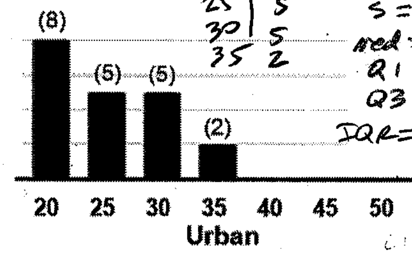
AP Statistics Semester 1 Free-Response Final REVIEW

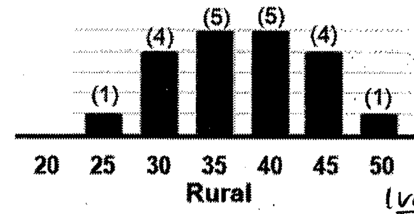
#1. The goal of a nutritional study was to compare the caloric intake of adolescents living in rural areas of the United States with the caloric intake of adolescents living in urban areas of the United States. A random sample of ninth-grade students from one high school in a rural area was selected. Another random sample of ninth graders from one high school in an urban area was also selected. Each student in each sample kept records of all the food he or she consumed in one day.

The bar charts for each region are shown to the right displaying the number of calories of food consumed per kilogram of body weight for each student on that day (the number of students in each column is shown in parentheses above the column)



L1	L2
20	8
25	5
30	5
35	2

1 var stats L1, L2
$\bar{x} = 25.20$
$s = 5.20$
med = 25
$Q1 = 20$
$Q3 = 30$
IQR = 30 − 20 = 10



L1	L2
25	1
30	4
35	5
40	5
45	4
50	1

1 var stats L1, L2
$\bar{x} = 37.5$
$s = 6.59$
med = 37.5
q1 = 32.5
q3 = 42.5
IQR = 42.5 − 32.5 = 10

(a) Write a few sentences comparing the distribution of the daily caloric intake of ninth-grade students in the rural high school with the distribution of the daily caloric intake of ninth-grade students in the urban high school.

(shape) Urban caloric intake is skewed right, but rural is symmetrical.

(center) Typical urban calorie values are lower than rural (median 25 cals/kg urban vs 37.5 cals/kg for rural).

(spread) The variability is about same for both urban and rural with IQR of 10 cals/kg.

(outliers) There are no outliers in either distribution.

urban upper fence:
$Q3 + 1.5\,IQR$
$30 + 1.5(10) = 45$

*SOCS

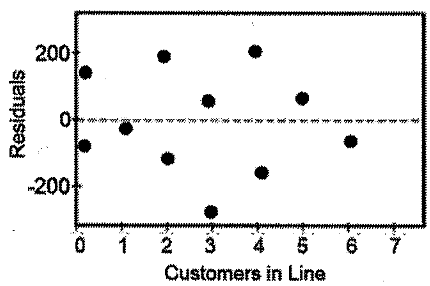
(b) Is it reasonable to generalize the findings of this study to all rural and urban ninth-grade students in the United States? Explain.

No, These samples are only from one rural and one urban high school so it is not reasonable to expect these students to be representative of all 9th graders in the U.S.

(c) What percentage of all the students studied consumed 35 calories per kilogram of body weight?

$$\frac{5+2}{8+5+5+2+1+4+5+5+4+1} = \frac{7}{40} = .175 \quad \boxed{17.5\%}$$

#2. The manager of a grocery store selected a random sample of 11 customers to investigate the relationship between the number of customers in a checkout line and the time to finish checkout. As soon as the selected customer entered the end of a checkout line, data were collected on the number of customers in line who were in front of the selected customer and the time, in seconds, until the selected customer was finished with the checkout. A linear regression was performed on the data and the regression output and residual plot are shown below:



Predictor	Coef	SE Coef	T	P
Constant	72.95	110.36	0.66	0.525
Customers in line	174.40	35.06	4.97	0.001
S = 200.01		R-Sq = 73.33%		R-Sq (adj) = 70.37%

(a) Is a linear model appropriate for modeling these data? Clearly explain your reasoning.

yes, there is no pattern in the residuals.

(b) Write the LSRL for the association between customers in line and time to finish checkout (be sure to define your variables)

$\hat{y} = 72.95 + 174.40x$

x: customers in line
y: time to finish checkout (seconds)

(c) What is the predicted time to check out if there are 4 customers in line?

$\hat{y} = 72.95 + 174.40(4) = \boxed{770.55 \text{ seconds}}$

(d) Interpret the value of the slope of the LSRL in the context of this problem.

For every additional 1 person in line, the time to finish checkout increases by 174.40 seconds, on average.

(e) Interpret the value of the y-intercept of the LSRL in the context of this problem.

If there were no customers in front of you in line, the model predicts time to checkout will be 72.95 seconds.

(f) Interpret the value of r^2 in the context of this problem.

About 73.33% of the variation in time to checkout is explained by the LSRL which relates time to checkout to number of customers in line.

#3. A coed youth sports league includes multiple teams each with roughly the same number of players. For advertising purposes, the organization which manages the league wants to produce a brochure which contains a picture taken by a professional photographer of a 'typical' team – which represents the youth league's participants well. The photographer is tasked with selecting which players should appear in the photograph. She is considering two sampling strategies: a cluster sample, and a stratified random sample.

(a) Explain the difference between cluster sampling and stratified random sampling in this context.

Cluster: randomly select one or more teams and include <u>all</u> the players on these teams in the sample.

Stratified: Select an SRS of players from each team to include in the sample.

(b) If cluster sampling is used, and the sample turns out to be biased, explain why this would <u>not</u> be called voluntary response bias.

Voluntary response bias requires that players volunteer to be in the picture (which is not what is happening here).

(c) If cluster sampling is used, and the sample turns out to be biased, what would the correct term be for the type of bias which may occur?

Undercoverage bias.

#4. Airlines routinely overbook flights because they expect a certain number of no-shows. An airline runs a 5 P.M. commuter flight from Washington, D.C., to New York City on a plane that holds 38 passengers. Past experience has shown that if 41 tickets are sold for the flight, then the probability distribution for the number who actually show up for the flight is as shown in the table below.

Number who actually show up	36	37	38	39	40	41
Probability	0.46	0.30	0.16	0.05	0.02	0.01

Assume that 41 tickets are sold for each flight.

(a) There are 38 passenger seats on the flight. What is the probability that all passengers who show up for this flight will get a seat?

$P(\text{everyone gets a seat}) = P(X \leq 38) = .46 + .30 + .16 = \boxed{.92}$

(b) What is the expected number of no-shows for this flight?

$EV = \mu = (36)(.46) + (37)(.3) + (38)(.16) + (39)(.05) + (40)(.02) + (41)(.01)$

or

can put 36, 37 .. into L1
.46, .30 ... into L2

and run 1varstats L1, L2 $\bar{X} = 36.9$

this is the number who show up
so the expected number of no-shows is $41 - 36.9 = \boxed{4.1}$

(c) Given that not all passenger seats are filled on a flight, what is the probability that only 36 passengers showed up for the flight?

$P(X = 36 \mid \text{not all of the 38 seats are full}) = P(X = 36 \mid X < 38)$

$= \dfrac{.46}{.46 + .30} = \boxed{.6053}$