STA 200 Statistics II

Midterm Exam #3

Summer 2025

Please circle the correct answer. If you circle more than one answer, only the first one will be counted. Due to rounding error in some questions, please always choose the one that is closest to the answer.

Problem 1.

What is the primary purpose of randomization in experimental design?

A) To eliminate all variability in the data
B) To ensure that each experimental unit has an equal chance of receiving any treatment
C) To guarantee that the experiment will have significant results
D) To reduce the number of replicates needed

Answer: B) To ensure that each experimental unit has an equal chance of receiving any treatment

Problem 2.

In a Completely Randomized Design (CRD), how are experimental units assigned to treatments?

A) Based on pre-existing groupings
B) Randomly, without any blocking
C) By matching similar units together
D) Sequentially in the order they are observed

Answer: B) Randomly without any blocking

Problem 3.

Which of the following is an advantage of a Randomized Block Design (RBD) over a CRD?
A) It requires fewer experimental units
B) It eliminates all variability in the response
C) It controls for known sources of variability by grouping similar units
D) It does not require randomization

Answer: C) It controls for known sources of variability by grouping similar units

Problem 4.

When should a Randomized Block Design (RBD) be used instead of a CRD?

A) When there is no known source of variability among experimental units
B) When the experimental units can be grouped based on a known nuisance variable (e.g., age, location)
C) When the sample size is very small
D) When the treatments are not randomly assigned

Answer: B) When the experimental units can be grouped based on a known nuisance variable

Problem 5.

In an RBD with replicates, what does a "block" represent?

A) A single experimental unit
B) A group of homogeneous experimental units subjected to all treatments
C) A type of treatment
D) A random error term

Answer: B) A group of homogeneous experimental units subjected to all treatments

Problem 6.

Which experimental design is most appropriate if there is no significant source of extraneous variation among experimental units?

A) Randomized Block Design (RBD)
B) Completely Randomized Design (CRD)
C) Latin Square Design
D) Factorial Design

Answer: B) Completely Randomized Design (CRD)

7. What is the key assumption of a CRD?
A) The experimental units are heterogeneous
B) The treatments are assigned in a specific order
C) The experimental units are homogeneous, or variability is accounted for by randomization
D) The blocks must be of equal size

Answer: C) The experimental units are homogeneous, or variability is accounted for by randomization

Problem 8.

In an RBD without replicates, how many times is each treatment applied per block?

A) Once
B) Twice
C) Depends on the number of blocks
D) Equal to the number of experimental units

Answer: A) Once

Problem 9.

Which of the following data table layouts represents an RBD with 3 blocks and 4 treatments (without replicates)?

Block	Treatment 1	Treatment 2	Treatment 3	Treatment 4
1	Y_{11}	Y_{12}	Y_{13}	Y_{14}
2	Y_{21}	Y_{22}	Y_{23}	Y_{24}
3	Y_{31}	Y_{32}	Y_{33}	Y_{34}

A) CRD layout
B) RBD layout without replicates
C) RBD layout with replicates
D) Factorial design layout

Answer: B) RBD layout without replicates

10. What is the main disadvantage of a CRD when experimental units are heterogeneous?

A) Increased precision in treatment comparisons
B) High variability within treatment groups due to uncontrolled factors
C) Difficulty in randomization
D) Requires larger blocks

Answer: B) High variability within treatment groups due to uncontrolled factors

Problem 11.

In an RBD, what is the purpose of the blocking variable?

A) To introduce additional variability
B) To group similar experimental units to reduce within-block variability
C) To replace randomization
D) To increase the number of treatments

Answer: B) To group similar experimental units to reduce within-block variability

Problem 12.

Which of the following is a key difference between CRD and RBD?

A) CRD always has more replicates than RBD
B) RBD controls for a known source of variability, while CRD does not
C) CRD requires blocking, while RBD does not
D) RBD cannot be used with quantitative data

Answer: B) RBD controls for a known source of variability, while CRD does not

Problem 13.

What is the correct null hypothesis for testing treatment effects in an RBD?

A) All block means are equal
B) All treatment means are equal
C) The interaction between blocks and treatments is significant
D) The variances across blocks are unequal

Answer: B) All treatment means are equal

Problem 14

Suppose we want to know whether or not three different exam prep programs lead to different mean scores on a certain exam. To test this, we recruit 30 students to participate in a study and split them into three groups.

```
group1 <- c(85, 86, 88, 75, 78, 94, 98, 79, 71, 80)
group2 <- c(91, 92, 93, 85, 87, 84, 82, 88, 95, 96)
group3 <- c(79, 78, 88, 94, 92, 85, 83, 85, 82, 81)
```

Perform one-way ANOVA analysis using R. Which of the following conclusions is correct based on your R output ANOVA table?

A) Reject the null hypothesis that three prep programs are the same at significance level 0.05.
B) Reject the null hypothesis that three prep programs are the same at significance level 0.10.
C) Conclude the null hypothesis that three prep programs are the same at level 0.10
D) Conclude the null hypothesis that three prep programs are the same at level 0.15.

Answer: C

```
group1 <- c(85, 86, 88, 75, 78, 94, 98, 79, 71, 80)
group2 <- c(91, 92, 93, 85, 87, 84, 82, 88, 95, 96)
group3 <- c(79, 78, 88, 94, 92, 85, 83, 85, 82, 81)
group <-c(rep("group1", 10), rep("group2", 10), rep("group3",10))
grades <- data.frame(
       grade = c(group1, group2, group3),
       group = group
    )
summary(aov(grade ~ group, data = grades))

           Df Sum Sq Mean Sq F value Pr(>F)
group       2  192.2   96.10   2.358  0.114
Residuals  27 1100.6   40.76
```

Problem 15

The following is a partial ANOVA table based on a CRD experiment.

Source	SS	df	MS	F	P
Treatment	192.2	2		?	
Error	1100.6	27			
Total	1292.8	29			

Based on the given sum of squares in the above partial table, what is the F value?

A 2.156
B 0.175
C 2.358
D 0.914

Answer C

Problem 16

An ANOVA test is conducted to compare mean weights of chicks fed different diets. The Tukey HSD test shows a p-value of 0.027 between "soybean" and "linseed." What does this mean?

a) No significant difference between soybean and linseed diets.
b) Soybean diet leads to significantly higher chick weights than linseed.
c) The ANOVA assumptions are violated.
d) The p-value must be below 0.001 to be significant.

Correct Answer: b)

Problem 17

An ANOVA test compares the effect of two drugs on extra sleep hours. The p-value is 0.003. Assuming significance level 0.05, what is the correct interpretation?

a) Both drugs have the same effect.
b) At least one drug significantly affects sleep.
c) The data is not normally distributed.
d) The sample size is insufficient.

Correct Answer: b)

Problem 18

The data at the following URL were collected from an experiment assessing plant growth rate (measured by weight) using different soil treatments (control, treatment 1, and treatment 2). The data were organized in a long-format CSV file: https://pengdsci.github.io/STA200/dataset/oneWayPlantGrowth.csv

Using appropriate R commands, read the data file from the URL, perform an ANOVA test, and rank the treatments based on the output. [*Hint*: *you can use either* `TukeyHSD()` or `aggregate()` *for ranking treatment*]

A. ANOVA test shows that the treatment effect is significant (p = 0.0120). The treatment ranking is: trt2 > trt1 > ctr.

B. ANOVA test shows that the treatment effect is significant (p = 0.0159). The treatment ranking is: `trt2 > trt1 > ctr.`

C. ANOVA test shows that the treatment effect is significant (p = 0.0159). The treatment ranking is: `trt2 > ctr > trt1.`

D. ANOVA test shows no treatment effect (p=0.1980). The treatment ranking is: `trt2 > ctr > trt1.`

Answer: C

```
> plant =
read.csv("https://pengdsci.github.io/STA200/dataset/oneWayPlantGrowth.csv")
> names(plant)
[1] "weight" "group"
> plant.aov <- aov(weight ~group, data = plant)
> summary(plat.aov)
            Df Sum Sq Mean Sq F value Pr(>F)
group        2  3.766  1.8832   4.846 0.0159 *
Residuals   27 10.492  0.3886
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> aggregate(weight ~ group, data = plant, FUN = mean)
  group weight
1  ctrl  5.032
2  trt1  4.661
3  trt2  5.526
```

Problem 19

A pharmaceutical company conducts an experiment to test the effect of a new cholesterol medication. The company selects 15 subjects randomly from a larger population. Each subject is randomly assigned to one of three treatment groups. Within each 7treatment group, subjects receive a different dose of the new medication. In Group 1, subjects receive 0 mg/day; in Group 2, 50 mg/day; and in Group 3, 100 mg/day.

Dosage		
Group 1, 0 mg	Group 2, 50 mg	Group 3, 100 mg
210	210	180
240	240	210
270	240	210
270	270	210
300	270	240

Using R to perform an ANOVA analysis to assess the treatment effect. Based on your output ANOVA table, which of the following statements is correct?

A. The F-value (4.16) is smaller than the cutoff value (5), so we reject the null hypothesis that treatment effects are different at different dosage levels.
B. The P-value (0.04) is smaller than the significance level (0.05), so we reject the null hypothesis that treatment effects are different at different dosage levels.
C. The P-value (0.04) is smaller than the significance level (0.05), so we reject the null hypothesis that treatment effects are the same at different dosage levels.
D. The F-value (4.16) is smaller than the cutoff value (5), so we fail to reject the null hypothesis that treatment effects are different at different dosage levels.

Answer B

```
grp1 <- c(210, 240, 270, 270, 300)
grp2 <- c(210, 240, 240, 270, 270)
grp3<- c(180, 210, 210, 210, 240)
group <-c(rep("grp1",5), rep("grp2",5), rep("grp3",5))
cholestrol <- data.frame(
    treat = c(grp1, grp2, grp3),
    group = group
)
aov.model <- aov(treat ~ group, data = cholestrol)
summary(aov.model)


            Df Sum Sq Mean Sq F value Pr(>F)
group        2   6240    3120    4.16 0.0424 *
Residuals   12   9000     750
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Problem 20

Consider an experiment examining the effect of Vitamin C dosage (dose) and delivery method (supp) on tooth growth in guinea pigs. The data is organized in long format and saved as a CSV file. The URL of the data file is provided below.

https://pengdsci.github.io/STA200/dataset/oneWayToothGrowth.csv

Perform a two-way ANOVA analysis. Based on your R output, which of the following statements is correct? [Hint: The variable **dose** in the data set is given in the numerical form. You should use `factor(dose)` in the model formula to convert the numerical variable to a factor variable. The formula should be something like `len ~ supp * factor(dose)`]

A). `supp`, `factor(dose)`, and their **interaction** are significant at the level of 0.05.
B). Both `supp` and `factor(dose)` are significant, but not their **interaction** at the level of 0.05.
C). Only `factor(dose)` is significant at the level of 0.05.
D). Only `supp` is significant at the level of 0.05.

Answer: A

```
> tooth.growth = 
read.csv("https://pengdsci.github.io/STA200/dataset/oneWayToothGrowth.csv")
> names(plant)
[1] "weight" "group"
> tooth.aov <- aov(len ~ supp * factor(dose), data = tooth.growth)
> summary(tooth.aov)
                  Df Sum Sq Mean Sq F value   Pr(>F)    
supp               1  205.4   205.4  15.572 0.000231 ***
factor(dose)       2 2426.4  1213.2  92.000  < 2e-16 ***
supp:factor(dose)  2  108.3    54.2   4.107 0.021860 *  
Residuals         54  712.1    13.2                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Problem 21

The following **reaction time dataset** was collected from a **replicated randomized block design (RBD) experiment** to perform a comprehensive **ANOVA analysis and post hoc tests.** The objective of this replicated RBD was to assess the effects of **beer** and **caffeine** on the **reaction time** of the experimental units.

	no caffeine	caffeine
no beer	2.24, 1.62, 1.48, 1.70, 1.06, 1.39,	0.62, 1.72, 1.75, 1.84, 1.30, 1.52,
beer	1.71, 2.19, 2.27, 2.35, 2.47, 2.07,	2.05, 1.51, 1.65, 2.68, 2.06, 1.80,

Using R, create a long-format data frame with three columns: time, beer (yes/no), and caffeine (yes/no). Perform a two-way ANOVA. At a significance level of 0.05, which of the following statements is correct?

A). Both caffeine and beer affect the reaction time significantly, but not the interaction effect.
B) Only caffeine affects the reaction time significantly.
C) Only beer affects the reaction time significantly.
D) Caffeine, beer, and their interaction affect the reaction time significantly.

Answer: C

```
## first, define 4 column vectors
Caff.N = c(2.24, 1.62, 1.48, 1.70, 1.06, 1.39, 1.71, 2.19, 2.27, 2.35, 2.47,
2.07)
Caff.Y = c(0.62, 1.72, 1.75, 1.84, 1.30, 1.52, 2.05, 1.51, 1.65, 2.68, 2.06,
1.80)
caff = c(rep("caff.N",12), rep("caff.Y",12))
beer = rep(c(rep("beer.N", 6), rep("beer.Y",6)),2)
reactionData <- data.frame(
            time = c(Caff.N, Caff.Y),
            caffeine = caff,
            beer = beer
)
summary(aov(time ~ caffeine * beer, data = reactionData))
```

Problem 22
A botanist wants to know whether different levels of sunlight exposure and watering frequency affect plant growth. She plants 40 seeds and lets them grow for one month under different conditions for sunlight exposure and watering frequency. The following table shows the partial results of the two-way ANOVA:

Source of Variation	SS	df	MS	F	p-value
Watering Frequency	0.0003	1	0.0003	F_A	
Sunlight Exposure	18.7648	3	6.2549	F_B	
Interaction	1.0108	3	0.3369	F_C	
Within	8.684	32			
Total	28.4597	39			

A). $F_A = 0.00024, F_B = 5.0382, F_C = 0.2713$
B). $F_A = 0.00000145, F_B = 2.1608, F_C = 0.1163$
C). $F_A = 0.0011, F_B = 23.0489, F_C = 1.2416$
D). $F_A = 0.0000015, F_B = 0.27138, F_C = 0.01177$

Answer: C

Problem 23
An ANOVA test is run to compare yield across different fertilizer treatments. The TukeyHSD results show a p-value of 0.12 between treatments N and P. What should you conclude?

a) N and P have significantly different yields.
b) N and P do not have significantly different yields.
c) The ANOVA assumptions are violated.
d) The F-statistic is too low.

Answer: b)

Problem 24
In a two-way ANOVA, a significant interaction effect implies that:

A) The effect of one factor depends on the level of the other factor
B) Both factors have independent effects on the dependent variable
C) Neither factor has a significant main effect
D) The assumptions of homogeneity of variance are violated

Answer: A) The effect of one factor depends on the level of the other factor

Problem 25
Which of the following is NOT an assumption of two-way ANOVA?
A) Homogeneity of variance (equal variances across groups)
B) Normality of residuals
C) Independence of observations
D) All group sizes must be equal

Answer: D)

Problem 26
What is the null hypothesis in a one-way ANOVA?

A) All group means are equal.
B) At least one group mean is different.
C) Group variances are equal.
D) The data follows a normal distribution.

Answer: A

Problem 27
When analyzing the InsectSprays dataset with ANOVA, the null hypothesis is that all sprays have the same mean insect count. If the p-value is 0.0001, what should you do?

a) Accept the null hypothesis.
b) Reject the null hypothesis and conclude at least one spray differs.
c) Perform a t-test instead.
d) Increase the sample size.

Answer: b)

Problem 28

When should you use a one-way ANOVA instead of a two-way ANOVA?
A. When you have more than two independent variables
B. When you want to analyze the interaction effect between two factors
C. When you have one categorical independent variable and one continuous dependent variable
D. When you are analyzing only two groups

Answer: C

Problem 29
You conducted a one-way ANOVA in R and obtained a significant result. What should you do next?

A. Accept the null hypothesis
B. Perform a post hoc test (e.g., Tukey HSD)
C. Collect more data
D. Run a t-test for each pair of groups without adjustment

Answer: B

Problem 30
ANOVA is used to compare breaks across different wool types (A and B). The p-value is 0.059. At α = 0.05, what should you conclude?

a) Reject the null hypothesis; wool types differ.
b) Fail to reject the null hypothesis; no significant difference.
c) The data is not normally distributed.
d) The variances are unequal.

Answer: b)

Week #4 Exam Summary

The class boundary is: 60,70,80,90,100

cut.data.freq	Freq	midpts	rel.freq	cum.freq	rel.cum.freq
[60,70]	1	65.00	0.06	1	0.06
(70,80]	9	75.00	0.56	10	0.62
(80,90]	5	85.00	0.31	15	0.94
(90,1e+02]	1	95.00	0.06	16	1.00

The five-number summary of this given data set is:

stats	value
Min.	66.70
1st Qu.	74.58
Median	79.15
3rd Qu.	83.30
Max.	93.30