Sampling & Standard Error

PSY 3307

Jonathan A. Pedroza, PhD

Cal Poly Pomona

2022-02-17

Making Predictions

- we can use the mean and the sum of squared errors/sum of squares to see the best fit of the model
    - if we wanted to, we could guess and test/predict a number we believe would result in a best fit
- using the mean, it will provide a better fitting model, not a good fit, but better than randomly choosing

Standard Error

- when interested in how our sample is representative of the data, we have to use the **standard error**
- when we take a sample from the population, we are taking a sample of many possible samples of that population
  - Ex: 300 students from CPP (~26,000)
- if possible, we could get the population mean (mu), which is the parameter (population) that we are trying to estimate

$$\mu$$

Standard Error

- when using a sample, we are estimating the population mean from a sample mean (X bar)

- if we keep sampling from the same population, we get a different value, this is due to **sampling variation/variability**

    - another 300 students from 26,000
    - samples vary from one another in the population

Standard Error

```r
library(tidyverse)
set.seed(2172022)
random_data <- rnorm(26000, mean = 25, sd = 4.6)
random_data <- as_tibble(random_data)
random_data$value[1:100]
```

```
##   [1] 30.85734 18.72664 30.04358 23.24918 26.52253 25.74549 19.36471 22.317
##   [9] 24.05391 32.78262 23.62539 31.24831 22.24254 25.72913 25.07690 29.513
##  [17] 20.89285 29.48269 24.94836 22.27650 26.72770 22.05872 20.72287 27.483
##  [25] 15.21358 29.05867 28.01164 18.10329 20.30965 19.59180 24.99068 18.844
##  [33] 26.08931 24.50260 15.25826 26.07062 18.43159 32.95596 28.66938 27.646
##  [41] 31.09853 27.28187 19.40615 27.85964 24.37489 18.58874 18.24395 25.381
##  [49] 22.69417 26.36015 25.18411 31.02851 25.19970 35.37291 26.79003 13.995
##  [57] 32.60972 22.40780 30.54944 27.41149 25.96758 25.65285 21.97346 27.116
##  [65] 19.00856 17.53705 21.98350 34.76241 19.19105 21.03400 18.49232 18.782
##  [73] 22.63451 27.04031 22.44856 20.75731 19.94058 24.61535 22.76806 26.187
##  [81] 26.95327 27.67388 19.65787 26.35765 18.32739 34.94460 25.52944 25.609
##  [89] 21.30055 29.50874 28.53688 25.53970 28.73167 20.69251 24.16129 27.562
##  [97] 24.41874 20.45842 25.49285 23.71957
```

Standard Error

```
random_data %>%
  sample_frac(.01) %>%
  summarize(mean = mean(value))
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1  25.2
```

```
random_data %>%
  sample_frac(.01) %>%
  summarize(mean = mean(value))
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1  24.9
```

Standard Error

- **sampling distribution** is the distribution of sample means from the population
    - no longer are we focused on values for each individual
    - values are for a sample of scores for a sample (group of individual responses)
    - the mean of the sampling distribution (of all the samples = population mean)
- **standard error** is the standard deviation of the sample means
    - also known as the standard error of the mean
    - **central limit theorem** states that if you have a sampling distribution with samples > 30, you'll have a normal distribution

Standard Error

$$\sigma_{\overline{X}} = \frac{S}{\sqrt{N}}$$

- we can approximate the standard error (SE), using the pouplation standard deviation, but since we don't tend to know that, we can use the sample standard deviation to estimate

You okay JP?

Ex: Let's talk foin flips

- say you flip a code 20 times
- 8 times it is heads
- 12 times it is tails

- **Probability** = how many times is heads going to occur out of all the total number of events
    - 8/20 heads probability
    - 12/20 tails probability

- **Odds** = how many times is heads going to occur compared to how many times is tails going to occur
    - the odds of getting heads is 8 to 12
    - or if you divide by the common denominator, you get odds of 2 to 3 odds of getting heads

Confidence Intervals

- since we are estimating, we can never be 100%
- we use **confidence intervals** to have some confidence in what we are estimating by stating that whatever value we get, we are __% confident that the true value lies within the intervals
    - our estimate, **the point estimate** is whatever value we have for our sample
    - **confidence intervals** are the upper and lower limits around the sample value (point estimate) as a midpoint
- we typically use 95% confidence intervals
    - 95% confident that our value (population mean) is within the limits

Setting up Sampling Distribution

- **Criterion** is the probability that defines whether a sample is reflective of a population or not
    - .05 or .01 is often used in the social sciences
- Sample means within the 5% or 1% of the sampling distribution are unlikely to occur and we reject that our sample is representative of the population
- If we are talking about scores above or below our criterion then we look at both tails
    - if we are only interested in scores above or below then we focus on extreme values in one tail of the sampling distribution

Identifying the Critical Value

- We'll first do this with z-scores
- **Critical value** is the score that marks the inner edge of the region of rejection in a sampling distribution
  - values outside this critical value are rejected as being representative of the population
- We get our critical value by looking at the criterion of .05 (for most cases)
- For our z-scores, we see that .025 of the distribution is beyond the z-score of **1.96**
- We reject that our sample is representative of the population if the sample mean is greater than the absolute value of 1.96

Z-scores & Confidence Intervals

- to find confidence intervals (CI) when using z-scores, we want to know what are the corresponding values that go along with covering 95% of the distribution
    - let's look at our z-table
- the value that we end up with is +-1.96
- we can then use the z-score formula

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

$$1.96 = \frac{X - \overline{X}}{S}$$

$$-1.96 = \frac{X - \overline{X}}{S}$$

Z-scores & Confidence Intervals

$$\overline{X} + (1.96 * S) = X$$

$$\overline{X} - (-1.96 * S) = X$$

Z-scores & Confidence Intervals

$$Upper\ Limit = \overline{X} + (1.96 * S)$$

$$Lower\ Limit = \overline{X} - (1.96 * S)$$

Calculating Confidence Intervals

```
## [1] 2 5 6 3 4
```

```
## [1] 4
```

```
## [1] 1.581139
```

- mean of 4
- standard deviation of 1.58
    - we'll use this for the standard error calculation

Calculating Confidence Intervals

```
1.58/sqrt(5)
```

```
## [1] 0.7065975
```

```
# se = .71

4 - (1.96*.71)
```

```
## [1] 2.6084
```

```
# lower is 2.61

4 + (1.96*.71)
```

```
## [1] 5.3916
```

```
# upper is 5.39
```

Calculating Other Confidence Intervals in Small Samples

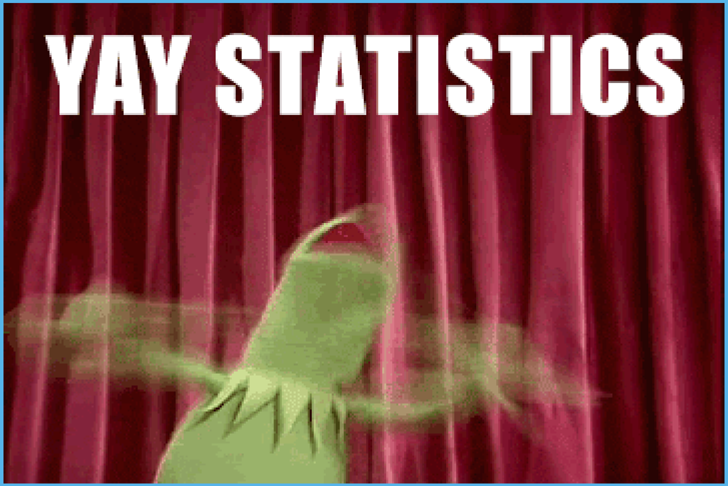
$$Lower = \overline{X} - (t_{n-1} * SE)$$

$$Upper = \overline{X} + (t_{n-1} * SE)$$

- we will talk about showing confidence intervals in visualizations over SPSS
- final thing to know about confidence intervals
  - if your CI don't overlap, then you have found a significant difference/finding
  - if they do overlap slightly, there is a chance of a significant finding
  - if they overlap for a majority of the CI, then no difference is found

Null-hypothesis Significance Testing (NHST)

- all inferential statistics we talk about will be using NHST
- the process revolves around the magic p value (probability) and hypotheses



Fisher's p value

- scientists tend to use 5% as a threshold for confidence
    - 5% chance of getting the result we hypothesized OR that the average score we found was extreme
    - statistics are based on probabilities (p values)
- does not mean we found a true effect/finding, just stating that in our sample, we found a value that was outside of the rest of the distribution
    - for z-scores this is a z score that is greater or less than 1.96 (95%)

Example of Probabilities

- 52 playing cards
  - 26 black
  - 26 red
  - 13 hearts
  - 13 clubs
  - 13 diamonds
  - 13 spades

Types of Hypotheses

- the hypothesis we create when looking for a difference/a relationship is present
  - **alternative hypothesis**, for experiments this would be the **experimental hypothesis**
  - denoted as H1
- the opposite hypothesis is the hypothesis that supports that there is no relationship/difference
  - **null hypothesis**
  - denoted as H0

Example of Hypotheses

H1: More sleep will lead to being happier

H1: More sleep will lead to a change in happiness

H0: More sleep will not change your happiness

NHST Process

$$outcome_i = (b_0 + b\, predictor_i) + error_i$$

- this example, we are seeing if the relationship is *significantly* different from zero
    - null = no relationship = 0

Test Statistic

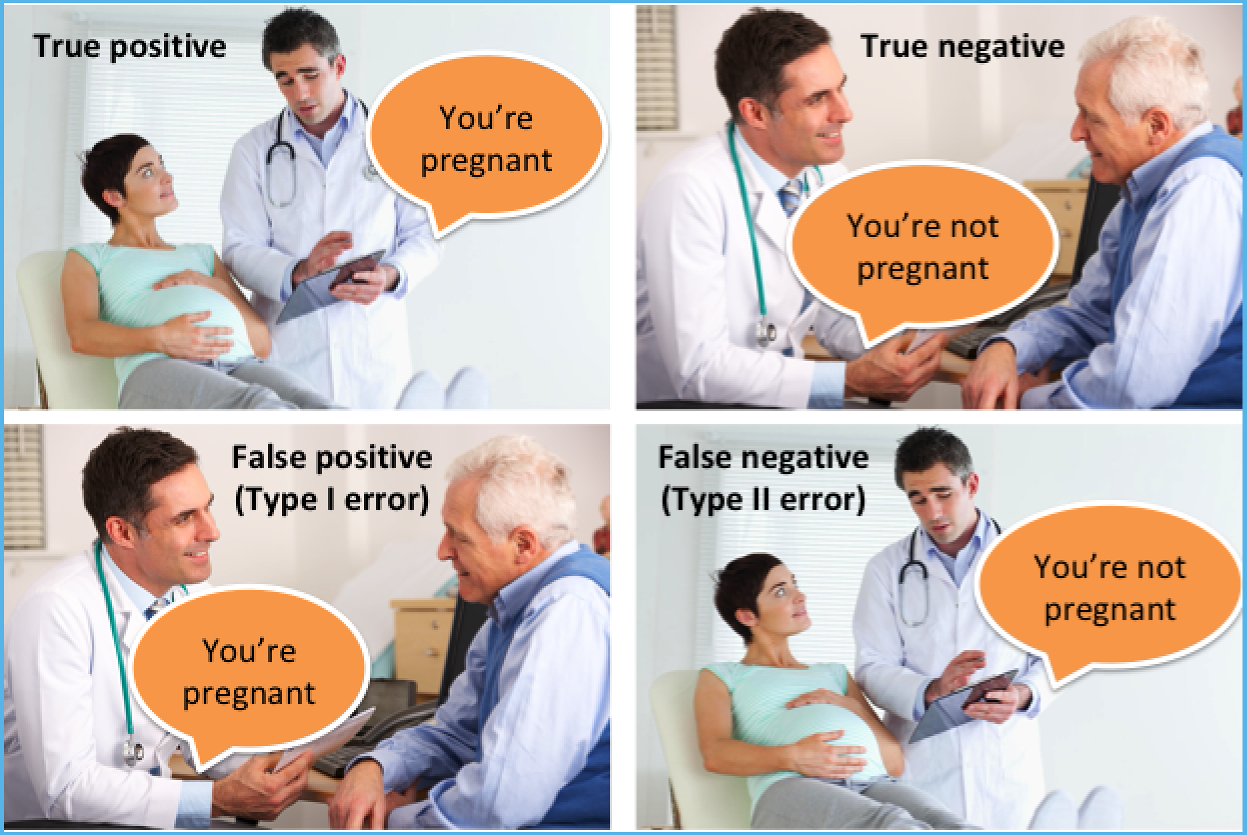
- **systematic variation** is the variation that you can explain with your model and hypothesis you're testing
- **unsystematic variation** is variation that cannot be explained by your model
- we will be learning about several test statistics
  - t and F will be the most common
- if our test statistic falls outside of the middle 95% of the distribution, then we report that we have a statistically significant finding

$$test\ statistic = \frac{signal}{noise} = \frac{variance\ explained}{variance\ not\ explained} = \frac{effect}{error}$$

One- and Two-tailed Tests

- a directional hypothesis = **one-tailed test**
- a non-directional hypothesis = **two-tailed test**

Type I & Type II Errors



Type I ;& Type II Errors

- **type I errors** is when you think there is an **effect** in the population, but there is no actual effect
  - the 5% we use for probability/significance level is referred to as alpha
  - if we took 100 samples, we would incorrectly support the alternative hypothesis 5 times

$$\alpha$$

- **type II errors** are when you think there is **no effect** in the population, but there is actually an effect
  - it is accepted from statisticians that an acceptable probability for type II error is 20% or .20
  - this is referred to as beta
  - if we took 100 samples, we would fail to detect an effect 20 times

$$\beta$$