Introductory Statistics - Midterm # 1
Answer Key

1. The Lumina Foundation and Gallup would like to investigate how important are the laws in the state where the college is located on access to reproductive health services is in your decision to stay enrolled until you have graduated. Results were based on web survey responses, conducted Oct. 26 - Nov. 17, 2022, from 6,008 U.S. adults currently enrolled in an associate degree, bachelor's degree, certificate or certification program. It is reported that nearly three-quarters of currently enrolled college students (72%) report that the reproductive health laws in the state where their college is located are at least somewhat important to their decision to stay enrolled.[1]
   a. What is the population of interest?
      The population of interest in this study is U.S. adults currently enrolled in an associate degree, bachelor's degree, certificate, or certification program.

   b. What is the sample in the given scenario?
      The sample in this study conducted by the Lumina Foundation and Gallup consists of 6,008 U.S. adults who are currently enrolled in an associate degree, bachelor's degree, certificate, or certification program.

   c. What is the observational unit of the study?
      The observational unit in this study is an individual U.S. adult who is currently enrolled in an associate degree, bachelor's degree, certificate, or certification program at a college or educational institution.

   d. Does 72% represent a parameter or statistics? Explain.
      The 72% reported in the study is a statistic.
      In this case, the 72% represents the proportion of a sample (6,008 U.S. adults currently enrolled in college programs) who reported that reproductive health laws in their state are at least somewhat important to their decision to stay enrolled. This percentage is based on data from the sample and is used to estimate or describe the corresponding parameter in the broader population of all U.S. adults enrolled in similar programs.

   e. What are, if any, the potential sources of bias in their sampling method?
      In this study, the researchers collected data through a web survey, which relies on individuals voluntarily participating by responding to the survey questions. A potential source of bias in the sampling methods used for this study is non-response bias. It could be that it is such a sensitive topic that participants are reluctant to fill out the survey.

[1] https://news.gallup.com/poll/474365/reproductive-health-laws-factor-college-decisions.aspx

Not all individuals have equal access to the internet or may face technical difficulties in completing online surveys. The results might not represent the true population because students who are more politically active or have strong opinions on reproductive health laws are more likely to respond, the results may overstate the importance of these laws in the population.

2. In 2020, researchers worked to develop a safe and effective vaccine against SARS-CoV-2, the coronavirus that causes COVID-19. A clinical trial was conducted with more than 30,000 adult volunteers nationwide. Participants were 18 years of age or older with no known previous SARS-CoV-2 infection. Volunteers were randomly assigned to receive either two doses of the investigational vaccine (100 micrograms each) or two shots of a saline placebo.[2]
   a. Explain why this is an experiment and not an observational study.
      It is an experiment because of multiple factors, such as: in this clinical trial, participants were randomly assigned to one of two groups, controlled intervention existed in this clinical trial that allows researchers to assess the impact of the vaccine on the participants' health outcomes, the use of a double-blind design, it is designed to establish causal relationships between variables, that is, determining whether the vaccine causes a reduction in infections, and researchers have control over the design, implementation, and manipulation of variables.

      In contrast, observational studies primarily involve the passive observation of individuals or groups without direct intervention by the researcher, which is not the case in this scenario.

   b. What is the explanatory variable in this study? What type of variable is the explanatory variable, categorical or quantitative?
      The explanatory variable is the "type of treatment" or "vaccine assignment." Specifically, it refers to whether a participant was assigned to receive the investigational vaccine (100 micrograms each) or the saline placebo. Therefore, it is categorical.

   c. What is the response variable in this study? What type of variable is the response variable?
      The response variable is typically whether or not a participant contracts a SARS-CoV-2 infection. This response variable is categorical, as it involves outcomes like "infected" or "not infected."

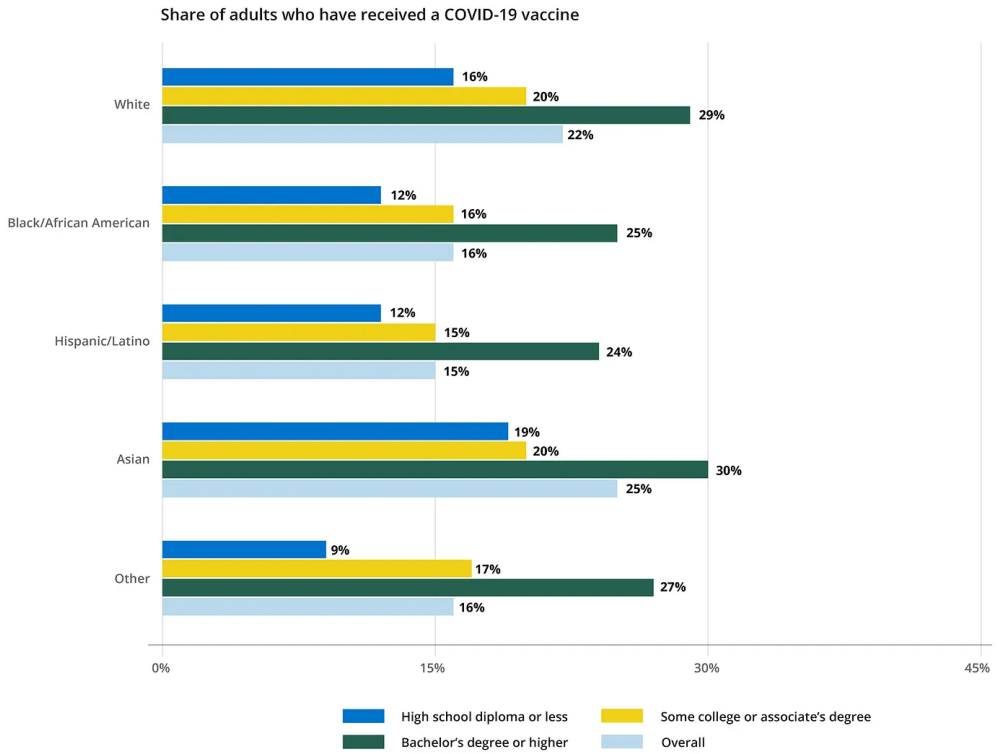
   d. What are the nuisance factors in the experiment?
      There are many nuisance factors in this experiment, such as, age, gender, geographic location, preexisting health conditions, environmental conditions, etc.

[2] https://www.nih.gov/news-events/nih-research-matters/experimental-coronavirus-vaccine-highly-effective

e. Which group is the experimental group? Which group is the control group?
Experimental group consists of volunteers who received two doses of the investigational vaccine (100 micrograms each). They are the group that researchers are primarily interested in studying to assess the vaccine's safety and effectiveness.
Control group consists of volunteers who received two shots of a saline placebo and is used as a baseline for comparison.

f. Which form of replication is demonstrated in the experiment?
Replication was achieved by including large numbers of participants in the clinical trial.

3. Many Americans are eager to get back out into the world after a year of being cooped up at home because of the pandemic. The development, approval, and administration of vaccines will be major factors in the lifting of restrictions on people's activities. Progress is slowly but steadily occurring on this front: 32 percent of US adults (84 million adults) had received at least one COVID-19 vaccine dose by March 23, 2021. However, the vaccine rollout continues to encounter problems. One is ensuring that underserved communities have fair access to vaccinations, and another is overcoming some groups' entrenched distrust of the vaccines.[3]
Use the side-by-side bar graph to answer the questions below.



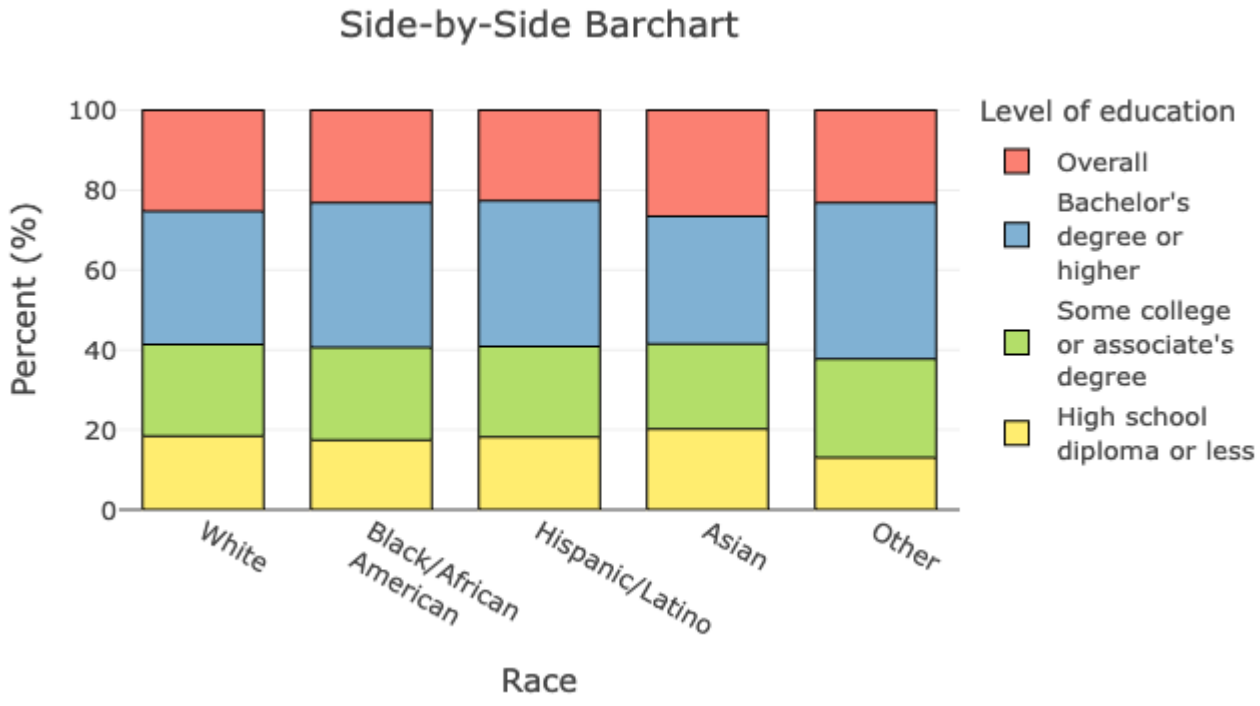
3 https://medium.com/georgetown-cew/vaccinations-have-ramped-up-but-has-distribution-been-equitable-ac76d91b6bfd

a. Based on the overall percentage, provide the list from the least likely to most likely groups of people to have received a COVID-19 vaccine.
Hispanic/Latino, Other, Black/African American, White, Asian

b. Based on the bar graph above, which level(s) of education were more likely to have received a vaccine during the early phases of the distribution?
Bachelor's degree or higher

c. Based on the bar graph above, for those who have a Bachelor's degree or higher level of education, which group were least likely to have received a vaccine during the early phases of the distribution?
Hispanic/Latino

d. Create a stacked bar graph for the data set.



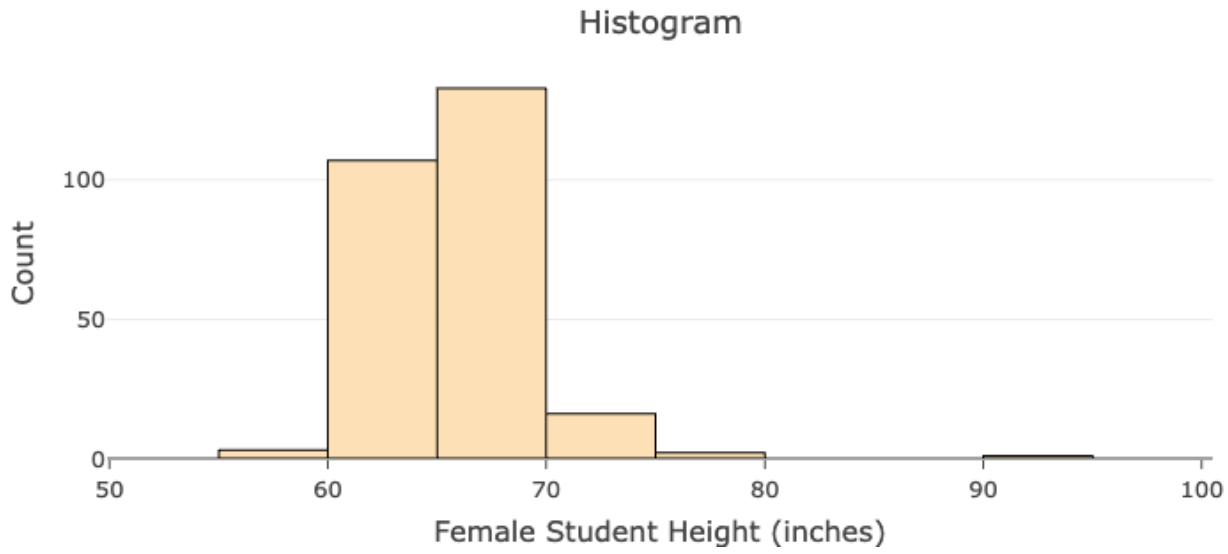
e. What would be the benefit of using the stacked bar graphs rather than the side-by-side bar graph presented above?
Stacked bar graphs are useful when you want to emphasize the total value and show the composition of that total by different categories. It also allows us to easily compare the relative proportions of different categories within each total.

f. The title of the article where this graph came from is "Vaccinations Have Ramped Up, But Has Distribution Been Equitable?" What is the answer to the question based on the

bar graph provided? Are there additional explanatory/extraneous variables we should consider to answer the question?
I do not think that the distribution has been equitable. Overall, Hispanic/Latino were least likely to have received a vaccine during the early phases of the distribution.

Yes, when analyzing the factors affecting vaccine rollout and distribution, there are several additional explanatory and extraneous variables to consider. Factors such as age, gender, race, ethnicity, and socioeconomic status can play a significant role in vaccine distribution and acceptance. Geographic location, including urban vs. rural areas and regional disparities, can impact vaccine distribution. The availability and capacity of healthcare facilities, including hospitals, clinics, and pharmacies, can affect the speed and efficiency of vaccine distribution. The way information about vaccines is communicated to the public can influence trust and vaccine acceptance. Public behavior, such as adherence to preventive measures (e.g., mask-wearing, social distancing) and vaccine hesitancy, can be influenced by various factors, including political beliefs and misinformation.

4. The height of 262 female students (in inches) are plotted in the histogram below.



a. Describe the shape of the distribution of the data set.
The distribution of the data set is skewed to the right.

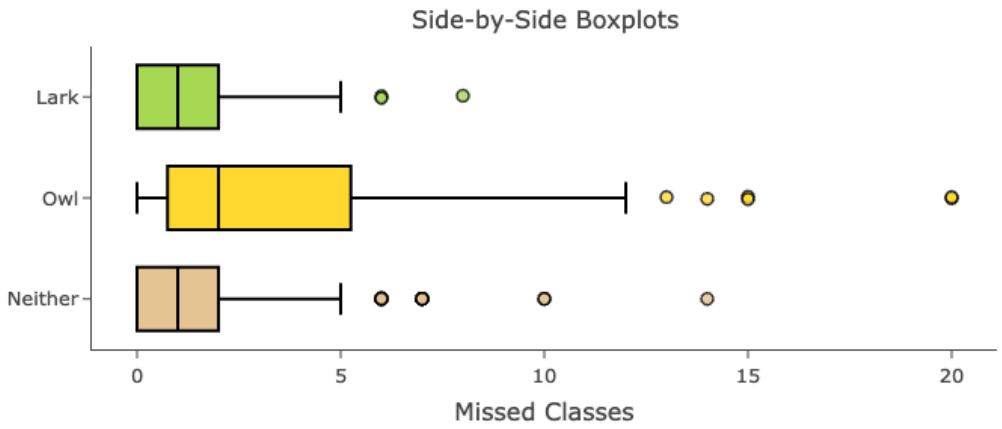
b. What is the approximate mean and median of the data set? In a few sentences, interpret your findings.
The mean is approximately 66 inches and the median is approximately 65 inches. The mean height represents the average height of female students in inches. In this scenario, on average, female students in the sample have a height of 66 inches.

The median height is the middle value when all the heights are arranged in ascending order. It's the height at which half of the female students are taller, and half are shorter.

c. Which measure of center represents the data set best? Explain your choice.
Because the data set is skewed to the right, the median represents the data set best because the median is not affected by extreme values or outliers in the data set.

d. Instead of a histogram, you can represent the data set as a dotplot. What would be the advantages and disadvantages of having a dotplot rather than a histogram?
Dotplots are particularly effective when dealing with small data sets. You can see the exact values of each data point and it is easy to compare individual data points. However, when the data set is larger, dotplots can become cluttered and less informative. Some data might not be suitable for dotplot. Additionally, the clarity of the distribution might depend on the spacing of the dots.

5. The sleep study data set investigated whether college students' chronotypes tend to be larks (morning people) or owls (night people) and measured the number of classes they missed in one semester.
The boxplots are displayed below.



a. Use the boxplots to describe the difference in variability between the groups.
Based on the dotplots, college students who identify as Owl have the largest variability, followed by those who identify as Neither, and then those who identify as Lark.

b. The mean and standard deviation of the "Lark" category is 1.56 and 2, respectively. Write a sentence describing the standard deviation in context of the problem.
In the context of the sleep study data set, the standard deviation of 2 for the "Lark" category means that, on average, the number of classes missed by college students who are morning people tends to vary from the mean of 1.56 by approximately 2 classes.

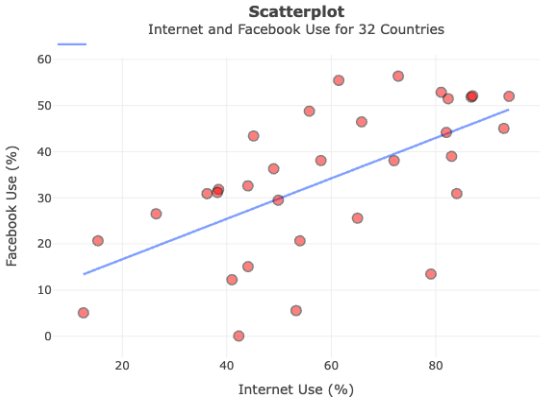
c. List the five-number summary for students who are neither lark nor owl.
0, 0, 1, 2, 14

d. For the “Neither” category, what is the IQR? How many missed classes would be considered a lower outlier and upper outlier?
IQR = 2
Lower Outlier: 0 - 1.5 * 2 = -3. So, any data point with a number of missed classes less than -3 would be considered a lower outlier.
Upper Outlier: 2 + 1.5 * 2 = 5. So, any data point with a number of missed classes greater than 5 would be considered an upper outlier.

e. For the “Lark” category, __50__ % of students missed more than 2 classes.

f. For the “Owl” category, __50__ % of students missed less than 2 classes.

6. Let's investigate the relationship between the Internet and Facebook usage from 32 countries in 2012. The scatterplot is provided below.



a. Select the correlation coefficient that best represents the data set.
-0.614 -0.425 -0.213 0. 213 0.425 0.614

b. Interpret the correlation coefficient.
The correlation coefficient "r" of 0.614 indicates a moderately strong positive linear relationship between Internet usage and Facebook usage across the 32 countries in 2012.

c. The line of best fit: $\hat{y} = 7.9 + 0.439x$.
   i. Identify the slope and interpret it in context of the problem.
   Slope is 0.439.
   For every percentage increase in Internet usage, there is a corresponding increase of approximately 0.439% in Facebook usage on average among these countries.

   ii. Identify the y-intercept and interpret it in context of the problem.

y-intercept: 7.9.
If Internet usage were non-existent (i.e., zero %) in these countries, the expected level of Facebook usage would still be approximately 7.9%.

d. The internet usage in the U.S. is approximately 81% in the year of 2012. Based on this data set, predict the Facebook usage in the U.S. in 2012. Round your answer to 3 decimal places.
Predicted Facebook usage in the U.S. in 2012 = 7.9 + 0.439(81)= 7.9 + 35.559 = 43.459%.

e. Facebook usage in Brazil is approximately 29.5% in the year of 2012. Based on this data set, predict the internet usage in Brazil in 2012. Round your answer to 3 decimal places.
Facebook Usage = 7.9 + 0.439(Internet Usage)
29.5 = 7.9 + 0.439(Internet Usage)
Predicted Internet Usage in Brazil in 2012= (29.5-7.9)/0.439 = 49.201%.

7. Gallup tracks Americans' views about global warming each March as part of its annual Environment poll. Below are the results of the latest survey conducted on March 1-23, 2023.[4]

Worry: Global Warming	18-34 years old	35-54 years old	55+ years old
A great deal	134	112	145
A fair amount	49	77	88
Only a little	38	60	83
Not at all	53	65	87

a. If one adult American is chosen at random, what is the probability that he/she/they worry a great deal about global warming?
391/991

b. If one adult American is chosen at random, what is the probability that he/she/they is 35-54 years old?
314/991

c. If one adult American is chosen at random, what is the probability that he/she/they worry a great deal about global warming and are 35-54 years old?
112/991

[4] https://news.gallup.com/poll/474542/steady-six-say-global-warming-effects-begun.aspx

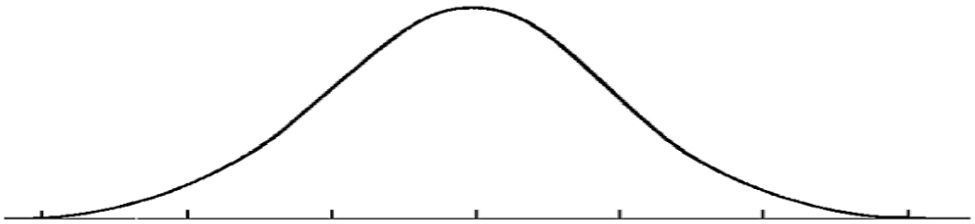
d. If one adult American is chosen at random, what is the probability that he/she/they worry a great deal about global warming or are 35-54 years old?
593/991

e. Given that he/she/they is 35-54 years old, what is the probability that he/she/they worry a great deal about global warming?
112/314

f. Is "35-54 years old" and "worry a great deal about global warming" mutually exclusive? Justify your answer.
They are not mutually exclusive because there are 35-54 years old who worry about

g. Is "35-54 years old" and "worry a great deal about global warming" independent? Justify your answer.
They are not independent because P(35-54 years old)*P(worry a great deal about global warming) = 314/991 * 391/991 does not equal to 112/991 = P(35-54 years old AND worry a great deal about global warming).

8. SAT scores are typically normally distributed. In 2021, the distribution of scores in the math section of the SAT had a mean of 528 and a standard deviation 120.[5]
   a. On the following graph, label each tick mark on the x-axis.



168 288 408 528 648 768 888

   b. Fill-in-the blanks below using the Empirical Rule.

   About 68% of the scores are between ___408___ and ___648____.

   About 95% of the scores are between ___288___ and ___768____.

   About 99.7% of the scores are between ___168___ and ___888___.

   c. Using the empirical rule, answer the following.

   $P(x > 528)$ = ___50%__

[5] https://nces.ed.gov/programs/digest/d21/tables/dt21_226.40.asp

$P(408 < x < 528)$ = ___34%___

$(408 < x < 768)$ = ___81.5%__

$P(x > 888)$ = ___0.15%___

d. Calculate and interpret the z-value for the SAT math section score of 600. Then, use the statistical tool to calculate the probability that an SAT test taker will score above 600.
z = (600-528)/120 = 0.6
P(z>0.6) = 27.43%

e. Use the statistical tool to answer: What SAT math section score corresponds to the 95th percentile?
725.3824

9. In 2021, 62% of U.S. citizens who were 18 to 29 years old did not have retirement savings.[6] Suppose that you want to test this theory and decide to pool your introductory statistics class. There are 30 students in the class who are 18 to 29 years old.
   a. Explain why this is a binomial distribution and not a normal distribution.
   Because it involves a discrete set of independent trials, each with only two possible outcomes: either a student has retirement savings (success) or does not have retirement savings (failure). A normal distribution is typically used when dealing with continuous data, which in this case, we do not have.

   b. What are the probability of success (p) and the probability of failure (1–p) in this scenario?
   The probability of success (not having retirement savings) is p = 0.62, and the probability of failure (having retirement savings) is (1-p) = 0.38.

   c. Use the binomial distribution statistical tool to determine the probability that 15 out of 30 students in the class who are 18 to 29 years old did not have retirement savings.
   0.0041

   d. Use the binomial distribution statistical tool to determine the probability more than 20 students in the class who are 18 to 29 years old did not have retirement savings.
   0.8773

   e. Would you be surprised to find that more than half of your classmates did not have retirement savings? Justify your answer using probability.

[6] https://www.statista.com/statistics/1273812/adults-with-no-retirement-savings-by-age-us/

No. Because the probability that more than 15 out of 30 of my classmates did not have retirement savings is 0.9975.

f. Would you be surprised to find that less than a quarter of your classmates did not have retirement savings? Justify your answer using probability.
Yes. Because the probability that less than 7.5 out of 30 of my classmates did not have retirement savings is 0.