# Review Questions

Grinnell College

April 8, 2024

# Major Topics

No multiple testing

- Study design and bias
- Sampling distributions
  - CLT and bootstrapping
  - Confidence Intervals
- Hypothesis Testing
  - p-values
  - Type I and Type II errors (tradeoff)

# Study design and bias

Questions to ask yourself:

1. What is my population?
2. What is my research question (i.e., statement about parameter)
3. How was my sample data collected?
   - Is my sample representative?
   - Was my method of collection associated with the outcome of interest?
   - Do I have enough observations?
4. Bias describes the *process*, not the sample

# Example

Researchers are interested in investigating the reading habits of college-aged (18-22) adults in the United States. As the researchers are located in Iowa, they decide to sample 100 students each from Grinnell College, the University of Iowa, and Iowa State. To avoid bias in their collection, they are sure to sample an equal number of men and women from each cohort.

1. What is the population in this study?
2. How was the sample collected?
3. Is this sample representative of the population they intended to study?

1. College-aged adults in US
2. random sampling from 3 Iowa colleges
3. No, only Iowa, only college students, colleges themselves not representative

# Example

We are interested in determing the the average length of turtle shells for the turtles living down by Arbor lake. To ensure that our sample is not biased, as we collect more observations, we keep track of those that we have already seen; that is, if we notice that the last couple turtles that we have collected have been large, we make an effort to collect more small turtles, and vice versa. We continue until we have collected 60 turtles

1. What is the population in this study?
2. How was the sample collected?
3. Is this sample representative of the population they intended to study?

1. Arbor lake turts
2. iterative sampling based on previous observations
3. Likely no, biased by collection method

# Sampling Distribution

Questions to ask yourself:

1. What is a sampling distribution, and what does it tell me?
2. How can I estimate the standard deviation of a sampling distribution ? (two ways)
3. When does the CLT apply?
4. When should we use a t-distribution, how is it impacted by sample size, and when can we get away with using a normal approximation?
5. When should I use bootstrapping instead, and how can I use this to find confidence intervals?

1. Distribution of statistic
2. Standard error $\hat{\sigma}/\sqrt{n}$ or standard deviation of bootstrapped distribution
3. Estimating mean or proportion
4. When sample size smaller, larger $n$ means more normal
5. Bootstrapping always valid, use percentiles to find CI

# Central Limit Theorem

What does the central limit theorem state?

How does this justify the use of 68-95-99 rule?

How will bootstrapping the sample mean differ from using CLT?

How is the CLT related to the t-distribution?

1. Samling distribution of mean or proportion is normal
2. 68-95-99 applies to normal distribution
3. Should be identical when finding mean or proportion
4. t distribution is modified normal using estimated $\hat{\sigma}$

# Notation

X is normally distributed with mean $\mu$ and s.d. $\sigma$

$$\Leftrightarrow$$

$$X \sim N(\mu, \sigma)$$

---

$$\overline{X} \sim N(\mu, \sigma/\sqrt{n})$$

$$\hat{p} \sim N\left(p,\ \sqrt{\frac{p(1-p)}{n}}\right)$$

# Hypothesis Testing

Questions to ask yourself:

1. What is my null hypothesis?
2. What constitutes a type I error? A type II error?
3. What is a $p$-value, and how is it used?
4. How are these related to confidence intervals?

# Terms

For $\alpha$:

- Type I error rate
- Significance level (because check $p < \alpha$)

For $1 - \alpha$

- Coverage probability
- Confidence level

# A note on confidence intervals

Keep in mind:

1. The coverage probability of a CI describes the *process* not the interval itself
2. All values in a CI are equally likely to be true estimate

# Example

In April of 1940, British mathematician John Kerrich was visting in-laws in Copenhagen when it was invaded by German forces. Caught up in the invasion, he was interned in Midtjylland where he empirically tested the laws of probability by flipping a coin 10,000 times and tabulating the result. His final tally was 5067 heads and 4933 tails.

1. What is the t-statistic associated with this value?
2. Is this value large enough to use a normal approximation?
3. Approximately how many standard deviations away from the expected value is the observed value?
4. Create an approximately 95% confidence interval
5. Based on this, is there evidence to suggest this was not a fair coin?

1. $t = 1.34$
2. Yes, 10,000 observations
3. 1.34 since approximately normal, t statistic is like a z score
4. $\hat{p} \pm 2 \times SE = (0.4967, 0.5167)$
5. No, $p = 0.5$ in interval

# Example cont.

Flipped coin 10,000 times, 5067 heads, 4933 tails

1. The $p$-value associated with this experiment was $p = 0.18$. What does that mean in this context?
2. Suppose that instead Kerrich had flipped his coin 100,000 times, getting 50,670 heads and 49,330. Without computing it, how do you think the associated t-statistic will change from the original?
3. The $p$-value associated with 100,000 flips is $p = 0.000023$. Why has this changed so drastically?

1. Probability of observing deviation of 67 heads *or greater* away from 5,000 is 0.18 if true probability is 0.5
2. $n$ increase so t-statistic is larger
3. Larger $t$ statistic, greater degrees of freedom, more evidence, etc.,