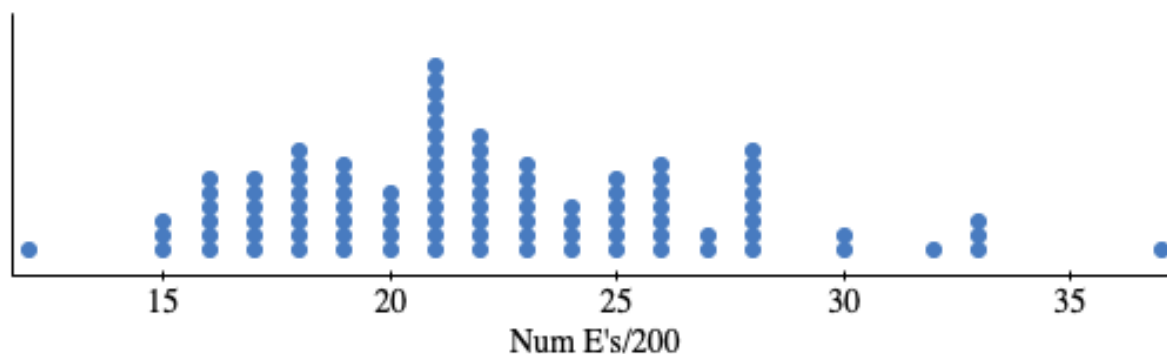# AP Stats Mock Exam
2020

1. The letter E appears in 11% of all words in the English language. However, a curious stats student is suspicious that textbooks are different. After all, academic textbooks are not known for being readable! He takes a random sample of 200 words from one of his history textbooks. He finds that 30 out of 200 words contain the letter E.

    a) Name the procedure that the student could use to examine if his entire history textbook is different than the reported proportion.
    b) Write the appropriate hypotheses that would be used to carry out this procedure.
    c) If the 11% proportion was exactly true for this random sample of 200 words, you would expect there to be 22 words with the letter E and 178 words without. Describe why these values are important and what condition they verify for this procedure.
    d) The *P*-value for this procedure is calculated to be 0.071. Write the appropriate conclusion for this procedure. Use a significance level of 0.05. Also, interpret the *P*-value in context.
    e) If an error was made in this procedure, would it have been a Type I or Type II error? Describe a consequence of this error.

    Given access to some software, the student decides to run a simulation. The software simulates 200 words with each word having an 11% chance of containing the letter E. Then the total number of words with the letter E are counted. Below are the results of running the simulation 100 times.

Num E's/200

    f) Describe the distribution of the number of words with the letter E out of 200.
    g) Without performing calculations, do you expect the mean of this simulation to be similar to the median, less than the median, or greater than the median? Explain.
    h) Is it appropriate to say that this simulation is a simulated binomial distribution? Why or why not?
    i) How do the results of this simulation compare to the *P*-value provided in part (d)? Explain.
    j) Based on the results of this simulation, how many words would you need to see in a sample of 200 with the letter E to convince you that a book has fewer words with the letter E than the 11% that is claimed?
    k) Explain how you could use this simulation to estimate the Interquartile Range of number of words with the letter E out of 200. Include your estimate of this value in your answer.

2. Poke is a bowl filled with various sushi-like ingredients to make a type of a salad called a Poke Bowl. When customers order, they select between a large variety of choices to make their Poke Bowl taste just the way they like it best. Some ingredients are classified as Premium and cost extra when added to any bowl. Customers may choose as many (or none) of the Premium ingredients as they would like.

One Poke restaurant has tracked the number of Premium ingredients that customers chose. This tracking is important as it helps with ordering inventory and also greatly affects the average cost and profit of the orders. Here is the distribution of the number of Premium ingredients selected.

# of Premium Ingredients	Probability
0	0.42
1	0.33
2	0.14
3	0.08
4	0.03

a) What is the probability that a randomly selected customer selects two or more Premium Ingredients?
b) If a randomly selected customer does select at least one Premium Ingredient, what are the chances that he selects three or more?
c) What is the shape of this distribution? Explain how you arrived at your decision.

This distribution has a mean, $\mu = 0.97$ Premium Ingredients and a standard deviation, $\sigma = 1.072$ Premium Ingredients.

d) If 16 customers are chosen at random from all the orders taken on a certain Saturday night, describe the shape of the sampling distribution of the sample mean number of Premium Ingredients. Explain how you arrived at your decision.
e) A shift leader has suggested that she wants to use the random sample from Saturday night to plan orders for the next week. Describe a statistical problem with this decision. Also explain whether you think this decision will result in an under- or over-estimate of number of Premium Ingredients needed for the next week.
f) The shift leader isn't done working hard yet! She has calculated that Friday nights typically have 300 Poke Bowl orders. If each Premium Ingredient has a profit of $2.00, what is the expected profit for a typical Friday night? Note: You may round the values to the nearest dollar to make the calculation easier to complete without a calculator.

Finally, the shift leader takes a larger random sample. She takes each of the 50 weeks of the year and randomly selects two orders from each week.

g) Describe the sampling distribution of the sample mean number of Premium Ingredients, for this sample of size 100.
h) Name this sampling method and describe a statistical advantage to using this method instead of simply taking a simple random sample.

i) Business on Fridays, Saturdays, and Sundays is very different than during the rest of the week. Describe and name a sampling method that would adjust for this observation <u>and</u> explain a statistical advantage created by this method.

Solutions and scoring guidelines

1. The letter E appears in 11% of all words in the English language. However, a curious stats student is suspicious that textbooks are different. After all, academic textbooks are not known for being readable! He takes a random sample of 200 words from one of his history textbooks. He finds that 30 out of 200 words contain the letter E.

   a) <u>Name</u> the procedure that the student could use to examine if his entire history textbook is different than the reported proportion.

   1-sample z test for a proportion

   b) Write the appropriate hypotheses that would be used to carry out this procedure.

   Ho: p = 11%; Ha: p =/ 11%; p = % of words that contain the letter E in the entire history textbook.
   3 components: correct symbol, correct direction (2-sided), definition of parameter

   c) If the 11% proportion was exactly true for this random sample of 200 words, you would expect there to be 22 words with the letter E and 178 words without. Describe why these values are important and what condition they verify for this procedure.

   Because 22 and 178 are greater than 10, we know that the sample size is sufficient for the sampling distribution of the sample proportion to be approximately normal.
   Note: E cannot be earned without use of the phrase "sampling distribution"

   d) The *P*-value for this procedure is calculated to be 0.071. Write the appropriate conclusion for this procedure. Use a significance level of 0.05. <u>Also</u>, interpret the *P*-value in context

   Since 0.071 > 0.05, we fail to reject Ho. We fail to find significant evidence that the true percentage of words that contain the letter E in this history book is different than 11%.
   If 11% of the words in this text contained the letter E, then we would observe a statistic like this (30/200), or more extreme, 7.1% of the time just by chance.

   e) If an error was made in this procedure, would it have been a Type I or Type II error? Describe this error.

   Type II. We may have concluded that this book is not different than other books, when in fact it really is.

   Given access to some software, the student decides to run a simulation. The software simulates 200 words with each word having an 11% chance of containing the letter E. Then the total number of words with the letter E are counted. Below are the results of running the simulation 100 times.

Num E's/200
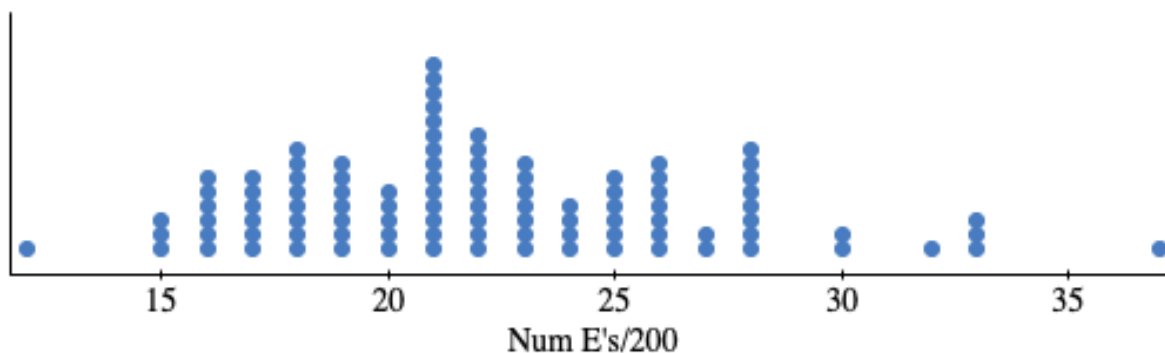
f) Describe the distribution of the number of words with the letter E out of 200.
The distribution of the number of words that have the letter E out of 200 is roughly symmetrical, is centered around 21 to 23 and has a spread from about 12 to 37.
Note: describing the shape as approximately normal is acceptable, as per part (c) above.
Center, shape, spread, and context required for E.

g) Without performing calculations, do you expect the mean of this simulation to be similar to the median, less than the median, or greater than the median? Explain.
Because this distribution is symmetrical (approximately normal, even!) we would expect the mean and median to be very close to each other, around 22.
Note: A response that states too strongly that the mean and median are exactly the same cannot receive an E.

h) Is it appropriate to say that this simulation is a simulated binomial distribution? Why or why not?
Yes. We have success or failure (words contain E or don't), it is believed that 11% probability is fixed throughout, the "words" are selected randomly in the simulation, so they are independent, and we are counting the number of successes out of a fixed number of trials (200).
Note: Any response lacking context will be deducted by one level. Responses with 2 or 3 of the 4 components receive a P.

i) How do the results of this simulation compare to the *P*-value provided in part (d)? Explain.
The simulated two-sided P-value is 8%. This is very close to the P-value of 7.1%. By the law of large numbers we expect these results to be reasonably close.
Note: A simulated 1-sided P-value of 7% receives a score of P. Yes, this is tricky!

j) Based on the results of this simulation, how many words would you need to see in a sample of 200 with the letter E to convince you that a book has fewer words with the letter E than the 11% that is claimed?
In order to get a P-value less than 5%, one would need to observe 15 words or less. Thus 15 or less than 15/200 would cause suspicion that the percentage is less than 11%.
Note: If a different, reasonable significance level is used, that is acceptable. But the answer should explicitly or implicitly use a reasonable alpha in the conclusion.

k) Explain how you could <u>use this simulation</u> to estimate the Interquartile Range of number of words with the letter E out of 200. Include your estimate of this value in your answer.
Counting in from the max and the min 25 values, we estimate that Q1 is about 18 and Q3 is about 26, giving an IQR of 8 words.

Solutions and scoring guidelines

2. Poke is a bowl filled with various sushi-like ingredients to make a type of a salad called a Poke Bowl. When customers order, they select between a large variety of choices to make their Poke Bowl taste just the way they like it best. Some ingredients are classified as 2 Premium and cost extra when added to any bowl. Customers may choose as many (or none) of the Premium ingredients as they would like.

   One Poke restaurant has tracked the number of Premium ingredients that customers chose. This tracking is important as it helps with ordering inventory and also greatly affects the average cost and profit of the orders. Here is the distribution of the number of Premium ingredients selected.

# of Premium Ingredients	Probability
0	0.42
1	0.33
2	0.14
3	0.08
4	0.03

   a) What is the probability that a randomly selected customer selects two or more Premium Ingredients?
   0.14 + 0.08 + 0.03 = 0.25

   b) If a randomly selected customer does select at least one Premium Ingredient, what are the chances that he selects three or more?
   0.11/0.58 = 19.0%

   c) What is the shape of this distribution? Explain how you arrived at your decision.
   The distribution of of the number of Premium Ingredients is skewed right. We see that the higher numbers are chosen less frequently and the lower numbers are chosen more frequently.
   Note: Answers must be in context and clearly explain that higher values are less frequent to receive an E.

   This distribution has a mean, μ = 0.97 Premium Ingredients and a standard deviation, σ = 1.072 Premium Ingredients.

   d) If 16 customers are chosen at random from all the orders taken on a certain Saturday night, describe the shape of the sampling distribution of the sample mean number of Premium Ingredients. Explain how you arrived at your decision.
   Because the population is skewed to the right and because 16 is small (< 30), the sampling distribution of this sample mean will also be skewed right.
   Note: If the provided answer is (approx) normal, the score is I.

e) A shift leader has suggested that she wants to use the random sample from Saturday night to plan orders for the next week. Describe a statistical problem with this decision. Also explain whether you think this decision will result in an under- or over-estimate of number of Premium Ingredients needed for the next week.

We would expect that Saturday nights are different and that people are in the mood to celebrate and spend extra money. Therefore, a sample from Saturday night will probably be higher than the rest of the week and will over-estimate the number of Premium Ingredients needed.

Note: Correct answers should include a reasonable characteristic of how Saturday night customers are different.

f) The shift leader isn't done working hard yet! She has calculated that Friday nights typically have 300 Poke Bowl orders. If each Premium Ingredient has a profit of $2.00, what is the expected profit for a typical Friday night? Note: You may round the values to the nearest dollar to make the calculation easier to complete without a calculator.

*$1*2*300 = $600; $582 is the exact answer.*

Note: $ required for full credit.

Finally, the shift leader takes a larger random sample. She takes each of the 50 weeks of the year and randomly selects two orders from each week.

g) Describe the sampling distribution of the sample mean number of Premium Ingredients, for this sample of size 100.

The mean is 0.97 Premium Ingr.; sd = 1.072/10 = 0.1072 Prem Ing; shape is approx normal.

Note: 4 components: center, spread, and shape. Shape must included approx.

h) Name this sampling method <u>and</u> describe a statistical advantage to using this method instead of simply taking a simple random sample.

This is a stratified random sample where the strata are the weeks. Because the weeks of the year are different (Decembers are probably busier and people spend more, for example), the stratified sample makes sure to account for the week-to-week variability. A simple random sample might accidentally result in a sample that is not evenly spread throughout the year and may not include the full variation in a year.

i) Business on Fridays, Saturdays, and Sundays is very different than during the rest of the week. Describe and name a sampling method that would adjust for this observation <u>and</u> explain a statistical advantage created by this method.

A stratified random sample would be important. Then the sample would be <u>sure</u> to include some weekend days and some weekdays.

Teacher Notes:

This exam was configured with lots parts based on Trevor Packer's increasingly strong assertions that students will be writing until time is completed and will have have time to finish the entire exam.

What is a passing score? I tell my students to give themselves 1 point for every E and 1/2 point for every P. If they get half the points, they're passing. Three-fourths of the points for a 5. That has historically worked for me. Hopefully this year is no different.

I weighted this practice exam a bit heavier towards probability. It has been stated recently that probability provides a strong predictor for student success on the exam. And Trevor Packer said something about using stronger predictors due to the shortened exam.

I don't have any inside information. These are guesses as to the nature of the exam. As my friend Corey says, the best way to help students pass the AP Statistics exam is to help students understand statistics. My biggest goal was to write questions where students need to demonstrate statistical understanding in order to answer the questions.

As much as possible, I avoided questions where the answer could be given using a memorized phrase. This proved to be a difficult goal!

My thanks to Jeff Eicher for editorial input and for some of the ideas he gave me for #2 (he used the old # of pets FRQ). It was Jeff's idea that an 11/9 split is exactly 55/45 and would really keep kids writing the <u>whole</u> time.