Final Review STAB27 2024 Winter

Yushu Zou

March 2024

1 Hypothesis Testing

1.1 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

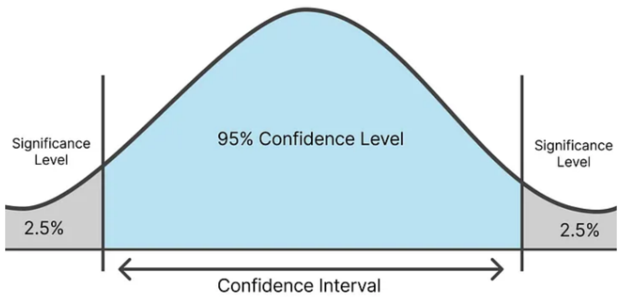

Figure 2: Two sided 95% Confidence Interval

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.

2 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;

2.1 Assumptions

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

3 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**.

4 Two sample T-test

4.1 Assumptions

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

4.2 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:
$$(\overline{x}_1 - \overline{x}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

4.3 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:
$$(\overline{x}_1 - \overline{x}_2) \pm t_{\frac{\alpha}{2}, df} \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

4.3.1 Adds-on Assumption

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

4.4 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:
$$(\overline{x}_1 - \overline{x}_2) \pm t_{\frac{\alpha}{2}, df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

5 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:
$$(\overline{x}_1 - \overline{x}_2) \pm t_{\frac{\alpha}{2}, df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

6 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.

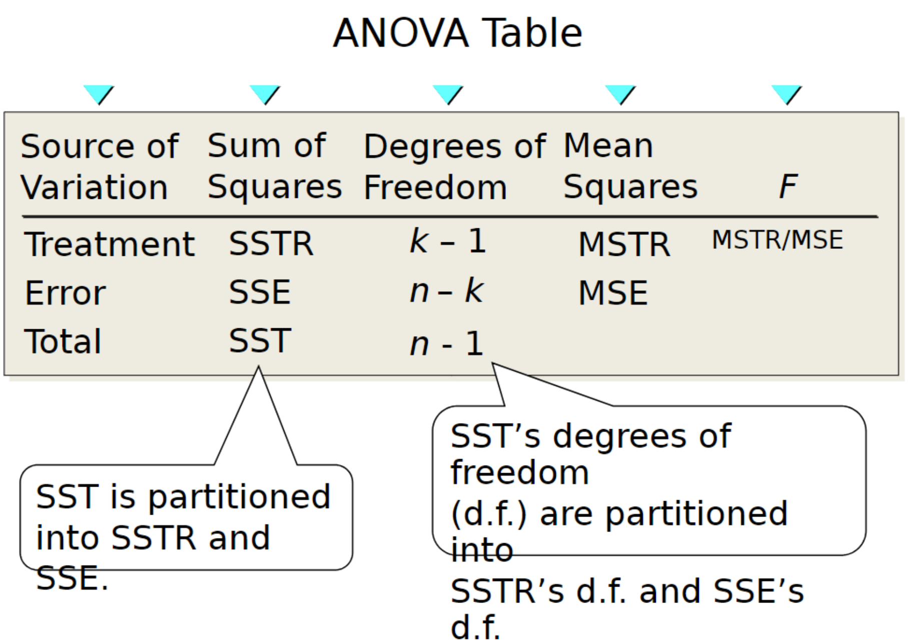

Figure 3: ANOVA Table Coefficients Part I

y = observed data value

$\bar{y}$ = mean of all sample scores combined (overall mean)

k = number of population means being compared

n_i = number of values in the ith sample

$\bar{y}_i$ = mean of values in the ith sample

s_i^2 = variance of values in the ith sample

Index **i** goes from 1 to k: **i = 1…k**

Figure 4: ANOVA Table Coefficients Part II

Recall :
Sample Variance: $s^2 = \dfrac{\sum (y-\bar{y})^2}{n-1}$

➢Total Variation = Total Sum of Squares (SS_{total}) = $\sum (y-\bar{y})^2$

➢Sum of Squares of Between Groups (SSB) =
Sum of Square of Treatments (SSTr) = $\sum_i n_i(\bar{y}_i-\bar{y})^2$

➢Sum of Squares of Within Groups (SSB) =
Sum of Square of Error (SSE) = $\sum_i (n_i-1)s_i^2$

SS(Total) = SSB + SSW or
SST = SSTr + SSE

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.

6.1 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**.

7 Multiple Comparisons of Means

7.1 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.

7.2 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.

8 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: $E(y) = \beta_0 + \beta_1 * x$; β_0 - intercept: predicted y value when x = 0; β_1 - slope: increase in predicted y for every unit increase in x.

8.1 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

8.2 Assumption

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

8.3 Application

8.3.1 How to calculate the coefficients of the linear model?

- Slope for the Estimated Regression Equation

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

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

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

Figure 7: Steps to calculate the slope

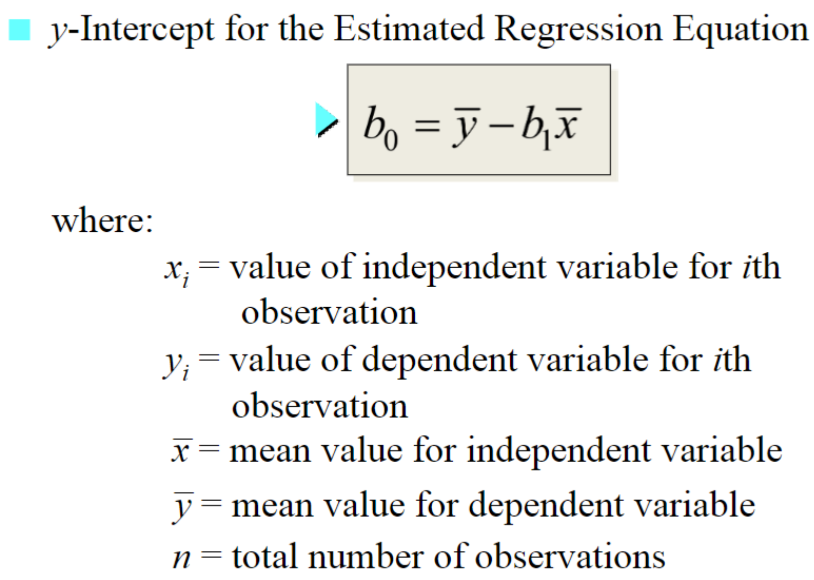

Figure 8: Steps to calculate the intercept

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

8.3.3 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

8.3.4 Residual Analysis

Based on the assumption of linear regression.

- Plot scatter plot of residual, see if they are **randomly distributed around 0**
- Run Normal test on residual (shapiro.test) to see if it is normal distributed: p¡0.05, not normal

8.4 Correlation Coefficient

Measures only the **linear** relationship between two quantitative variables

$$r = \frac{SS_{xy}}{\sqrt{SS_x}\sqrt{SS_y}} = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{(\sum(x-\overline{x})^2}\sqrt{(\sum(y-\overline{y})^2}} = \frac{Cov(X,Y)}{S_x S_y} \in [-1,1]$$

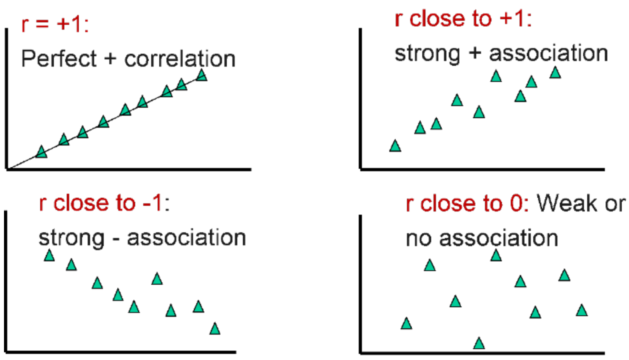

Figure 9: Correlation Coefficient

Correlation Coefficient (absolute value)	Interpretation
up to $\|0.2\|$	very low correlation
up to $\|0.5\|$	low correlation
up to $\|0.7\|$	moderate correlation
up to $\|0.9\|$	high correlation
above $\|0.9\|$	very high correlation

Figure 10: Inference on correlation coefficient

8.4.1 Lurking variable

hidden variable that may misdirect the association between variables.

8.4.2 Hypothesis Testing on Correlation

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

1. Test Statistic: $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}(\frac{\alpha}{2})$

8.5 Prediction Interval and Confident Interval

8.5.1 Confidence Interval for $E(y)$

1. $S_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$
2. Marginal Error $= \sqrt{\frac{1}{n} + \frac{(x_0+\overline{x})^2}{SS_{xx}}}$
3. Confidence Interval for $E(y)$ for$x = x_0$: $Y \pm t_{n-2}(\alpha/2)S_e \times$ Marginal Error
4. Conclusion: We are (1-α)% confident that the mean Y at X is x_0 will be between (xxx) and (xxx)

8.5.2 Prediction Interval for $\hat{y}$

1. $S_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$
2. Marginal Error $= \sqrt{1 + \frac{1}{n} + \frac{(x_0+\overline{x})^2}{SS_{xx}}}$
3. Prediction Interval for $\hat{y}$ for$x = x_0$: $\hat{y} \pm t_{n-2}(\alpha/2)S_e \times$ Marginal Error
4. Conclusion: We are (1-α)% confident that a single y when X is x_0 will be between (xxx) and (xxx)

9 Multiple Linear Regression

Expresses a linear relationship between a dependent variable y and two or more independent variables $(x_1, \cdots, x_n)$
Similar to simple linear regression, but for the inference on the coefficients:

- $E(y) = \beta_0 + \beta_1 * x_1 + \cdots + \beta_k * x_k$
- b_i represents an estimate of the change in y corresponding to a 1-unit increase in x_i **when all other independent variables are held constant**

9.1 Multiple Coefficient of Determination

Refer to Figure 3, the relationship between coefficients in ANOVA table, $R^2 = SSR/SST$

- Adjusted Coefficient of Determination $= 1 - \frac{n-1}{n-k-1}(1 - R^2)$

9.2 F test & t test

9.2.1 F test (test for overall significance)

Determine whether a significant relationship exists between the dependent variable and the set of **all** the independent variables

- Null Hypothesis: $b_1 = \cdots = b_k = 0$ vs Alternative Hypothesis: one of $b_i \neq 0$

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

9.2.2 t test(test for individual significance

Determine whether each of the individual independent variables is significant.

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

1. T-test: $t = \frac{b_1}{s_{b_1}} = \frac{b_1}{\sqrt{\frac{MSE}{SS_{xx}}}} \sim t_{n-k-1}$

9.3 Interaction Model

$E(y) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_1 x_2$
Since X_1 and X_2 are interactive. Slope for x_1: $\beta_1 + \beta_3 x_2$. The change in E(y) for every 1-unit change increase in x_1, holding x_2 fixed

9.4 Quadratic Model

$E(y) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_1^2 + \beta_3 * x_1^3$

9.5 Models with Qualitative Independent Variables

some of independent variable are discrete variables.

- $E(y) = \beta_0 + \beta_1 * x_1 + \cdots + \beta_k * x_k$

9.5.1 Dummy variables

9.6 Nested F test (Comparing Models)

Assume we have two models, they have many shared variables, but just minor different:

- Complete case model (with $\beta_1, \cdots, \beta_k$)
- Reduced model (with $\beta_1, \cdots, \beta_j, j < k$)
- Null Hypothesis: $\beta_j = \cdots \beta_k = 0$ vs Alternative Hypothesis: $\beta_i \neq 0$

1. F-test: $F = \frac{(SSE_R - SSE_c)/(k-j)}{MSE_C} \sim F_{k-j,n-k}$

10 Multicollinearity

- VIF $= \frac{1}{1-R^2}$; if VIF > 10 then model suffers from multicollinearity.