🎉 STAT 350 Help Session for Final Finale 🎉

Part I: Midterms 1 & 2 Topics

FALL 2024 Final

1.2. Let X follows a binomial distribution with fixed p and sufficiently large number of trials n. The estimator $\hat{p} = \frac{X}{n}$, representing the sample proportion of successes, is derived from the random variable X, which is the sum of n independent and identically distributed Bernoulli trials,

Ⓣ or Ⓕ the sampling distribution of $\hat{p} = \frac{X}{n}$ is approximately normal.

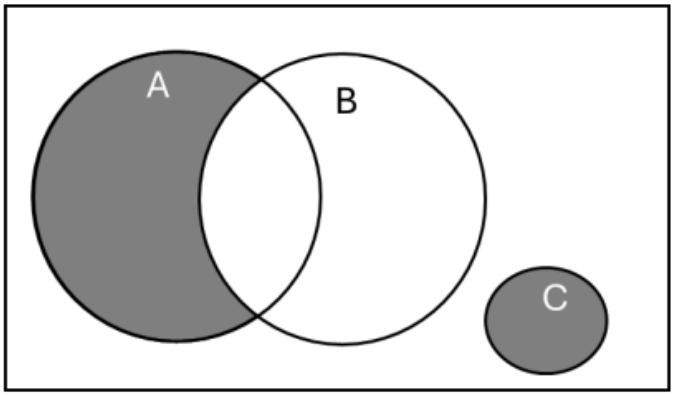
1.4. The time it takes for a customer to complete a transaction at a store is uniformly distributed between 2 and 10 minutes.

Ⓣ or Ⓕ The probability that a transaction lasts between 1 and 6 minutes is greater than the probability that a transaction lasts between 6 and 10 minutes.

1.5. A research team investigates whether consuming a spoonful of apple cider vinegar before meals prevents blood sugar spikes. They selected 30 pairs of identical twins, randomly assigning each twin in a pair to one of two groups.

Ⓣ or Ⓕ In this scenario, a two-sample independent procedure is appropriate to compare the groups.

2.1. Select the expression that does **NOT** correctly represent the probability of the colored area in the Venn diagram shown below.



Ⓐ $P(A \cup B) - P(B) + P(C)$

Ⓑ $P(A) - P(A \cap B) + P(C)$

Ⓒ $P(A) - P(B) + P(C)$

Ⓓ $P(A \cup C) - P(A \cap B)$

Ⓔ $P(A \cup B \cup C) - P(B)$

2.2. Fréchet distribution is a heavily skewed, right-tailed continuous distribution that is used for modeling extreme events such as earthquake magnitudes, daily rainfall totals, and large insurance claims. Which of the following statements is TRUE about Fréchet distribution?

Ⓐ The mean is the largest among the measures of central tendency, followed by the mode and the median.

Ⓑ A small sample size is adequate to apply the central limit theorem to the distribution of the sample mean.

Ⓒ For samples from this distribution, the median and variance are recommended measures of central tendency and spread, respectively.

Ⓓ The interquartile range (IQR) is preferred for describing the spread of a population with a Fréchet distribution because it is less sensitive to outliers.

Ⓔ None of the above statements are TRUE for the Fréchet distribution.

2.3. Suppose $\boldsymbol{X}$ is a random variable with $\boldsymbol{E[2^X] = 16, Var(X) = 32}$, and $\boldsymbol{E[3X + 2] = 8}$. Let a new random variable $\boldsymbol{Y}$ be defined as $\boldsymbol{Y = 2^X - \frac{1}{4}X^2}$. What is $\boldsymbol{E[Y]}$?

Ⓐ **0**

Ⓑ **7**

Ⓒ **8**

Ⓓ **36**

Ⓔ None of the above

SPRING 2025 Final

1.1. Suppose $\boldsymbol{X}$ and $\boldsymbol{Y}$ are two random variables with a large covariance

$$\boldsymbol{COV(X, Y) = 100,000},$$

and the individual standard deviation $\boldsymbol{\sigma_X}$ and $\boldsymbol{\sigma_Y}$ are unknown but finite.

Ⓣ or Ⓕ From this information, we can conclude that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are strongly correlated.

1.3. Bayes' Theorem is often used for revising probabilities based on new evidence.

Ⓣ or Ⓕ Bayes' Theorem applies when the events of interest, $A_1, A_2, \ldots, A_k$ are mutually exclusive (disjoint), and the evidence event B has positive probability.

2.1. Suppose A, B, C and D are non-empty events in the sampe sample space where $P(C) > P(A \cup B) > P(D) > 0.7$. Which of the following statements is **TRUE**?

Ⓐ D must be a subset of C.

Ⓑ $P(A' \cap B') < P(C') < 0.3$ holds.

Ⓒ The two events C and D can be mutually exclusive.

Ⓓ If $A \cap C = \emptyset$, then $B \cap C$ must be a non-empty set.

Ⓔ The two events A and B are independent.

2.2. Which statement regarding the properties of common random variables is **FALSE**?

Ⓐ A **Poisson random variable** counts the number of events occurring in a fixed interval of time, area, or space.

Ⓑ A **Uniform random variable** assigns equal probability density across its entire support.

Ⓒ A **Binomial random variable** counts the number of independent trials required to achieve a specified number of successes.

Ⓓ An **Exponential random variable** measures the waiting time between consecutive independent events.

Ⓔ None of the statements listed above are false.

Part II: Post-Midterm 2 Topics

Chapter 12: The Analysis of Variance (ANOVA)

Definitions

- Factor: a variable that makes the population groups distinguishable
- Level: the number of different categories or conditions in a factor (= # of populations)

ANOVA Assumptions

1. Simple random sample (SRS) from each of the k populations
2. The responses in each group are independent of those in the other groups.
3. The sample mean of each population is (approx..) normally distributed. ➔ Histograms/Normal Probability Plots
4. Population variances are the same (homogeneity of variance). ➔ $\frac{\max sample\ SD}{\min sample\ SD} \leq 2$

ANOVA Model: $X_{ij} = \mu_i + \epsilon_{ij},\ \ \epsilon_{ij} \sim_{iid} N(0, \sigma^2)$

- X_{ij} – the jth observation in a group i (data)
- μ_i – the population mean of a group i (group mean)
- ϵ_{ij} – the unexplained error by μ_i (error)

Hypotheses

H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$
H_1: at least one μ_i is different from the rest

ANOVA table

- $SST = SSA + SSE$ and $df_T = df_A + df_E$

Source	df	Sum of Squares	Mean Squares	F
Factor A (between)	$k-1$	$SSA = \sum_{i=1}^{k} n_i(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2$	$MSA = \frac{SSA}{k-1}$	$F_{ts} = \frac{MSA}{MSE}$
Error (within)	$n-k$	$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i\cdot})^2$	$MSE = \frac{SSE}{n-k}$	
Total	$n-1$	$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{\cdot\cdot})^2$		

- Variance estimate: $\hat{\sigma} = s = \sqrt{MSE}$
- P-value $= P(F \geq F_{ts}) = pf(F_{ts}, df_A, df_E, lower.tail = FALSE)$
- Conclusion:
  - If p-value ≤ α, we reject H_0. The data provides strong evidence that at least one population mean of (context) is different from the rest.
  - If p-value > α, we fail to reject H_0. There is not enough evidence to show that at least one population mean of (context) is different from the rest.
- Special case of # groups $= 2$ ➔F_{ts} for ANOVA = t_{ts}^2 for two-sample indep, two-sided hypothesis with $\Delta_0 = 0$

Pairwise comparison: number of distinct pairs out of k groups: $c = \frac{k(k-1)}{2}$

Goal: control type I error to test multiple hypotheses at once
Visualization: Line Plot (put a line under the means that are statistically same)
Common structure for CI: only critical value changes

$$\bar{x}_{i\cdot} - \bar{x}_{j\cdot} \pm (critical\ value) * \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

1. Bonferroni: evenly share the significance level $\boldsymbol{\alpha_{each}} = \frac{\alpha_{overall}}{c}$ ➔ critical value $= t_{\alpha_{each}/2, n-k}$
   😊 simple construction ☹ α_{each} becomes too small for large c (conservative)
2. Tukey: establish a distribution on $\bar{X}_{max} - \bar{X}_{min}$ ➔ critical value $= \frac{Q_{\alpha,k,n-k}}{\sqrt{2}}$ = qtukey(1-alpha, k, n-k)/sqrt(2)
   😊 more powerful method for large c ☹ complicated theoretical construction
3. Dunnett ☹ compares multiple treatments to a control group (only $k-1$ pairwise comparisons)

Fall 2023

1.3. Both **Bonferroni** and **Tukey's** method are statistical techniques used to control the **Family-Wise Error Rate (FWER)** in multiple comparison procedures.

Ⓣ or Ⓕ Tukey's method is generally less conservative than Bonferroni's method in controlling the **FWER**.

1.6. Consider a random variable ***X*** that follows an F distribution with numerator and denominator degrees of freedom equal to 5 and 15, respectively.

Ⓣ or Ⓕ In this context, it is theoretically possible for the random variable ***X*** to take negative values.

2.1. In the context of a researcher conducting an ANOVA analysis to compare the population means of 4 populations, and all necessary ANOVA assumptions have been met, and the ANOVA procedure has resulted in statistical significance, how many total pairwise comparisons should the researcher conduct as a follow-up?

Ⓐ **2**

Ⓑ **4**

Ⓒ **6**

Ⓓ **10**

Ⓔ **24**

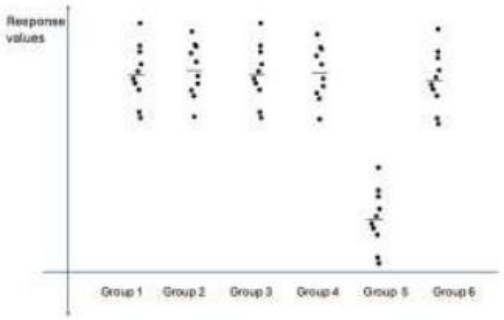
Fall 2024

1.6. In a one-way ANOVA analysis for a factor with nine levels, the F-test resulted in rejection of the null hypothesis. If all possible pairs of levels are to be compared,

Ⓣ or Ⓕ the Multiple Comparisons step would involve 72 paired comparisons.

Spring 2025

1.4. The plot below shows the response values of an experiment, organized by different treatment groups.



Ⓣ or Ⓕ According to the plot, an ANOVA F-test for this dataset will most likely result in a failure to reject H_0 because most groups are shaped very similarly.

2.6. A researcher performs ANOVA to analyze a dataset, and they mistakenly used the entire sample size $\boldsymbol{n}$ instead of $\boldsymbol{n_i}$ when calculating group variances $\boldsymbol{s_i^2}$. Assuming the formula below was used for calculating the $\mathbf{MS_E}$ value, which of the following statements is **TRUE**?

$$\mathbf{MS_E} = \frac{\sum_{i=1}^{k}(\boldsymbol{n_i} - \mathbf{1})\boldsymbol{s_i^2}}{\boldsymbol{n} - \boldsymbol{k}}$$

Ⓐ The $\mathbf{MS_E}$ value is overestimated, so it increases the chance of rejecting $\boldsymbol{H_0}$.

Ⓑ The $\mathbf{MS_E}$ value is overestimated, so it decreases the chance of rejecting $\boldsymbol{H_0}$.

Ⓒ The $\mathbf{MS_E}$ value is overestimated, but the ANOVA results remain the same.

Ⓓ The $\mathbf{MS_E}$ value is underestimated, so it increases the chance of rejecting $\boldsymbol{H_0}$.

Ⓔ The $\mathbf{MS_E}$ value is underestimated, so it decreases the chance of rejecting $\boldsymbol{H_0}$.

Ⓕ The $\mathbf{MS_E}$ value is underestimated, but the ANOVA results remain the same.

Chapter 13. Simple Linear Regression

Response variable Y: a *random* variable whose changes are being studied.
Explanatory variable X: a *fixed* variable which potentially explains the changes in the response variable.

Assumptions

1. SRS with each pair of observations independent of other pairs
2. The relationship between X and Y is linear in the population ➔ check before/after the analysis
3. The errors have an iid normal distribution (i.e., equal variance of errors) ➔ check after the analysis

Before analysis: use a scatterplot and sample correlation r to check the form, direction, strength of the relationship

- Form: linear ($r \neq 0$) or nonlinear/no relationship ($r = 0$)
- Direction: positive ($r > 0$) or negative ($r < 0$)
- Strength: strong ($|r| \geq 0.8$), moderate ($0.5 \leq |r| < 0.8$), weak ($|r| < 0.5$)

Regression model: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $\epsilon_i \sim_{iid} N(0, \sigma^2)$ in the population

- Estimation of unknown parameters by minimizing sum of errors: $\hat{\beta}_0, \hat{\beta}_1 = \arg\min \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$
- $\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$ ➔ the value of (y context) is expected to go (up/down) by $|\hat{\beta}_1|$ as (x context) goes up by one.
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ ➔ the value of $\hat{y}$ when $x = 0$ (meaningless in most cases)
- Estimated regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * x_{context}$
- Residuals: $e_i = y_i - \hat{y}_i$

Check assumptions using diagnostic plots

- Normality of errors: Normal prob plot, histogram
- Constant variance of errors & linearity & outlier detection: Scatterplot, residual plot

Hypothesis testing

- F-test ➔ H_0: there is NO linear association btw X and Y vs H_a: there is a linear association btw X and Y

Source	df	SS	MS	F
Regression	$df_r = 1$	$\text{SSR} = \Sigma(\hat{y}_i - \bar{y})^2$	$MSR = \frac{SSR}{1}$	$\frac{MSR}{MSE}$
Error	$df_e = n - 2$	$\text{SSE} = \Sigma(y_i - \hat{y}_i)^2$	$MSE = \frac{SSE}{n-2}$	
Total	$df_t = n - 1$	$\text{SST} = \Sigma(y_i - \bar{y})^2$		

  - p-value = $P\left(F_{df_r, df_e} > F_{ts}\right) = pf(F_{ts}, df_r, df_e, lower.tail = FALSE)$
  - Estimated variance of errors: $\hat{\sigma} = \sqrt{MSE}$
- t-test (two-sided case, you may expand it to upper- or lower-tail hypothesis)
  - $H_0: \beta_1 = \beta_{10}$ vs $H_a: \beta_1 \neq \beta_{10}$ ➔ $t_{ts} = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MSE}{S_{xx}}}} \sim t(df = n-2)$ & $CI: \hat{\beta}_1 \pm t_{\alpha/2, n-2}\sqrt{MSE/S_{XX}}$
- Relationship between t and F: $F_{ts} = t_{ts}^2$ when t-test is two-sided with $\beta_{10} = 0$ (i.e., the same p-values)

Other statistics

- Sample correlation $r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$ or $r = sign(\hat{\beta}_1) * \sqrt{R^2}$ ➔ how strong a linear association is
- Coefficient of determination $R^2 = \frac{SSR}{SST} = r^2$ ➔ % of the total variation in Y explained by the linear relationship

Prediction

- Confidence Interval (mean response/observed value) ➔ narrower than PI

$$\hat{\beta}_0 + \hat{\beta}_1 x_i \pm t_{\alpha/2, n-2}\sqrt{MSE\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{XX}}\right)}$$

- Prediction Interval (single response/new observation) ➔ wider than CI

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2, n-2}\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}\right)}$$

- Interpolation 😊 (within the data region) vs extrapolation ☹ (outside of the data region)

Fall 2023

1.2. A least squares regression is conducted, and the estimated slope of the mean response, denoted as $\widehat{\boldsymbol{\beta}}_{\mathbf{1}}$, is found to have a value close to zero in magnitude.

Ⓣ or Ⓕ It follows that the sample Pearson correlation must also be close to zero in magnitude.

1.5. A regression analysis between weight **(y kg)** and height **(x cm)** resulted in the following least-squares regression line: $\hat{y} = -5 + 0.4x$.

Ⓣ or Ⓕ In this context, the estimated value of the slope ($b_1 = 0.4$) indicates that if the height is increased by **1 cm**, the weight will exactly increase by **0.4 kg**.

Fall 2024

1.7. The **98%** lower confidence bound on the true slope of a simple linear regression line, $\boldsymbol{\beta}_{\mathbf{1}}$, gives the value of **-0.43**. Then

Ⓣ or Ⓕ there is **0.98** probability that $\boldsymbol{\beta}_{\mathbf{1}}$ is greater than **-0.43**.

1.8. In simple linear regression,

Ⓣ or Ⓕ all influential points must be outliers.

1.9. In simple linear regression,

Ⓣ or Ⓕ prediction intervals are wider than confidence intervals for the mean response at the same value of the predictor variable.

2.4. An ANOVA F-test was performed on a dataset with two treatment levels ($k = 2$), resulting in a test statistic of $f_{ts} = 3.28$. If the same dataset is used for a hypothesis test on the difference of means of the two levels, then $t_{ts}^2 = 3.28$ only if

Ⓐ The null value, Δ_0 is 0.

Ⓑ The hypothesis test is two-tailed t-test.

Ⓒ The observations within and across the two levels are assumed to be independent.

Ⓓ The population variances are assumed to be equal, and the pooled variance estimate is used to construct the test statistic.

Ⓔ All of the above must hold simultaneously.

Ⓕ We cannot be certain because the $\mathbf{F_{TS}}$ and $\mathbf{T_{TS}}$ are test statistics for two different hypothesis tests, each with distinct assumptions and interpretations.

1.1. Suppose X and Y are two random variables with a large covariance

$$\mathbf{COV}(X, Y) = \mathbf{100,000},$$

and the individual standard deviation σ_X and σ_Y are unknown but finite.

Ⓣ or Ⓕ From this information, we can conclude that X and Y are strongly correlated.

1.5. Let $\widehat{\beta}_1$ and $\widehat{\beta}_0$ be the slope and intercept of the regression line computed from a dataset $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$. Suppose a function l is defined as

$$l(a, b) = \sum_{i=1}^{n} (y_i - (a + bx_i))^2.$$

Ⓣ or Ⓕ Then l attains the smallest value possible by plugging in $\widehat{\beta}_0$ for a and $\widehat{\beta}_1$ for b.

2.4. In a simple linear regression, the predictor variable is recorded in pounds (lbs). What restriction, if any, does this put on the units of the response variable?

Ⓐ It must be measured in some other unit of mass or weight.

Ⓑ It cannot be expressed in pounds or any other weight units.

Ⓒ It may be measured in whatever units are appropriate for the outcome.

Ⓓ It must be measured in pounds exactly, matching the predictor units.

Ⓔ Each of the statements given above is simultaneously valid and applicable.