Midterm Review STAB27 2024 Winter

Yushu Zou

March 2024

1 Recap for the previous knowledge

- Parameter vs. Statistics
- Central Limit Theorem

2 Hypothesis Testing

2.1 General Steps to Perform the Hypothesis Testing

- State Null Hypothesis and Alternative Hypothesis

$H_0 : \mu = k$	$H_0 : \mu \leq k$	$H_0 : \mu \geq k$
$H_A : \mu \neq k$	$H_A : \mu > k$	$H_A : \mu < k$

Figure 1: Two-sided test vs. One-sided Test

- Calculate the statics of suitable tests and compare with critical value

  Will be discussed later for the suitable test;

  Conclusion:

  - Since test statistics is greater than the **critical value**, we can conclude that there exists a statistically significant difference ...; therefore, we **reject** H_0 (you need to state what is your H_0)
  - Since test statistics are smaller than the **critical value**, we can not conclude that there exists a statistically significant difference ...; therefore, we **fail to reject** H_0 (you need to state what is your H_0)

- Draw Confidence Interval and make Conclusions

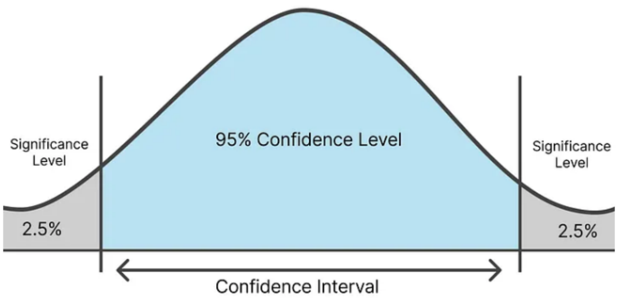


Figure 2: Two sided 95% Confidence Interval

Conclusion:

- Since ... is within the interval, we are (1-α)% confident that the difference ... (population parameter) is between ... (Lower bound) and ... (Upper bound)
- There's a xxx% chance that this xxx% confidence interval construction procedure captured the true parameter value
- A larger confidence level (e.g., instead of 95%, use 98%) would ensure that we capture the population parameter in more samples. This would give a wider confidence interval.

3 One sample T-test

Statistics:

$$t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

$\overline{x}$ - sample mean; μ_0 - stated by null hypothesis; s - sample standard deviation; n - sample size;

3.1 Assumptions

Random & Independent sample of size < 10% of population size, coming from approx. **unimodal & symmetric (bell-shaped) population distribution**.

4 Paired Sample vs. Not Paired Sample

Dependent samples are **paired measurements** for one set of items. Independent samples are measurements made on two **different sets of items**.

5 Two sample T-test

5.1 Assumptions

- The two samples are **independent**.
- The two samples are randomly selected form normally distributed populations.

5.2 Case 1: Large sample sizes($n_1 \geq 30$ and $n_2 \geq 30$)

1. Statistics:

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

2. Test: Z-table (Normal Distribution)
3. Confidence Interval:

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

5.3 Case 2: Small sample sizes from *normal* populations ($\sigma_1^2 = \sigma_2^2$)

Pooled T-test

1. Statistics:

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

   Where,

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} = MSE$$

2. Test: T-table with $df = n_1 + n_2 - 2$
3. Confidence Interval:

$$(\overline{x}_1 - \overline{x}_2) \pm t_{\frac{\alpha}{2}, df} \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

5.3.1 Adds-on Assumption

- At least one of the two sample sizes is small: $n < 30$
- $\sigma_1^2 = \sigma_2^2$

5.4 Case 3: Small sample sizes from *normal* populations ($\sigma_1^2 \neq \sigma_2^2$)

Welch's Approximation

1. Statistics:
$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
2. Test: T-table with
$$df = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2 - 1}}$$
3. Confidence Interval:
$$(\overline{x}_1 - \overline{x}_2) \pm t_{\frac{\alpha}{2}, df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

6 Paired T test

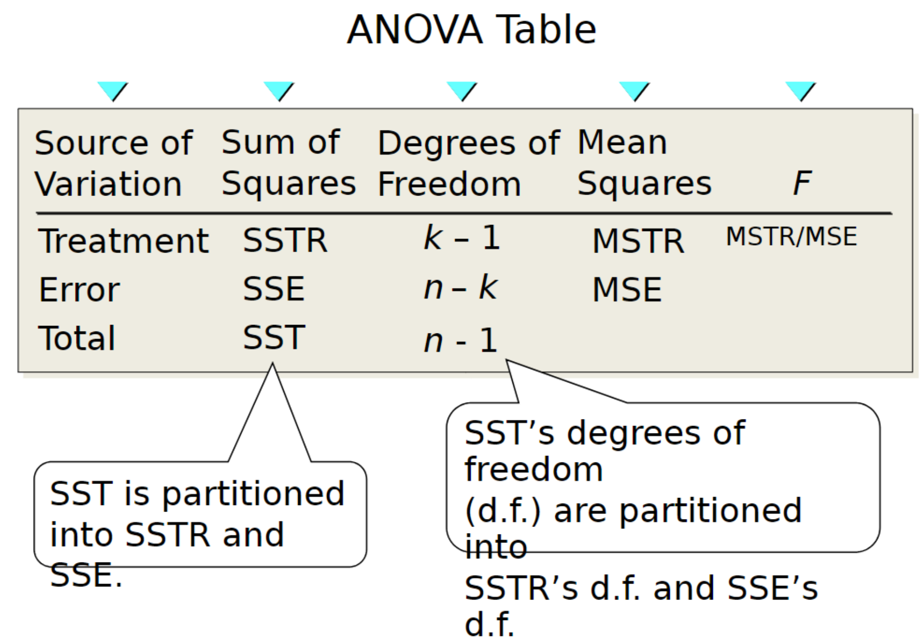
Let $d_i = x_i - y_i$, calculate

1. the mean value - $\overline{d}$;
2. standard deviation of the d for the paired sample data - s_d;
3. number of pairs of data in the sample - n

1. Statistics:
$$t = \frac{\overline{d} - \mu_d}{\sqrt{\frac{s_d^2}{n}}}$$
2. Test: T-table with $df = n - 1$
3. Confidence Interval:
$$(\overline{x}_1 - \overline{x}_2) \pm t_{\frac{\alpha}{2}, df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

7 ANOVA Test: for more than 3 groups comparison

The more variation there is between the groups, the more likely we are to conclude that there is a significant difference between the means of at least two of these groups.



ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F
Treatment	SSTR	$k - 1$	MSTR	MSTR/MSE
Error	SSE	$n - k$	MSE	
Total	SST	$n - 1$		

Figure 3: ANOVA Table Coefficients Part I

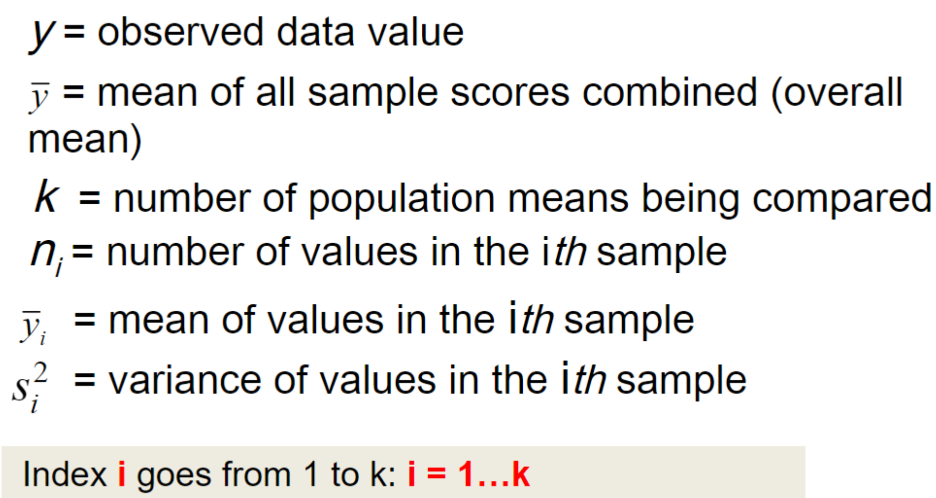


Figure 4: ANOVA Table Coefficients Part II

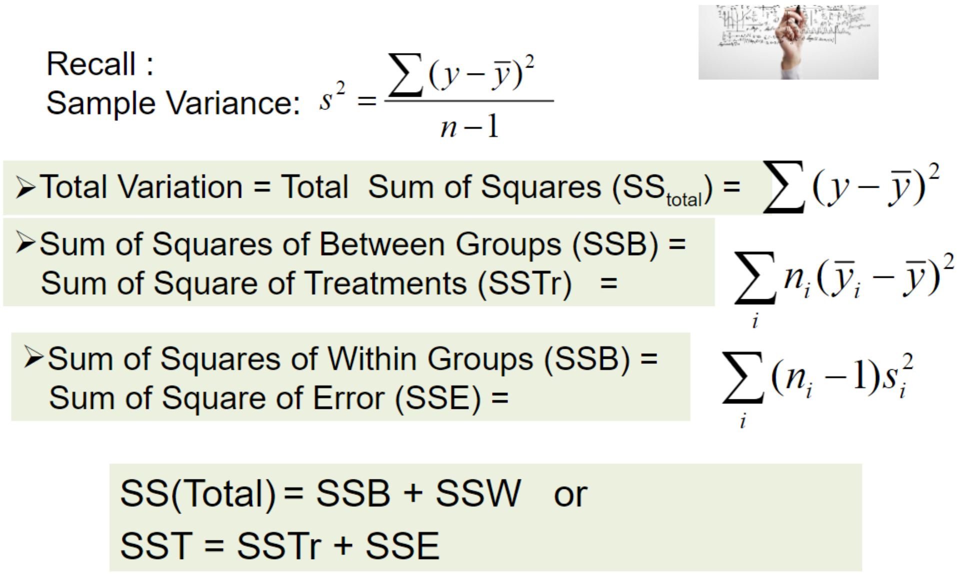


Figure 5: ANOVA Table Coefficients Part III

1. Hypotheses: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_n$ vs. H_a: Not all population means are equal. At least two population means are different

2. Statistics:
$$F = \frac{MSTR}{MSE} = \frac{MS(\text{between})}{MS(\text{within})} = \frac{SSTR/(k-1)}{SSE/(n-k)} = t^2$$

3. Test: F distribution with $k-1$ numerator d.f. and $n-k$ denominator d.f.

7.1 Assumptions

- The populations have approximately **normal** distributions.
- The populations have the **same variance** σ^2.
- The samples are **random** and **independent** of each other.
- The different samples are from populations that are categorized in only **one way**.

8 Multiple Comparisons of Means

8.1 Turkey's Test

1. Confidence Interval:
$$|\overline{x}_1 - \overline{x}_2| \pm \frac{q_\alpha(k,v)}{\sqrt{2}}\sqrt{\frac{MSE}{n_i} + \frac{MSE}{n_j}}$$
Note: in q-table, $p = k$ and $v = n - k$ If 0 falls outside the confidence interval, then we can conclude that and are significantly different between two population means.

8.2 Bonferroni's Test

1. Number of pair-wise comparisons- $C = \frac{k(k-1)}{2}$, where k is the number of groups

2. Set $\alpha = \frac{\alpha_E}{C}$, where α_E is the true probability of making at least one Type I error (called **experiment wise Type I error**).

3. Statistics:
$$\frac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{MSE(\frac{1}{n_1} + \frac{1}{n_2})}}$$

4. Test: T-distribution, df $= n - k$, with $\frac{\alpha}{2} = \frac{\alpha_E}{2C}$

5. Conclusion: if

$$|\overline{x}_1 - \overline{x}_2| > t_{\frac{\alpha_E}{2C}} \sqrt{MSE(\frac{1}{n_1} + \frac{1}{n_2})}$$

then there is a significant difference between two population means.

9 Simple Linear Regression: Available data on *two or more variables*

- Y - Dependent Variable/Response Variable;
- $(x_1, x_2, \cdots x_k)$ - Independent Variables/ Explanatory Variable
- Estimated Regression Line: $y = b_0 + b_1 * x$; b_0 - intercept: predicted y value when x = 0; b_1 - slope: increase in predicted y for every unit increase in x.

9.1 Simple Linear Regression Model

- First order model: $y_i = \beta_0 + \beta_1 * x_i + \epsilon_i$ to describe each data point
- Simple Linear Equation: $E(y) = \beta_0 + \beta_1 * x$
- Positive Linear Relationship: $\beta_1 > 0$; Negative Linear Relationship: $\beta_1 < 0$; No Relationship: $\beta_1 = 0$

Notation for Regression Equation

	Population Parameter	Sample Statistic
y-intercept of regression equation	β_0	b_0
Slope of regression equation	β_1	b_1
Equation of the regression line	$E(y) = \beta_0 + \beta_1 x$	$\hat{y} = b_0 + b_1 x$

Figure 6: Linear Regression Notation

9.2 Assumption

1. The error term $\epsilon \sim N(0, \sigma^2)$
2. $Var(Y) = Var(\epsilon) = \sigma^2$
3. ϵ are iid distributed

9.3 Application

9.3.1 How to calculate the coefficients of the linear model?

- Slope for the Estimated Regression Equation

$$\hat{\beta}_1 = b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

$$SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^{\,2} - n(\bar{x})^2$$

Figure 7: Steps to calculate the slope

- y-Intercept for the Estimated Regression Equation

$$b_0 = \bar{y} - b_1\bar{x}$$

where:

x_i = value of independent variable for ith observation

y_i = value of dependent variable for ith observation

$\bar{x}$ = mean value for independent variable

$\bar{y}$ = mean value for dependent variable

n = total number of observations

Figure 8: Steps to calculate the intercept

9.3.2 Coefficient of Determination

$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ - about $R^2\%$ of variation in Y is explained by this regression model with X

- Note: $F = \frac{R^2}{1-R^2}(n-2)$

9.3.3 Hypothesis Test on b_1

- Null Hypothesis: $b_1 = 0$ vs Alternative Hypothesis: $b_1 \neq 0$

1. F-test: $F = \frac{MSR}{MSE} \sim F_{1,n-2}$

2. T-test: $t = \frac{b_1}{s_{b_1}} = \frac{b_1}{\sqrt{\frac{MSE}{SS_{xx}}}} \sim t_{n-2}$

3. Confidence Interval: $b_1 \pm t_{n-2}(\alpha/2)s_{b_1}$; Reject H_0 if 0 is not included in the confidence interval for b_1