MA5701: Statistical Methods

Summary

Kui Zhang, Mathematical Sciences

Final Exam

- **Time and Date:** 12:45pm, Wednesday, April 23, 2025
- **Room:** Rekhi 214
- **Requirements:**
  - Two Hours
  - Pens, One Calculator, 4 letter size one-sided note
  - Covers contents from Chapter 1 to Chapter 5
  - Problems will be similar with homework problems

Chapter 1 – Variables

- **Qualitative (Categorical) variable** - is a variable that is not numerical. It describes data that fits into categories.
  - The **ordinal scale** distinguishes between measurements Generally, the relative amounts of some characteristic they process.
  - The **nominal scale** identifies observed values by name or classification.
- **Quantitative Variable** – is a variable that is measured on a numeric scale for which meaningful arithmetic operations make sense.
  - A **discrete** variable can assume only accountable number of values.
  - A **continuous** variable is one that can take any one of an uncountable number of values in an interval.

Chapter 1 – Descriptive Statistics

- Let $y_1, \cdots, y_n$ denote a sample of interest.
- **Sample Mean:** $\bar{y} = \frac{1}{n}(y_1 + y_2 + \cdots + y_n)$
- **Sample Variance:**

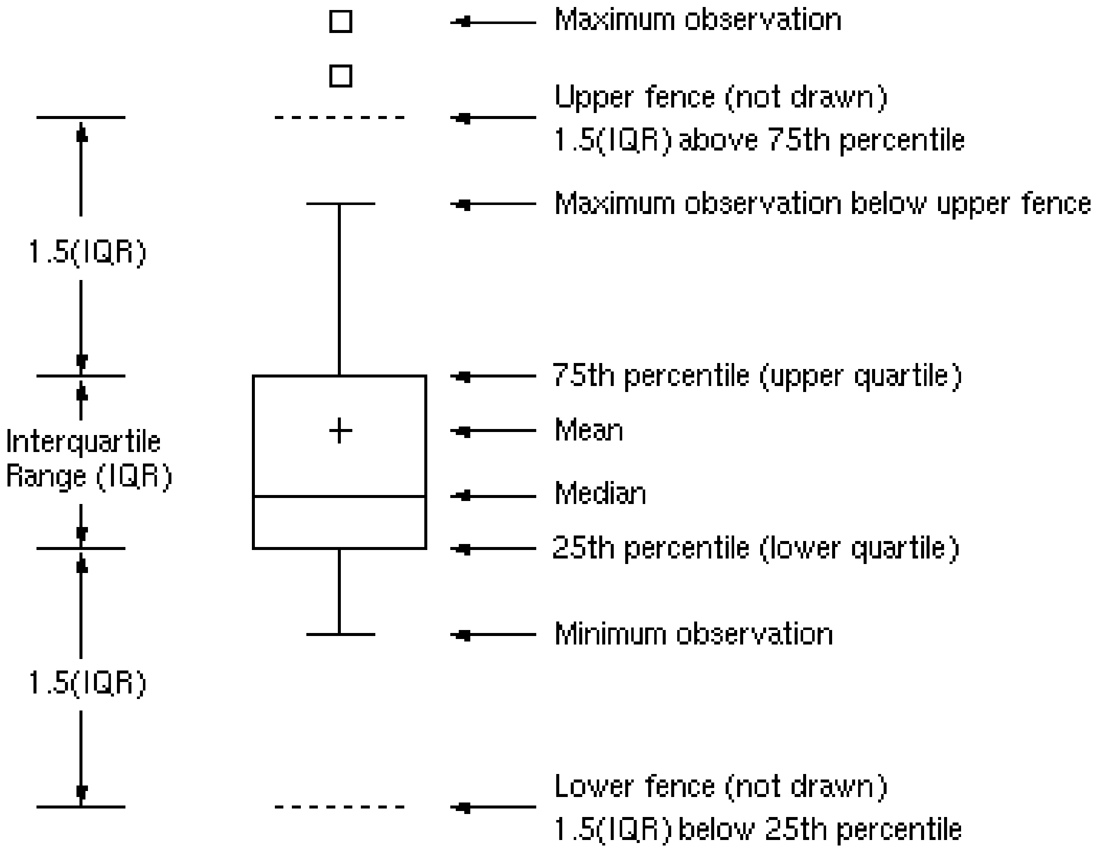
$$s^2 = \frac{1}{n-1}[(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2] = \frac{1}{n-1}(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)$$

- Sample Median: The **median** of a set of observed values is defined to be the middle value when the measurement are arranged from lowest to the highest. **Need to know how to find it.**
  - Median $\tilde{y} = y_{(\frac{n+1}{2})}$ if n is odd; $\tilde{y} = \frac{y_{(\frac{n}{2})} + y_{(\frac{n}{2}+1)}}{2}$ if n is even

Chapter 1 – Descriptive Statistics

- **Quartiles**, 25%, 50%, 75% percentile
  - 25% percentile – lower quartile, first quartile (Q_1)
  - 50% percentile – median, second quartile
  - 75% percentile – upper quartile, third quartile (Q_3)
- The **interquartile range** is the length of the interval between the 25th and 75th percentiles.

Chapter 1 – Schematics of Boxplot



Chapter 2 – Probability Calculation

- **Always True:**

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

- **If A and B are mutually exclusive**

$$\Pr(A \cup B) = \Pr(A) + \Pr(B)$$
$$\Pr(A \cap B) = 0$$

- **If A and B are independent**

$$\Pr(A \cap B) = \Pr(A) * \Pr(B)$$

- **For any event A,**

$$\Pr(A) + \Pr(A^c) = 1$$

Chapter 2 – Probability Calculation

- **If $A_1, \cdots, A_n$ are mutually exclusive**

$$\Pr(A_1 \cup \cdots \cup A_n) = \Pr(A_1) + \cdots + \Pr(A_n)$$

- **If $A_1, \cdots, A_n$ are independent**

$$\Pr(A_1 \cap \cdots \cap A_n) = \Pr(A_1) * \cdots * \Pr(A_n)$$

- **If A, B, C are independent, then**

$$\Pr(A \cap B \cap C) = \Pr(A) * \Pr(B) * \Pr(C)$$
$$\Pr(A^c \cap B \cap C) = \Pr(A^c) * \Pr(B) * \Pr(C)$$
$$\Pr(A \cap B^c \cap C^c) = \Pr(A) * \Pr(B^c) * \Pr(C^c)$$

Chapter 2 – Random Variables

- A **discrete random variable** is one that can take on only a countable number of values.
  - It has a probability mass function: $f(y) = \Pr(Y = y)$
  - $\Pr(Y \leq y) = \sum_{x \leq y} f(x)$
- A **continuous random variable** is one that can take on any value in an interval.
  - It has a probability density function: $f(y)$ (and $\Pr(Y = y) = 0$)
  - $\Pr(Y \leq y) = \int_{-\infty}^{y} f(x)\, dx$

Chapter 2 – Discrete Random Variable

For a discrete random variable y with the pmf $f(y)$

- **Expected Value**

$$\mu = E[Y] = \sum_y f(y) * y$$

- **Population variance** of Y, denoted by σ^2, is:

$$\sigma^2 = \text{var}(Y) = \sum_y f(y) * (y - \mu)^2$$

- **Population standard deviation** of Y, denoted by σ, is $\sigma = \sqrt{\sigma^2}$, the square root of the population variance.

Chapter 2 – Bernoulli and Binomial R.V.s

- A random variable Y has a ***Bernoulli(p)*** distribution if $f(1) = \Pr(Y = 1) = p$ and $f(0) = \Pr(Y = 0) = 1 - p$.
  - Population mean: $\mu = E[Y] = p$.
  - Population variance: $\sigma^2 = \text{var}(Y) = p(1 - p)$.
- A random variable Y has a ***Binomial distribution***, Binomial(n, p), if its pmf is

$$f(y) = \Pr(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y}, y = 0, \cdots, n.$$

  - If Y is the number of successes from n independent identical Bernoulli trials, then Y has Binomial(n, p).
  - Population mean: $\mu = E[Y] = np$.
  - Population variance: $\sigma^2 = \text{var}(Y) = np(1 - p)$.

Chapter 2 – Normal Distribution

- The random variable has the following pdf:

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right), -\infty < y < \infty$$

- Many times, we use the simple notation: $Y \sim N(\mu, \sigma^2)$
- Population mean is: μ.
- Population variance and standard deviation are: σ^2 and σ.
- If $Y \sim N(\mu, \sigma^2)$ and $\mu = 0$ and $\sigma^2 = \sigma = 1$, then Y has a standard normal distribution.

Chapter 2 – Normal Table

- Need to know how to use the normal table to find:
- If $Z \sim N(0,1)$,

$$\Pr(Z > a)\,,\Pr(Z < b)\,,\Pr(a < Z < b)$$

- If $Y \sim N(\mu, \sigma^2)$

$$\Pr(Y > a)\,,\Pr(Y < b)\,,\Pr(a < Y < b)$$

- Find Z_α such as $Z_{0.05}$, $Z_{0.02}$, etc.

Chapter 2 – Distribution of Sample Mean

- If the sample is a random sample from the normal distribution $N(\mu, \sigma^2)$, then sample mean $\sim N(\mu, \frac{\sigma^2}{n})$
- **Central Limit Theorem** - If random samples of size n are taken from any distribution with mean μ and variance σ^2, sample mean will have a distribution approximately normal with mean μ and variance σ^2/n.
- If $Y \sim Binomial(n, p)$, then approximately

$$\frac{\frac{Y}{n} - p}{\sqrt{p(1-p)/n}} = \frac{Y - np}{\sqrt{np(1-p)}} \sim N(0,1)$$

Chapter 3 – Hypothesis Testing

- **Alternative hypothesis (H_a) -** a statement that contradicts the null hypothesis. This hypothesis is accepted if the null hypothesis is rejected. The alternative hypothesis is often called the ***research hypothesis*** because it usually implies that some action is to be performed, some money spent, or some established theory overturned.

Chapter 3 - Possible Errors in Hypothesis Testing

- A **type I error** occurs when we incorrectly reject H_0, that is, when H_0 is true, and our sample-based inference procedure rejects it.
- A **type II error** occurs when we incorrectly fail to reject H_0, that is, when H_0 is not true, and our inference procedure fails to detect this fact.

	In the Population	
The Decision	H_0 **is True**	H_0 **is Not True**
H_0 **is Not Rejected**	Correct	Type II Error
H_0 **is Rejected**	Type I Error	Correct

Chapter 3 – Rejection Region

- The **rejection region** (also called the **critical region**) is the range of values of a sample statistic that will lead to rejection of the null hypothesis.
- R: rejection region and W is your test statistic
- Probability of making a type I error

$$\alpha = \Pr(W \in R | H_0\ is\ true)$$

- Probability of making a type II error

$$\beta = \Pr(W \in R^c | H_a\ is\ true)$$

- $1 - \beta$: the power of test.

Chapter 3 – 5 Steps

- **Step 1:** Specify H_0, H_a, and α
- **Step 2:** Define test statistic
- **Step 3:** Determine rejection region
- **Step 4:** Calculate test statistic based on data
- **Step 5:** State conclusions
- **Alternative approach:** p-value
- **Alternative approach:** confidence interval

Chapter 3 – Interval Estimators

How to interpret an interval estimator? For example:

Interpretation of 95% confidence interval, $(7.792, 7.988)$. Which statement is correct:

1. $\Pr(7.792 < \mu < 7.988) = 0.95$.
2. 95% of all weights between 7.792 and 7.988.
3. We sampled 95% of all weights.
4. We know that $7.792 < \mu < 7.988$.
5. We are 95% confident that the true population mean is between 7.792 and 7.988.

Chapter 4 – One Sample $\boldsymbol{t}$-test

- $H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$
- **Test statistic is always:** $T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}}$
- **Rejection Region:** $|t| > t_{n-1,\alpha/2}$
- ***p*-value:** $2\Pr(T_{n-1} > |t|)$
- **Two-sided CI:** $\left(\bar{y} - t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{y} + t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right)$

Chapter 4 – Test for One Proportion

- $H_0: p = p_0$ versus $H_a: p > p_0$
- **Test statistic is always:** $Z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$
- **Rejection Region:** $|z| > z_\alpha$
- ***p*-value:** $\Pr(Z > z)$
- **Lower CI:** $\left(\hat{p} - z_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, 1\right)$

Chapter 5 – Two Sample $\boldsymbol{t}$-test

- $H_0\colon \mu_1 - \mu_2 = \delta_0$ versus $H_a\colon \mu_1 - \mu_2 \neq \delta_0$.
- **Test statistic is always:** $T = \dfrac{(\bar{y}_1 - \bar{y}_2) - \delta_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
- Where $s_p = \sqrt{s_p^2}$ and $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \dfrac{SS_1 + SS_2}{n_1 + n_2 - 2}$
- **Rejection Region:** $|t| > t_{n_1+n_2-2,\alpha/2}$
- ***p*-value:** $2\Pr(T_{n_1+n_2-2} > |t|)$
- **Two-sided CI:** $\left(\bar{y}_1 - \bar{y}_2 \pm t_{n_1+n_2-2,\alpha/2}\, s_p\sqrt{1/n_1 + 1/n_2}\right)$

Chapter 5 – Paired t-test

- Paired t-test is just one sample t-test for difference of paired data.

Chapter 5 – Inference for Two Proportions

- $H_0: p_1 - p_2 = 0$ versus $H_a: p_1 - p_2 < 0$
- **Test statistic** is: $Z = \dfrac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$
- Where $\hat{p} = \frac{y_1 + y_2}{n_1 + n_2}$ is the pooled estimate of proportion.
- **Rejection region** is: $z < -z_\alpha$
- ***p*-value** is: $\Pr(Z < z)$
- **Upper CI**: $\left(-1, \hat{p}_1 - \hat{p}_2 + z_\alpha \sqrt{\hat{p}_1(1-\hat{p}_1)/_{n_1} + \hat{p}_2(1-\hat{p}_2)/_{n_2}}\right)$

Chapter 5 – Assumptions

- Assumptions for t-test (one-sample, two-sample, paired)
  - Data is normal or approximate normal
  - Check with Q-Q plot
- Assumptions for inference of proportions
  - One proportion: $np_0 \geq 5$ and $n(1-p_0) \geq 5$
  - Two proportions:

$$n_1\hat{p}_1 \geq 5; n_1(1-\hat{p}_1) \geq 5$$
$$n_2\hat{p}_2 \geq 5; n_2(1-\hat{p}_2) \geq 5$$