Lecture 16: Sample Size and Power

Chapter 4.6

Last Time: Reedie Sleep Example

Tested number of hours of sleep:

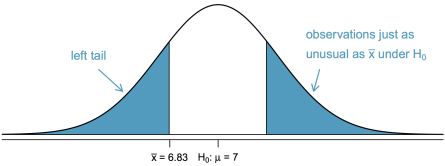
- $H_0 : \mu = 7$
- $H_A : \mu > 7$

Two-Sided Alternative Hypothesis

Say instead we had a two-sided alternative hypothesis:

- $H_0 : \mu = 7$
- $H_A : \mu \neq 7$

The the p-value would be double: $2 \times 0.007 = 0.014$. Picture:



Setting α

Say Dr. Quack is conducting a hypothesis tests. They start with $\alpha = 0.05$.

They conduct the test and get p-value $= 0.09$. They then declare "having used an $\alpha = 0.10$, we reject the null hypothesis and declare our results to be significant."

What's not honest about this approach?

Ronald Fisher, the creator of p-values, never intended for them to be used this way: http://en.wikipedia.org/wiki/P-value#Criticisms

Goals for Today

- More in depth discussion of
  - 10% sampling rule
  - Skew condition to check to use the normal model
- How big a sample size do I need?
- Statistical power
- Statistical vs practical significance

10% Sampling Rule

Question: Why do we set n to be less than 10% of the population size N?

Intuition: Shouldn't we always sample as many people as we can?

Answer: Yes, if we only care about the mean. If we also care about the SE, then we need to be careful.

Explanation: Recall from HW5 Q1, sampling without replacement from a rooms that are half male/female but with $N = 10$ and $N = 10000$.

Finite Population Correction

Sampling

We can tie the conceptual and mathematical notions of sampling:

Conceptual: If we sample everybody, we know the true μ.

and

Mathematical: If $n = N$ then $FPC = \sqrt{\frac{N-n}{N-1}} = 0$ then $SE = \frac{\sigma}{\sqrt{n}} \times FPC = 0$

i.e.

- the sampling distribution is just one point: the true μ.
- if we repeat this procedure many times, we get the same value each time: 0 variability.

Sampling and the SE

Question: Why do we care that our SE is correct?

Answer: If not

- the *SE* in confidence intervals is off
- the z-scores of $\overline{x}$ have the wrong denominator

Skew Condition to Check to Use Normal Model

Throughout the book, they talk about the condition for $\overline{x}$ being nearly normal and using s in place of σ in $SE = \frac{\sigma}{\sqrt{n}}$:

- On page 164: the population distribution is not strongly skewed
- On page 167: the data are not strongly skewed
- On page 168: the distribution of sample observations is not strongly skewed
- On page 185: the population data are not strongly skewed

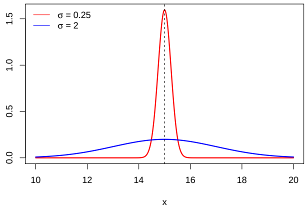
Skew Condition to Check to Use Normal Model

However, they all mean the same thing:

1. The true population distribution from which you are drawing your sample observations/data $x_1, \ldots, x_n$ is not too skewed.
2. The histogram (visual estimate) of the sample observations/data $x_1, \ldots, x_n$ is not too skewed.

Sample Size: Thought Experiment

Say you have two distributions with $\mu = 15$ but different σ.



Which of the two distributions do you think will require a bigger n to estimate μ "well"?

Margin of Error

Recall our formula for a 95% confidence interval:

$$\left[\overline{x} - 1.96\frac{s}{\sqrt{n}},\ \overline{x} + 1.96\frac{s}{\sqrt{n}}\right]$$

The margin of error is half the width of the CI.

Say we knew the true standard deviation σ, then

$$\text{Margin of Error} = 1.96\frac{\sigma}{\sqrt{n}}$$

Identify n for a Desired Margin of Error

Identify n for a Desired Margin of Error

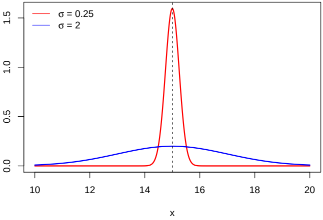
Since

So

- As σ goes up, you need more n
- As z^* goes up, i.e. higher confidence level, you need more n
- As the desired margin of error goes down, you need more n

Back to Thought Experiment

For the same desired maximal margin of error m and same confidence level, we need a larger n to estimate the mean of the blue curve:



Type II Error Rate and Power

Type II Error Rate and Power

Practical vs Statistical Significance

When rejecting the null hypothesis, we call this a statistically significant result. But statistically significant results aren't always practically significant.

Example: say we are comparing the average exam score of men μ_M and women μ_W. We can do a two-sample test:

- $H_0 : \mu_M - \mu_F = 0$ (same average exam score)
- $H_A : \mu_M - \mu_F \neq 0$ (different average exam score)

Practical vs Statistical Significance

Say for very large n_M & n_F we observe $\overline{x}_M = 19.0002$ and $\overline{x}_F = 19.0001$.

The point estimate of $\mu_M - \mu_F$ is $\overline{x}_M - \overline{x}_F = 0.0001$. This difference is near negligible, it is still possible to "reject H_0 at an α-significance level."

However, the 95% confidence interval on the difference might look like

$$[0.00005, 0.00015]$$

Practical vs Statistical Significance

Moral of the story

- Hypothesis tests with "rejections of H_0" focus almost entirely on statistical significance.
- Confidence intervals allow you to also focus on practical significance.