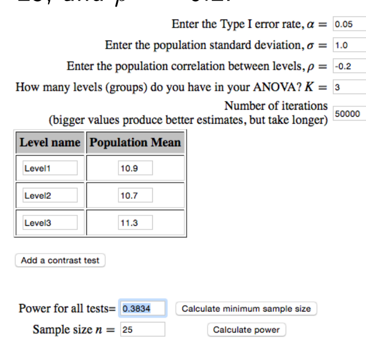
- We estimate the power to be 1.0

DEPENDENT POWER CALCULATOR

- Positive correlations are easy to imagine:
- e.g., reading scores in third, fourth, and fifth grades are positively correlated with each other
- It is plausible that the correlations are (nearly) the same for all variables
- Negative correlations for all variables would be weird
- e.g., GPA for athletes over three different sports seasons
- kind of suggests multiple factors influencing behavior
- for three measures, a negative correlation can be no stronger than $\rho = -0.5$
- the calculator will take your negative correlation and try to do something, but be skeptical about the results

NEGATIVE CORRELATIONS

- Consider a situation with $K = 3$ dependent means (all different from each other), $n = 25$, and $\rho = -0.2$:



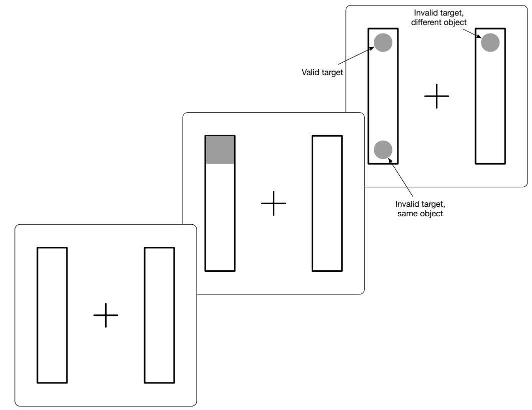
- We estimate the power to be 0.38 (worse than when $\rho = 0$)

OBJECT BASED ATTENTION

- The eye is (kind of) like a camera, with photoreceptors that are similar to pixels in a camera
- However, what people *see* corresponds to objects that are somehow "grouped" together
- We can *select* and *attend* some objects to the exclusion of other objects
- One way of measuring this property of visual perception is to study the "object based attention" effect

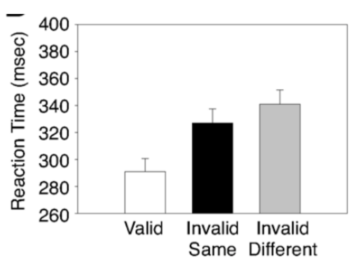
OBJECT BASED ATTENTION

- Measure reaction time (RT) to the target
- Dependent design: each subject provides data for 3 types of targets



PREVIOUS DATA

- A study by Marrara & Moore (2003) found the following results for $n = 19$:



- ANOVA finds: $F_{2,36} = 100.63$, $p \approx 0$
- Contrast for RT on valid trials vs. RT on invalid-same trials: $t_{36} = 11.76$
- Contrast for RT on invalid-same trials vs. RT on invalid-different trials: $t_{36} = 4.13$ (this is the object based attention effect)

REPLICATION

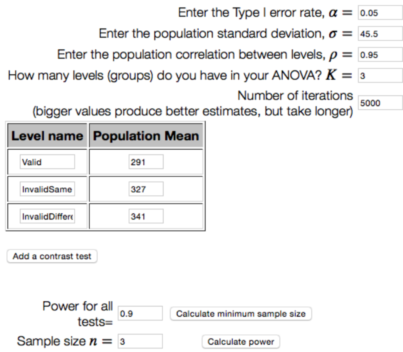
- The study was done 15 years ago (before everyone spent all day staring at a phone). You might want to repeat it with current students to make sure the object based attention effect still exists.
- To design your experiment, you can use the original data to do a power analysis. It takes a bit of effort, but you find that the data has:

$$\overline{X}_{\text{Valid}} = 291, \quad \overline{X}_{\text{InvalidSame}} = 327, \quad \overline{X}_{\text{InvalidDifferent}} = 341$$

$$s = 45.5, \quad r = 0.95$$

POWER CALCULATOR

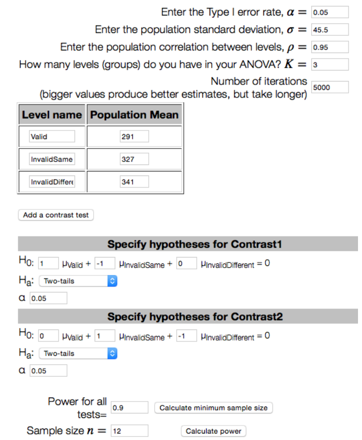
- For just the ANOVA



- To have 90% power, we need only 3 subjects (if the effects are similar to the original study)

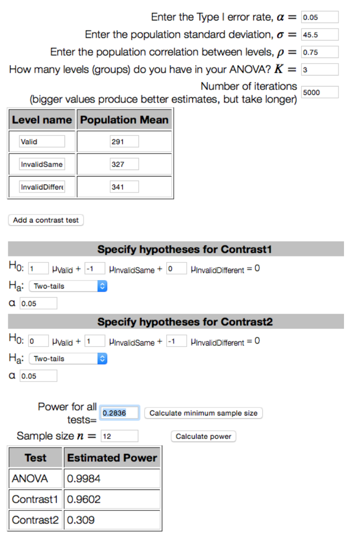
POWER CALCULATOR

- We can add the contrasts:
- Now, we need $n = 12$ to get 90% power



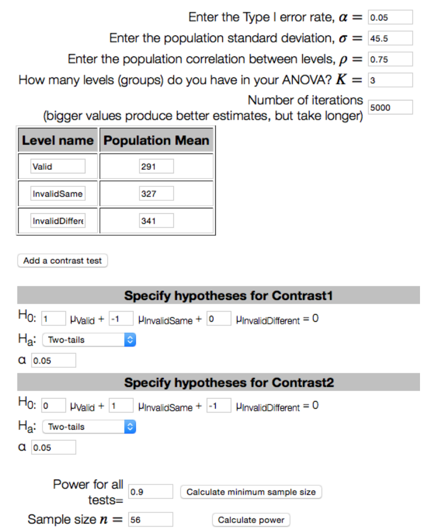
CORRELATION

- The original study found $r = 0.95$, which seems rather high. Maybe we think it should be smaller, say $r = 0.75$.
- What is the impact on power if we use $n = 12$?
- Power drops to 0.28!



CORRELATION

- The original study found $r = 0.95$, which seems rather high. Maybe we think it should be smaller, say $r = 0.75$. What sample size do we need to have 90% power?
- Need $n = 56$! The correlation makes a *big* difference in dependent means experiments



CONCLUSIONS

- power for dependent ANOVA
- power for contrasts

NEXT TIME

- Catch all
- Challenges with hypothesis testing
- Questionable Research Practices

Tell the truth!

PSY 201: Statistics in Psychology

Lecture 37

Catch all

Tell the truth!

Greg Francis

Purdue University

Fall 2023

HONESTY

- It is important that you are honest about what happened in an experiment and in its analysis
  - To yourself
  - To other researchers

HONESTY

- "The first principle is that you must not fool yourself – and you are the easiest person to fool." Richard Feynman
- Let's look at several ways you can fool yourself
  - Optional stopping
  - Pilot studies
  - HARKing

OPTIONAL STOPPING

- Hypothesis testing is a *procedure* that controls the Type I error rate
- It works when we know the sampling distribution for the null hypothesis
- The sampling distribution depends on the sample size, so we have to know that
  - Seems trivial, just see how many subjects you have
  - But no.
  - The sample size that matters is how many subjects you *would* run if you repeated the experiment many times
- Surprisingly, many people do not know how many subjects they would run if the experiment were repeated

OPTIONAL STOPPING

- Suppose you run a between subjects test of means with $n_1 = n_2 = 25$

$$t = 2.0$$

  which gives

$$p = 0.0512$$

- Some people call this a "marginally significant" result, meaning it is close to the $\alpha = 0.05$ criterion
  - This is nonsense, what you have is a non-significant result
  - You do not get to conclude anything
- But you might think that the results are suggestive, so you run 10 more subjects in each group. Now, with $n_1 = n_2 = 35$, you get

$$t = 2.2$$

$$p = 0.0312$$

- What is the Type I error rate for this kind of procedure?

OPTIONAL STOPPING

- The Type I error rate has to be bigger than $\alpha = 0.05$ because your first test had that error rate
- The second test (with added subjects) has some unknown additional Type I error rate
- What would you have done if the second test produced?:

$$t = 1.99$$

$$p = 0.0506$$

- If you would have run even more subjects, then you need to consider those steps as also being part of the Type I error rate of your *procedure* (even if they were not actually done in your particular situation)

OPTIONAL STOPPING

- In fact, you have to know what you would do for every (infinitely many) possible situations
- If you are willing to keep adding subjects until you get a significant result, your Type I error rate is 1.0!
- The best way out of this problem is to *fix* a sample size and stick to it. This is best done with a good power analysis before gathering any data.
- If that is not possible, then honestly describe what was done. Describe each test and explain why subjects were added.

PILOT STUDIES

- When investigating a new topic, it is common to run multiple experiments while identifying what to measure and how to do it
- For example, suppose you want to study the effect of eating bananas on recall of words
- There are lots of variables to consider
  - How many bananas?
  - How long after eating do you study words?
  - How long after eating do you test?
  - What kind of words do you use?
- You can explore hundreds of these variables to find a combination that shows an effect
- It might be tempting to use statistical significance to decide whether a study "works"
- Don't do it!

PILOT STUDIES

- What sometimes happens is people interpret the difference between significant and non-significant results as indicating *methodological* differences:
  - One banana does not improve memory (warning: accepting the null!)
  - Two bananas does improve memory
- But both of these studies also involve random sampling
  - A null effect might produce a Type I error
  - A real effect might not produce a significant result (Type II error)
- It is dishonest to run tests like these and not report the results, even if you think you can explain why a study "failed"

PILOT STUDIES

- If you only report successful studies (publication bias), it becomes impossible for other scientists to interpret the Type I error rate of your results
- They do not know if you are reporting the only result you tested for
- Or if you are reporting one study out of dozens of others that did not work
- Your best bet is to figure out how to run a good study and then do it once.
  - Easier said than done
  - You might spend years figuring out how to run a good study

HARKing

- Hypothesizing After the Results are Known
- You sometimes learn a *lot* after looking at your data
- Sometimes scientists look at the data then identify hypotheses that match the results
- Sometimes scientists then "pretend" that they predicted the outcome and write up their paper accordingly
- Don't do this. It is fraud.
- Just be honest and explain that you learned from your findings.

HARKing

- Be careful about what you learn, though
- You might run a study on eating bananas and word memory and sift through a large set of subjects to find a subset that shows an effect
  - Age: Young, Middle, Old
  - Sex: male, female
  - Socioeconomic status: quartile
  - Religious affiliation: Christian, Muslim, Jewish, Atheist, Buddhists, Other
- Maybe you find:
  - A significant improvement for Young, male, 25th percentile SES, Buddhists
  - A significant decrement for Young, female, 75th percentile SES, Christians
  - A significant increment for Young, female, 10th percentile SES, Atheists
  - A significant decrement for Old, female, 25th percentile SES, Jews

HARKing

- Sometimes these kinds of conclusions feel like *aha!* moments, where you suddenly have deep insight into what is going on
- Based on other research, you realize all the increments are for people with "high sense of self" while the decrements are for people with "low sense of self" (I'm just making these terms up)
- You are very possibly seeing "signal" in pure "noise"
- What you are doing is exploratory work
- It is (maybe) good for coming up with ideas, but you cannot use one set of data both to identify ideas and to test them
- In a follow-up test you need to measure whatever other variables you think really matter (e.g., "sense of self")

IN THE WILD

- Scientists do these kinds of "questionable research practices" all the time
  - Often unintentionally
  - They just do not know any better
- This is why you hear so much conflicting advice on some topics
  - Chocolate is good for you / chocolate is bad for you
  - A glass of wine a day is good for you / no it's not
  - Take statins to improve your health / they seem to do nothing
- This is why you sometimes see nonsense published in journals
  - People can get information from the future
  - Eating breakfast makes a woman more likely to have a boy baby
  - Women find men wearing red shirts to be more attractive

STATISTICS LIMITS

- Hypothesis testing (and statistics in general) is not synonymous with science
- Science is about identifying mechanisms to explain why things happen the way they do
- Hypothesis testing (at best) prevents misinterpretations of signal for noise, but that is not enough to identify mechanisms
- At best, statistics is a check on interpreting noise as if it were signal
- At worst, statistics is a way of "validating" noise as if it were a signal
- In some sense, the best science does not require statistics

CONCLUSIONS

- Good science is difficult to do well
- There are lots of ways to "cheat" hypothesis testing
- People actually do cheat
- Be skeptical about published work
- Use some common sense!

NEXT TIME

- Review for the final exam