STAT 408
Applied Regression Analysis

Miles Xi

Department of Mathematics and Statistics

Loyola University Chicago

Fall 2022

Review of Probability and Statistics

Part 3: Statistical Inference

Point Estimation

- Point estimation uses sample data to calculate a single value to serve as a "best guess" of an unknown population parameter

- Example
  - Use the average GPA in this class to estimate the mean GPA for Loyola students

- The symbol $\hat{\theta}$ is to denote the point estimator of θ from a given sample

Unbiased Estimator

- Any estimator $\hat{\theta}$ is a random variable and a function of sample X_i 's

- There are always gaps between estimator $\hat{\theta}$ and true parameter θ

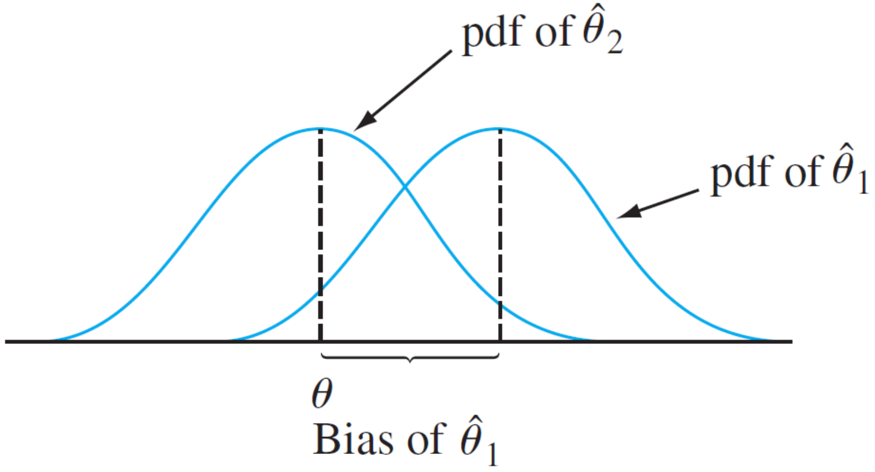
$$\hat{\theta} = \theta + \text{error of estimation}$$

- The unbiased estimator is defined as

$$E(\hat{\theta}) = \theta$$

- The average of unbiased estimator is equal to the true parameter

Unbiased Estimator



The pdfs of a biased estimator $\hat{\theta}_1$ and an unbiased estimator $\hat{\theta}_2$

Some Unbiased Estimators

- Sample mean $\bar{X}$ is an unbiased point estimator for population mean

- Sample variance S is an unbiased point estimator for population variance

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then the estimator

$$\hat{\sigma}^2 = S^2 = \frac{\sum (X_i - \overline{X})^2}{n-1}$$

is unbiased for estimating σ^2.

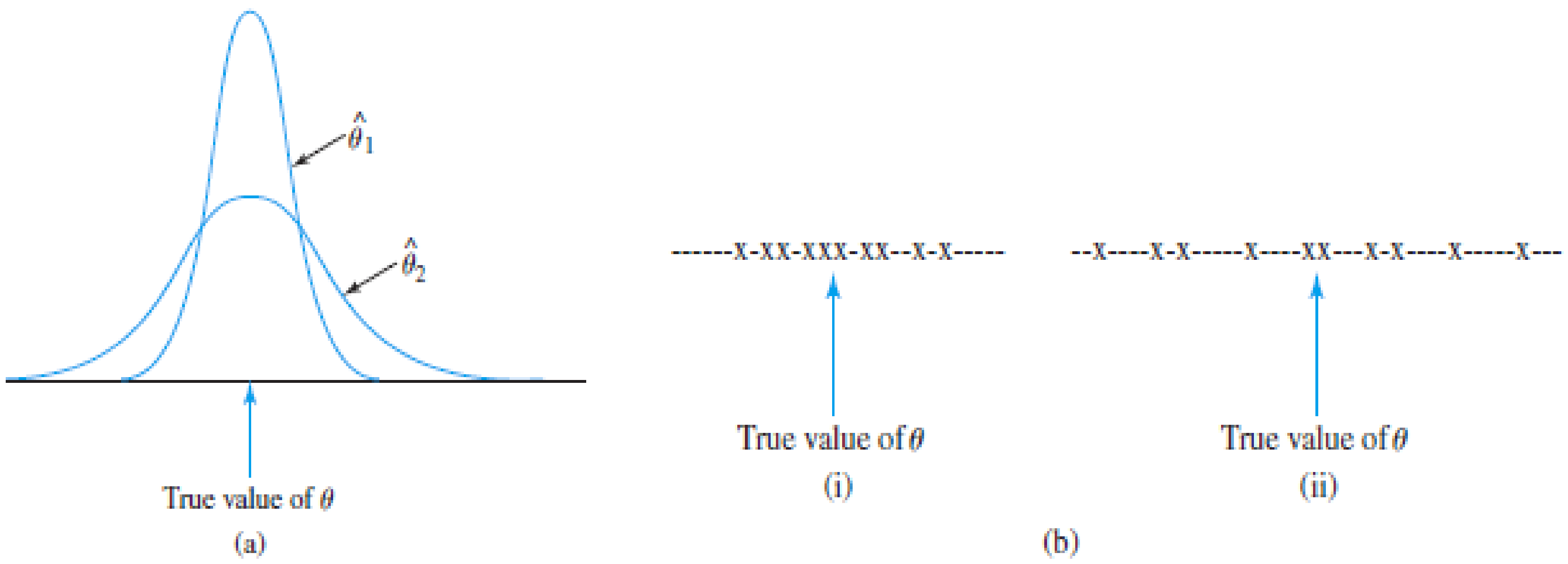
The Variance of Estimator

- What if we have two unbiased estimators? Which one we should use?

- Bias only focuses on the “average” of one estimator, without considering its variability

- Even though the “average” of the estimator is equal to the true parameter, one specific estimation may be far from the truth

The Variance of Estimator



- Between $\hat{\theta}_1$ and $\hat{\theta}_2$, which one should we prefer?

Estimators with Minimum Variance

Principle of Minimum Variance Unbiased Estimation

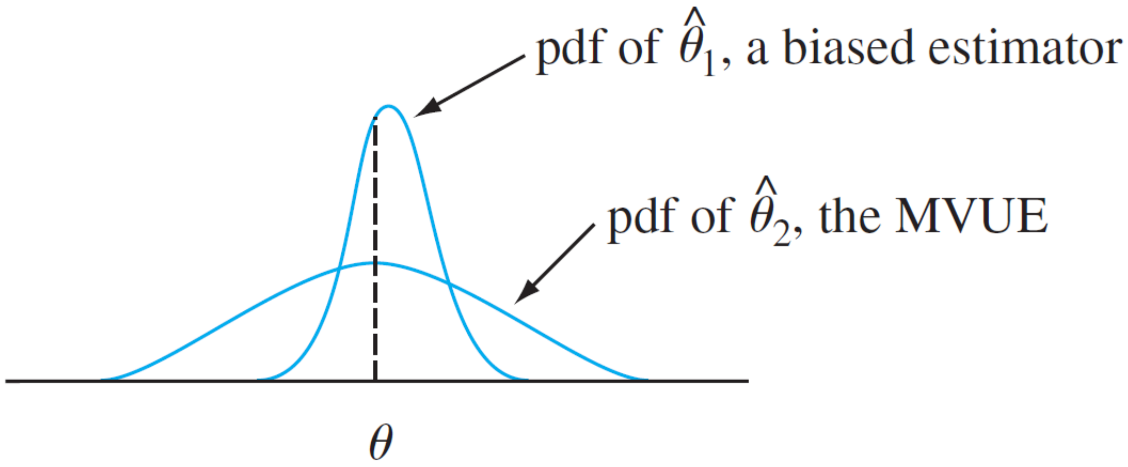
Among all estimators of θ that are unbiased, choose the one that has minimum variance. The resulting $\hat{\theta}$ is called the **minimum variance unbiased estimator (MVUE)** of θ.

- Estimators are random variables which rely on the sample
- There is no guarantee that any specific value of estimator is equal to the true parameter

Estimators with Minimum Variance

- It is possible to obtain an estimator with some bias but small variance

- Such biased estimator may be better than one unbiased estimator with large variance



- But most time we still use unbiased estimator with small variance

Confidence Interval

- The point estimation $\hat{\theta}$ gives a “best guess” for the population parameter θ

- But it doesn’t tell us about the true value of θ

- Confidence interval is an interval that we are highly confident that the true θ falls into:

$$\text{Probability (lower bound} < \theta < \text{upper bound)} \approx 1$$

Confidence Interval

- Example: confidence interval for population mean

- Suppose we have a random sample X_1, …, X_n from a normal distribution with mean μ and standard deviation σ

- Two facts:
  1. $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$
  2. $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$

Confidence Interval

- According to standard normal distribution

$$P\left(-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = .95$$

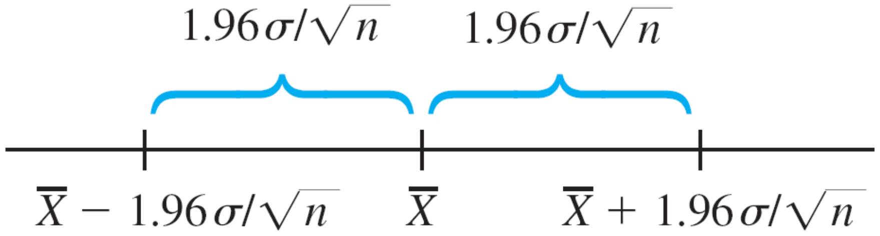
- Moving $\overline{\mathrm{X}}$ and $\sigma/\sqrt{\mathrm{n}}$ to both sides gives an interval for μ

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = .95$$

- This is a 95% confidence interval for population mean μ

Confidence Interval

- The 95% is the confidence level



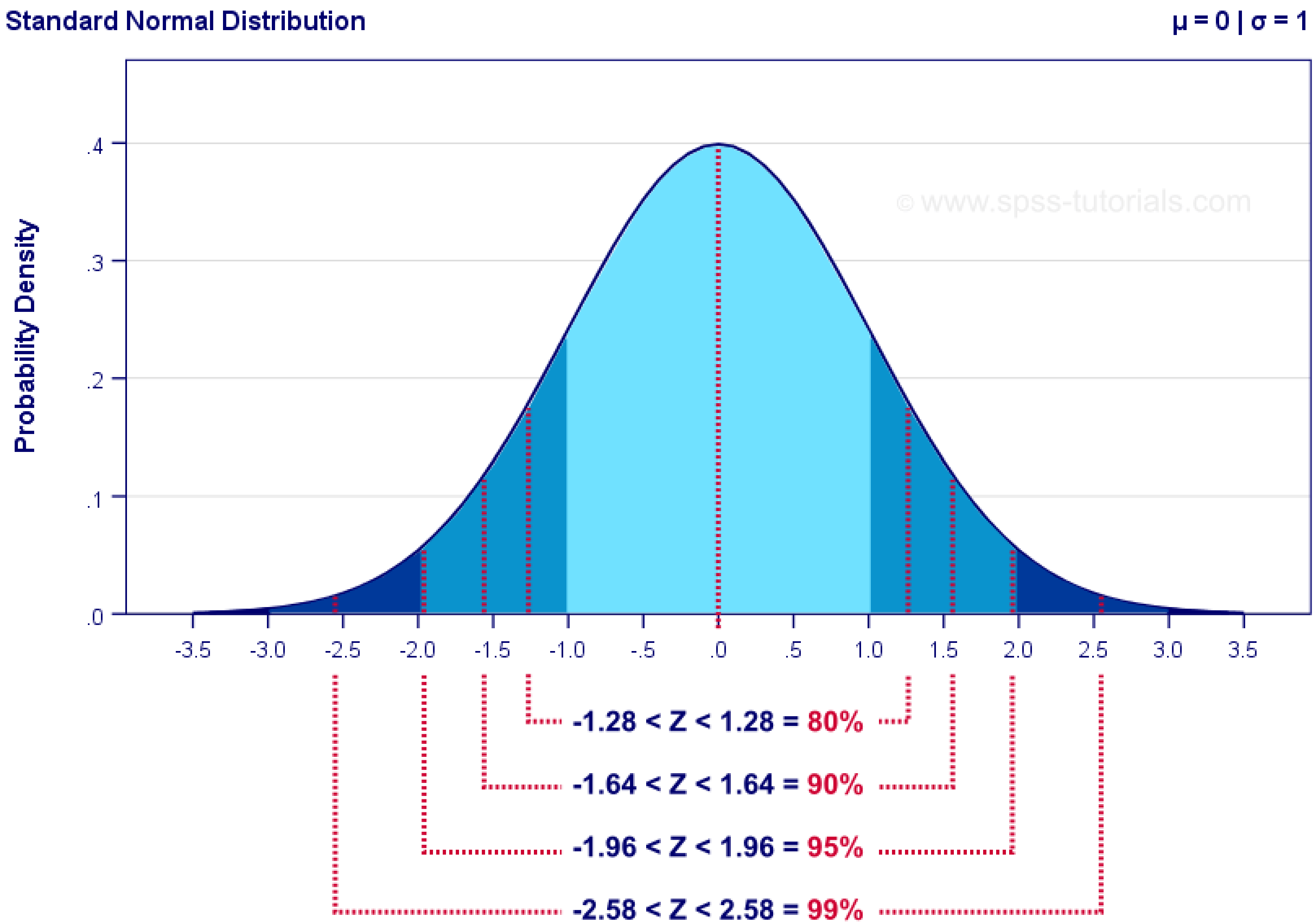
- The CI for population mean is symmetric around sample mean

- The width is determined by sample size and confidence level (How?)

Confidence Interval

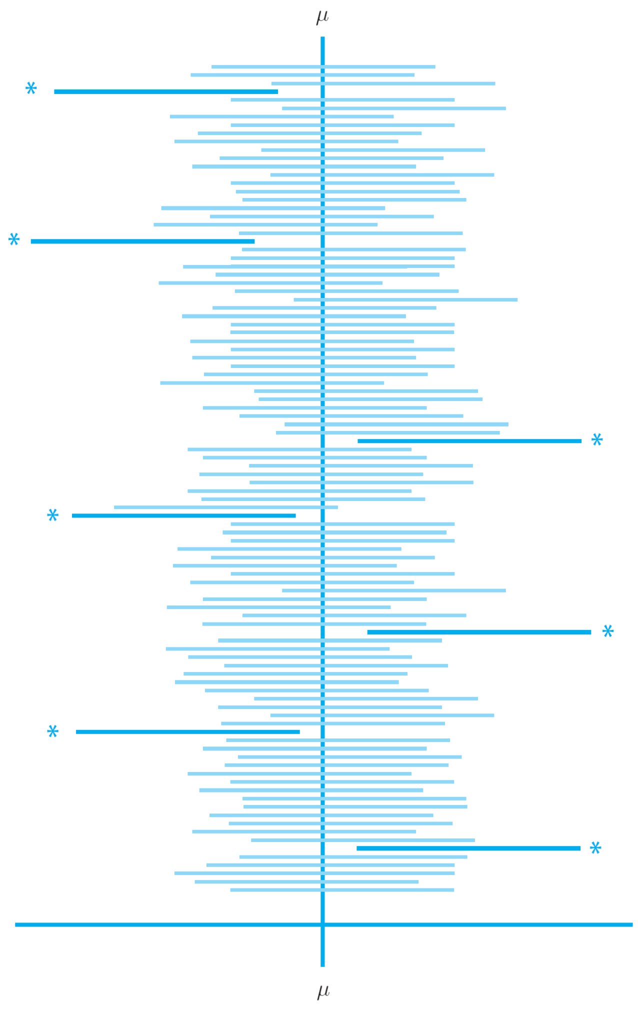
- Question
  - In the last example, what is the 99% confidence interval? How about 90%?

Confidence Interval



Interpretation of Confidence Interval

- The vertical line shows the true population mean μ
- Each horizontal line is one confidence interval constructed by one sample
- Among all such confidence intervals, about 95% of them cover the true μ



Hypothesis Test

- A statistical hypothesis is a claim about:
  1. The value of a single parameter
  2. The values of several parameters
  3. The form of an entire probability distribution

- For example
  1. $\mu = 0.75$
  2. $\mu_1 = \mu_2$
  3. The population is normal

Hypothesis Test

- In hypothesis test, there are two contradictory hypotheses under consideration

- Hypothesis test is to decide which of the two hypotheses is correct based on sample information

- The null hypothesis**,** denoted by H_0, is the claim that is initially assumed to be true

- The alternative hypothesis**,** denoted by H_a, is the assertion that is contradictory to H_0

Hypothesis Test

- Examples
  - $H_0 : \mu = 0.75$, $H_a : \mu \neq 0.75$
  - $H_0 : p = 0.5$, $H_a : p < 0.5$

- In general, we set the null hypothesis $H_0 : \theta = \theta_0$ and set alternative hypothesis H_a as the one of the following three forms:
  1. $H_a : \theta > \theta_0$
  2. $H_a : \theta < \theta_0$
  3. $H_a : \theta \neq \theta_0$

Test Procedure

1. Write down the H_0 and H_a

2. Find a test statistic X as a function of sample data

3. Assume H_0 is true, identify the distribution of test statistics X

4. Calculate the probability we observe such a value of X (p-value)

5. If p-value is less than a threshold (<u>significance level α</u>), the possibility of observing such test statistic is rare, we reject H_0

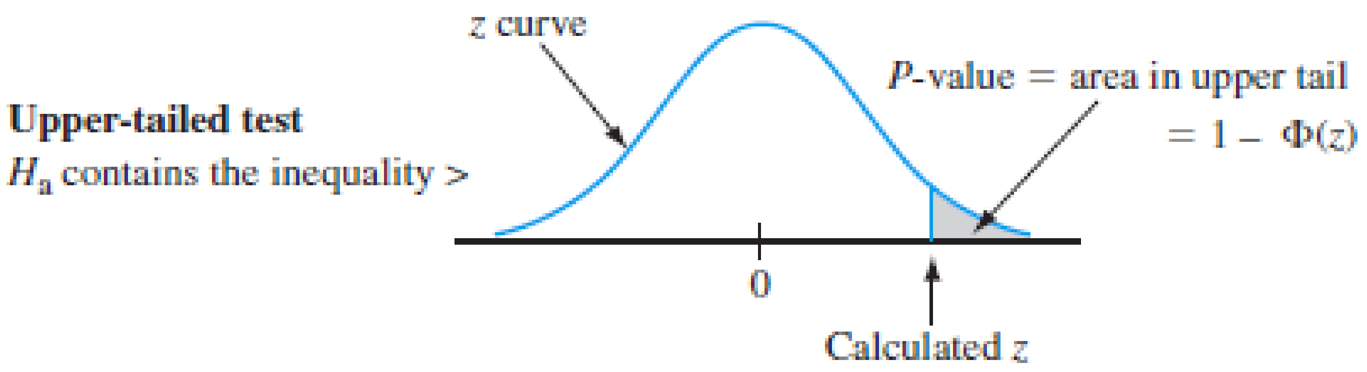
6. Otherwise, we fail to reject H_0

Test Procedure

- For example, we want to test if a normal distribution has a mean μ_0 or larger
- H_0: $\mu = \mu_0$ $vs.$ H_a: $\mu > \mu_0$
- Let $X_1,\ldots, X_n$ represent a random sample of size n from this normal population
- Under H_0 , the sample mean $\overline{\mathrm{X}} \sim \mathrm{N}(\mu, \sigma/\sqrt{\mathrm{n}})$ and Z = $\frac{\overline{\mathrm{X}}-\mu}{\sigma/\sqrt{\mathrm{n}}} \sim$ N(0, 1)
- We use Z as test statistic

Test Procedure

- Since H_0: $\mu = \mu_0$ and H_a: $\mu > \mu_0$, then the p-value is the probability of observing the value of test statistics Z or larger



- The evidence against H_0 lies on the upper tail of standard normal distribution

Test Procedure

- Question
  - What if the alternative hypothesis is H_a: $\mu < \mu_0$ or $\mu \neq \mu_0$?

Type I Error

- In hypothesis test, the p-value shows the probability we observe such a sample and test statistics if H_0 is true
- If the significant level is 0.05, and the p-value is 0.04, then we think observing such a test statistics is “rare”
- The valid explanation is that “H_0 is false”, so we reject it

Type I Error

- However, what if H_0 is correct and we do observe such a “rare” event? Because we do have 4% chance for such observation under H_0

- In such case, we made a <u>type I error</u>

- Type I error is the error in which we falsely reject H_0 , in other words, H_0 is correct but we reject it

- Type I error means we are too “aggressive” to reject a true statement: it is a false alarm

Type II Error

- Similarly, if H_0 is wrong but we fail to reject it, then we made a <u>type II error</u>

- Type II error means we are too “conservative” to miss the wrong statement: it is a missing alarm

Type I and Type II Error

Null hypothesis is...	True	False
Rejected	**Type I error** False positive Probability = α	**Correct decision** True positive Probability = $1-\beta$
Not rejected	**Correct decision** True negative Probability = $1-\alpha$	**Type II error** False negative Probability = β

The Probability of Type I Error

- Suppose the significant level of hypothesis test $\alpha = 0.05$

- Under H_0 , we calculate the probability of observing test statistics, which is p-value; If $p < 0.05$, then we reject H_0

- However, if H_0 is true, we would falsely reject H_0 in those 5% cases;

- <u>The probability of making type I error is the significant level α</u>