Review of Basic Statistics

Cheng Peng

Contents

1 Introduction

This module reviews the basic inferential statistics covered in the previous class. The major topics are sampling distribution of sample means and sample proportions, constructing confidence intervals for population means, and testing hypotheses of population means. Details will not be covered in this note. Some of the examples will be explained using software programs such as R and SPSS.

2 A Review of Normal and t Distribution

The following YouTube video from **Ace Tutors** demonstrates how to use the standard normal table to find probabilities of general normal distributions.

3 Sampling distribution

The general estimation method in statistics is to estimate the population parameters such as population mean (usually denoted by the Greek letter μ) and population proportion (denoted by p).

3.1 Sampling Distribution of Sample Means

Assume that a random sample $\{x_1, x_2, \cdots, x_n\}$ is from a population with unknown population μ, then the estimated population mean is given by

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Since the sample is random, therefore, the estimated sample mean $\bar{x}$ is also random. **The question is what is the distribution of** $\bar{x}$

We break down the discussion of the sampling distribution of $\bar{x}$ in three different scenarios.

- **Scenario #1: Large Sample Size**: if the sample size $n > 30$, using the CLT, we have

$$\bar{x} \to N\left(\mu, \frac{s}{\sqrt{n}}\right) \quad \text{or equivalently} \quad \frac{\bar{x} - \mu}{s/\sqrt{n}} \to N(0, 1).$$

Note that, in this scenario, the assumption of a large sample size is crucial to guarantee a good normal approximation. Although there is no theoretically recommended threshold of the sample size to determine whether a sample is large, we use an operational threshold of $n = 30$. That is, in this course, if $n > 30$, the sample is considered a large sample, and the above sampling distribution applies.

- **Scenario #2: Normal Population with Known Standard Deviation** In this scenario, we assume that the population is normally distributed and the population standard deviation (σ) is known (may be based on the historical information). In other words, we will estimate the population variance. With these assumptions, we have

$$\bar{x} \to N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \text{or equivalently} \quad \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \to N(0, 1).$$

- **Scenario #3: Normal Population with Unknown Standard Deviation** The basic assumption is that the population is normal and the population standard deviation is **unknown** - that is, we need to estimate the population standard deviation from the given random sample, denoted by s. In this scenario, the standardized score in the following is a t distribution with $n - 1$ degrees of freedom.

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \to t_{n-1}.$$

A t distribution with df is a symmetric distribution with mean 0 but variance $df/(df - 2) > 1$. If degrees of freedom get bigger, the t distribution approaches standard normal distribution; consequently, we can simply use the central limit theorem to claim the sampling distribution of $\bar{x}$ to be normal as discussed in **scenario #1**. However, under this scenario, if the sample size n is small, the t distribution with $df = n - 1$ must be used to characterize the sampling of

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \to t_{n-1}.$$

3.2 Sampling distribution of Sampling Proportions

The discussion of the sampling distribution of sample proportions is relatively straightforward. Without loss of generality, a binary population has two distinct values: **success** and **failure**. They assume the population proportion size is N, the population proportion.

$$p = \frac{sucesses}{N}.$$

Consider a random sample $\{x_1, x_2, \cdots, x_n\}$ taken from a binary population with distinct values **success** and **failure**. Let $X =$ number of successes in the sample, then the sample proportion is defined to be $\hat{p} = X/n$.

Based on the same logic in the sample mean, $\hat{p}$ is random. Its distribution can be approximated by a normal distribution if the following conditions are satisfied.

$$n\hat{p} \geq 10 \quad \text{and} \quad n(1-\hat{p}) \geq 10.$$

To be more specific, under the above assumptions,

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

4 Confidence Interval

We will review one-sample and two-sample confidence intervals separately.

4.1 One-sample Confidence Intervals

The confidence interval of population mean (μ) and proportion (p) are based on the corresponding sample mean $\bar{x}$ and sample proportion $\hat{p}$ and their associated sampling errors. The explicit form of the confidence interval of μ and p is given respectively as

$$\mu \in (\bar{x} - E_{\bar{x}}, \bar{x} + E_{\bar{x}}) \text{ and } p \in (\hat{p} - E_{\hat{p}}, \hat{p} + E_{\hat{p}}).$$

Where E is the margin of error associated with the sample mean ($\bar{x}$) and proportion ($\hat{p}$) are defined respectively by

$$E_{\bar{x}} = \text{CV}\frac{s}{\sqrt{n}} \quad \text{and} \quad E_{\hat{p}} = \text{CV}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

The critical value CV is a table value (either from the t- or normal table based on the given confidence level).

Before working on examples, let's watch the following YouTube video that provides an example of constructing a normal confidence interval of the population mean.

Example 1. Suppose we want to estimate, with 95% confidence level, the mean (average) length of all walleye fingerlings in a fish hatchery pond. A random sample of 100 fingerlings was selected. The average length is 7.5 inches, and the standard deviation is 2.3 inches. That is, $\bar{x} = 7.5$, $s = 2.3$, and $n = 100$.

Solution: Since $n = 100$, by scenario #1, this means the sampling distribution of sample mean $\bar{x}$ approximately normally distributed. The critical value corresponding $1 - \alpha = 95\%$ confidence interval is given by $Z_{\alpha/2} = Z_{0.025} = 1.96$. The margin of error E is

$$E = 1.96 \times \frac{2.3}{100} = 0.045.$$

The 95% confidence interval for the population mean is $(7.5 - 0.045, 7.5 + 0.045) = (7.05, 7.95)$.

Example 2. In a survey of 1219 U.S. adults, 354 said that their favorite sport to watch is football. Construct a 95% confidence interval for the proportion of adults in the United States who say that their favorite sport to watch is football.

Solution: First of all, the sample proportion $\hat{p} = 354/1219 \approx 0.29$ and sample size $n = 1219$. Since $n\hat{p} = 354 > 5$ and $n(1 - \hat{p}) = 1219 \times 0.71 = 865 > 5$, the sampling distribution of $\hat{p}$ is normally distributed. The critical value corresponding to a 95% confidence level is $Z_{0.025} = 1.96$. The margin of error

$$E = 1.96 \times \sqrt{\frac{0.29(1 - 0.29)}{1219}} \approx 0.0255$$

Therefore, the 95% confidence interval is given by $(0.29 - 0.0255, 0.29 + 0.0255) = (0.2645, 0.2155)$.

Example 3. . Estimating Car Pollution - In a sample of seven cars, each car was tested for nitrogen-oxide emissions (in grams per mile) and the following results were obtained: `0.06, 0.11, 0.16, 0.15, 0.14, 0.08, 0.15` (based on data from the Environmental Protection Agency). Assuming that this sample is representative of the cars in use. Further, the amounts of nitrogen oxide emissions for all cars are normally distributed. Construct a 98% confidence interval estimate of the mean amount of nitrogen oxide emission for all cars.

Solution We first calculate the sample mean and sample standard deviation using the formulas introduced in the note on descriptive statistics.

$$\bar{x} = 0.1214, \quad s = 0.0389.$$

Since this small sample was taken from a normal population with an unknown standard deviation. The critical value should be based on the t-distribution with $7 - 1 = 6$ degrees of freedom. The 98% critical value $t_{6,0.01} = 3.143$. Therefore, the margin of error is

$$E = 3.143 \times 0.0389/\sqrt{7} \approx 0.0462.$$

The resulting 98% confidence interval is given by

$$(0.1214 - 0.0462, 0.1214 + 0.0462) = (0.0752, 0.0752).$$

4.2 Two-sample Confidence Intervals

The information of two-sample inference is the sampling distribution of the difference between the two population parameters, such as means and proportions. The following YouTube video discusses the sampling distribution of the difference of two sample means.

We will only review the confidence intervals for the difference of two population means from **two independent random samples**. The basic settings are summarized in the following table.

	Population #1	Population #2
sample mean	$\bar{x}_1$	$\bar{x}_2$
(Sample) Standard Deviation	s_1 or σ_1	s_2 or σ_2
Sample size	n_1	n_2

Based on different given conditions, we outline the confidence intervals in three different cases:

Case 1: Both Sample sizes Are Large If both sample sizes are large, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is given by

$$\bar{x}_1 - \bar{x}_2 \to N\left(\mu_1 - \mu_2, \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right).$$

The $100(1-\alpha)\%$ confidence interval is explicitly given by

$$(\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Case 2: Both Populations Are Normal and Standard Deviations (σ_1, σ_2) Are Known In this case, there is no restriction on the sample sizes. The sampling distribution of $\bar{x}_1 - \bar{x}_2$ is given by

$$\bar{x}_1 - \bar{x}_2 \to N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right).$$

The corresponding confidence $100(1-\alpha)\%$ confidence interval of $\mu_i - \mu_2$ is explicitly given by

$$(\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Case 3: Both Populations Are Normal and Standard Deviations (σ_1, σ_2) Are Unknown but Equal In this case, we need to combine the two samples to estimate the **common** variance. If the descriptive statistics are given, the **common** variance of the pooled samples is

$$\sigma_{\text{pool}}^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}.$$

The sampling distribution **related to** $(\bar{x}_1 - \bar{x}_2)$ is given by

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_{\text{pool}}^2/n_1 + \sigma_{\text{pool}}^2/n_2}} \to t_{n_1+n_2-2}.$$

Using t-critical value, we write the $100(1-\alpha)\%$ confidence interval of $\bar{x}_1 - \bar{x}_2$ as

$$(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2,\alpha/2}\sqrt{\frac{\sigma^2_{\text{pool}}}{n_1} + \frac{\sigma^2_{\text{pool}}}{n_2}}.$$

The following YouTube video summarizes the two-sample t confidence interval with an example.

5 Testing Hypothesis

We've reviewed confidence intervals (CI), which estimate a population parameter with a range of plausible values. Let's focus on the logic of hypothesis testing, the other major type of statistical inference. While CIs provide a range, hypothesis testing evaluates a specific claim about a population parameter. We primarily focus on testing hypotheses about one and two population means.

Hypothesis testing follows a structured process to determine whether sample data provides sufficient evidence to reject a default assumption (null hypothesis) in favor of an alternative claim. Here's the step-by-step reasoning:

- Identify the claim about the population.
  - For testing a population mean, the claim **must** be **one** of the 6 forms: $\mu > \mu_0, \mu < \mu_0, \mu \neq \mu_0, \mu = \mu_0, \mu \geq \mu_0, \mu \leq \mu_0$.
- State the Hypotheses
  - **Null Hypothesis (H_0)**: It is either the **identified claim** or the **opposite of the identified claim** that **must include an equal sign ("=")**.
  - **Alternative Hypothesis (H_a or H_1)**: The opposite of the **null hypothesis**. That is, **the alternative hypothesis must NOT include an equal sign ('=').**
- Choose Significance Level (α)
  - The threshold for deciding whether to reject H_0 (common: $\alpha = 0.05$, 0.01).
  - The default significance level $\alpha = 0.05$.
- Select the Test Statistic
  - Depends on the data type, sample size, and whether population parameters (e.g., σ) are known:
    - **z-test**: For large samples ($n > 30$) **OR** a normal population with a known σ.
    - **t-test**: For small samples ($n \leq 30$) **AND** a normal population with unknown σ.
- Calculate the Test Statistic
  - Quantifies how far the sample result deviates from Ho, scaled by standard error.
    - z-test:
    $$TS = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$
    - t-test:
    $$TS = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$
    where $\bar{x}$ = sample mean, μ_0 = hypothesized mean, s = sample standard deviation.
- Determine the p-value or Critical Value
  - **Option 1**: p-value Approach
    - **p-value**: Probability of observing the test statistic (or more extreme) if Ho is true.
    - **Decision**: Reject H_0 if p-value $\leq \alpha$.
  - **Option 2**: Critical Value (CV) Approach - compare the test statistic to a critical value from tables (e.g., z-table, t-table) based on α and tails. Rejection Rule:
    - **Two-tailed**: |TS| > CV.

        * **One-tailed**: TS > CV (right) or < CV (left).
- Make A Statistical Decision
  - **Reject** H_0: If p-value $\leq \alpha$ or test statistic falls in the critical region. **Conclusion**: *Statistically significant evidence for* H_a
  - **Fail to Reject** H_0: If p-value $> \alpha$ or test statistic does not fall in the critical region. **Conclusion**: *Insufficient evidence to support* H_a.
- Interpret the Result
  - Avoid saying "accept H_0"; instead, say "fail to reject H_0" (absence of evidence $\neq$ evidence of absence). The interpretation is something like *at* $\alpha = 0.05$*, we reject* H_0*, suggesting the mean IQ differs significantly from 100 (*$p = 0.004$*).*

The following YouTube video from Ace Tutors summarizes the basic logic and process of performing hypothesis testing.

5.1 One Sample Tests for Population Means

This type of inference about the population mean is to verify a claim about the population mean, which has one of the aforementioned six forms. Based on the available amount of information, we discuss the test in the following scenarios.

- **Large Sample Normal Test** - The only assumption is to have a large sample size (the conventional large sample size $n > 30$). In this case, the test statistic is

$$TS = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \to N(0, 1).$$

The critical value and p-value are based on the standard normal distribution.

- **Normal Population with Known Standard Deviation** - If the normal population standard deviation σ is given, the the test statistic

$$TS = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \to N(0, 1).$$

In this case, the critical value and p-value are based on the standard normal distribution.

- **Normal Population with Unknown Standard Deviation** - If the normal population standard deviation σ is **not** given, the the test statistic

$$TS = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \to t_{n-1}.$$

The following YouTube video explains when a Z-test or t-test should be used.

5.2 Two-sample Test of Population Means

This type of inference focuses on comparing the difference between two population means $\mu_1 - \mu_2$. The definition of the test statistic of this hypothesis testing is dependent on the given conditions.

Both Sample Sizes Are Large: If both sample sizes are large (say $n_1 > 30$ **and** $n_2 > 30$), the test statistic is defined to be

$$TS_1 = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

where s_1^2 and s_2^2 are sample variances. If both population variances σ_1^2 and σ_2^2 are given, replace the two sample variances in the above formula with the given population variances. Using the Central Limit Theorem (CLT), $TS_1 \to N(0,1)$

Both Populations Are Normal and Corresponding Variances Are Given: In this case, we don't have a restriction on the sample sizes. The test statistic is defined to be

$$TS_2 = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \to N(0,1).$$

Both Populations Are Normal and Population Variances Are Unknown But Equal: In this case, we need to pool the two samples to estimate the **common** variances. For given descriptive statistics (i.e., sample sizes, sample means, sample variances), the estimated **common** variance is given by

$$s_{\text{pool}}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

The test statistic is defined as

$$TS_3 = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\text{pool}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \to t_{n_1+n_2-2}.$$

Note that the condition of **equal variances** is crucial to guarrantee the t distribution with $n_1 + n_2 - 2$ degrees of freedom. If this **equal variances** is not satisfied, TS_3 is not a t distribution. However, we can define the test statistic and approximate a t distribution.

The next YouTube video explains the pooled sample t-test.

6 Simple Linear Regression and Correlation

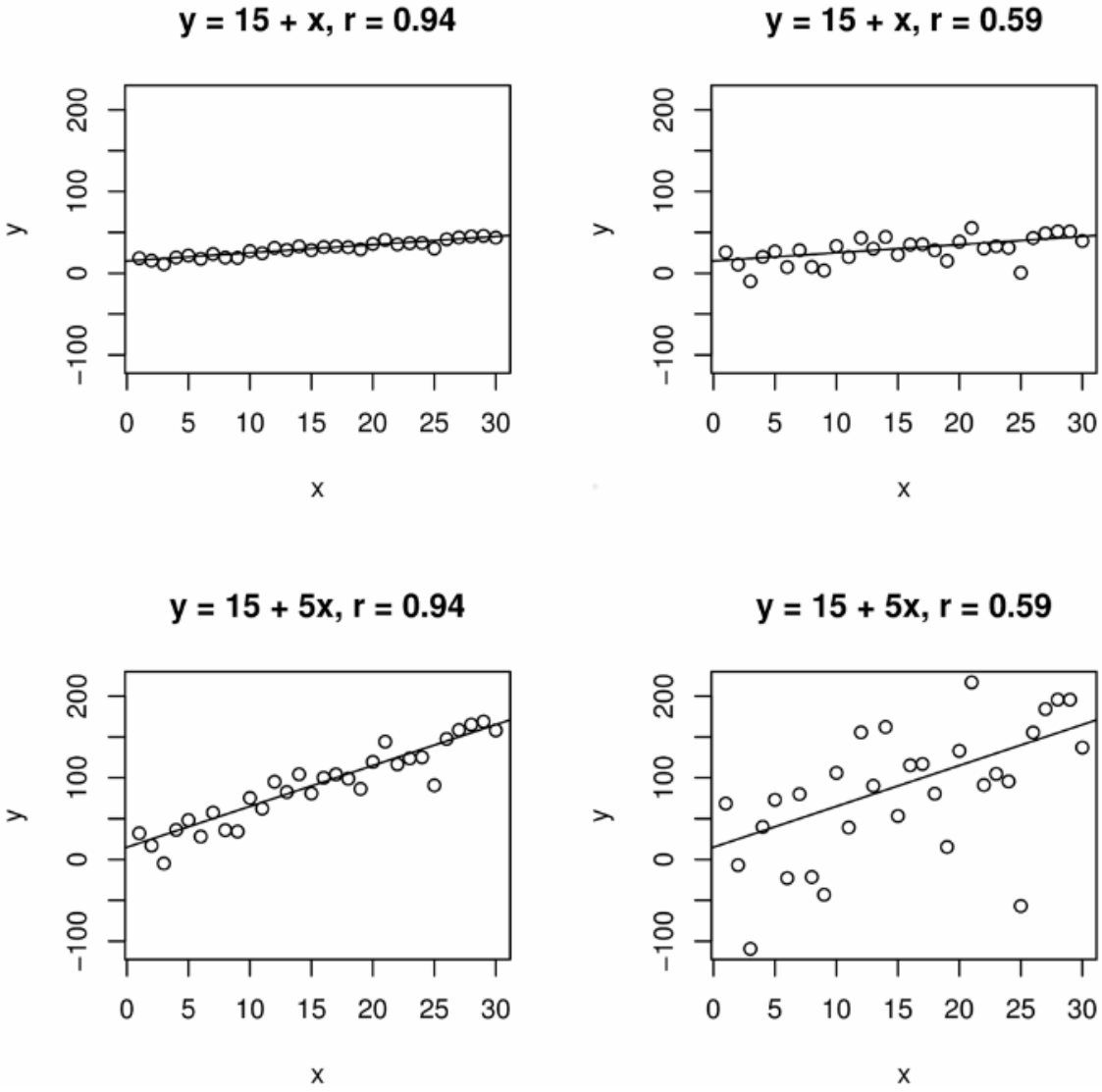
In introductory statistics, we also learned how to characterize the relationship between two numerical variables using the correlation coefficient and least squares regression.

6.1 Pearson Correlation Coefficient

The Pearson correlation coefficient measures the magnitude and the direction of the linear **relationship**. The formula for the sample correlation coefficient (Pearson) is given by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where x and y are two numerical variables. Although r carries the magnitude and direction of the linear relationship, it does not indicate how one variable influences the other variable.



6.2 Linear Regression

Simple linear regression (SLR) explicitly shows the influence of one variable on the other. A simple linear regression (SLR) is defined as

$$y = \beta_0 + \beta_1 x + \epsilon,$$

Where y is called the response variable (also called the dependent variable), x is called the predictor variable (also called the independent variable or explanatory variable). ϵ is a random variable which is called **residual**. The basic assumptions on SLR are

- y and x are linearly related (i.e., linear pattern)
- y is random and x is **not** random.
- $\epsilon \to N(0, \sigma^2)$ where σ is a constant.

β_0 and β_1 are intercept and slope parameters in the SLR. The slope parameter β_1 reflects the change of y when x increases by one unit.

Relationship between Pearson Correlation Coefficient (r) and Slope parameter (β_1): For SLR, both regression coefficients (β_0 and β_1) can be explicitly based on the sample data.

obs	1	2	3	...	$n-1$	n
x	x_1	x_2	x_3	...	x_{n-1}	x_n
y	y_1	y_2	y_3	...	y_{n-1}	y_n

Let

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad \text{and} \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the estimated values of β_0 and β_1 and can be expressed by

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \times \bar{x}.$$

Note that

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \left(\frac{S_{xy}}{S_{xx}}\right)\left(\frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}}\right) = \hat{\beta}_1\sqrt{\frac{S_{xx}}{S_{yy}}}.$$