STAT 113
Statistical Power and Effect Size

Colin Reimer Dawson

Oberlin College

November 21, 2017

Outline

Targeting a Specific Margin of Error

- Recall that a Confidence Interval consists of a *point estimate* together with a *margin of error* (the number we add and subtract from the point estimate).
- When designing a study, we may desire a specific margin of error.
- E.g., estimate the mean mercury level within a particular number of parts per million.
- How can we decide how much data we need to get that level of precision?

Outline

Sample Size Calculations

- For a proportion:

$$CI : \hat{p} \pm Z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- For a mean:

$$CI : \bar{x} \pm T^* \cdot \sqrt{\frac{s^2}{n}}$$

- Goal: set the MoE (in blue) to the target value, and solve for n.

Sample Size Calculations

For a proportion:

$$\text{MoE} = Z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\text{MoE}^2 = (Z^*)^2 \frac{\hat{p}(1-\hat{p})}{n}$$

$$n = \left(\frac{Z^*}{\text{MoE}}\right)^2 \hat{p}(1-\hat{p}) \leq \frac{1}{4}\left(\frac{Z^*}{\text{MoE}}\right)^2$$

where we can be conservative/pessimistic and set $\hat{p} = 1 - \hat{p} = \frac{1}{2}$ to give the largest possible standard error.

Example: Polling

We want to estimate the extent of popular support for marijuana legalization in Ohio with a 95% margin of error of no more than 2%. How many people do we need to poll?

$$n = \left(\frac{Z^*}{\text{MoE}}\right)^2 \hat{p}(1-\hat{p}) \leq \frac{1}{4}\left(\frac{Z^*}{\text{MoE}}\right)^2$$

Sample Size Calculations

For a mean:

$$\begin{aligned}
\text{MoE} &= T^* \cdot \sqrt{\frac{s^2}{n}} \\
\text{MoE}^2 &= \frac{(T^* \cdot s)^2}{n} \\
n &= \left(\frac{T^* \cdot s}{\text{MoE}}\right)^2 \\
&\approx \left(\frac{Z^* \cdot \tilde{\sigma}}{\text{MoE}}\right)^2
\end{aligned}$$

where we sub the smaller Z^* for T^* because it does not depend on the sample size, and our "best guess", $\tilde{\sigma}$, for the population variance, erring high to be safe, and to compensate for using Z^*.

Example: Arsenic

We want to estimate the mean level of Arsenic in chicken from a supplier with a 99% margin of error of no more than 5 ppb. We estimate that the standard deviation across chickens is about 25 ppb. How many chickens do we need to test?

$$n \approx \left(\frac{Z^* \cdot \tilde{\sigma}}{\text{MoE}} \right)^2$$

Sample Size Calculations

For a difference of proportions:

$$\text{MoE} = Z^* \cdot \sqrt{\frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}}$$

$$(\text{MoE})^2 = (Z^*)^2 \left(\frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B} \right)$$

Assuming $n_A = n_B =: n_{each}$ and setting $p_A = p_B = 0.5$ to be conservative:

$$(\text{MoE})^2 \approx (Z^*)^2 \left(2\frac{\frac{1}{2}(\frac{1}{2})}{n_{each}} \right) = \frac{(Z^*)^2}{2n_{each}}$$

$$n_{each} \approx \frac{1}{2} \left(\frac{Z^*}{\text{MoE}} \right)^2$$

Sample Size Calculations

For a difference of means of independent groups:

$$\begin{aligned}
\text{MoE} &= T^* \cdot \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} \\
&\approx Z^* \cdot \sqrt{\frac{2\tilde{\sigma}_{each}^2}{n_{each}}} \\
n_{each} &\approx 2\left(\frac{\tilde{\sigma}_{each} Z^*}{\text{MoE}}\right)^2
\end{aligned}$$

Sample Size Calculations

For a correlation:

$$\begin{aligned} \text{MoE} &= T^* \cdot \sqrt{\frac{1-r^2}{n-2}} \\ &\approx Z^* \cdot \sqrt{\frac{1}{n-2}} \\ n &\approx 2 + \left(\frac{Z^*}{\text{MoE}}\right)^2 \end{aligned}$$

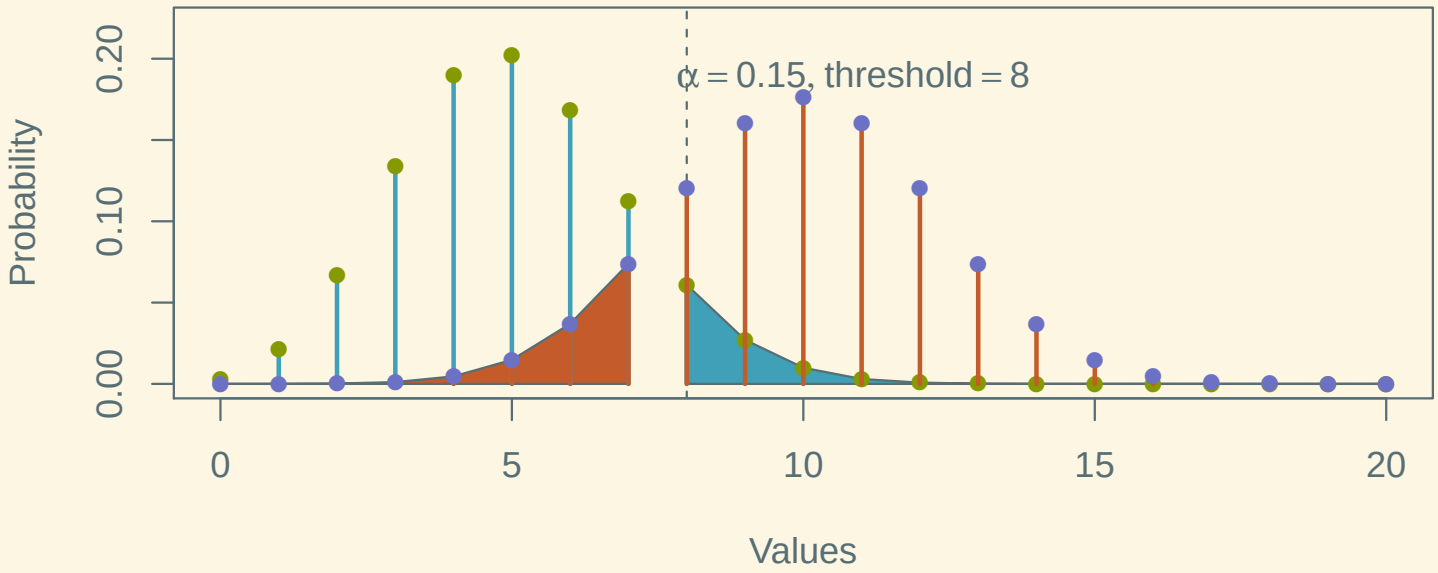
Outline

Targeting a Type II Error Rate

- If we are doing a test, then we might want to choose n in order to control our Type II Error rate. Larger $n \rightarrow$ lower error rate. (Why?)
- The **power** of a test $= 1 -$ Type II Error Rate. Depends on
  1. significance level, α
  2. the distance between μ_{true} and μ_0 (or whatever parameter)
  3. population variability (via standard error)
  4. sample size (via standard error)

Type I vs. Type II Errors

We retain H_0 when we do not exceed the threshold. But if H_1 is correct, this is a Type II Error. More stringent threshold $\rightarrow$ more missed discoveries.

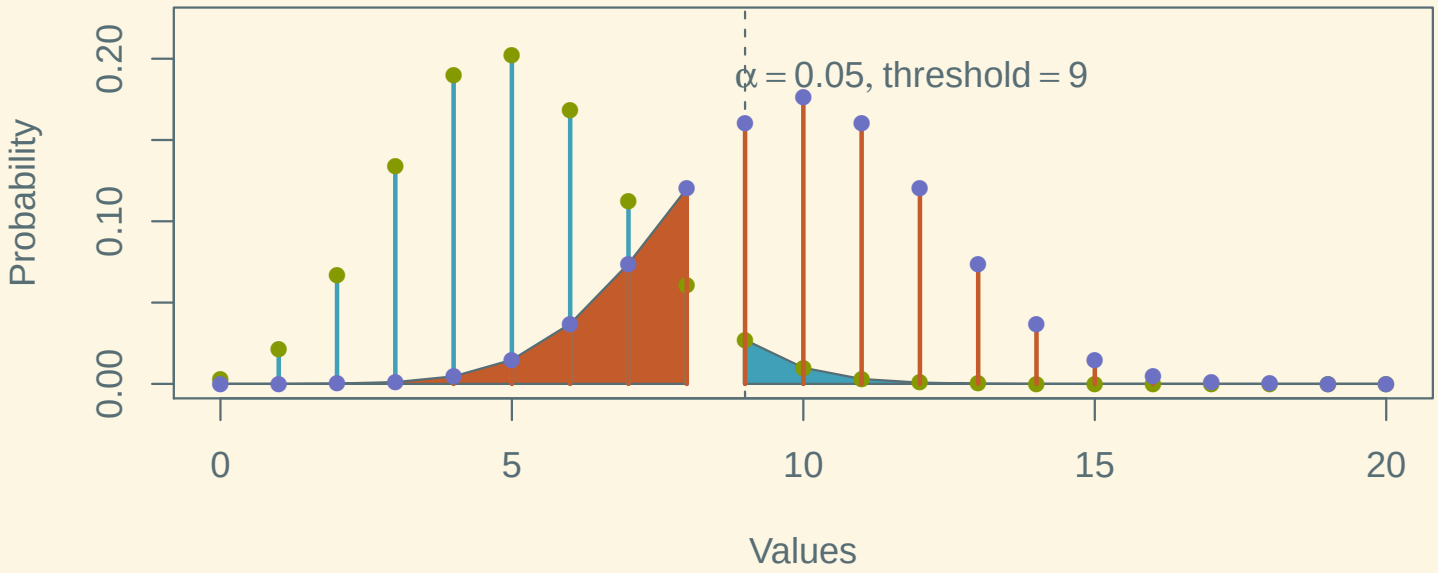


Blue spikes: Distribution of outcomes if H_0 is true

Orange spikes: Distribution of outcomes for one possible parameter value under H_1.

Type I vs. Type II Errors

We retain H_0 when we do not exceed the threshold. But if H_1 is correct, this is a Type II Error. More stringent threshold $\rightarrow$ more missed discoveries.

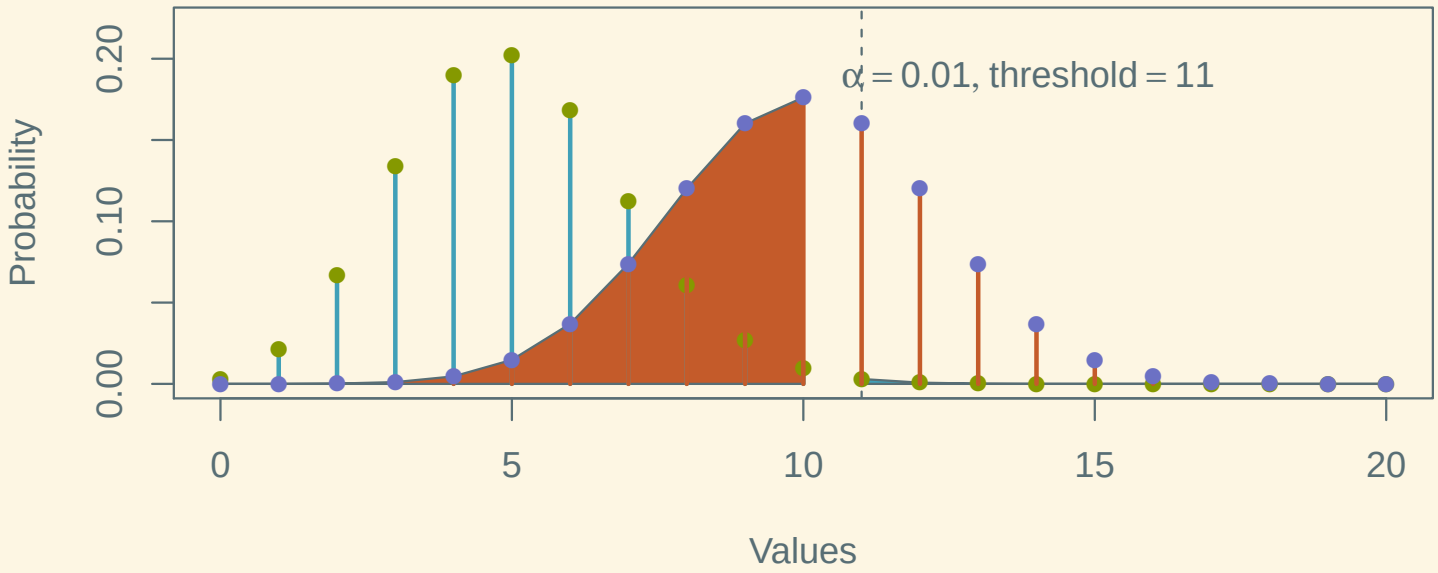


Blue spikes: Distribution of outcomes if H_0 is true

Orange spikes: Distribution of outcomes for one possible parameter value under H_1.

Type I vs. Type II Errors

We retain H_0 when we do not exceed the threshold. But if H_1 is correct, this is a Type II Error. More stringent threshold $\rightarrow$ more missed discoveries.



Blue spikes: Distribution of outcomes if H_0 is true

Orange spikes: Distribution of outcomes for one possible parameter value under H_1.

Estimating Power

- We can calculate a **power function**: for each combination of parameter value and sample size, find power using chosen α and estimated pop. variability.
- At the board

Outline

Statistical Significance Vs. Practical Significance

- With large enough n, any trivial discrepancy between null param. and truth can be detected.
- We may not always care about every difference.
  - A new insulin pump leads to a highly statistically significant increase in life expectancy for diabetics ($P < 0.001$).
  - ... The estimated mean increase over the state of the art is 0.02 years.
  - ... And it costs $12000

Outline

Effect Size

- Always a good idea to report confidence intervals (for, e.g., $\mu_A - \mu_B$) along with tests. Don't just report that "the difference is statistically significant".
- Also a good idea to report an **effect size**, putting the difference from H_0 in the context of (estimated) population variability.
- Regression/correlation: R^2
- Mean(s) / Proportion(s): **Cohen's** d

$$\text{Cohen's } d = \frac{\text{Observed Difference}}{\text{Within Group St. Dev.}}$$

  - Tells us the *number of standard deviations the sample stat is from the null param.*

Cohen's d Calculation

- Single proportion:

$$\text{Cohen's } d = \frac{|\hat{p} - p_0|}{\sqrt{p_0(1-p_0)}}$$

- Single mean:

$$\text{Cohen's } d = \frac{|\bar{x} - \mu_0|}{s}$$

- Two proportions:

$$\text{Cohen's } d = \frac{|\hat{p}_A - \hat{p}_B|}{\sqrt{\hat{p}_{combined}(1-\hat{p}_{combined})}}$$

- Two means:

$$\text{Cohen's } d = \frac{|\bar{x}_A - \bar{x}_B|}{\sqrt{\frac{n_A s_A^2 + n_B s_B^2}{n_{combined}}}}$$

Cohen's d Interpretation

$$\text{Cohen's } d = \frac{\text{Observed Difference}}{\text{Within Group Variability}}$$

Tells us the *number of standard deviations the sample stat is from the null param.*

- $d \approx 0.2$ considered a "small effect"
- $d \approx 0.5$ considered a "medium effect"
- $d \approx 0.8$ considered a "large effect"