PSY 201: Statistics in Psychology

Lecture 20
Power
Plan ahead!

Greg Francis

Purdue University

Fall 2019

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:

$$H_a : \mu = 20.5$$

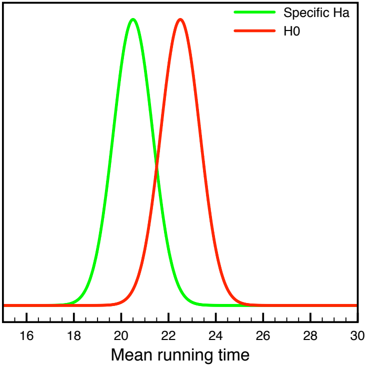
- If she runs the experiment, she will typically get $\overline{X}$ close to 20.5 minutes, but it will vary from sample to sample

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

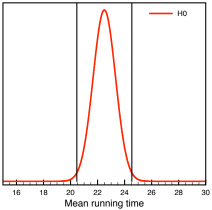
- When doing the hypothesis test, the runner will compute a t-value and see if it is more extreme than

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

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

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

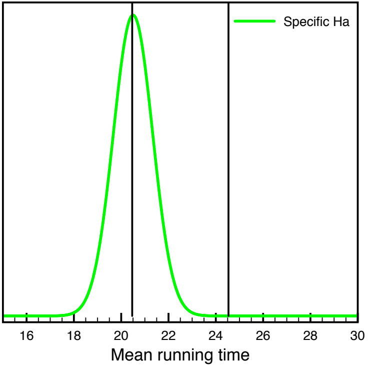
$$\overline{X}_{upper} = \mu + s_{\overline{X}} t_{cv} = 22.5 + (0.8315)(2.44691) = 24.535$$



- Note the region of rejection!

POWER

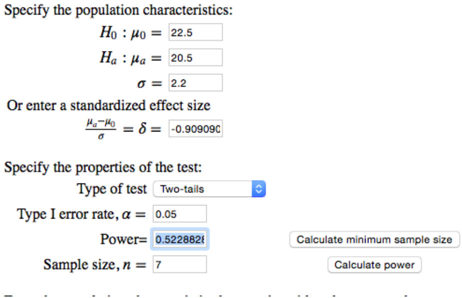
- If the specific H_a is true, then the probability of rejecting the null is the area under the distribution for the specific H_a over the region of rejection



- Looks pretty close to 0.5

CALCULATOR

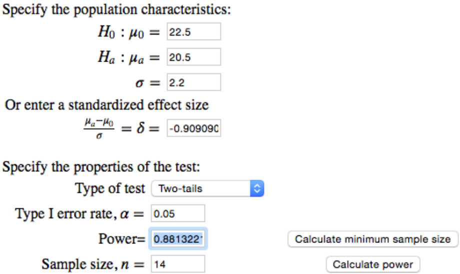
- Online power calculator does the work for you



- 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:

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: Two-tails

Type I error rate, $\alpha =$ 0.05

Power= 0.8 [Calculate minimum sample size]

Sample size, $n =$ 12 [Calculate 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:

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: Two-tails

Type I error rate, $\alpha =$ 0.05

Power= .95 — Calculate minimum sample size

Sample size, $n =$ 18 — Calculate power

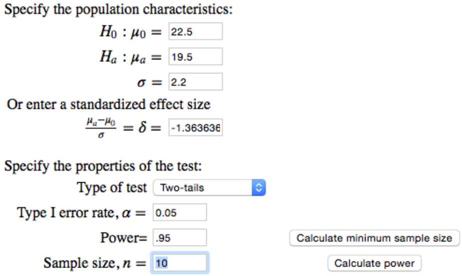
- the runner needs to gather data over 18 runs

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

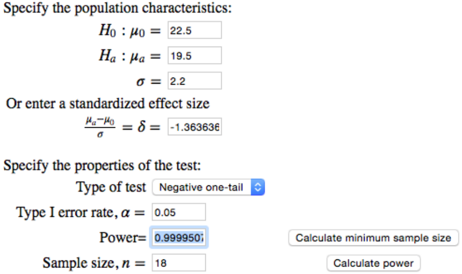
- Power increases as the difference between μ_0 and μ_a increases.
  - Bigger signal is easier to detect.
- Suppose the runner used a "smallest meaningful" effect of 3 minutes, so $\mu_a = 19.5$ compared to the null hypothesis, $\mu_0 = 22.5$



- 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 $\mu_a = 19.5$ 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



DIRECTIONAL HYPOTHESIS

- One-tailed tests are more powerful than two-tailed tests, provided the effect is in the correct tail
- Using $\mu_a = 19.5$ 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

$H_a : \mu_a =$ 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: Negative one-tail

Type I error rate, $\alpha =$ 0.05

Power= 0.95 — Calculate minimum sample size

Sample size, $n =$ 8 — Calculate power

α CRITERION

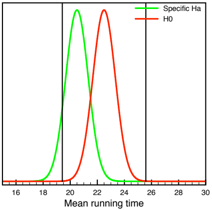
- Suppose the runner plans to use a criterion of $\alpha = 0.01$. Then, when doing a two-tailed hypothesis test with $n = 7$, the runner will compute a t-value and see if it is more extreme than

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

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

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

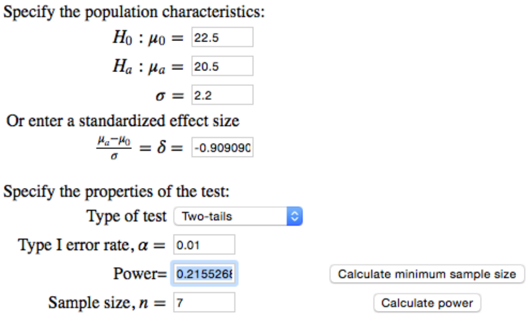
$$\overline{X}_{upper} = \mu + s_{\overline{X}} t_{cv} = 22.5 + (0.8315)(3.70743) = 25.583$$



- Power will be smaller!

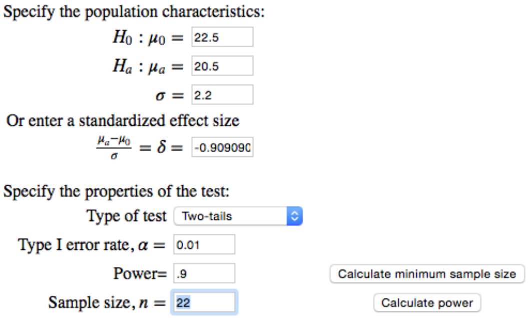
α CRITERION

- 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?