STA 200 Statistics II

Midterm Exam #2

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.

The Chi-square distribution is primarily used for:
a) Modeling continuous symmetric data
b) Testing independence in contingency tables and goodness-of-fit
c) Analyzing linear regression coefficients
d) Computing probabilities for normal distributions

Answer: b) Testing independence in contingency tables and goodness-of-fit

Problem 2.

In R, which function calculates the right-tail probability of a Chi-square distribution with df and quantile x?
a) pchisq(x, df, lower.tail = TRUE)
b) qchisq(x, df, lower.tail = TRUE)
c) pchisq(x, df, lower.tail = FALSE)
d) qchisq(x, df, lower.tail = FALSE)

Answer: c) pchisq()

Problem 3.

The **qchisq()** function in R is used to:

a) Generate random Chi-square variates
b) Compute the probability density at a given quantile
c) Find the quantile (critical value) for a given cumulative probability
d) Estimate the mean of the Chi-square distribution

Answer: c) Find the quantile (critical value) for a given cumulative probability

Problem 4:

If $X \sim \chi^2(df=5)$, what R command gives $P(X \leq 3.2)$?

a) pchisq(3.2, df=5, lower.tail = TRUE)
b) qchisq(3.2, df=5, lower.tail = TRUE)
c) pchisq(3.2, df=5, lower.tail = FALSE)
d) qchisq(3.2, df=5, lower.tail = FALSE)

Answer: a) pchisq(3.2, df=5, lower.tail = TRUE)

Problem 5:

Which of the following is true about the Chi-square distribution?

a) It is symmetric around its mean
b) Its shape depends only on the mean
c) It is always right-skewed
d) It can take negative values

Answer: **c)** It is always right-skewed

Problem 6.

The Chi-square goodness-of-fit test is used to:

a) Compare means of two independent groups
b) Test if observed frequencies match expected frequencies under a given distribution
c) Assess the correlation between two continuous variables
d) Evaluate variance homogeneity

Answer: **b)** Test if observed frequencies match expected frequencies under a given distribution

Problem 7

In R, which function performs a Chi-square goodness-of-fit test?

a) `chisq.test()` with specified `p` argument
b) `t.test()`
c) `anova()`
d) `fisher.test()`

Answer: a) `chisq.test()` with specified `p` argument

Problem 8

In R, how do you perform a goodness-of-fit test for observed counts **`c(20, 30, 50)`** with expected probabilities **`c(0.2, 0.3, 0.5)`**?

a) `chisq.test(c(20, 30, 50), p = c(0.2, 0.3, 0.5))`
b) `t.test(c(20, 30, 50))`
c) `chisq.test(c(20, 30, 50))`
d) `anova(c(20, 30, 50))`

Answer: **a)** `chisq.test(c(20, 30, 50), p = c(0.2, 0.3, 0.5))`

Problem 9.

The null hypothesis in a Chi-square goodness-of-fit test states that:

a) The sample means are equal
b) The observed frequencies differ from expected frequencies
c) The observed frequencies match expected frequencies
d) The variances are unequal

Answer: c) The observed frequencies match expected frequencies

Problem 10.

If the Chi-square test statistic is much larger than the critical value, we:

a) Fail to reject the null hypothesis
b) Accept the null hypothesis
c) Reject the null hypothesis
d) Conclude the data is normally distributed

Answer: c) Reject the null hypothesis

Problem 11.

If chisq.test() gives a warning "Chi-squared approximation may be incorrect," it likely means:

a) The sample size is too large
b) Some expected frequencies are too small
c) The data is normally distributed
d) The test is invalid for categorical data

Answer: b) Some expected frequencies are too small

Problem 12

A cohort study differs from a case-control study because:

A) It starts with exposed and unexposed groups and follows them for outcomes
B) It starts with diseased and non-diseased groups and looks back at exposures
C) It is always a follow up study
D) It cannot calculate relative risk

Answer A

Problem 13.

Which study design is best for studying rare diseases with a long latency period (e.g., taking 20 years to develop)?

A) Cohort study
B) Case-control study
C) Cross-sectional study
D) Randomized controlled trial

Answer: B

Problem 14.

If the odds ratio (OR) is 3.0, this indicates:

A) No association
B) A protective effect
C) A positive association (Increased risk with exposure)
D) A causal relationship

Answer: C

Problem 15.

A study compares disease rates between two groups (such as exposed group vs unexposed group) over time. This is a:
A) Case-control study
B) Cross-sectional study
C) Prospective cohort study
D) Ecological study

Answer: C

Problem 16.

Consider the following 2-by-2 contingency table obtained based on a **prospective cohort study design**.

	Diseased (+)	**Disease-free (-)**
Diagnostic Test (+)	a	b
Diagnostic Test(-)	c	d

The relative risk (RR) is calculated as:

A) (a/(a+b)) / (c/(c+d))
B) (a/(a+c)) / (b/(b+d))
C) (a×d) / (b×c)
D) (a/(a+b)) – (c/(c+d))

Answer A

Problem 17.

Consider the following 2-by-2 contingency table obtained based on a **prospective cohort study design**.

	Diseased (+)	**Disease-free (-)**
Exposed (+)	a	b
Not Exposed(-)	c	d

What is odds ratio of diseased exposed v.s. unexposed groups?

A) (a/(a+b)) / (c/(c+d))
B) (a/(a+c)) / (b/(b+d))
C) (a×d) / (b×c)
D) (a/(a+b)) – (c/(c+d))

Answer C

Problem 18.

What is the null hypothesis in a Chi-square test of independence?

a) The two variables are dependent.
b) The two variables are independent.
c) The means of the two variables are equal.
d) The variances of the two variables are equal.

Answer: b)

Problem 19.

Which of the following assumptions is required for a valid Chi-square test of independence?

a) Normally distributed data
b) Homogeneity of variance
c) Expected cell counts ≥ 5 (or similar guidelines)
d) Continuous variables

Answer: c)

Problem 20.

If a Chi-square test on a 3×4 table yields a p-value of 0.01, what should we conclude?

a) The variables are independent.
b) The variables are dependent.
c) The test assumptions are violated.
d) The sample size is insufficient.

Answer: b)

Problem 21.

What is the degrees of freedom for a Chi-square test on a 2×3 contingency table?

a) 1
b) 2
c) 5
d) 6

Answer: b) (df = (rows-1) × (cols-1) = 1×2 = 2)

Problem 22.

A survey of 200 people compares gender (Male/Female) and movie preference (Action/Comedy/Drama):

	Action	Comedy	Drama
Male	30	25	15
Female	20	45	65

Which R code correctly analyzes this data?

a) `chisq.test(c(30, 25, 15, 20, 45, 65))`

b) `chisq.test(matrix(c(30, 25, 15, 20, 45, 65), nrow = 2, byrow= TRUE))`

c) `t.test(movie ~ gender)`

d) `fisher.test(table(gender, movie))`

Answer: b)

Problem 23.

Given the following R code:

```
obs <- c(40, 35, 25)
chisq.test(obs, p=c(0.4, 0.4, 0.2))
```

Which null hypothesis is being tested?

A. The proportions are 40%, 35%, 25%
B. Observed data follow the distribution (0.4, 0.4, 0.2)
C. The categories are independent
D. The proportions are equal across categories

Correct Answer: B

Problem 24.

Consider the chi-square test of independence between shopping time and shopping preference. The following is the observed contingency table.

		Time customers shop (hours)			Total:
		<10	10<Time<20	>20	
Where customers shop	Online	53	50	64	167
	In-Store	44	65	224	333
Total:		97	115	288	500

Under null hypothesis **Ho**: **shopping time and shopping preference are independent**

		Time customers shop (hours)			Total:
		<10	10<Time<20	>20	
Where customers shop	Online				167
	In-Store		@		333
Total:		97	115	288	500

What is the expected frequency of the cell labeled with **@** ?

a). 65/333
b). 65/500
c). 65/115
d).115x333/500

Answer: d)

Problem 25.

The following contingency table gives the results of a survey of male and female respondents with regard to their preferred brand of notebook.

	Preferred Brand			
	HP	**Lenovo**	**Dell**	**Total**
No. of Females	110	60	70	**240**
No. of Males	40	50	70	**160**
Total	**150**	**110**	**140**	**400**

Perform a chi-square test of **Ho: gender and brand preference are independent** using an R function.

The resulting p-value is

a). 0.000379976
b). 0.005506867
c). 0.000105796

Answer b)

```
c.table <- matrix(c(110, 60,70,40,50,70), nrow = 2, byrow = TRUE)
chisq.test(c.table)
```

Problem 26

A local college claims that 32% of its graduates are in humanities, 28% in liberal arts, 26% in the biological sciences, and 14% in the physical sciences. A random sample of 385 students finds 148 humanities graduates, 112 liberal arts graduates, 89 biological science graduates, and 36 physical science graduates. Perform a hypothesis test on the college's claimed proportions. That is, the observed frequency table is

Humanities	Liberal arts	Biological Sci	Physical Sci
148	112	89	36

Perform a chi-square test for the claim:

Ho: p_1=0.32, p_2=0.28, p_3=0.26, and p_4=0.14

What is the p-value? *[Hint: using R function to test the null hypothesis*]

a). 0.0151
b). 0.0063
c). 0.0021

Answer b): chisq.test(c(148,112,89,36), p=c(0.32,0.28,0.26,0.14))

Problem 27

The Car Evaluation Database was derived from a simple hierarchical decision model, directly relating to the CAR to the six input attributes: buying, maintenance, doors, persons, luggage boot, safety, and evaluated result (class). Because of known underlying concept structure, this database may be particularly useful for testing constructive induction and structure discovery methods. The database is hosted at

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

Using R to load the data and perform a chi-square test at a significant level of 0.05 to see whether the maintenance and safety are independent. Based on your R output, which iof the following is correct?

a) P-value = 0.07, df = 6
b) P-value = 1, df = 6
c) P-value = 0, df = 3
d) P-value =0.76, df = 4

Answer: B

```
car <- read.csv("https://pengdsci.github.io/STA200/dataset/CarEvalData.csv")
chisq.test(car$maint, car$safety)
```

Problem 28.

The Car Evaluation Database was derived from a simple hierarchical decision model, directly relating to the CAR to the six input attributes: buying, maintenance, doors, persons, luggage boot, safety, and evaluated result (class). Because of known underlying concept structure, this database may be particularly useful for testing constructive induction and structure discovery methods. The database is hosted at

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

Using R to load the data and construct a frequency table of the buying price. Which of the following frequency tables is identical to what you obtained using R function `table()`

a).

```
 high   low   med vhigh
  340   187  160   432
```

b).

```
high   low   med vhigh
 187   432   340   160
```

c).

```
high   low   med vhigh
 160   340   432   187
```

d).

```
high   low   med vhigh
 160   432   340   187
```

Answer D

```
car <- read.csv("https://pengdsci.github.io/STA200/dataset/CarEvalData.csv")
table(car$buying)
```

Problem 29

The Car Evaluation Database was derived from a simple hierarchical decision model, directly relating to the CAR to the six input attributes: buying, maintenance, doors, persons, luggage boot, safety, and evaluated result (class). Because of known underlying concept structure, this database may be particularly useful for testing constructive induction and structure discovery methods. The database is hosted at

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

Using R to load the data and perform a chi-square test at a significant level of 0.05 to see whether the class (evaluated categories of the car) has following distribution.

acc	good	unacc	vgood
20%	5%	70%	5%

Based on your R output, which of the following is correct?

a). TS = 11.2432 df = 3 p-value = 0.03487
b) . TS = 8.615, df = 3, p-value = 0.03487
c) . TS = 8.615, df = 4, p-value = 0.04387
d) . TS = 6.6715, df = 4, p-value = 0.4387

Answer B

```
car <- read.csv("https://pengdsci.github.io/STA200/dataset/CarEvalData.csv")
chisq.test(table(car$class), p=c(0.2, 0.05, 0.7, 0.05))
```

Summary of Weekly #2

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]	2	75.00	0.12	3	0.19
(80,90]	10	85.00	0.62	13	0.81
(90,1e+02]	3	95.00	0.19	16	1.00

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

stats	value
Min.	69.00
1st Qu.	86.20
Median	86.20
3rd Qu.	89.70
Max.	93.10

The boxplot is a geometric representation of the five-number summary. The boxplot of the given data set is given below.