Question **1**

All hypothesis tests wind up producing a P-value. The P-value represents

Select one:

a. the probability that the null hypothesis is false.
b. the probability that the null hypothesis is true.
c. the relative frequency of values in the null distribution that are as extreme or more so than the test statistic.
d. the position of the test statistic in the null distribution.
e. the population proportion.

Question **2**

Read each description of a dataset, and imagine a research question, along with an inference procedure one might reasonably use on the data to answer that question. Choose, from the list of procedures provided, the letter for one that is most applicable to the setting. (You may use the same procedure twice if you think it appropriate.)

categorical w/ 3 values

A random sample of American adults is asked about marital status ("currently married", "previously married", and "never married") and the amount (in milligrams) of caffeine consumption.

1-way ANOVA — quantitative measure of caffeine

Answer 1Choose...goodness-of-fit test1-way ANOVAtwo-proportionmatched-pairs tone-proportionchi-square test for associationModel Utility Test2-sample mean1-sample mean

quantitative response

The daily energy expenditure in kJ (kilojoules) is measured for men of two groups, one consisting of a sample of obese men, the other consisting of "lean" men.

2-sample mean or 1-way ANOVA — 2 groups, a binary categorical variable, independent samples

Answer 2Choose...goodness-of-fit test1-way ANOVAtwo-proportionmatched-pairs tone-proportionchi-square test for associationModel Utility Test2-sample mean1-sample mean

quantitative — matched pairs?

The amount of energy intake (from food) in kJ (kilojoules) is measured for adult women both just prior to and just after their menstrual cycles.

same women in both groups — paired data

Answer 3Choose...goodness-of-fit test1-way ANOVAtwo-proportionmatched-pairs tone-proportionchi-square test for associationModel Utility Test2-sample mean1-sample mean

Like above, but now caffeine is made categorical — χ^2-test for association

A random sample of American adults is asked about marital status ("currently married", "previously married", and "never married") and the amount of caffeine consumption. Instead of recording actual caffeine values, the researcher logs one of these (categorical) values for each case, along with marital status: "0 mg", "1-150 mg", "151-300 mg", and "more than 300 mg".

Answer 4Choose...goodness-of-fit test1-way ANOVAtwo-proportionmatched-pairs tone-proportionchi-square test for associationModel Utility Test2-sample mean1-sample mean

Question **3**

The HELP study (Health Evaluation and Linkage to Primary Care) recruited patients with no primary care physician. Among the variables recorded for each patient

are `racegrp` (race/ethnicity) and `i1`, (the average number of drinks consumed per day in the past 30 days). Some results for `i1` broken down by `racegrp` are displayed in the table

Results from favstats(i1 ~ racegrp)

ratio of $\frac{\text{largest}}{\text{smallest}} = \frac{21.442}{18.461} < 2$

racegrp	min	Q1	median	Q3	max	mean	sd	n
black	0	3	10.0	23.0	134	15.711	19.041	211
hispanic	0	0	11.0	26.0	73	16.700	18.705	50
other	0	1	9.5	23.0	67	14.923	18.461	26
white	0	6	17.0	31.5	142	21.530	21.442	166

k = # of groups = 4

only one sample size < 30

These observations are relevant for Part H.

A partial ANOVA table appears below. Some entries have letters in them instead of numbers.

Result from anova(lm(i1 ~ racegrp))

Source	df	SS	MS	F
Groups	(a) = 3	(c) = 3511	(d) = 1170.3 (f)	$F = \frac{MSG}{MSE} = 2.96$
Error	(b) = 449	177655	(e) = 395.67	
Total	452	181166		

Fill out the missing entries of the ANOVA table as indicated by letter:

$SSG = SST - SSE = 3511$

- (a) Answer $= k - 1 = 3$
- (b) Answer $= 452 - 3 = 449$
- (c) Answer
- (d) Answer $= SSG/df1 = 3511/3$
- (e) Answer
- (f) Answer (Round this answer to 2 decimal places)

(g) Write an RStudio command that will produce the corresponding P-value as computed from a theoretical F distribution. Use numbers from your filled-out table as appropriate.

Your R command: Answer `1 - pf(2.96, df1 = 3, df2 = 449)`

Write out your answers to the remaining parts of this question **on your handwritten pages**, taking note of the question and letter.

- H. Does it appear that conditions are met justifying the use of a theoretical F distribution to compute a P-value? Explain why or why not.

Independent samples? As best we can tell.
Normal? sample sizes are ~30 or larger
Same population s.d? $\frac{\text{largest } s}{\text{smallest } s} < 2$.

- I. Suppose we reject the null hypothesis. What null hypothesis has been rejected? Write it.
- J. What is the alternative hypothesis? Write it.

H_0: Mean drinks per day is same in all 4 groups $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a: At least two of these

population means are unequal

- K. Does the output from the TukeyHSD() command below allow you the ability to conclude anything more? Explain who or why not?

Result from TukeyHSD(aov(i1 ~ racegrp))

	diff.	lwr	upr
hispanic-black	0.9891	-7.079	9.057
other-black	-0.7878	-11.449	9.873
white-black	5.8192	0.498	11.141
other-hispanic	-1.7769	-14.179	10.626
white-hispanic	4.8301	-3.445	13.105
white-other	6.6070	-4.212	17.426

This interval does not contain 0 — a significant difference in pop. means for blacks vs. whites.

Question **4**

If you need it, the **formula sheet** is available.

The number of men and women among professors in Math, Physics, Chemistry, Linguistics, and English departments from an SRS of small colleges were counted, and the results are shown in the table below.

Dept.	Math	Physics	Chemistry	Linguistics	English	Total
men	48	29	41	31	37	186
women	6	3	5	15	27	56
total	54	32	46	46	64	242

(a) Determine the amount of contribution to the overall $\chi 2$-statistic coming from the "Math--women" cell. Answer observed: 6, expected = (54)(56)/242 = 12.496

contribution from that cell: $\frac{(6-12.496)^2}{12.496} = 3.377$

(b) Find the cell which has the smallest expected count, and **give that expected count**. Answer This cell is the one at the intersection of row and column w/ smallest totals: physics — women

That cell's expected count: $\frac{(32)(56)}{242} = 7.41$

C. Write null and alternative hypothesis that one might use this contingency table's $\chi 2$-statistic to test. [Write your answer on your paper.] H_0: No association between gender and field

H_a: There is an association

D. Suppose you plan to use a theoretical chi-square distribution to obtain a P-value, and that the $\chi 2$-statistic is 27.07. Write an R command that directly gives you this P-value. [Write your answer on your paper.] # dfs = (2-1)(5-1) = 4

Right tailed: 1 - pchisq(27.07, df=4)

E. Is it justified to use a theoretical chi-square distribution to obtain a P-value? Why or why not? [Give your answer and explanation on your paper.] Yes. Smallest expected count = 7.41 is larger than 5

Question **5**

To investigate the daily expenses of summer tourists in Vienna, a survey of 43 tourists is conducted. The results of the sample show that the tourists spend on average 177.7 EUR. The sample standard deviation s is equal to 13.98.

$SE_{\bar{x}} = \frac{13.98}{\sqrt{43}} = 2.132$

Estimating both μ and σ:

Employ t-distribution w/ df = 43 - 1

Critical t^*-value comes from qt(.975, df=42) = 2.018

95% CI has

lower bound: 177.7 - (2.018)(2.132)

upper bound: 177.7 + (2.108)(2.132)

Determine a 95% confidence interval for the average daily expenses (in EUR) of a tourist, giving your answer in interval notation (using square brackets [lowerBound, upperBound]). Available sites include StatKey, RStudio, and the formula sheet.

Answer:

Question **6**

Identify whether the paradigm in the study is that of **independent samples** or **matched-pairs**.

independent samples

The average box-office receipts on opening weekend of 10 movies released within a week of Christmas is compared with the opening weekend box-office receipts of 10 movies released in early June, to see if there is a difference in average ticket sales.

Answer 1Choose...matched-pairsindependent samples

matched pairs (cases are married couples)

The age at the time of the wedding for 20 husbands is compared with the age of their wives on the same date, to see if there is a difference in mean age-at-marriage between men and women.

Answer 2Choose...matched-pairsindependent samples

20 wine connoisseurs are asked to rate, on an 11-point scale ranging from 0 to 10, the taste of two different chardonnay wines, to see if there is a difference in average rating.

Answer 3Choose...matched-pairsindependent samples

– matched pairs: each connoisseur contributes a number to both samples: of interest is their difference.

Question **7**

z^*-value for proportion at 90% confidence: qnorm(0.95) = 1.645

You may, if you choose, access these sites StatKey, RStudio, and the formula sheet.

(a) Suppose you want to construct a 90% confidence interval for the proportion of people who have learned how to swim by age 4. If you want the entire width of the interval (from lower bound to upper bound) to be no more than 0.04, estimate the minimum sample size you need to include in your sample. **Answer** $n \geq$ Answer $n \geq \left(\frac{1.645}{.02}\right)^2 \cdot (.5)(1-0.5) = 1691.27$ So, a sample size at least 1692.

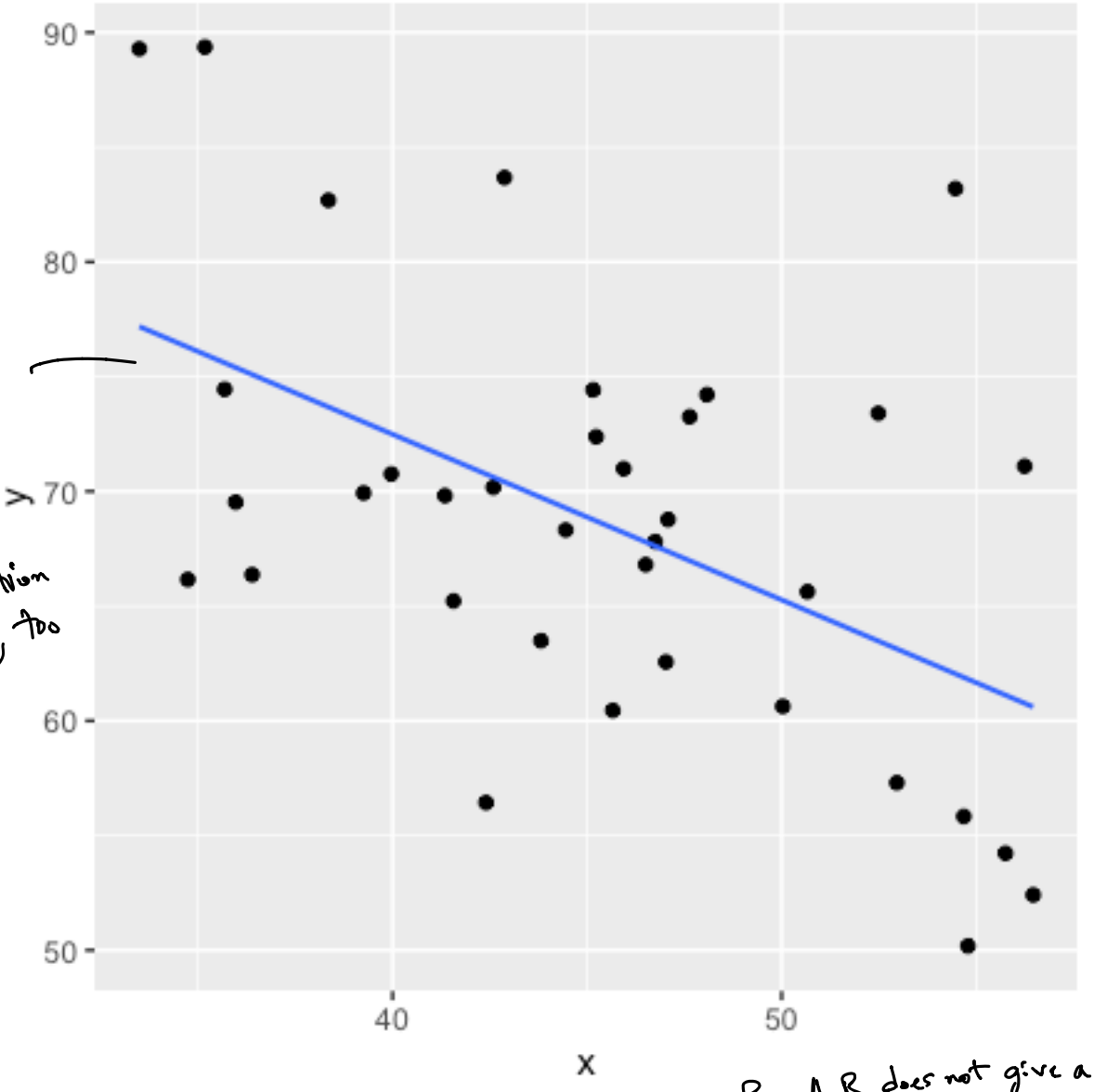
(b) Suppose, from a random sample of $n=152$ fourteen-year-old American boys, you have constructed a 95% confidence interval for the mean number of hours of screen time. Having done so, you wish you could redo the study so the margin of error is one-tenth as large. What is the least number of boys you should include in the sample for your follow-up study? **Answer** $n \geq$ Answer

We know spread (SE) is proportional to $\frac{1}{\sqrt{n}}$

Cut margin of error to $\frac{1}{10}$ its current size means increasing sample size $10^2 = 100$ times to 15200.

Question **8**

The picture gives a scatterplot of y on x and the least-squares regression line. Below it there is an ANOVA table computed from the same data. Answer the questions based on this output.



slope of regression line is negative, so correlation is negative, too

Recall R does not give a total row.

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## x          1  823.12  823.12  12.009 0.001453
## Residuals 34 2330.47   68.54
```

model row

error row

(a) Is there significant evidence at the 1% level to conclude an association exists between x and y? On your paper writeup, give support for your answer.
AnswerYesNo Yes, since $P = 0.001453$ which is smaller than 0.01.

(b) What percentage of the variation in y-values is explained by the linear (blue-line) model? Answer

(c) What is the sample correlation? Answer $\text{correlation} = -\sqrt{0.261} = -0.511$

(d) How many cases are in the dataset? In your write-up provide evidence, but do not use a count of dots as your evidence, as one dot may lie on top of others. Answer

(b) is asking for coefficient of determination, R^2.

$$R^2 = \frac{SSM}{SST} = \frac{SSM}{SSE + SSM} = 0.261$$

If n = # of cases, then

$df_{total} = 1 + 34 = n - 1$

$\Rightarrow n = 36$

$= \frac{823.12}{823.12 + 2330.47}$

or 26.1 percent

That is extrapolation.

(e) If one uses the model to predict the value of y when x=78, that is called AnswerExtrapolationIntrepidationInterpolation

(f) A confidence interval for the mean y-value occurring when x=24 is Answerwiderless wide than a prediction interval for the next observed y-value when x=42.

Predicting a single value is always subject to more variability than estimating the mean response. Thus, a CI for the mean response is less wide than a prediction interval.