STA 103 (SS2 2025): Problem Set 3
Instructor: Wookyeong Song
Institution: UC Davis

TOPICS COVERED

Large n inference for ...

1. One-sample population mean μ;
2. One-sample population proportion p;
3. A difference in population means $\mu_X - \mu_Y$;

For general population parameter, we denote θ, (i.e., θ is either μ, p, or $\mu_X - \mu_Y$.

In each case you will base your inference on a test statistic $\hat{\theta}$ which an estimator of a population parameter and you will be asked to compute the following:

- Compute the Z-score for the observed value of your test statistic $\hat{\theta}$:

$$Z = \frac{\hat{\theta} - \theta}{\text{SE}(\hat{\theta})}$$

  assuming the given probability model for the data is correct, i.e. values assumed under the null hypothesis H_0.

- Determine if your observed Z-score is consistent with a typical random fluctuations in Z-scores. Use the CLT to quantify how rare the observed Z-score is, i.e. the p-value.

- Produce $100(1-\alpha)\%$ approximate **Confidence Intervals** for the unknown population parameter:

$$\hat{\theta} \pm z_{1-\frac{\alpha}{2}} \, \text{SE}(\hat{\theta}),$$

  where formula for $\text{SE}(\hat{\theta})$ can be computed but values for $\text{SE}(\hat{\theta})$ are the “plug-in" estimates from your data or “conservative" estimates for one-sample proportion estimate $\hat{p}$.

- Two-sided Hypothesis Testing:

$$H_0 : \theta = \theta_0 \quad H_a : \theta \neq \theta_0.$$

- One-sided Hypothesis Testing “$>$":

$$H_0 : \theta \leq \theta_0 \text{ (equivalently, } \theta = \theta_0) \quad H_a : \theta > \theta_0.$$

- One-sided Hypothesis Testing “$<$":

$$H_0 : \theta \geq \theta_0 \text{ (equivalently, } \theta = \theta_0) \quad H_a : \theta < \theta_0.$$

- The null H_0 must have the equality $=$. If the statement we want to prove is inequality $>$ or $<$, then we put it on alternative H_a.

- Perform hypothesis test at significance level α for the unknown population parameter θ,
  - Step 1, find the Z-score.
  - Step 2, find the p-value following the direction of H_a.
  - Step 3, compare p-value with significance level, decide whether we reject the null H_0 or not.

- The population correlation coefficient ρ (or ρ_{XY}), defined as

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

- Given paired data $(X_1, Y_1), \ldots, (X_n, Y_n)$, the sample correlation coefficient is defined as:

$$\hat{\rho} = r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}.$$

- Simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2), \text{ i.i.d.}$$

- The fitted (predicted) response is given by

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,$$

  where

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

- The variance σ^2 is estimated by

$$s^2 = \hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2.$$

- Sampling variability of $\hat{\beta}_1$ given $X_1, \ldots, X_n$ are fixed. We expect it to be

$$\hat{\beta}_1 \sim N(E(\hat{\beta}_1), \text{Var}(\hat{\beta}_1)),$$

  where $E(\hat{\beta}_1) = \beta_1$, and

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \approx \frac{s^2}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

PRACTICE PROBLEMS

Lecture 9–10

1. A business analyst collects daily sales data from a sample of 36 coffee shops. The sample mean is \$2,860 and the sample standard deviation is \$450.

   - What is the point estimate for the population mean?
   - Estimate the standard error.
   - Construct a 95% confidence interval for the mean daily sales.
   - Interpret the result in the business context.

2. A company surveyed 250 customers and found 205 were satisfied with a recent product upgrade.

   - Calculate the sample proportion.
   - Construct a 95% confidence interval for the satisfaction rate. (Use both plug-in approach and conservative approach to estimate standard error (SE).
   - Interpret the interval.

3. A restaurant manager wants to estimate the average wait time for customers during lunch hour. From a random sample of 200 customers, the mean wait time is 12.4 minutes with a sample standard deviation of 2.6 minutes.

   - State the appropriate distribution to use for constructing a confidence interval.
   - Construct a 95% confidence interval for the population mean wait time.
   - Interpret the interval in the context of customer service.

4. **(Comparing Employee Training Outcomes)** A company rolls out two different training programs for two regional sales departments. After the training, employees complete a standardized performance test. The results are summarized as follows:

$$\text{Department A } (n_A = 25) : \bar{x}_A = 82.4, s_A = 6.2$$
$$\text{Department B } (n_B = 28) : \bar{x}_B = 79.1, s_B = 7.1$$

   - Construct a 95% confidence interval for the difference in mean test scores ($\mu_A - \mu_B$).
   - Interpret your results. What can the company conclude about the relative effectiveness of the two programs?

5. **(Regional Price Sensitivity Study)** A market research team investigates how much customers are willing to pay for a new subscription service in two different regions. The results:

$$\text{Region West } (n_W = 40) : \bar{x}_W = 48.3, s_W = 5.0$$
$$\text{Region East } (n_E = 35) : \bar{x}_E = 45.8, s_E = 6.3$$

   - Construct a 90% confidence interval for the difference in mean test scores ($\mu_W - \mu_E$).
   - What do your results suggest about regional pricing strategies?

Lecture 11

1. A tech company claims average tech support resolution time is 4 hours. A manager samples 35 support tickets and finds a sample mean of 3.7 hours with a standard deviation of 0.8 hours. Test at the 5% significance level whether resolution time has decreased.

   - State the hypotheses.
   - Compute the Z-score.
   - Estimate the p-value and make a decision.
   - State your conclusion in business terms.

2. A firm estimates the average delivery time using a 95% CI of (2.8, 3.2) days. The company's advertised delivery time is 3.0 days.

   - Should the company's claim be rejected at the 5% level? Justify using the confidence interval.
   - Would your answer change if the interval were (2.6, 2.9)?

3. A snack company tests whether a new eco-friendly packaging design influences consumer satisfaction. Two groups of consumers are randomly assigned:

   - Group A (Traditional packaging): $n = 40$, $\bar{x}_A = 7.8$, $s_A = 1.2$.
   - Group B (Eco-friendly packaging): $n = 38$, $x_B = 7.3$, $s_B = 1.0$.

   At the 5% significance level, test whether there is a difference in average satisfaction scores between the two packaging styles.

4. An employment researcher compares starting salaries for MBA graduates from two different programs:

- Program X: $n = 30$, mean salary $= 112,000$, $s = 9,800$.
- Program Y: $n = 35$, mean salary $= 105,500$, $s = 8,700$

At the 1% significance level, is there evidence that Program X graduates earn more on average than Program Y graduates?

5. Wikipedia claims that among the population of Russia only 46% are male. To test this claim suppose you randomly sample 1000 people from Russia and find that, among the sample, 49.5% are male. Develope a null test of Wikipedia's claim.

6. The famous probabilist, Persi Diaconis, claims to be able to flip a fair coin and make it land heads with probability 0.8. To test this claim I asked him to flip a fair coin 100 times and watched him get 72 heads. Is this evidence he is able make a fair coin land heads with probability greater than 1/2?

7. Suppose you flip a coin 100 times and get 75% heads. Develop a null test that the coin is fair based on the percent of heads observed.

8. Suppose you are comparing two populations and get the following statistics from random sampling:

population 1	population 2
$\overline{X}_1 = -.5$	$\overline{X}_2 = .5$
$\hat{\sigma}_1 = 10.4$	$\hat{\sigma}_2 = 20.5$
$n_1 = 35$	$n_2 = 100$

If μ_1 and μ_2 denote the population averages for the two populations, respectively, which sentence best describes this data:
(a) Strong evidence that $\mu_1 \neq \mu_2$.
(b) The data is consistent with $\mu_1 = \mu_2$.

9. Consider a box of red and blue tickets. Suppose I told you the proportion of red tickets in the box, p, is 0.4. Now suppose you randomly sampled, with replacement, 100 tickets from this box and got a sample proportion $\hat{p} = 0.42$ of red tickets.

   - Find the Z-score of $\hat{p}$.
   - Is this observed Z-score rare or typical? Use the CLT to approximate the chance of observing more extreme Z-score greater (and less) than the one you observed, i.e. $P(|Z| > z)$?
   - Find 99.7% confidence intervals for p.

10. You are VP at Apple Computers in charge of the new OLED displays coming out next year. Samsung has promised its production process has a defect rate $p = 1/50$.

    In the first shipment of $n = 400$ displays, after assembly and testing, you find $\hat{p} = \frac{9}{400}$ defective displays. Based on this sample, construct an argument to Samsung that they are in breach of contract (i.e. that there true production defect rate is larger than what they promised).

    - Find the Z-score of $\hat{p}$.
    - Is this observed Z-score rare or typical? Use the CLT to approximate the chance of observing more extreme Z-score greater (and less) than the one you observed, i.e. $P(|Z| > z)$?
    - Find 68% confidence intervals for p. (Use conservative approach to estimate standard error (SE))

11. Suppose you have following data sampled from two different populations:

    Samples from Population A: $X_1, \ldots, X_{300}$,
    Samples from Population B: $Y_1, \ldots, Y_{300}$,

    which have the following statistics

$$\overline{X} = -36.444$$
$$\overline{Y} = -35.631$$
$$s_X = 1.238$$
$$s_Y = 1.215$$

    Let μ_A and μ_B denote the population averages from group A and B, respectively.

    Analyze $\hat{\theta} = \hat{\mu}_A - \hat{\mu}_B = \overline{X} - \overline{Y}$ to determine if the data is consistent with $\theta = \mu_A - \mu_B = 0$ or if there is evidence that $\theta = \mu_A - \mu_B \neq 0$.

    - Find the Z-score of $\hat{\theta}$.
    - Is this observed Z-score rare or typical? Use the CLT to approximate the chance of observing more extreme Z-score greater (and less) than the one you observed, i.e. $P(|Z| > z)$?
    - Find 95% confidence intervals for $\theta = \mu_A - \mu_B$.

Lecture 12

1. Suppose X and Y are random variables with joint PMF given by:

$X \diagdown Y$	-1	0	1
-1	0.1	0.3	0.1
0	0.1	0	0.1
1	0.1	0.1	0.1

Find the correlation between X and Y.

2. Given the joint probability table

$x\backslash y$	1	2	3
1	1/9	2/9	1/3
2	2/9	1/9	0

Find the correlation ρ between X and Y.

3. A real estate analyst models house price (y, in \$1000s) as a function of square footage (x, in 100s of square feet). The regression yields:

$$\hat{y} = 45 + 25x \quad \text{with} \quad \text{SE}(\hat{\beta}_1) = 5.0,\ n = 20.$$

- Test whether square footage is a significant predictor of house price.
- Construct and interpret a 95% confidence interval for the slope β_1.

4. A company investigates the relationship between the amount it spends on online advertising (in \$1,000s) and weekly product sales (in units). Data from 6 recent weeks are recorded below:

Week	Ad Spending (X)	Sales (Y)
1	2	45
2	4	52
3	3	48
4	6	60
5	5	54
6	7	65

- Calculate the sample means $\bar{X}$ and $\bar{Y}$, and compute the correlation coefficient r.
- Compute the slope and intercept of the least squares regression line, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$.
- Use the regression equation in to predict weekly sales if the company spends \$8,000 on advertising, make sure you use the correct unit of measurements.
- Suppose the correlation coefficient between spending (X) and weekly sales (Y) is found to be r= 0.98. An analyst concludes: "This high correlation proves that increasing advertising causes an increase in sales." Is this conclusion correct? Explain why or why not.