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 you.

Problem 1.

How do you access the column "age" from a data frame with multiple columns named "mydata"?

a) mydata (age)

b) mydata $age

c) mydata ["age"]

d) mydata ["age",]

Answer: b

Problem 2.

What function is used to select the first few rows of a data frame " mydata "?

a) head[mydata]

b) header[mydata]

c) head(mydata)

d) row(mydata)

Answer: C

Problem 3.

How do you select the 3rd column of a data frame "mydata "?

a) mydata [, 3]

b) mydata [3,]

c) col(mydata [3])

d)mydata[col = 3]

Answer: A

Problem 4:

How do you select rows from the data frame **mydata** where "gender" is "Female" and "age" is above 30?

a)

```
id <- which(mydata $gender == "Female" & mydata $age > 30)
mydata[id, ]
```

b)

```
id <- which(mydata $gender = "Female" & mydata $age > 30)
subset(data, row = id)
```

c)

```
id <- which(gender == "Female" & age > 30)
mydata[id, ]
```

d)

```
id <- which(mydata $gender == Female & mydata $age > 30)
mydata[id, ]
```

Answer: A

Problem 5:

What does the following code return?

```
age.id <- which(mydata$age > 25)
mydata[age.id, "name"]
```

a) Names of all rows where age > 25
b) All columns for rows where age > 25
c) The entire data frame filtered by age
d) An error

Answer: a) Names of all rows where age > 25

Problem 6.

What is the primary purpose of a one-sample t-test?

A) To compare the means of two independent groups
B) To test whether the mean of a single sample differs from a known population mean
C) To compare the variances of two samples
D) To test the median of a sample

Answer: B

Problem 7

Which of the following is an assumption of the one-sample t-test?

A) The data must be normally distributed
B) The sample size must be greater than 100
C) The population variance must be known
D) The data must be ordinal

Answer: A

Problem 8

Which R function can perform a sign test?
A) t.test()
B) wilcox.test()
C) binom.test()
D) prop.test()

Answer: C

Problem 9.

The sign test is a non-parametric alternative to the one-sample t-test. What does it test?
A) The mean of the sample
B) The median of the sample
C) The variance of the sample
D) The proportion of successes in the sample

Answer: B) The median of the sample

Problem 11.

Which of the following is a non-parametric alternative to the one-sample t-test?
A) One-sample z-test
B) One-sample sign test
C) Paired t-test
D) Chi-square test

Answer: B) One-sample sign test

Problem 12

In a one-sample sign test, what is the null hypothesis typically formulated as?
A) The sample median is equal to a specified value
B) The sample mean is greater than the population mean
C) The sample variance differs from the population variance
D) The data follows a normal distribution

Answer: A) The sample median is equal to a specified value

Problem 13.

Suppose you run a sign test in R and get the output:

```
Exact binomial test
number of successes = 8, number of trials = 10, p-value = 0.1094
What is the correct interpretation at α = 0.05?
```

A) Reject H_0, median is different
B) Fail to reject H_0, insufficient evidence
C) The test is invalid
D) The sample size is too small

Answer: B) Fail to reject H_0, insufficient evidence

Problem 14.

The sign test is based on:
A) The sum of ranks
B) The number of positive and negative differences from the hypothesized median
C) The mean difference
D) The standard deviation

Answer: B) The number of positive and negative differences from the hypothesized median

Problem 15.
How is a one-sample t-test formulated as a regression model with only an intercept?
A) y ~ 0 (no intercept)
B) y ~ 1 (intercept-only model)
C) y ~ x (simple linear regression)
D) y ~ x + 0 (no intercept with predictor)

Answer: B
Explanation: The formula y ~ 1 fits a regression model with only an intercept, which estimates the mean of y. The one-sample t-test compares this mean to a hypothesized value.

Problem 16.

To perform a one-sample t-test for H_0: μ=10 using regression in R, which of the R code reflects the correct model formula that is equivalent to the original test?

A) `lm(I(y - 10) ~ 1).`
B) `lm(y - 10 ~ 1).`
C) `lm(y ~ 10).`
D) `lm(y~ 1).`

Answer: A
Explanation: To test H_0 : μ=10, fit `lm(I(y - 10) ~ 1).` The t-test for the intercept now evaluates if the shifted mean differs from zero, equivalent to the original test.

Problem 17.

A clinical study measures the systolic blood pressure (mmHg) of 20 patients after administering a new drug.

```
113, 103, 125, 121, 122, 120, 123, 116, 144, 100,
127, 122,  115, 115, 135, 126, 129, 124, 115, 108
```

The expected population mean under the null hypothesis is Ho: μ0 = 120 mmHg. We test whether the observed blood pressure differs from 120 mmHg using a regression approach. The following is the output.

```
Call:
lm(formula = I(blood_pressure - 120) ~ 1)

Residuals:
   Min     1Q Median     3Q    Max
-20.15  -5.15   1.35   5.10  23.85

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.150      2.293    X     0.949

Residual standard error: 10.26 on 19 degrees of freedom
```

What is the value of **x**?

A). **x** = (0.15 – 0)/ 2.293 = 0.065

B). **x** = (0.15 – 0)/ (2.293/$\sqrt{20}$) = 0.2907

C). **x** = (0.15 – 120)/ 2.293= 91.65

D). **x** = (0.15 –12 0)/ (2.293$\sqrt{20}$) = 20.453

Answer A TS = (0.15 – 0)/ 2.293 = 0.065

Problem 18.

Which of the following is an assumption of the two-sample t-test?
A) The samples must be paired.
B) The populations must have equal variances.
C) The data must be normally distributed in both populations.
D) The sample sizes must be equal.

Answer: C) The data must be normally distributed in both populations.
(The two-sample t-test assumes normality, but equal variances and sample sizes are not strictly required, especially with Welch's correction.)

Problem 19.

What is the sampling distribution of the test statistic in a two-sample t-test under the null hypothesis?
A) Standard normal (Z)
B) Chi-square
C) t-distribution
D) F-distribution

Answer: C) t-distribution

(The test statistic follows a t-distribution with degrees of freedom dependent on sample sizes and variance assumptions.)

Problem 20.

Which R function is used for a two-sample t-test assuming equal variances?

A) `t.test(x, y, paired = FALSE, var.equal = TRUE)`
B) `t.test(x, y, var.equal = FALSE)`
C) `t.test(x, y, paired = TRUE)`
D) `var.test(x, y)`

Answer: A) `t.test(x, y, paired = FALSE, var.equal = TRUE)`
(Setting `var.equal = TRUE` performs a classic two-sample t-test.)

Problem 21.

What is the null hypothesis of the two-tailed Wilcoxon rank-sum test?
A) The two population means are equal.
B) The two population distributions are identical.
C) The two population variances are equal.
D) The two samples are paired.

Answer: B) The two population distributions are equal.

Problem 22.

Suppose we run the following in R:

```
x <- c(10, 12, 15, 20, 25)
y <- c(5, 8, 9, 13, 18, 22)
wilcox.test(x, y, alternative = "two.sided", paired = FALSE)
```

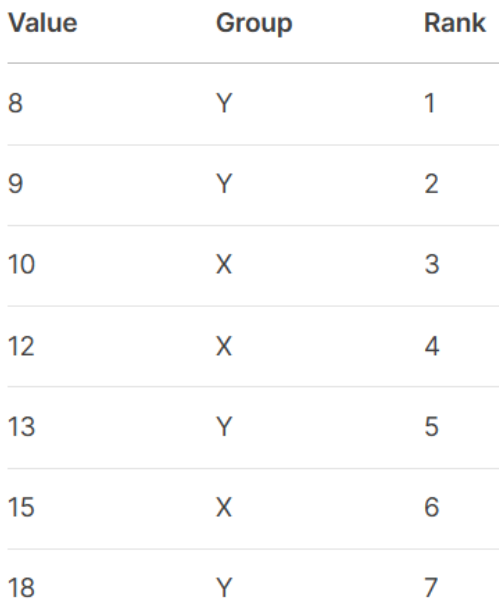
The test compares:

A) Means of x and y.
B) Medians of x and y.
C) Whether values in x and y come from the same distribution.
D) Variances of x and y.

Answer: C) The test checks whether two populations are identical, not just medians.)

Problem 23.

Consider the following data table.



Value	Group	Rank
8	Y	1
9	Y	2
10	X	3
12	X	4
13	Y	5
15	X	6
18	Y	7

What is the test statistic of the Wilcoxon Rank Sum test?

A). 13
B). 15
C). 14
D). 2

Answer A) The smaller rank sum of the two groups.

Problem 24.

For the data [10, 15, 15, 15, 20], the ranks are:

A) [1, 2, 3, 4, 5]
B) [1, 3, 3, 3, 5]
C) [1, 4, 4, 4, 5]
D) [1, 3.5, 3.5, 5, 6]

Answer: B) [1, 3, 3, 3, 5]

Explanation: The three 15s occupy positions 2, 3, 4 ➔ average rank = (2+3+4)/3 = 3.

Problem 25.

Data analysis: Consider the Pima Indian Diabetes data set, which is at

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

Perform a two-sample t-test using a regression approach to see whether the mean glucose levels in the diabetes and diabetes-free groups. Based on the output of the regression model, which of the following is correct? [*Hint: x variable must be a factor. R function* `factor()` *to a variable to a factor.]*

A). $\hat{y}_{pos} - \hat{y}_{neg} = 111.431, TS = 1.636,$ df = 390, p-value ≈ 0.2659
B). $\hat{y}_{pos} - \hat{y}_{neg} = 2.84, TS = 11.89,$ df = 390, p-value ≈ 0
C). $\hat{y}_{pos} - \hat{y}_{neg} = 111.431, TS = 1.636,$ df = 390, p-value ≈ 0.2641
D). $\hat{y}_{pos} - \hat{y}_{neg} = 33.761, TS = 11.89,$ df = 390, p-value ≈ 0

Answer: D.

```
diabetes <-
read.csv("https://pengdsci.github.io/STA200/dataset/PimaIndiaDiabetes.csv")
model <- lm(glucose ~ factor(diabetes), data = diabetes)
summary(model)
```

Problem 26

Data analysis: Consider the Pima Indian Diabetes data set, which is at

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

Perform a non-parametric two-sample Wilcoxon Rank Sum test to assess the difference between glucose levels in diabetes and diabetes-free groups. Based on the output of the Wilcoxon test, which of the following is correct? [*Hint: need to use normal approximation by specifying* `correction = TRUE`]

A). W = 390 (sample size -2) and p-value = 0.002
B). W = 2745 and p-value = 2.2e-16
C). W = 130 (number of diabetes) and p = 0
D. W = 262 (number of non diabetes) and p ≈ 0.

Answer: B.

Problem 27

A study compares the weights of 15 individuals before and after a diet program. The differences are normally distributed. What test should be used?

A) Wilcoxon signed-rank test
B) Paired t-test
C) Mann-Whitney U test
D) Kruskal-Wallis test

Answer: B) Paired t-test

Problem 28.

In R, if before = c(70, 75, 80, 85, 90) and after = c(68, 72, 78, 82, 88), what is the correct way to run a paired t-test?

A) t.test(before, after, paired = FALSE)
B) t.test(before, after, paired = TRUE)
C) wilcox.test(before, after, paired = TRUE)
D) t.test(after - before)

Answer: B) t.test(before, after, paired = TRUE)

Problem 29.

For the Wilcoxon signed-rank test, what is the null hypothesis?
A) The means of the two groups are equal
B) The median difference between pairs is zero
C) The distributions are identical for the before and after groups

Answer: C (The Wilcoxon test assumes symmetry and tests the median difference)

Problem 30.

The anorexia dataset is a classic dataset used in statistics to evaluate the effectiveness of different treatments for **anorexia nervosa** (an eating disorder). It contains weight measurements of young female patients **before** and **after** receiving one of three treatments ("Cont" – control, "CBT" – cognitive and behavioral treatment, and "FT" =family treatment). A copy of this dataset is at https://pengdsci.github.io/STA200/dataset/anorexia.csv

The anorexia data frame has 72 rows and 3 columns: Treat, Prewt, and Postwt. The objective is to assess the treatment effect of **Family Therapy (FT)** on weight gain in anorexia patients. Specifically, we test whether **FT significantly increases** the average post-treatment weight compared to the pre-treatment weight using the Wilcoxon signed-rank test. Which of the following code snippets is incorrect?

A).

```
anorexia <-
read.csv("https://pengdsci.github.io/STA200/dataset/anorexia.csv")
anorexia$diff <- anorexia$Postwt - anorexia$Prewt
anorexia.ft <- anorexia[which(anorexia$Treat == "FT"), ]
wilcox.test(anorexia.ft$diff, alternative = "greater", paired = FALSE,
correct = FALSE)
```

B).

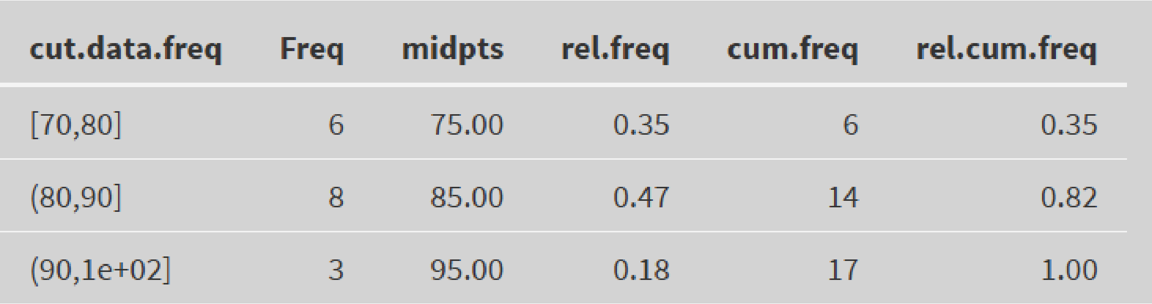
```
anorexia <-
read.csv("https://pengdsci.github.io/STA200/dataset/anorexia.csv")
anorexia.ft <- anorexia[which(anorexia$Treat == "FT"), ]
wilcox.test(anorexia.ft$Postwt, anorexia.ft$Prewt, alternative =
"greater", paired = TRUE, correct = FALSE)
```

C).

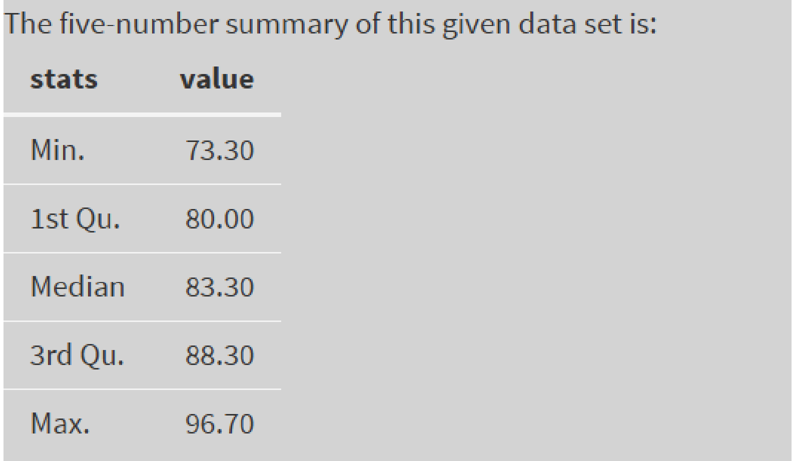
```
anorexia <-
read.csv("https://pengdsci.github.io/STA200/dataset/anorexia.csv")
anorexia.ft <- anorexia[which(anorexia$Treat == "FT"), ]
wilcox.test(anorexia.ft$Prewt, anorexia.ft$Postwt,  alternative =
"greater", paired = TRUE, correct = FALSE)
```

Answer B

Example #3 Summary

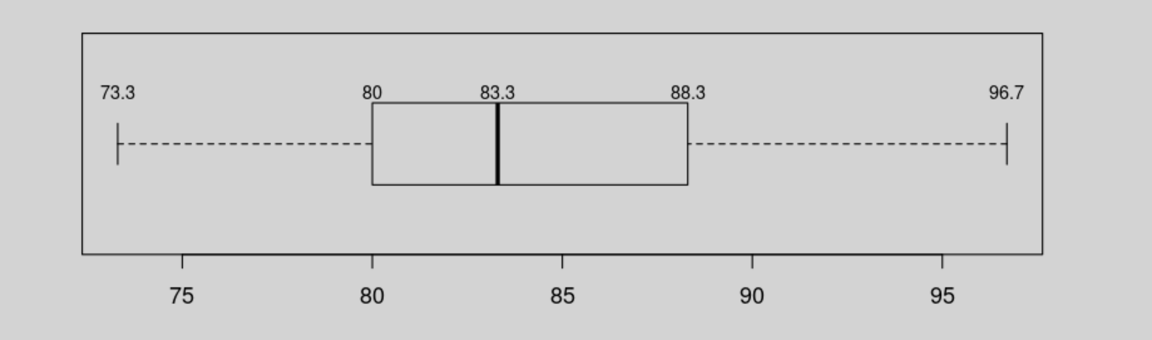


cut.data.freq	Freq	midpts	rel.freq	cum.freq	rel.cum.freq
[70,80]	6	75.00	0.35	6	0.35
(80,90]	8	85.00	0.47	14	0.82
(90,1e+02]	3	95.00	0.18	17	1.00



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

stats	value
Min.	73.30
1st Qu.	80.00
Median	83.30
3rd Qu.	88.30
Max.	96.70


73.3
80
83.3
88.3
96.7
75
80
85
90
95