Class 16

DATA1220-55, Fall 2024

Sarah E. Grabinski

2024-10-04

Recap: The Central Limit Theorem (CLT)

- We can use properties of the normal distribution to calculate the probability of observing a given value or range of values
- ***The Central Limit Theorem***: The probability distribution of means for multiple samples of the same size n from the same population approximates a normal distribution as the n increases
- We can combine these principles to create point estimates and confidence intervals for population parameters from our observed sample statistics

Recap: Point Estimates & Confidence Intervals

- A ***point estimate*** describes the ***location*** of an estimate or distribution
- A ***confidence interval*** describes the ***scale*** or ***precision*** of an estimate or distribution
- The ***confidence threshold*** or ***confidence level*** describes our uncertainty regarding the ***accuracy*** of our estimates

Recap: Z-Scores & Confidence Intervals

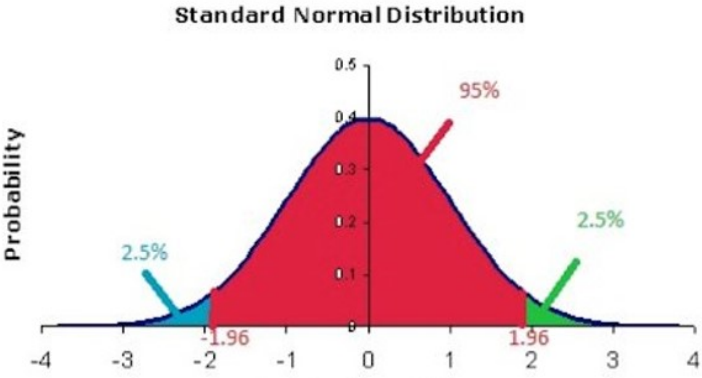

Figure 1: We use Z-Scores from the standard normal distribution to calculate the boundaries of our confidence interval.

Recap: Assumptions

1. A random process follows a known distribution which we can use to model that process and draw inferences about our population.
2. Your data is ***reliable***, so your sample statistics are ***reliable*** estimations of your sample population distribution.
3. Your data is ***valid***, so a sampling distribution based on your sample statistics is a ***valid*** estimation of the "true" distribution in the study population.
4. Your data is ***generalizable***, so your estimated sampling distribution for your study population is ***generalizable*** as the "true" sampling distribution for your target population

Statistical Inference and Hypothesis Testing

- We use sample statistics to describe sample populations and estimate the parameters of the study population's sampling distribution
- We also describe the variability of our measure and quantify our uncertainty regarding our estimate
- We use the overlap between theoretical distributions to decide how meaningful the differences between groups are

Hypothesis Testing Framework

- $\mathbf{H_0}$: The "Null" Hypothesis
  - Represents a position of skepticism
  - "There is *not* an association between process A and B"
  - Variables are *independent*
- $\mathbf{H_A}$: The "Alternative" Hypothesis
  - Represents the complement of H_0, that *something* is happening here
  - "There *is* an association between process A and B"
  - Variables are *dependent*

Conducting a hypothesis test

- Begin by *assuming* H_0 is the "true" state
- Calculate the probability that you would see results *as extreme or more extreme* than what you saw in your study
- The lower the probability, the less likely it is that we would see these results if H_0 was the "true" state of our pouplation
- If the probability is sufficiently low, we ***reject*** $\mathbf{H_0}$ and ***accept*** $\mathbf{H_A}$

Recap: Calculating a Z-Score

A ***Z-score*** indicates how many standard deviations σ away from the mean μ a given observation is.

$$
\begin{aligned}
Z &= \frac{\mathrm{observed value} - \mathrm{mean}}{\mathrm{standard deviation}} \\
&= \frac{x - \mu}{\sigma}
\end{aligned}
$$

Example: Calculating & Interpreting a Z-Score

The retirement age of NFL players follows the distribution $N(\mu = 34, \sigma = 3)$, and Aaron Rodgers, the quarterback for the New York Jets, is 40 years old. How unusual is it for Aaron Rodgers to still be playing?

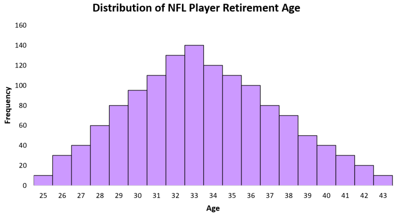

Figure 2: Histogram of the ages at which NFL players retire, which approximates the normal distribution $N(34, 3)$.

Example: Z-Score Calculation by Hand

$$
\begin{aligned}
Z &= \frac{\text{observedvalue} - \text{mean}}{\text{standarddeviation}} \\
&= \frac{x - \mu}{\sigma} \\
&= \frac{40 - 34}{3} \\
&= 2
\end{aligned}
$$

Example: Z-Score Calculation in R