CollegeBoard
Advanced Placement
Program

AP® Statistics
2005 Sample Student Responses

The College Board: Connecting Students to College Success

The College Board is a not-for-profit membership association whose mission is to connect students to college success and opportunity. Founded in 1900, the association is composed of more than 4,700 schools, colleges, universities, and other educational organizations. Each year, the College Board serves over three and a half million students and their parents, 23,000 high schools, and 3,500 colleges through major programs and services in college admissions, guidance, assessment, financial aid, enrollment, and teaching and learning. Among its best-known programs are the SAT®, the PSAT/NMSQT®, and the Advanced Placement Program® (AP®). The College Board is committed to the principles of excellence and equity, and that commitment is embodied in all of its programs, services, activities, and concerns.

Visit the College Board on the Web: www.collegeboard.com.
AP Central is the official online home for the AP Program and Pre-AP: apcentral.collegeboard.com.

1A

STATISTICS
SECTION II
Part A
Questions 1-5

Spend about 65 minutes on this part of the exam.

Percent of Section II grade—75

Directions: Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy of your results and explanation.

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 back-to-back stemplot below displays the number of calories of food consumed per kilogram of body weight for each student on that day.

Urban		Rural
99998876	2	
44310	3	2334
97665	3	56667
20	4	02224
	4	56889
	5	1

Stem: tens
Leaf: ones

(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.

The distribution of the daily caloric intake of ninth-grade students in the rural high school is approximately symmetric while the distribution of the daily caloric intake of ninth-grade students in the urban high school is skewed right. Neither distributions have an outlier. The median$^{(32)}$ of the urban distribution and its $Q_1^{(21)}$ and $Q_3^{(36)}$ are all less than that of the rural distribution. The rural distribution has a spread of 19 larger than the urban distribution spread of 16. The IQR of the urban distibution is 5 which is less than the IQR of the Rural distribution which is 10.

which has a median of 41 Q_1 equal to 35.5 and Q_3 equal to 45.

GO ON TO THE NEXT PAGE.

(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, there are many different rural and urban areas in the united states, this sample only encompases students from two schools. There are many more students who don't go to highschool, or many schools may even administer healthier lunches. There are just too many confounding or lurking variables to generalize the findings of this study to ALL rural and urban ninth-grade studen ts in the United states.

(c) Researchers who want to conduct a similar study are debating which of the following two plans to use.

Plan I: Have each student in the study record all the food he or she consumed in one day. Then researchers would compute the number of calories of food consumed per kilogram of body weight for each student for that day.

Plan II: Have each student in the study record all the food he or she consumed over the same 7-day period. Then researchers would compute the average daily number of calories of food consumed per kilogram of body weight for each student during that 7-day period.

Assuming that the students keep accurate records, which plan, I or II, would better meet the goal of the study? Justify your answer.

Plan II would be better because it encompasses a 7-day period which would average out any days that a student might have eaten an extremel large amount or a very small amount. In plan I the amount of food is only during one day, and in that one day there might be variables which would cause that person to eat more or less then they normally would.

GO ON TO THE NEXT PAGE.

1B

STATISTICS
SECTION II
Part A
Questions 1-5

Spend about 65 minutes on this part of the exam.

Percent of Section II grade—75

Directions: Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy of your results and explanation.

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 back-to-back stemplot below displays the number of calories of food consumed per kilogram of body weight for each student on that day.

Urban		Rural	
99998876	2		
44310	3	2334	
97665	3	56667	median 41
20	4	02224	mean : 40.45
median 34	4	56889	n=20
n= 20 mean 32.6	5	1	

Stem: tens
Leaf: ones

30chs

(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.

the Urban sample distribution has a center of approximately 33 and is skewed towards lower caloric intakes a spread of ~~2~~ ~~96 to~~ 26 to 42 with no outliers

the rural sample distribution is roughly symmetrical arround a center of 41 cal per kg of body weight a spread of 32 to 51 with no outliers

GO ON TO THE NEXT PAGE.

1B

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

It is not reasonable to generalize these findings to all rural and urban 9th grade students because the sample size is so small, and only includes two schools

these two schools could have very different meal plans provided during the school day than the average plans of highschools in the nation

Only using two schools leaves room for confounding variables

(c) Researchers who want to conduct a similar study are debating which of the following two plans to use.

Plan I: Have each student in the study record all the food he or she consumed in one day. Then researchers would compute the number of calories of food consumed per kilogram of body weight for each student for that day.

Plan II: Have each student in the study record all the food he or she consumed over the same 7-day period. Then researchers would compute the average daily number of calories of food consumed per kilogram of body weight for each student during that 7-day period.

Assuming that the students keep accurate records, which plan, I or II, would better meet the goal of the study? Justify your answer.

plan II, it would give a more realistic view of the students daily caloric intake by averaging over a seven day period. this would help reduce the impact of unusually high or low days (such as a party or a day in which a meal was missed) these averages would more realistically represent each student

1C

STATISTICS
SECTION II
Part A
Questions 1-5

Spend about 65 minutes on this part of the exam.

Percent of Section II grade—75

Directions: Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy of your results and explanation.

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 back-to-back stemplot below displays the number of calories of food consumed per kilogram of body weight for each student on that day.

Urban		Rural
9 9 9 9 8 8 7 6	2	
4 4 3 1 0	3	2 3 3 4
9 7 6 6 5	3	5 6 6 6 7
2 0	4	0 2 2 2 4
	4	5 6 8 8 9
	5	1

42

Stem: tens
Leaf: ones

(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.

Students in rural high school has higher median and range Compared to students in urban high school. Almost half of the data for urban school is in 20's, but the data of rural is well distributed. Shape of the distribution for urban is skewed to the right, and shape of the distribution for rural is nearly symmetric. There is no gap in either school.

GO ON TO THE NEXT PAGE.

1C

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

It is not reasonable to generalize the findings of this study because the Sample size is small, and there could be some confounding variables that controls the calories of students in each area.

(c) Researchers who want to conduct a similar study are debating which of the following two plans to use.

Plan I: Have each student in the study record all the food he or she consumed in one day. Then researchers would compute the number of calories of food consumed per kilogram of body weight for each student for that day.

Plan II: Have each student in the study record all the food he or she consumed over the same 7-day period. Then researchers would compute the average daily number of calories of food consumed per kilogram of body weight for each student during that 7-day period.

Assuming that the students keep accurate records, which plan, I or II, would better meet the goal of the study? Justify your answer.

Plan II would better meet the goal of the study because the study is trying to find out the general idea of difference in consumption of calories of adolscence between urban and rural area. If researchers study a week period of food he or she consume, the data would be more reliable. If researchers only study one day of the food, they can't get the general idea of the number of calories.

GO ON TO THE NEXT PAGE.

2A

2. Let the random variable X represent the number of telephone lines in use by the technical support center of a software manufacturer at noon each day. The probability distribution of X is shown in the table below.

x	0	1	2	3	4	5
$p(x)$	0.35	0.20	0.15	0.15	0.10	0.05

(a) Calculate the expected value (the mean) of X.

$$\mu_X = \Sigma(X \cdot P_X) = (0 \cdot .35) + (1 \cdot .20) + (2 \cdot .15) + (3 \cdot .15) + (4 \cdot .10) + (5 \cdot .05)$$

$$\boxed{\mu_X = 1.6}$$

(b) Using past records, the staff at the technical support center randomly selected 20 days and found that an average of 1.25 telephone lines were in use at noon on those days. The staff proposes to select another random sample of 1,000 days and compute the average number of telephone lines that were in use at noon on those days. How do you expect the average from this new sample to compare to that of the first sample? Justify your response.

The average of the samlle of 1,000 ~~Both~~ should be closer to the true mean of 1.6 than the sample of 20. This is due to the fact that the variability of a sample mean decreases as sample size increases, by the equation $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$. In this case, since the standard deviation will be lower, values further from 1.6 become much more unlikely, and the sample mean should fall within a much closer range to the true mean.

(c) The median of a random variable is defined as any value x such that $P(X \leq x) \geq 0.5$ and $P(X \geq x) \geq 0.5$. For the probability distribution shown in the table above, determine the median of X.

$P(0) + P(1) = .55 \rightarrow P(x \leq 1) = .55 \rightarrow P(x \geq 1) = .65$

$P(1) + \dots P(5) = .65$

$\boxed{\text{median} = 1}$

(d) In a sentence or two, comment on the relationship between the mean and the median relative to the shape of this distribution.

The distribution is skewed to the right, which pulls the mean toward the higher values. Medians are more resistant to skewing than means, so the median here is lower than the mean.

GO ON TO THE NEXT PAGE.

2B

2. Let the random variable X represent the number of telephone lines in use by the technical support center of a software manufacturer at noon each day. The probability distribution of X is shown in the table below.

x	0	1	2	3	4	5
$p(x)$	0.35	0.20	0.15	0.15	0.10	0.05

(a) Calculate the expected value (the mean) of X.

$\mu = .35(0) + .20(1) + (.15)2 + .15(3) + .10(4) + .05(5)$

$\boxed{\mu = 1.6}$

(b) Using past records, the staff at the technical support center randomly selected 20 days and found that an average of 1.25 telephone lines were in use at noon on those days. The staff proposes to select another random sample of 1,000 days and compute the average number of telephone lines that were in use at noon on those days. How do you expect the average from this new sample to compare to that of the first sample? Justify your response.

$n = 20$ $\bar{x} = 1.25$

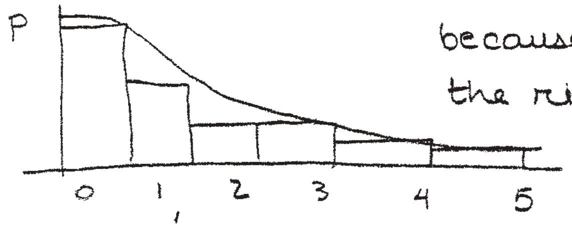
$n = 1{,}000$

I expect the average from the new sample will be closer to the expected value (1.6). This is because the Law of Large Number says that the larger my sample size is, the closer my mean will come to the true or expected value (1.6). Thus, $\bar{x}$ for $n = 1000$ should be slightly larger than $\bar{x}$ for $n = 20$.

(c) The median of a random variable is defined as any value x such that $P(X \leq x) \geq 0.5$ and $P(X \geq x) \geq 0.5$. For the probability distribution shown in the table above, determine the median of X.

median = 1

(d) In a sentence or two, comment on the relationship between the mean and the median relative to the shape of this distribution.



The mean is larger than the median, because the distribution is skewed to the right.

GO ON TO THE NEXT PAGE.

2c

2. Let the random variable X represent the number of telephone lines in use by the technical support center of a software manufacturer at noon each day. The probability distribution of X is shown in the table below.

x	0	1	2	3	4	5
$p(x)$	0.35	0.20	0.15	0.15	0.10	0.05

(a) Calculate the expected value (the mean) of X.

Expected value $X = (.35)(0) + (.20)(1) + (.15)(2) + (.15)(3) + (.10)(4) + (.05)(5)$

$= \boxed{1.6}$

(b) Using past records, the staff at the technical support center randomly selected 20 days and found that an average of 1.25 telephone lines were in use at noon on those days. The staff proposes to select another random sample of 1,000 days and compute the average number of telephone lines that were in use at noon on those days. How do you expect the average from this new sample to compare to that of the first sample? Justify your response.

The average from the new sample would be closer to the mean of x (1.6) than the first sample because the number of days is multiplied by 50 and this is an example of the law of large numbers.

(c) The median of a random variable is defined as any value x such that $P(X \leq x) \geq 0.5$ and $P(X \geq x) \geq 0.5$. For the probability distribution shown in the table above, determine the median of X.

median of $X = P(X \leq x) \geq 0.5 + P(X \geq x) \geq 0.5$

median of $X = \boxed{1.3}$

(d) In a sentence or two, comment on the relationship between the mean and the median relative to the shape of this distribution.

The relationship between the mean/median is that the median is smaller than the mean because the above distribution is right skewed.

GO ON TO THE NEXT PAGE.

3A

```
The regression equation is
Fuel Consumption = 10.7 + 2.15 Railcars

Predictor     Coef        StDev          T        P
Constant      10.677      5.157         2.07     0.072
Railcar       2.1495      0.1396       15.40     0.000

S = 4.361  R-Sq = 96.7%  R-Sq(adj) = 96.3%
```

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

Yes, a linear model is appropriate because the original data appears linear and the residuals appear randomly distributed with no shape or bending.

(b) Suppose the fuel consumption cost is \$25 per unit. Give a point estimate (single value) for the change in the average cost of fuel per mile for each additional railcar attached to a train. Show your work.

fuel consumption = 10.677 + 2.1495 railcars

For every additional railcar, 2.1495 additional units of fuel/mile is used. $(2.1495)(25) = \$53.74$ increase in cost of fuel per mile for each additional car.

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

96.7% of the variation in fuel consumption is explained by the change in number of railcars.

(d) Would it be reasonable to use the fitted regression equation to predict the fuel consumption for a train on this route if the train had 65 railcars? Explain.

No, I do not believe this would be appropriate. Any extrapolation should be used with caution, and since 65 is so far away from the closest observed value (50) I do not feel we can assume it would be accurate.

GO ON TO THE NEXT PAGE.

3B

```
The regression equation is
Fuel Consumption = 10.7 + 2.15 Railcars
              ŷ=10.7+2.15x
Predictor     Coef       StDev        T        P
Constant     10.677      5.157       2.07    0.072
Railcar       2.1495     0.1396     15.40    0.000

S = 4.361  R-Sq = 96.7%   R-Sq(adj) = 96.3%
```

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

Yes, a linear model is appropriate for modeling these data. The residual plot shows no trends, and therefore does not suggest any other models. Additionally the r^2 value is very high, 96.3%, which means that 96.3% of the variation is accounted for.

(b) Suppose the fuel consumption cost is \$25 per unit. Give a point estimate (single value) for the change in the average cost of fuel per mile for each additional railcar attached to a train. Show your work.

$\hat{y} = 10.7 + 2.15x$

$2.1495 \times (\$25)$

$= 53.7375$

For each additional railcar attached to a train, fuel consumption increases by 2.15 units. This increase converts to a \$53.74 avg. cost increase per additional railcar attached

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

The value of r^2, according to the regression analysis output, is 96.3%. In the context of this problem, this means that 96.3% of the variation that occurs in the fuel consumption linear regression model is accounted for

(d) Would it be reasonable to use the fitted regression equation to predict the fuel consumption for a train on this route if the train had 65 railcars? Explain.

No, it would not be reasonable to use the fitted regression equation to predict the fuel consumption for a train on this route if the train had 65 railcars because the highest number of railcars used to create the regression line was 50 railcars. 65 railcars would just be too far off for us to know for sure if our regression line has accurated predicting the fuel consumption

3C

```
The regression equation is
Fuel Consumption = 10.7 + 2.15 Railcars

Predictor    Coef       StDev          T        P
Constant     10.677     5.157         2.07     0.072
Railcar      2.1495     0.1396       15.40     0.000

S = 4.361  R-Sq = 96.7%   R-Sq(adj) = 96.3%
```

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

The linear model is appropriate because the r^2 is high ($r^2 = .967$) and the residuals do not indicate any type of pattern.

(b) Suppose the fuel consumption cost is $25 per unit. Give a point estimate (single value) for the change in the average cost of fuel per mile for each additional railcar attached to a train. Show your work.

Fuel consumption= $10.7 + 2.15(1)$

For each additional railcar, the fuel consumption increases by 2.15.

$2.15 \times \$25 = \boxed{\$53.75}$

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

The r^2 shows how well the regression fits the data.

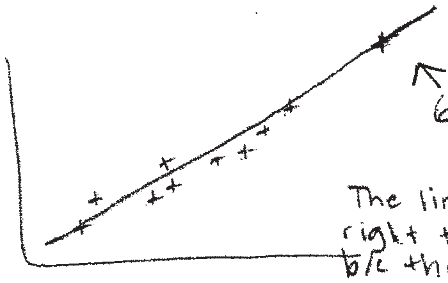
In this problem, there seems to be a linear correlation between fuel consumption and number of railcars because the r^2 is high.

(d) Would it be reasonable to use the fitted regression equation to predict the fuel consumption for a train on this route if the train had 65 railcars? Explain.

yes, it would be reasonable because the r^2 is high and the residuals are random so the data is reliable to predict from.

$F.C. = 10.7 + 2.15(65)$

$= \boxed{150.45}$



65 railcars

The line fits right through it b/c there is a high correlation btwn fuel consumption + number of railcars.

GO ON TO THE NEXT PAGE.

4A

4. Some boxes of a certain brand of breakfast cereal include a voucher for a free video rental inside the box. The company that makes the cereal claims that a voucher can be found in 20 percent of the boxes. However, based on their experiences eating this cereal at home, a group of students believes that the proportion of boxes with vouchers is less than 0.2. This group of students purchased 65 boxes of the cereal to investigate the company's claim. The students found a total of 11 vouchers for free video rentals in the 65 boxes.

Suppose it is reasonable to assume that the 65 boxes purchased by the students are a random sample of all boxes of this cereal. Based on this sample, is there support for the students' belief that the proportion of boxes with vouchers is less than 0.2 ? Provide statistical evidence to support your answer.

$\hat{p} = \frac{x}{n} = \frac{11}{65} = .1692$ $\sigma = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.20(1-.20)}{65}} = .0496$

spec: .20

I will do a proportion z test

Conditions

Assume sample of boxes is unbiased estimator of true proportion

Assume population $\geq 10n \rightarrow pop \geq 650$

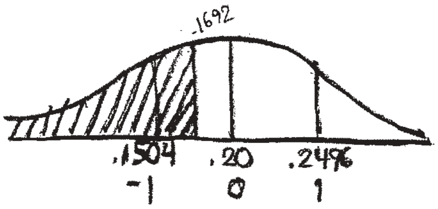
$np > 10$? $n(1-p) > 10$?

$65(.20) > 10$ ✓ $65(1-.20) > 10$ ✓

$n = 65$, $n > 30$

$H_0 : p = .20$

$H_a : p < .20$



← Assume H_0 for sampling distribution

z test stat $= \frac{\text{stat}-\text{parameter}}{\text{standard error}} = \frac{.1692-.20}{.0496} = \boxed{-.6210}$

$p(\hat{p} < .1692) = p(z < -.6210) = \text{normalcdf}(-10^{99}, -.6210) = .2673$

Assuming the proportion of vouchers found in boxes is .20, there is a .2673 chance of getting a sample with a proportion more extreme than .1692

Large p → Fail to reject H_0.

There is insufficient evidence to claim that the proportion of boxes with vouchers is less than 20 percent.

4B

4. Some boxes of a certain brand of breakfast cereal include a voucher for a free video rental inside the box. The company that makes the cereal claims that a voucher can be found in 20 percent of the boxes. However, based on their experiences eating this cereal at home, a group of students believes that the proportion of boxes with vouchers is less than 0.2. This group of students purchased 65 boxes of the cereal to investigate the company's claim. The students found a total of 11 vouchers for free video rentals in the 65 boxes.

Suppose it is reasonable to assume that the 65 boxes purchased by the students are a random sample of all boxes of this cereal. Based on this sample, is there support for the students' belief that the proportion of boxes with vouchers is less than 0.2 ? Provide statistical evidence to support your answer.

1 - proportion test

$\frac{11}{65} = .169 \quad \therefore \hat{p} = .169$

$H_o : p = .2$

$H_a : p < .2$

Assumptions

1. Random sample.
2. Population is sufficiently large.
3. $np > 10 \qquad n(1-p) > 10$

   $(65)(.2) > 10 \qquad (65)(.8) > 10$

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

$$= \frac{.169 - .2}{\sqrt{\frac{.2(1-.2)}{65}}} = -.6202$$

$p = .2676$

$\therefore$ We fail to reject Ho.

$\therefore$ The proportion of the boxes with vouchers is equal to .2

Therefore students' belief was wrong.

GO ON TO THE NEXT PAGE.

4C

4. Some boxes of a certain brand of breakfast cereal include a voucher for a free video rental inside the box. The company that makes the cereal claims that a voucher can be found in 20 percent of the boxes. However, based on their experiences eating this cereal at home, a group of students believes that the proportion of boxes with vouchers is less than 0.2. This group of students purchased 65 boxes of the cereal to investigate the company's claim. The students found a total of 11 vouchers for free video rentals in the 65 boxes.

Suppose it is reasonable to assume that the 65 boxes purchased by the students are a random sample of all boxes of this cereal. Based on this sample, is there support for the students' belief that the proportion of boxes with vouchers is less than 0.2 ? Provide statistical evidence to support your answer.

$H_0: p = 0.20$

$H_a: p < 0.20$

① sample size → $n = 65$

② $\hat{p} = \frac{11}{65} = 0.1692\ldots$

③ boxes are an SRS

Use a one porportion z-test

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \qquad z = \frac{\left(\frac{11}{65}\right) - 0.20}{\sqrt{\frac{0.20(0.80)}{65}}}$$

$z = -0.6202$

p-value of $z = 0.2676$

$C = 0.95$ (default) $\alpha = 0.05$

$0.2676 > 0.05$ → accept H_0

There is not enough evidence to reject H_0.

GO ON TO THE NEXT PAGE.

5A

5. A survey will be conducted to examine the educational level of adult heads of households in the United States. Each respondent in the survey will be placed into one of the following two categories:

- Does not have a high school diploma
- Has a high school diploma

The survey will be conducted using a telephone interview. Random-digit dialing will be used to select the sample.

(a) For this survey, state one potential source of bias and describe how it might affect the estimate of the proportion of adult heads of households in the United States who do not have a high school diploma.

There could be some sampling bias because of the way in which the sample is obtained. Not all households in the US have telephones, so the sample is only taken from the population of households with telephones, not all households. Since people with less education are less likely to have telephones, this may result in an estimate that is too low for the proportion of adult heads of households in the US who do not have a high school diploma.

(b) A pilot survey indicated that about 22 percent of the population of adult heads of households do not have a high school diploma. Using this information, how many respondents should be obtained if the goal of the survey is to estimate the proportion of the population who do not have a high school diploma to within 0.03 with 95 percent confidence? Justify your answer.

$$\text{Margin of Error} = z^* \sqrt{\frac{p(1-p)}{n}}$$

$$.03 \geq 1.95996 \sqrt{\frac{.22(.78)}{n}}$$

$$n \geq 732.4381$$

sample size should be at least 733 for a margin of error ≤ .03 with 95% confidence

GO ON TO THE NEXT PAGE.

SA

If you need more room for your work for part (b), use the space below.

(c) Since education is largely the responsibility of each state, the agency wants to be sure that estimates are available for each state as well as for the nation. Identify a sampling method that will achieve this additional goal and briefly describe a way to select the survey sample using this method.

The sample could be a stratified random sample with an SRS taken from each state. For each state, an SRS of household mailing addresses could be obtained, and a survey could be mailed to the desired number of respondents in each state. Nonresponse bias would be a concern when conducting a mail survey, which we would keep in mind when conducting the study.

GO ON TO THE NEXT PAGE.

5B

5. A survey will be conducted to examine the educational level of adult heads of households in the United States. Each respondent in the survey will be placed into one of the following two categories:

- Does not have a high school diploma
- Has a high school diploma

The survey will be conducted using a telephone interview. Random-digit dialing will be used to select the sample.

(a) For this survey, state one potential source of bias and describe how it might affect the estimate of the proportion of adult heads of households in the United States who do not have a high school diploma.

One source is bias in this case is undercoverage. This means that because the survey is conducted by phone, it leaves out all of the households without phones. This affects the estimate of the proportion of adult heads of households in the U.S. who do not have a high school diploma because, chances are if someone does not have a high school diploma, he is not well educated and does not make a lot of money. In that case, there is a high chance he won't have a phone in his household. Therefore, the survey will leave out a large portion of households that would, *

** otherwise increase the estimate of the proportion of adult heads of households who do not have a high school diploma.

(b) A pilot survey indicated that about 22 percent of the population of adult heads of households do not have a high school diploma. Using this information, how many respondents should be obtained if the goal of the survey is to estimate the proportion of the population who do not have a high school diploma to within 0.03 with 95 percent confidence? Justify your answer.

$$\text{margin of error} \leq z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$.03 \leq 1.96 \sqrt{\frac{(.22)(.78)}{n}}$$

$$.015306 = \sqrt{\frac{(.22)(.78)}{n}}$$

$$.000234 = \frac{(.22)(.78)}{n}$$

$$n = \frac{(.22)(.78)}{.000234}$$

$$= 733.333 \Rightarrow \boxed{734 \text{ respondents}}$$

GO ON TO THE NEXT PAGE.

5B

If you need more room for your work for part (b), use the space below.

734 respondents should be asked. This is because the surveyors want to estimate the proportion within .03. In this case, .03 is the margin of error. The margin of error is equal to the z^* value multiplied by the standard error which is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, where $\hat{p}$ is the sample proportion and n is the number of respondents. Thus, to calculate how many respondents are needed the margin of error (.03) has to be less than or equal to 1.96 (z^* for 95% confidence) times $\sqrt{\frac{(.02)(.98)}{n}}$. Solving for n, gives $n \approx 733.33$ respondents. This must be rounded up to 734 to ensure that $\frac{1}{3}$ of a respondent is included.

(c) Since education is largely the responsibility of each state, the agency wants to be sure that estimates are available for each state as well as for the nation. Identify a sampling method that will achieve this additional goal and briefly describe a way to select the survey sample using this method. To achieve this goal the agency should divide the United States into the 50 states. Then within each of the states, the agency should assign numbers to each household. Then using a random number generator, the agency should pick a random sample of 1000 households to receive the survey. For instance if there are 3000 households in a state, the agency would number them from 0001 to 3000 and then randomly generate 1000 4-digit #'s from 0001 to 3000. The agency would then call the household's selected with phones and visit those without phones to ask the question. This process would be repeated for each of the 50 states.

GO ON TO THE NEXT PAGE.

5C

5. A survey will be conducted to examine the educational level of adult heads of households in the United States. Each respondent in the survey will be placed into one of the following two categories:

- Does not have a high school diploma
- Has a high school diploma

The survey will be conducted using a telephone interview. Random-digit dialing will be used to select the sample.

(a) For this survey, state one potential source of bias and describe how it might affect the estimate of the proportion of adult heads of households in the United States who do not have a high school diploma.

One possible source of bias would be that a greater amount of households that are head by a person who does not have a high school diploma would not have phones. This would be undercoverage bias because they would not reach these people. A greater prop. of heads of houses will be in this sample than the true pop. prop

(b) A pilot survey indicated that about 22 percent of the population of adult heads of households do not have a high school diploma. Using this information, how many respondents should be obtained if the goal of the survey is to estimate the proportion of the population who do not have a high school diploma to within 0.03 with 95 percent confidence? Justify your answer.

$$.03 = 1.96 \cdot \sqrt{\frac{(.22)(.78)}{X}}$$

$$\frac{.03}{1.96} = \sqrt{\frac{(.22)\cdot(.78)}{X}}$$

$$.0153^2 = \sqrt{\frac{(.22)(.78}{X}}^2$$

$$.000234 = \frac{(.22)(.78)}{X}$$

GO ON TO THE NEXT PAGE.

5C

If you need more room for your work for part (b), use the space below.

$$\frac{X \cdot .000234}{.000234} = \frac{(.22)(.78)}{.000234} = 733.33$$

734 respondents should be obtained.

(c) Since education is largely the responsibility of each state, the agency wants to be sure that estimates are available for each state as well as for the nation. Identify a sampling method that will achieve this additional goal and briefly describe a way to select the survey sample using this method.

Blocking will be encorperated into the new study. First, each state will split into its own block. Then, a simple random sample will be taken from the entire population, not people who just have phones. Each home in the previous census would be alloted a number. Using the table of random digits each state would select a certain amount of people for the study. After the selection took place a surveyer would go to each house and ask about the educational level of the head of the house.

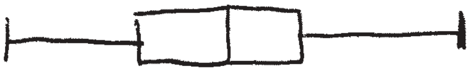
6A

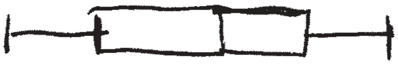
(a) Use a 95 percent confidence interval to estimate the difference in the mean amount of lead on a child's dominant hand after an hour of play inside versus an hour of play outside at urban day-care centers in this city. Be sure to interpret your interval.

Pop: children in urban day-care centers, playing inside or outside

Par: diff. in pop. means between amt. of lead (in mcg) inside (μ_1) or outside (μ_2).

↳ 2 sample t-interval → Assumptions: $\underset{\text{(both samples)}}{n \geq 40}$ → NO! (since $5 \leq n \leq 15$ in each sample, PLOT DATA!)

Inside: →  → looks approx. symm. (N. approx. is okay).

Outside: →  → VERY little skewness; approx. symm. (N. approx. is okay).

SRS → both given; σ is not known; samples are ind. → Yes! (what one child does outside doesn't affect child inside; samp.s not matched).

Using a 2-sample t-interval is OK.

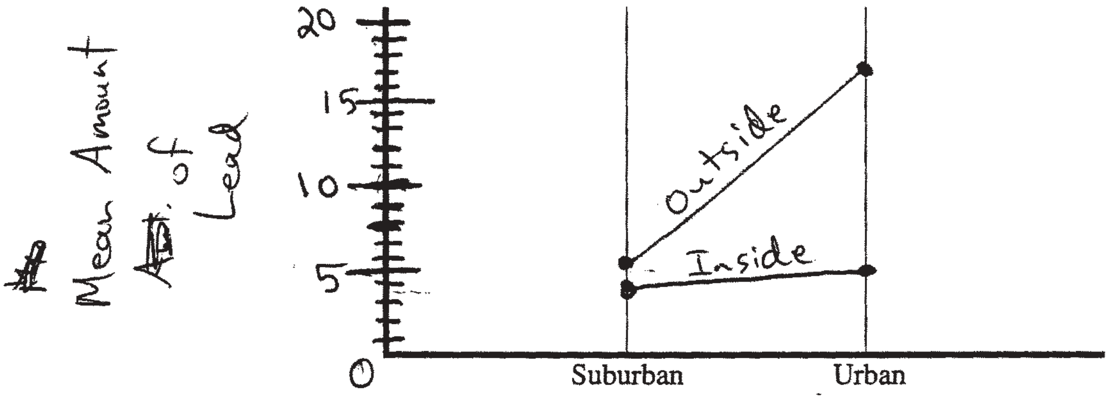
$CI_{95\%}$ for (diff. in pop. means) → ~~(-16.57, -9~~ → $\boxed{(-16.57, -9.434)}$

↳ ~~I~~ ~~we can be~~ I am 95% confident that ~~our~~ my CI, (-16.57 ~~↔~~ to -9.434), will capture the diff. in pop. means ~~between~~ of lead on each child's dominant hand between children playing inside and those playing outside (inside minus outside). ~~at urban day care centers in this city~~.

6A

(b) On the figure below,

- Using the vertical axis for the mean amount of lead, plot the mean for the amounts of lead on the dominant hand of children who played inside at the suburban day-care center and then plot the mean for the amounts of lead on the dominant hand of children who played inside at the urban day-care center.
- Connect these two points with a line segment.
- Plot the two means (suburban and urban) for the children who played outside at the two types of day-care centers.
- Connect these two points with a second line segment.



Pen → inside
Pencil → outside

[insert] measurements

6A

(c) From the study, what conclusions can be drawn about the impact of setting (inside, outside), environment (suburban, urban), and the relationship between the two on the amount of lead on the dominant hand of children after play in this city? Justify your answer.

B/c of a statistically significant difference between ~~st~~ suburban inside & outside ~~meas~~ ~~times~~ (using a test by CI @ .025 significance level) in which outside measurements of lead were higher as well as a statistically significant difference (also based on a CI~~s~~, using one-sided .025 significance level) between in & out measurements in urban areas (neither of these CIs captured zero, ~~meaning the outsi~~ and they were ONLY negative when subtracting the mean of inside measurements by those outside, meaning that there was more lead piled on outside than in). B/c I do not know ~~8~~ the sample standard deviation of suburban times, I cannot calculate whether the inside times (urban minus suburban) or outside times have a statistically significant difference. But I do see that the inside times have a small difference, while outside times have a HUGE difference, and can conclude that, when inside a day-care center, lead poisoning in ~~E~~ urban and suburban areas isn't much different, but when kids get outside, the amount of lead in urban areas is much higher, as proven by such high amounts of lead on urban children's dominant hands.

END OF EXAMINATION

THE FOLLOWING INSTRUCTIONS APPLY TO THE BACK COVER OF THE SECTION II BOOKLET.

- **MAKE SURE YOU HAVE COMPLETED THE IDENTIFICATION INFORMATION AS REQUESTED ON THE BACK OF THE SECTION II BOOKLET.**
- **CHECK TO SEE THAT YOUR AP NUMBER APPEARS IN THE BOX(ES) ON THE BACK COVER.**
- **MAKE SURE YOU HAVE USED THE SAME SET OF AP NUMBER LABELS ON ALL AP EXAMINATIONS YOU HAVE TAKEN THIS YEAR.**

6B

Part B

Question 6

Spend about 25 minutes on this part of the exam.

Percent of Section II grade—25

Directions: Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy of your results and explanation.

6. Lead, found in some paints, is a neurotoxin that can be especially harmful to the developing brain and nervous system of children. Children frequently put their hands in their mouth after touching painted surfaces, and this is the most common type of exposure to lead.

A study was conducted to investigate whether there were differences in children's exposure to lead between suburban day-care centers and urban day-care centers in one large city. For this study, researchers used a random sample of 20 children in suburban day-care centers. Ten of these 20 children were randomly selected to play outside; the remaining 10 children played inside. All children had their hands wiped clean before beginning their assigned one-hour play period either outside or inside. After the play period ended, the amount of lead in micrograms (mcg) on each child's dominant hand was recorded.

The mean amount of lead on the dominant hand for the children playing inside was 3.75 mcg, and the mean amount of lead for the children playing outside was 5.65 mcg. A 95 percent confidence interval for the difference in the mean amount of lead after one hour inside versus one hour outside was calculated to be (−2.46, −1.34).

A random sample of 18 children in urban day-care centers in the same large city was selected. For this sample, the same process was used, including randomly assigning children to play inside or outside. The data for the amount (in mcg) of lead on each child's dominant hand are shown in the table below.

Urban Day-Care Centers

Inside	6	5	4	4	4.5	5	4.5	3	5
Outside	15	25	18	14	20	13	11	22	20
Diffence I−O	−9	−20	−14	−10	−15.5	−8	−6.5	−19	−15

GO ON TO THE NEXT PAGE.

6B

(a) Use a 95 percent confidence interval to estimate the difference in the mean amount of lead on a child's dominant hand after an hour of play inside versus an hour of play outside at urban day-care centers in this city. Be sure to interpret your interval.

95% confidence interval:

$$C\text{-int} = \bar{x}_1 - \bar{x}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$n_1 = 9$

$n_2 = 9$

$\bar{x}_1$: mean amount of lead (in mcg) found on child's dominant hand after playing inside

Found on calculator:

$x_1 = 4.5556$

$x_2 = 17.5556$

$s_1 = 0.8457$

$s_2 = 4.6128$

x_2: mean amount of lead (in mcg) found on childs dominant hand after playing outside.

t^* for $p = 0.025$ and $9-1=8$ degrees of freedom is 2.306 (from table)

$$C\text{-int} = 4.5556 - 17.5556 \pm 2.306 \sqrt{\frac{0.8457^2}{9} + \frac{4.6128^2}{9}}$$

$$= -13 \pm 3.6048$$

$$= -16.6048 \text{ to } -9.3952$$

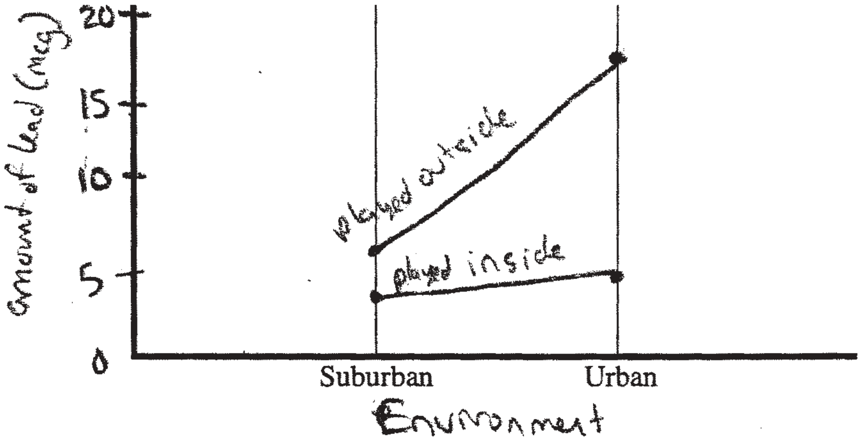
We can be 95% confidant that the difference between the mean amount of lead (in mcg) on a child's dominant hand after playing for one hour inside and the mean amount of lead (in mcg) on a child's dominant hand after playing for one hour outside is between -16.6048 and -9.3952.

GO ON TO THE NEXT PAGE.

6B

(b) On the figure below,

- Using the vertical axis for the mean amount of lead, plot the mean for the amounts of lead on the dominant hand of children who played inside at the suburban day-care center and then plot the mean for the amounts of lead on the dominant hand of children who played inside at the urban day-care center.
- Connect these two points with a line segment.
- Plot the two means (suburban and urban) for the children who played outside at the two types of day-care centers.
- Connect these two points with a second line segment.



Suburban: 3.75 mcg (inside), 5.65 mcg (outside)
urban: 4.5556 mcg (inside), 17.5556 mcg (outside)

GO ON TO THE NEXT PAGE.

6B

(c) From the study, what conclusions can be drawn about the impact of setting (inside, outside), environment (suburban, urban), and the relationship between the two on the amount of lead on the dominant hand of children after play in this city? Justify your answer.

The mean amount of lead on children who played outside was greater than the amount on children who played inside for both suburban and urban environments. This difference was greater for the urban setting, with a difference we can be 95% confident is between -16.6048 and -9.3952, as opposed to the suburban interval of -2.46 to -1.34. In the table of urban values all of the inside values were less than all of the outside values. This makes a strong case that in general, levels of Lead are higher outside than inside. Also, the mean value of lead in urban areas was higher than the mean value in suburban areas for both children who played inside and outside. Urban children had a mean of 17.5556 mcg outside, while soburban children only had a mean of 5.65 mcg outside, and urban children had a mean of 4.5556 mcg inside, while the suburban Children had a mean of 3.75 mcg inside. Based on this study, It appears that urban children playing outside will have more lead than suburban children playing inside, in general.

END OF EXAMINATION

THE FOLLOWING INSTRUCTIONS APPLY TO THE BACK COVER OF THE SECTION II BOOKLET.

- **MAKE SURE YOU HAVE COMPLETED THE IDENTIFICATION INFORMATION AS REQUESTED ON THE BACK OF THE SECTION II BOOKLET.**
- **CHECK TO SEE THAT YOUR AP NUMBER APPEARS IN THE BOX(ES) ON THE BACK COVER.**
- **MAKE SURE YOU HAVE USED THE SAME SET OF AP NUMBER LABELS ON ALL AP EXAMINATIONS YOU HAVE TAKEN THIS YEAR.**

6C

(a) Use a 95 percent confidence interval to estimate the difference in the mean amount of lead on a child's dominant hand after an hour of play inside versus an hour of play outside at urban day-care centers in this city. Be sure to interpret your interval.

$\bar{X}_{inside} = 4.556$ $\qquad$ $\bar{X}_{outside} = 17.556$

$$Sx = \sqrt{\frac{1}{n-1}\sum(X_i - \bar{X})^2}$$

$$\sum_{inside}(X_i - \bar{X})^2 = (6-4.556)^2 + 3(5-4.556)^2 + 2(4.5-4.556)^2 + 2(4-4.556)^2 + (3-4.556)^2$$

$$= 5.722$$

$$Sx = \sqrt{\frac{1}{8}(5.722)}$$

$$Sx_{inside} = .846$$

$$\sum_{outside}(X_i - \bar{X})^2 = (25-17.556)^2 + (22-17.556)^2 + 2(20-17.556)^2 + (18-17.556)^2 + (15-17.556)^2 + (14-17.556)^2 + (13-17.556)^2 + (11-17.556)^2$$

$$= 170.222$$

$$Sx = \sqrt{\frac{1}{8}(170.222)}$$

$$Sx_{outside} = 4.613$$

Simple Random Sample
Population assumed large

CI Interval $\qquad$ 2 Sample T interval

$-16.57 \leq \bar{X}_{inside} - \bar{X}_{outside} \leq -9.434$ where;

$\bar{X}_{inside}$ is the mean amount of lead on the dominant hand for children playing inside

$\bar{X}_{outside}$ is the mean amount of lead on the dominant hand for children playing outside

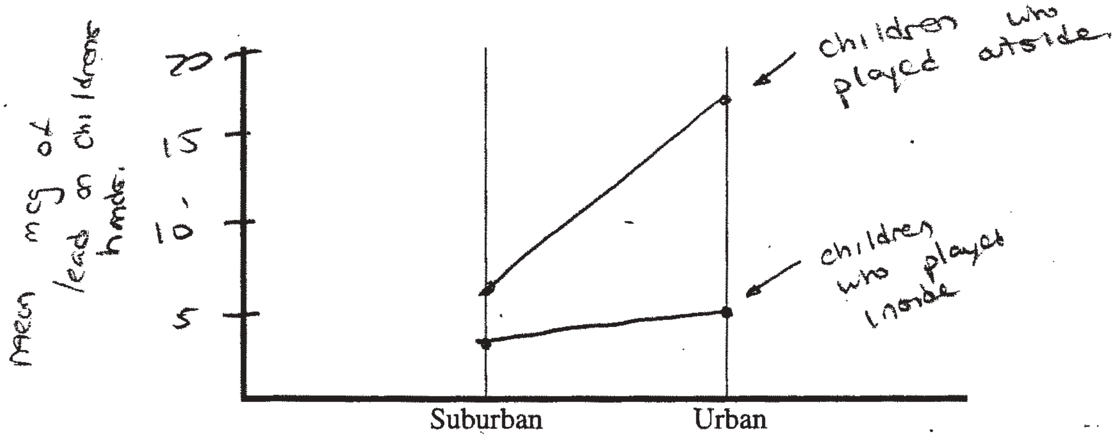
This interval means That $\mu_{inside} - \mu_{outside}$ (where μ_{inside} is the parameter of the mean of levels of mcg of lead on childrens hands who played inside; and $\mu_{outside}$ is the mean parameter of the mean of levels of mcg of lead on childrens hands who played outside) will fall between -16.57 and -9.434 95% of the time

GO ON TO THE NEXT PAGE.

6C

(b) On the figure below,

- Using the vertical axis for the mean amount of lead, plot the mean for the amounts of lead on the dominant hand of children who played inside at the suburban day-care center and then plot the mean for the amounts of lead on the dominant hand of children who played inside at the urban day-care center.
- Connect these two points with a line segment.
- Plot the two means (suburban and urban) for the children who played outside at the two types of day-care centers.
- Connect these two points with a second line segment.



GO ON TO THE NEXT PAGE.

6C

(c) From the study, what conclusions can be drawn about the impact of setting (inside, outside), environment (suburban, urban), and the relationship between the two on the amount of lead on the dominant hand of children after play in this city? Justify your answer.

From the study, it can be concluded that the urban environment had more lead for children to get their hands exposed to than in the suburban environment, as seen by a higher mean amount of micrograms of lead on the hands of children in urban day care centers than the mean amount of micrograms of lead on the hands of children in suburban day care centers. From this, it can be inferred that the urban environment has more lead than the suburban environment in terms of day care centers. It can also be inferred that inside (at both urban and suburban locations) there are less levels of lead for children to be exposed to. It appears that going outside increased the amount of lead children are exposed to relative to the inside, as seen by the differences in mean lead levels on childrens hands in the different environments.

END OF EXAMINATION

THE FOLLOWING INSTRUCTIONS APPLY TO THE BACK COVER OF THE SECTION II BOOKLET.

- **MAKE SURE YOU HAVE COMPLETED THE IDENTIFICATION INFORMATION AS REQUESTED ON THE BACK OF THE SECTION II BOOKLET.**
- **CHECK TO SEE THAT YOUR AP NUMBER APPEARS IN THE BOX(ES) ON THE BACK COVER.**
- **MAKE SURE YOU HAVE USED THE SAME SET OF AP NUMBER LABELS ON ALL AP EXAMINATIONS YOU HAVE TAKEN THIS YEAR.**