

Statistical Inference and Hypothesis Testing

Intermediate Data Analytics

Dr. Alex Marsella

March 20, 2024

Today

- I am going to teach you some stats in a way that is as accessible as possible to those who've not taken a stats class
- **Goal:** Understand the basics of statistical inference
- **Key Concepts:**
 - Bias, Noise, and Precision
 - Hypothesis Testing
 - Statistical Significance and p-value

Section 1

Statistical Inference

What is Statistical Inference?

- Making **decisions** or **predictions** about a population based on sample data
- Involves estimating population parameters using *sample statistics*
 - Information gathered from a sample is called a “statistic”
 - The “truth” about the population the sample is drawn from is called a “parameter”
- **Example:** Estimating the average height of Berry Students using a sample of 50 students.

Estimation

Every regression coefficient $\hat{\beta}$ you see is an *estimate* of a “true” relationship β , an **estimand**.

$$\text{Estimate} = \text{Estimand} + \text{Bias} + \text{Noise}$$

- **Estimand**: This is the “truth”; the population parameter; the thing in the world that we want to know about.
- **Estimate**: This is our best guess at the truth, it’s a sample statistic, the thing from our sample.
- **Bias** and **Noise** are the two things that cause $\text{Estimate} \neq \text{Estimand}$
 - By learning how to handle these, we can get good estimates.

A thought experiment

- Suppose we want to predict who will win the majority of votes in an election.
- We cannot ask all 160M registered voters, so we take a *sample*, by necessity.
- Imagine we poll 100 (draw a sample of 100) registered voters, and ask them “Will you support the Republican or the Democrat?”
 - Suppose “proportion answering Republican” is the thing we want to measure.
 - What is our estimate? How would it be calculated?
 - What is our estimand?
 - What mistake could we make in polling that would bias our estimate in one direction or another?

Section 2

Bias, Noise, and Precision

Bias

- **Definition:** Systematic error that leads to incorrect estimates of the population parameter
 - Not fixed by larger sample sizes, fixed by better sampling.
- **Example:** Pre-selection in surveys
 - Imagine a politician surveying their national approval rating by e-mailing people signed up for the mailing list.
 - In what direction would this bias their approval rating estimate?
 - Would the estimate be above or below the estimand (the truth)?
 - Imagine a 30 minute survey that pays you \$5 to complete it.
 - Would a certain type of person be more or less likely to answer such a survey?
 - What kind of questions in a survey like this would be subject to bias if we were trying to estimate something about the national population?

Thought Experiment about Bias

Suppose we want to estimate the average height (parameter) of berry students (population)

- What are some poor sampling strategies that would bias our estimate?
- How would you gather a sample of students that you think would give you a good estimate?

Noise

- Random error in measurements
- Imagine we poll 100 registered independent voters.
 - Good polling suggests about 45% lean R and 46% lean D.
 - Occasionally, we will poll 100 and 80 of them will say they're Democrat.
 - However, unbiased sampling of independents averages out to 45R/46D split.
- **As long as our estimate is unbiased, these extreme scenarios will balance each other out as we take many samples.**
 - Extreme results balance each other out and get averaged out to a good estimate.

Reducing Noise

- Take a larger sample size. Poll 10,000 people instead of 100, and your estimate (if unbiased) will be less noisy and closer to the estimand.
 - **Note: This is not the same as "taking more samples". Many noisy samples, if unbiased, can still average out to be close to the estimand.**


Precision

- The closeness of repeated measurements to each other
- Inverse relationship with noise

Unbiased and imprecise



Unbiased and precise



Visualization.

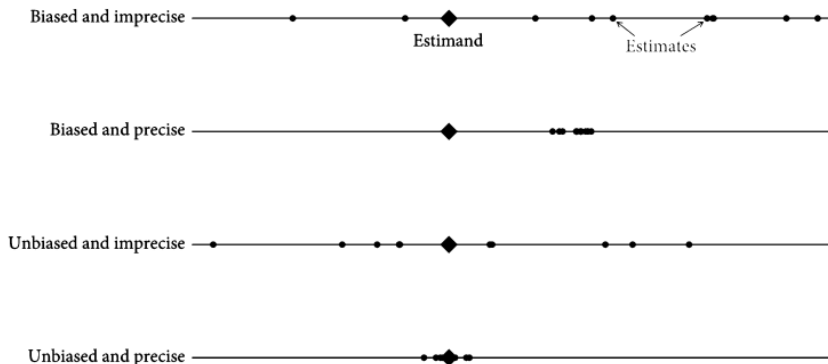


Figure 6.1. Understanding the difference between *unbiased* and *precise*.

It's up to you!

- Data analytics, research, statistics, etc. all involve us using sample data to make inference about how the world works.
- The computer just does math. You need to consider the way in which the data was gathered.
- Bias and Precision are things **you** need to understand about your sample before you even write the code.
- Context Dependent: Understand any limitations or problems with sampling in the context of the problem you want to solve.

Quantifying Precision

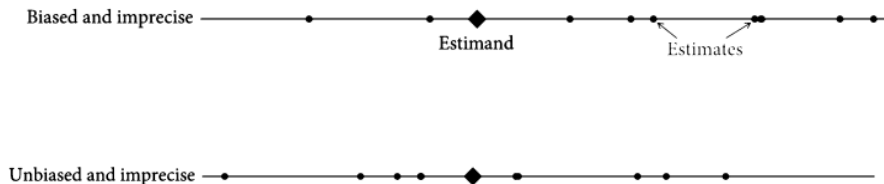
Standard error is how precision is quantified. Small standard errors relative to the estimate mean you have high precision.

- $S.E. = \frac{Std.Dev.}{\sqrt{n}}$ where n is the sample size.
- Notice that larger sample sizes lead to smaller standard errors!

Standard Error, in plain English

- Basically tells us, if we sampled from the same population an infinite number of times, how far each estimate would be from the average estimate.
 - **If the estimator is unbiased, then standard error tells us how far each estimate should be from the estimand.**
- Small standard errors relative to the estimate mean we expect to probably see a similar estimate if we kept sampling over and over again.
- Large standard errors relative to the estimate mean we expect to see estimates all over the place if we did it again and again.

Visualizing small and large standard errors.



Section 3

Hypothesis Testing and Statistical Significance

Hypothesis Testing

What is a Hypothesis?

- A statement that can be tested objectively.
- **Null Hypothesis (H_0):** No effect or difference
 - We go with this until we demonstrate otherwise.
- **Alternative Hypothesis (H_a):** Some effect or difference exists
 - The burden of proof lies on us to “prove” this one.

A hypothesis about age's effect on survival aboard the Titanic

We worked with a sample of ~700 Titanic passengers (There were over 2000 people in total.)

“Does Age affect survivability on the titanic?” is a question we asked.

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.05672	0.17358	-0.327	0.7438
Age	-0.01096	0.00533	-2.057	0.0397 *

A hypothesis test is implicitly being performed.

- **Null Hypothesis (H_0):** $\beta_{Age} = 0$ OR “There is no relationship between Age and Survival”.
- **Alternative Hypothesis (H_a):** $\beta_{Age} \neq 0$ OR “There is a relationship between Age and Survival”.
 - Note that β is our *estimand*, which is unknown: our *estimate* is $-.01096$

Standard Error, Z, and p-value

- Notice that the standard error is less than half the size of the estimate. Small standard error!
- If the estimand is equal to zero, as the null suggests, then our estimate is 2.057 standard errors BELOW the estimate of 0.
 - Z is the quantity of standard errors that an estimate lies from the null hypothesized estimand of 0.
 - $$Z = \frac{\text{Estimate} - \text{Estimand}_{H0}}{S.E.}$$
 -
 - $$Z = \frac{-.01096 - 0}{.00533} = -2.057$$

P-value

- The probability we would observe an estimate that far or farther (measured by quantity of standard errors, Z) from the null hypothesized estimand if the null hypothesized estimand were true.
 - Null hypothesized estimate in regression is “no relationship/effect”, aka $\beta = 0$
- **If age actually has no relation to survival on the titanic, and we took random identical samples of people from the Titanic over and over again, we would observe an estimate this large or larger 3.97% of the time.**

More on p-value

- **If p-value is less than 0.05, we reject the null hypothesis and conclude a relationship exists.**
- We call this “statistical significance”: a relationship is “statistically significant” when p-value is below a certain threshold.
 - The most common threshold is 5%.
- Statistical significance definition:
 - Basically: we think we have evidence that this relationship exists in the population from which the sample was drawn.
 - If the relationship doesn't actually exist, we'd only observe an estimate this large relative to standard error $< 5\%$ of the time.

Interpreting P-Value

- Low p-value ($< \alpha$): Reject H_0 , evidence for H_a
- High p-value ($\geq \alpha$): Fail to reject H_0

Conclusion

- Statistical inference allows us to make educated guesses about populations from samples.
- Hypothesis testing is a structured approach to test assumptions.
- Understanding bias, noise, and precision is crucial for accurate data analysis.

Next Class

- Friday's application will be more of a thought experiment/critical thinking puzzle, not coding.
- I will allow groups of up to three.

Exercise if we have time

Work with the people around you to determine the extent of bias and precision for each one of these:

- Surprising News Polls conducts large, representative polls, computes the average support for each candidate, and then flips a coin. If the coin is heads, they add 10 percent to Candidate A. If tails, they subtract 10 percent from Candidate A.
- Middle America Polling obtains a physical copy of the list of all registered voters, they flip to the middle page, and they contact and interview the ten individuals in the middle of that middle page.