Midterm 2 Review

Probability Distributions

A probability distribution is just a way to organize and show all possible outcomes and the chance of them occurring.

Sampling without replacement is the most common type of discrete probability distribution tested.

Creating a Probability Distribution

Step 1: Write a table with X on top and P(X) on the bottom.
Step 2: Fill in the remaining the possible outcomes X.
Step 3: Determine the probability of each outcome X and fill them into the table.
(Pay careful attention to whether you are dealing with sampling with or without replacement!)

Ex: A junk drawer in your house contains 15 old batteries, 5 of which are totally dead. You need two batteries for your remote control, so you select two batteries from the drawer.
Give a table with the probability distribution for X = the number of good batteries you get out of the drawer.

X	
P(X)	

Calculating the Mean and Standard Deviation for a Discrete Probability Distribution

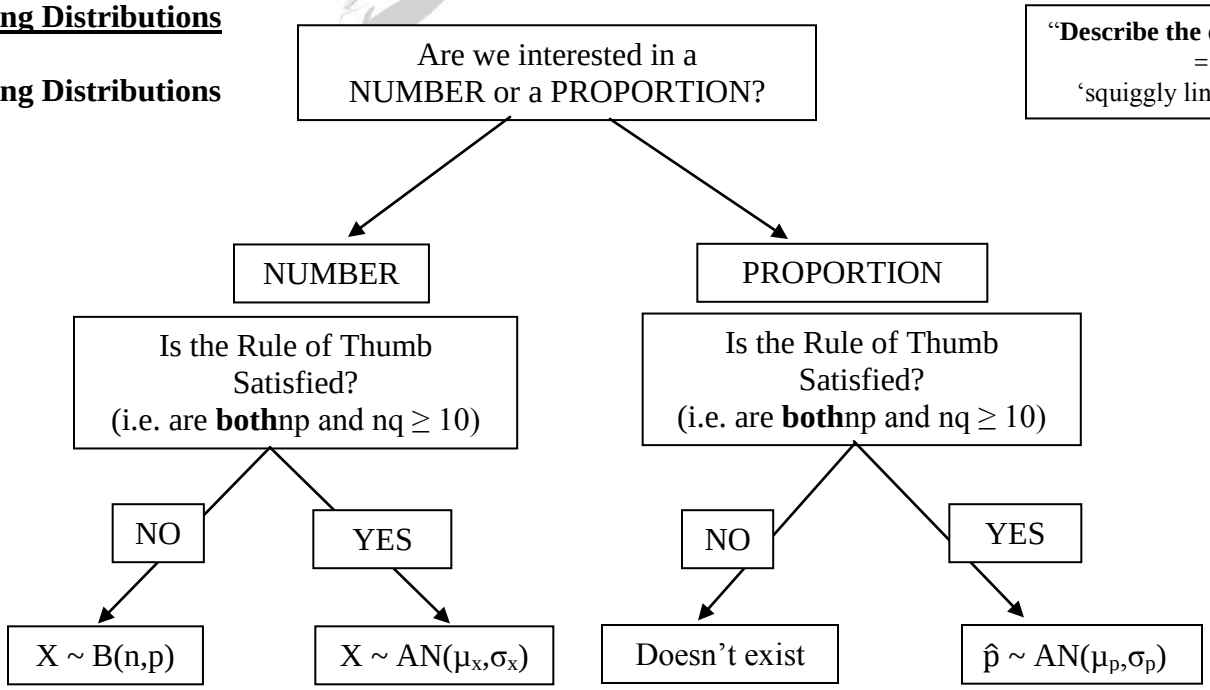
Step1: Enter all the X values in L1 and all the probabilities (as decimals) in L2.
Step 2: 1-VAR-STATS L1, L2 ** full details under calculator shortcuts on stat119review.com

Ex: Calculate the mean and standard deviation for the probability distribution we created above.

Sampling Distributions

Sampling Distributions

"**Describe the distribution**" = 'squiggly line' notation



Ex: Identifying Sampling Distributions

1. Consider rolling a die 20 times and recording the number of 2's rolled. What is sampling distribution of X, the number of 2's rolled?

Remember: "**proportion**" or "**number**"

2.An airline, knowing that about 5% of passengers fail to show up for flights, overbooks (sells more tickets than there are seats). Suppose the airline sells 275 seats for a flight. What is the sampling distribution of the number of passengers that will fail to show up?

For "number" questions, check the **Rule of Thumb**!

3. The Harvard College Alcohol Study finds that 67% of college students support efforts to "crack down on underage drinking." The administration of a college surveys 100 students and finds that 62% support a crackdown on underage drinking. Describe the distribution of the sample proportion of students who support a crackdown on underage drinking.

Be prepared for a question about assumptions or requirements!

Requirements & Assumptions of the Distributions

Properties of the Binomial Distribution

1. There are a fixed number (n) trials or observations.
2. The trials or observations are independent.
3. There are only two outcomes: Outcome of interest (success) and its complement (failure).
4. The probability of success is the same for each trial.

Properties of the Approximate Normal Distribution for a NUMBER

1. The above properties for a binomial must be satisfied.
2. The Rule of Thumb must be satisfied (np and nq $\geq$ 10).
3. Sample must come from SRS.

Properties of the Approximate Normal Distribution for a PROPORTION

1. Sampled values must be independent.
2. The Rule of Thumb must be satisfied (np and nq $\geq$ 10).
3. Sample must come from SRS.
4. Sample must be < 10% of population.

Same for inference except we don't know p, so we use p-hat in our Rule of Thumb!

Ex: There are approximately 1200 students enrolled in Stat119. To estimate the proportion of Stat119 students who voted in the election, a simple random sample of 200 students was taken from all students enrolled and it was found that 90 students voted.Are the necessary conditions met to create a confidence interval?

A. Yes, all of the requirements have been met.
B. No, $n\hat{q}$ is less than 10.
C. No, n is greater than 10% of the population.
D. No, the sample is not a simple random sample.

Ex: All of the following are requirements for using the normal approximation except:

A. Data must be collected using a simple random sample.
B. The sample size must be at least 10% of the population size.
C. np and n(1-p) must be greater than or equal to 10.
D. The sampled values must be independent from one another.

Properties of a Normal or Approximately Normal Distribution

There normally aren't many questions about the properties of the normal distribution, but you should know that its values follow a*symmetricbell-curve*.

Standardizing changes the mean to 0 and the standard deviation to 1. The shape of the distribution is not changed.

Determining Question Types

Before starting any question, you should determine what type of question it is. To help, here is a list of types of questions on Exam 2, not including those for inferential statistics (confidence intervals, hypothesis testing and sample size calculations) For greater detail, see the 'Sampling Distribution' handout on stat119review.com:

Normal Questions – any question that has the word "normal" in the question stem.

Direct – given a value, asked for a probability or a percent
Inverse – given a probability or a percent, asked for a value

Number / Proportion Questions – you'll be given n (usually a sample size) and p (as a decimal or percent, possibly even as a ratio.)

Number – find a probability that less than 30 people attended a concert
Binom – if np OR nq< 10
AN – if np AND nq ≥10
Proportion – find the probability that more than 52% of people use liquid handsoap

<u>Normal Distribution</u>

If the word "normal" is anywhere in the question, you'll be doing one of these types of questions!

A good first step is determining if it's a direct or inverse question!

Z-Scores / Standard Normal

A Z-score gives us the number of standard deviations away from the mean a value is, as well as if that value is below or above the mean. You can find the Z-score for a value with the given formula:

If you "standardized" your last test score and found you had a Z-score of 2, that would mean that yourscore is 2 standard deviations above the mean. A Z-score of -1 would mean that your score was 1 standarddeviation below the mean.

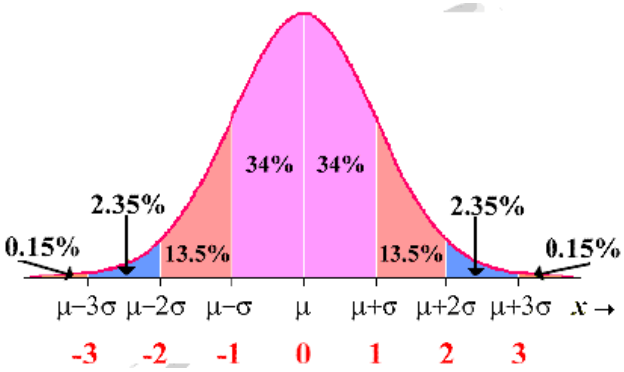
Two Types of Questions

Unusual: Scores further away from zero are more unusual, regardless of sign.
Better: Be careful about direction if asking who performed "better" – for a test, a higher Z-score would be better since scoring higher on the test is better; however, for a race, a lower Z-score is better since a smaller timeis actually a better performance.

The Empirical Rule

The Empirical Rule provides an estimate of the percent of data falling between certain values on a normal curve. Do NOT use the Empirical Rule unless the question expressly states to solve using the Empirical Rule.

Memorize the FOUR values corresponding to the amounts inside each section of half the curve.



Step 1: Draw your curve. Start with a line in the middle for your mean and then three additional lines to either side. Add standard deviations to the mean to get the values to the right. Subtract to get the values to the left. Then fill in the FOUR insidepercentages.

Step 2: Add up the percentages that correspond to portion of the curve you're being asked about.

Hints and Tips:

- If you really hate the Empirical Rule (I do!), then you could always solve these problems as if they were a regular normal problem. Your answer will be just slightly off from the estimate provided by the Empirical Rule. (Only works on multiple choice.)
- If you are looking for values that aren"t the ones you"ve written at the bottom of the curve, you"ve done something wrong. Either you"re using the Empirical Rule when the question didn"t say to OR you made a simple math error when constructing your curve.
- For questions that give you a percentage and ask what the value is, first determine if they are talking about a value on the right side or the left side, then tackle it in the forward direction using the three values on that side of the curve.
- You may still be asked a question about the other three values: 68-95-99.7%, but it would only be in a multiple choice or true/false question.

Ex: True or false. Approximately 95% of the data in a normal distribution falls within 2 standard deviationsof the mean.

Ex: Draw out the normal curve, using the Empirical rule, for a normally distributed variable with a mean of 100 and a standard deviation of 15.

Now, find the percent of values between 85 and 130.

And the percent of values greater than 145.

Find value such that 16% are higher.

Direct Calculations

This is when we're given a value and asked to find a probability or percent above or below that value (or between two values). Normally this is as easy as looking at the question, did they ask you: "what is theprobability…" or "what percent"?

These problems boil down to the following sequence:

$$X \rightarrow Z \rightarrow \%$$

Step 0: Draw your curve! This stops you from choosing the opposite area. There will be a keyword letting you know which side you're interested in.
Step 1: $\rightarrow Z$: Convert the given value to a Z-score using the **direct** formula in the formula sheet.
Step 2: $\rightarrow$ %: Convert that Z into a probability (area under the curve) by looking it up in the Z-table.Remembering that your table gives out the left side probability.

Calculator: normalcdf(lower, upper, mean, sd)

Ex: A pasta manufacturer has a machine that fills the boxes. The boxes are labeled "12 ounces" so the company wants to have that much pasta in each box. To prevent underweight boxes the manufacturer sets the mean fill higher than 12 ounces. Suppose the amount of pasta in the boxes follows a normal distribution with a mean fill of 12.4 ounces and a standard deviation of 0.24 ounces.

A. What percent of boxes weigh less than 12 ounces (i.e. are underweight)?

Draw your curve. It will save you from wrong answer traps on multiple choice questions!

B. What percent of boxes weigh more than 12.2 ounces?

C. What is the probability that a randomly selected box weighs between 12.5 and 12.8 ounces?

On the between questions, you will get two Z-scores, look them both up and subtract the smaller one fromthe larger one.

Inverse Calculations

If at any point in the question they give you a percentage, percentile or quartile, you're doing an inverse question.These questions will usually also ask for a value.

Do not get these confused with the sample proportion questions! You need the word NORMAL before you even consider an inverse question.

These problems boil down to the following sequence:

$$\% \rightarrow Z \rightarrow X$$

Step 0: Draw your curve! This stops you from choosing the opposite value. There will be a keyword letting you know which side you're interested in.
Step 1: $\% \rightarrow Z$: Convert the given area to a Z-score using**inv**Norm in your calculator.Beware,invNorm takes the left side probability!
Step 2: $\rightarrow X$: Convert that Z into Xusing the **inverse** formula on your formula sheet .

Calculator: invNorm(left area/probability, mean, sd)

There will definitely be one of the inverse questions where you need to solve for the mean or standard deviation. This requires you to use the steps, there is no calculator shortcut!

Ex: The diameters of a certain airplane tire (for landing gear) are normally distributed with an average of 24.2 inches, and a standard deviation of 0.15 inch.

A. What is the probability that the tire's diameter is above 23.98 inches?

B. How large would the diameter need to be for it to qualify as one of the 8% largest diameters?

C. What is the third quartile for these tire diameters?

Ex:According to the 2008 CIA World Factbook, the country with the world's longest life expectancy is Macau. Assume life expectancy is normally distributed and that 75% of people from Macau live to be at least 80 years old. If the standard deviation is 5 years, what is the mean life expectancy in Macau?

Number Questions

If you're given n and p, then asked about a **number**, you're doing one of these types of questions.

Your first step needs to be to check the Rule of Thumb, to see if you're doing a BINOMIAL or APPROX NORMAL question. You can get answers using the wrong method, but they'll be wrong answers…

Binomial Questions

You'll see two probability questions on the binomial on Midterm 2 – probably an exact question and one inequality question.

Exact Questions – will ask the probability that a single number is our outcome.

Calculator: binom**p**df(n, p, k) calculates the probability of getting exactly k successes.

Note: If the binomial question is a free response, you are expected to write the formula from the formula sheet (withvalues input) even if you're using binompdf.

InequalityQuestions – will ask the probability that our outcome is part of a range of values

Step 1: Write out all possible outcomes: 0, 1, 2…. N
Step 2: Circle the outcomes that we're interested in.
Step 3: Use either binompdf to add up all outcomes we're interested in OR use binomcdf.

Calculator: binom**c**df(n, p, k) adds up the probabilities from 0 up to and including k.

Note: It may be necessary to use the complement if it's easier to calculate. Be careful. Writing the list of outcomes will make sure you don't make a mistake.

Ex: Assume that 15% of people in the US are left-handed. If 10 people are selected at random (we can assume independence), find each of the following probabilities:
Probability that exactly 3 of the people selected are left-handed.

Probability that at least 3 people selected are left-handed.

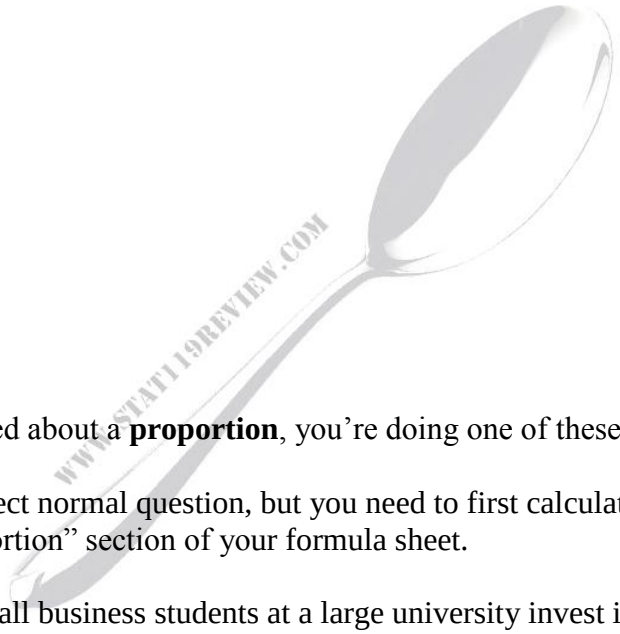
Probability that at most 8 people selected are left handed.

Approximate Normal Questions

These are the same as direct normal question, but you need to first calculate the mean and standard deviation from the "number" section of your formula sheet.

Ex. Assume that 15% of people in the US are left-handed. If 150 people are selected at random (we can assume independence), find each of the following probabilities:
A. Probability that at least 35 of the people selected are left-handed.

B. Probability that between 20 and 30 people selected are left-handed.



Proportion Questions

If you're given n and p, then asked about a **proportion**, you're doing one of these types of questions.

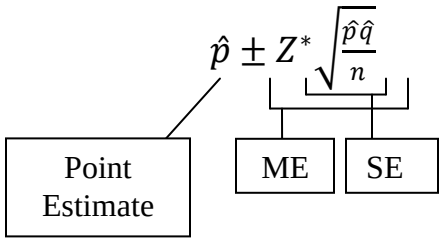
These are the same as direct normal question, but you need to first calculate the mean and standard deviation from the "proportion" section of your formula sheet.

Ex. Assume that 25% of all business students at a large university invest in the stock market. A random sample of 110 business students is selected from this university. What is the probability that more than 34% of this sample invests in the stock market?

Confidence Intervals

Formula

Remember, your formula decomposes into the parts of the confidence interval:



Finding Confidence Intervals

Plug into the formula and solve. Z* can be found on the t-table.

Calculator: 1-PropZInt

Note: Only works if given X, n. Can't use if we're given $\hat{p}$ and n.

Ex: In a random sample of 310 adults, 105 reported that they do not drink alcohol. Find 90% confidence interval for the true proportion of adults who do not drink alcohol.

Ex.A researcher is interested in the true proportion of adults with high blood pressure. A random sample of 73 adults finds 9 with high blood pressure. What is the standard error of our estimate?

Interpreting Confidence Intervals

We can be ____% (confidence level) confident that the true proportion of _____ (proportion we're looking for) is between _____ (lower confidence bound) and _____ (upper confidence bound).

Ex: Interpret interval obtained above.

Finding Values from Given Confidence Interval

Given a confidence interval: (lower, upper), you can obtain the point estimate and margin.

$$\hat{p} = \text{average} = \frac{\text{upper} + \text{lower}}{2}$$

$$\text{ME} = \text{half of width (range)} = \frac{\text{upper} - \text{lower}}{2}$$

Ex: A manufacturer creates the following 95% confidence interval for the true proportion of individuals who own a hybrid car: (0.47, 0.63). Find the point estimate and margin of error.

Ex: Which of the following will decrease the width of our confidence interval?

I. Decreasing the confidence level
II. Increasing the sample size
III. Decreasing the sample size

A. II only B. III only C. I and II D. I and III.

Sample Size

In questions asking "how many" people should be included in our sample, use the sample size formula.

- Remember, if a previous proportion or estimate isn't given, use $\hat{p} = 0.5$!

Ex: How many Americans would we need to sample in order to estimate the support for a particular ballot measure within 3% at a 95% confidence?

Ex: If we increased our confidence level, would that increase or decrease the required sample size?

Ex: If we increased our margin of error, would our required sample size increase or decrease?

Hypothesis Testing

Step 1: Identify Hypotheses

Always set H_0: $p = p_0$.

- $p_0$is the given proportion we're interested in testing.
- Usually easiest to determine $\hat{p}$, which means p_0 is the other proportion mentioned.
  - $\hat{p}$is either mentioned in the same sentence as the sample size, or we calculate it using x/n.

Read prompt to determine the alternative hypothesis.

- H_A: $p > p_A$ (increase, greater than, more than)
- H_A: $p < p_A$ (decrease, smaller than, less than)
- H_A: $p \neq p_A$ (changed, different than, not equal to)

If the prompt has at least (≥) or at most (≤), that is the null and its complement is the alternative!

Step 2: Calculate Test Statistic

Use the formula sheet to find the standardized value (z) for the given $\hat{p}$.

Step 3: Find the p-value

Whatever your test statistic, look up the negative Z in your table. This will correspond to the using the alternative todetermine the direction to look up in your normal table.

For two-sided tests(≠), multiply your pvalue by 2.

Your p-value should be less than .5!

Step 4: Make a conclusion

Memorize!!

- pvalue $\leq \alpha$ RTN!
- pvalue$>\alpha$ FTRN☹

Always remember, the **smaller** your pvalue the **more** evidence you have against the null.

Step 5: Interpret your decision

RTN: There is enough evidence to conclude the true proportion of _____ (whatever we're testing) is _____ (less than, greater than, not equal to – depends on alternative) ___ (p_0).

FTRN: There is **NOT** enough evidence to conclude the true proportion of _____ (whatever we're testing) is _____ (less than, greater than, not equal to – depends on alternative) ___ (p_0).

Remember, in multiple choice questions, you should approach any hypothesis testing questions as if you were doing a free response!

The interpretation should end with the ALTERNATIVE HYPOTHESIS. We cannot make any definitive conclusion about the null hypothesis.

Duality of Confidence Intervals and Hypothesis Testing

Two things to remember:

- The rule for decisions based on confidence intervals:
  - If our null value is OUTSIDE the confidence interval, we RTN.
  - If our null value is INSIDE the confidence interval, we FTRN.
- The relationship between α and the confidence level:
  - (1-α)100% = CL (i.e. 95% confidence level is equivalent to α = 0.05)

Note: You can make any question about confidence intervals and hypothesis testing easier by putting everything in the same context – I immediately change everything to hypothesis testing.

Ex.It has been reported that 30% of newly hired MBAs are confronted with unethical business practices during their first year of employment. One business school dean wondered if her MBA graduates had similar experiences. She surveyed recent graduates from her school's MBA program and found that 47 out of a random sample of 135 graduates had encountered unethical business practices in the past year. Is this sufficient evidence to conclude that the report of 30% being confronted with unethical business practices is incorrect? Use a level of significance of 0.05.

a.) State the appropriate null and alternative hypotheses.

b.) Calculate the test statistic.

c.) Calculate the corresponding p-value.

d.) Make a statistical decision using the level of significance of 0.05. Justify your decision.

e.) Interpret your decision in the context of this problem.

Ex: Suppose a 95% confidence interval for the true proportion of Stat119 students who pass the course is found to be (.7582, .8210). A TA wants to test the hypotheses H0: $p = .75$ and HA: $p \neq .75$. Which of the following can we conclude?

A. He would fail to reject the null at a significance level of 0.1
B. He would reject the null at a significance level of 0.1
C. He would fail to reject the null at a significance level of 0.05
D. He would reject the null at a significance level of 0.01

Ex: A 90% confidence interval of the proportion of migraine sufferers who use prescription medication is calculated to be (0.45,0.49). A company claims that 50% of migraine sufferers use prescription medication. Which of the following is correct interpretation?

A. There is enough evidence to conclude the company's claim is correct.
B. There is not enough evidence to conclude the company's claim is correct.
C. There is enough evidence to conclude the company's claim is incorrect.
D. There is not enough evidence to conclude the company's claim is incorrect.

Ex: In testing the hypotheses $H_0: p = 0.7$ $H_A: p \neq 0.7$, the test statistic is found to be 1.24. Which is the following is the correct p-value?

A. 1.7850 B. 0.2150 C. 0.8925 D. 0.1075 E. None of these

Statistical Significance

Things to know about statistical significance:

- Statistical significance is equivalent to rejecting the null.
- We have statistical significance when the difference between our sample proportion and null proportion is large enough that it could not have happened by chance.
- This is very different than practical significance! By taking a very large sample, I can usually make any difference statistically significant. However, if the magnitude of that difference is very small, it may not be meaningful in reality.
    - Consider a researcher who sampled 2000 sharks of a particular type to determine that they weigh, on average, 1 pound less than another type of shark. His findings may be statistically significant, but the difference is so small that it is relatively pointless.

Errors in Hypothesis Testing

RTN	FTRN
statistically significant p-value $\leq \alpha$ "There is enough evidence…" Type I error Power	p-value > α "There is not enough evidence…" Type II error

Remember, errors are errors because we made the wrong decision!

		Decision	
		FTRN (Null)	RTN (Alternative)
Truth	Null		
	Alternative		

Type I : $H_0: truth$ $H_A: decision$
Type II : $H_0: decision$ $H_A: truth$
Power : $H_0: -$ $H_A: decision\ \&\ truth$

Ex: Imagine a hypothesis test in the context of a pregnancy test, assume the null hypothesis is that the woman is NOT pregnant. Which of the following correctly describes a type II error in this example?

A. The pregnancy test says the woman is pregnant when in fact she is.
B. The pregnancy test says the woman is not pregnant when in fact she is.
C. The pregnancy test says the woman is pregnant when in fact she is not.
D. The pregnancy test says the woman is not pregnant when in fact she is not.

Ex: If you reduce your significance level, how does it affect Type I error, Type II error and power?

Ex: How can you reduce both Type I and Type II error? What does this do to power?

Definition of P-value

The definition of p-value is another conditional probability: Probability of getting a sample value as or more extreme than that obtained in our sample (or a test statistic as or more extreme than the one obtained from our sample), GIVEN that the true proportion is actually the null proportion (i.e. the null hypothesis is true.)

Ex: A statistician is interested in testing if the proportion of applicants to a particular university who are admitted is less than 20%. He takes a random sample of 170 applicants and finds 18% were admitted. Which of the following probability notations correctly demonstrates the definition of a p-value?

A. $P(\hat{p} < 0.20 \mid p = 0.20)$
B. $P(\hat{p} < 0.20 \mid p = 0.18)$
C. $P(p < 0.18 \mid \hat{p} = 0.20)$
D. $P(p < 0.18 \mid \hat{p} = 0.18)$