Name:____________________________________ **Date:**__________________

PSTAT 5A: Homework 03

Summer Session A 2023, with *Ethan P. Marzban*

Please make sure to periodically work on these problems throughout the week, so that you are well-prepared for the quiz on Friday!

1. Consider the random variable X with the following probability mass function (p.m.f.):

k	-2.4	-0.2	0	4
$\mathbb{P}(X = k)$	0.1	0.4	a	0.2

where a is an as-of-yet unknown constant.

(a) What is the value of a?

Solution: We know that the probability values in a p.m.f. must sum to 1. This means

$$0.1 + 0.4 + a + 0.2 = 1 \implies a = 1 - (0.1 + 0.4 + 0.2) = 0.3$$

(b) What is $\mathbb{P}(-2.4 \leq X < -0.2)$?

Solution:

$$\mathbb{P}(-2.4 \leq X < -0.2) = \mathbb{P}(X = -2.4) = 0.1$$

(c) What is $\mathbb{P}(X \geq 0)$?

Solution:

$$\mathbb{P}(X \geq 0) = \mathbb{P}(X = 0) = \mathbb{P}(X = 4) = 0.3 + 0.2 = 0.5$$

Or, using the complement rule,

$$\mathbb{P}(X \geq 0) = 1 - \mathbb{P}(X < 0) = 1 - [\mathbb{P}(X = -2.4) + \mathbb{P}(X = -0.2)] = 1 - [0.1 + 0.4] = 0.5$$

(d) If $F_X(x)$ denotes the cumulative distribution function (c.d.f.) of X at x, what is the value of $F_X(0)$?

Solution:

$$F_X(0) = \mathbb{P}(X \leq 0) = 1 - \mathbb{P}(X > 0) = 1 - \mathbb{P}(X = 4) = 1 - 0.2 = 0.8$$

(e) What is $\mathbb{E}[X]$?

Solution:

$$\begin{aligned}\mathbb{E}[X] &= \sum_{\text{all } k} k \cdot \mathbb{P}(X = k) \\ &= (-2.4) \cdot \mathbb{P}(X = -2.4) + (-0.2) \cdot \mathbb{P}(X = -0.2) + (0) \cdot \mathbb{P}(X = 0) + (4) \cdot \mathbb{P}(X = 4) \\ &= (-2.4) \cdot (0.1) + (-0.2) \cdot (0.4) + (0) \cdot (0.3) + (4) \cdot (0.2) = 0.48\end{aligned}$$

(f) What is $\mathrm{Var}(X)$?

Solution: If we use the second formula for variance, we first compute

$$\begin{aligned}\sum_{\text{all } k} k^2 \cdot \mathbb{P}(X = k) &= (-2.4)^2 \cdot \mathbb{P}(X = -2.4) + (-0.2)^2 \cdot \mathbb{P}(X = -0.2) \\ &\qquad + (0)^2 \cdot \mathbb{P}(X = 0) + (4)^2 \cdot \mathbb{P}(X = 4) \\ &= (-2.4)^2 \cdot (0.1) + (-0.2)^2 \cdot (0.4) + (0)^2 \cdot (0.3) + (4)^2 \cdot (0.2) = 3.792\end{aligned}$$

and so

$$\mathrm{Var}(X) = \left(\sum_{\text{all } k} k \cdot \mathbb{P}(X = k) \right) - (\mathbb{E}[X])^2 = 3.792 - (0.48)^2 = 3.5616$$

If we instead used the first formula for variance, we compute

$$\begin{aligned}\mathrm{Var}(X) &= \sum_{\text{all } k} (k - \mathbb{E}[X])^2 \cdot \mathbb{P}(X = k) \\ &= (-2.4 - 0.48)^2 \cdot \mathbb{P}(X = -2.4) + (-0.2 - 0.48)^2 \cdot \mathbb{P}(X = -0.2) \\ &\qquad + (0 - 0.48)^2 \cdot \mathbb{P}(X = 0) + (4 - 0.48)^2 \cdot \mathbb{P}(X = 4) \\ &= (-2.4 - 0.48)^2 \cdot (0.1) + (-0.2 - 0.48)^2 \cdot (0.4) + (0 - 0.48)^2 \cdot (0.3) \\ &\qquad + (4 - 0.48)^2 \cdot (0.2) \\ &= 3.5616\end{aligned}$$

2. In a large parking lot, 35% of cars are electric vehicles. A random sample of 200 cars is taken with replacement from this lot, and the number of electric vehicles in the sample is recorded.

(a) Define the random variable of interest and call it X.

Name:______________________________ **Date:**________________

Solution: Let X denote the number of electric vehicles in a random sample of 200 cars.

(b) Identify the distribution of X, taking care to list out any/all parameter(s)! Also, be sure to check and relevant conditions.

Solution: We have a suspicion that X is binomially distributed; to verify this, we check the 4 Binomial Criteria:

1) Independent Trials? Yes, since sampling is done without replacement.

2) Fixed Number of Trials? Yes; $n = 200$

3) Well-defined notion of 'success'? Yes; 'success' = 'electric vehicle'

4) Fixed probabiliy of success? Yes; $p = 0.35$.

Since all four conditions are satisfied, we conclude $X \sim \text{Bin}(200,\ 0.35)$.

(c) What is the probability that this sample contains exactly 72 electric vehicles?

Solution:

$$\mathbb{P}(X = 100) = \binom{200}{72}(0.35)^{72}(1 - 0.35)^{200-72} \approx 0.056 = 5.6\%$$

(d) What is the probability that between 70 and 73 of the cars in this sample (inclusive on both ends) are electric vehicles?

Solution:

$$\begin{aligned}\mathbb{P}(70 \leq X \leq 73) = \mathbb{P}(X = 70) + \mathbb{P}(X = 71) + \mathbb{P}(X = 72) + \mathbb{P}(X = 73)\\ = \binom{200}{70}(0.35)^{70}(1 - 0.35)^{200-70} + \binom{200}{71}(0.35)^{71}(1 - 0.35)^{200-71}\\ + \binom{200}{72}(0.35)^{72}(1 - 0.35)^{200-72} + \binom{200}{73}(0.35)^{73}(1 - 0.35)^{200-73}\\ \approx 0.227 = 22.7\%\end{aligned}$$

(e) What is the expected number of electric vehicles we expect to observe in this sample?

Solution:

$$\mathbb{E}[X] = np = (200)(0.35) = \text{70 cars}$$

(f) What is the standard deviation of the number of electric vehicles we expect to observe in this sample?

Solution:

$$\text{SD}(X) = \sqrt{np(1-p)} = \sqrt{(200)(0.35)(1-0.35)} \approx \text{6.74 cars}$$

3. The weight of a randomly-selected fish from *Lake Gaucho* (in bounds) is normally distributed with mean 4 lbs and standard deviation 1.3 lbs. A fish is selected at random from *Lake Gaucho* and its weight is recorded.

(a) Define the random variable of interest.

Solution: Let X = the weight of a randomly-selected fish from *Lake Gaucho*.

(b) What is the probability that this fish weighs less than 2 lbs?

Solution: We know that $X \sim \mathcal{N}(4,\ 1.3)$, and we seek $\mathbb{P}(X < 2)$; thus, we first standardize and then utilize our lookup table:

$$\mathbb{P}(X < 2) = \mathbb{P}\left(\frac{X-4}{1.3} < \frac{2-4}{1.3}\right) = \mathbb{P}\left(\frac{X-4}{1.3} < -1.53\right) = 0.0618$$

(c) What is the probability that this fish weighs more than 4.8 lbs?

Solution:

$$\begin{aligned}
\mathbb{P}(X > 4.8) &= \mathbb{P}\left(\frac{X-4}{1.3} > \frac{4.8-4}{1.3}\right) \\
&= 1 - \mathbb{P}\left(\frac{X-4}{1.3} \leq \frac{4.8-4}{1.3}\right) \\
&= 1 - \mathbb{P}\left(\frac{X-4}{1.3} < 0.62\right) = 1 - 0.7324 = 0.2676
\end{aligned}$$

(d) What is the probability that this fish weighs between 2.5 lbs and 3.8 lbs?

Solution: We seek $\mathbb{P}(2.5 \leq X \leq 3.8)$. Our strategy for computing quantities like this is always to convert everything to be in terms of left-tail areas, and then standardize:

$$\begin{aligned} \mathbb{P}(2.5 \leq X \leq 3) &= \mathbb{P}(X \leq 3.8) - \mathbb{P}(X \leq 2) \\ &= \mathbb{P}\left(\frac{X-4}{1.3} < \frac{3.8-4}{1.3}\right) - \mathbb{P}\left(\frac{X-4}{1.3} < \frac{2.5-4}{1.3}\right) \\ &= \mathbb{P}\left(\frac{X-4}{1.3} < -0.15\right) - \mathbb{P}\left(\frac{X-4}{1.3} < -1.15\right) \\ &= 0.4404 - 0.1251 = 0.3153 \end{aligned}$$

(e) Suppose, now, that a random sample of 10 fish is caught (assume that the weights of fish in *Lake Gaucho* are independent), and the number of fish that weight between 2.5 and 3.8 lbs is recorded. What is the probability that exactly 3 of these fish weigh between 2.5 and 3.8 lbs? **Hint:** you will need to define another random variable.

Solution: Following the hint, we let Y denote the number of fish, in a random sample of 10 fish caught from *Lake Gaucho*, that weigh between 2.5 and 3.8 lbs. Indeed, Y follows the Binomial distribution:

1) **Independent Trials?** Yes, since we are told to assume weights of different fish are independent.
2) **Fixed Number of Trials?** Yes; $n = 10$
3) **Well-defined notion of 'success'?** Yes; 'success' = 'weighing between 2.5 and 3.8 lbs'
4) **Fixed probabiliy of success?** Yes; $p = 0.3153$, as found in part (d) above.

Therefore, we have $Y \sim \mathrm{Bin}(10,\ 0.3153)$ and so

$$\mathbb{P}(Y = 3) = \binom{10}{3}(0.3153)^3(1-0.3153)^{10-3} \approx 0.265 = 26.5\%$$

4. The duration of a flight from SBA to SEA is uniformly distributed between 150 mins and 170 mins. A flight from SBA to SEA is selected at random, and its duration is recorded.

(a) Define the random variable of interest.

Solution: Let X denote the duration of a randomly-selected flight from SBA to SEA. Then,

$X \sim \text{Unif}(150,\ 170)$.

(b) What is the expected duration of a randomly-selected flight from SBA to SEA?

Solution:

$$\mathbb{E}[X] = \frac{150 + 170}{2} = \boxed{160 \text{ mins}}$$

(c) What is the standard deviation of the duration of a randomly-selected flight from SBA to SEA?

Solution:

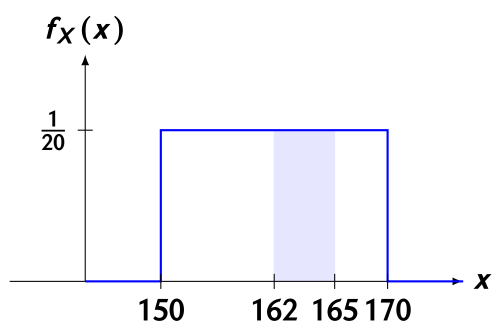
$$\text{SD}(X) = \frac{(170 - 150)}{\sqrt{12}} = \boxed{\frac{20}{\sqrt{12}} \approx 5.77 \text{ mins}}$$

(d) What is the probability that a randomly-selected flight from SBA to SEA will have a duration of 162 mins?

Solution: We seek $\mathbb{P}(X = 162)$. Since X is continuous, $\mathbb{P}(X = k) = 0$ for any value of k; hence the desired probability is $\boxed{0}$.

(e) What is the probability that a randomly-selected flight from SBA to SEA will have a duration of between 162 and 165 mins?

Solution: We seek $\mathbb{P}(162 \leq X \leq 165)$. Let's sketch a (not-to-scale) picture of the region of interest:



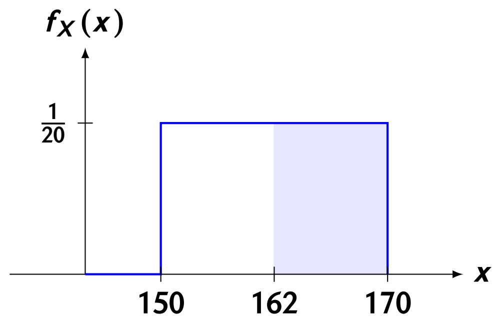
The area blue shaded region is precisely the probability we seek, which we can now see to be

$$\frac{1}{20}(165 - 162) = \boxed{\frac{3}{20}}$$

Name:____________________________________ **Date:**__________________

(f) What is the probability that a randomly-selected flight from SBA to SEA will have a duration of between 162 and 180 mins?

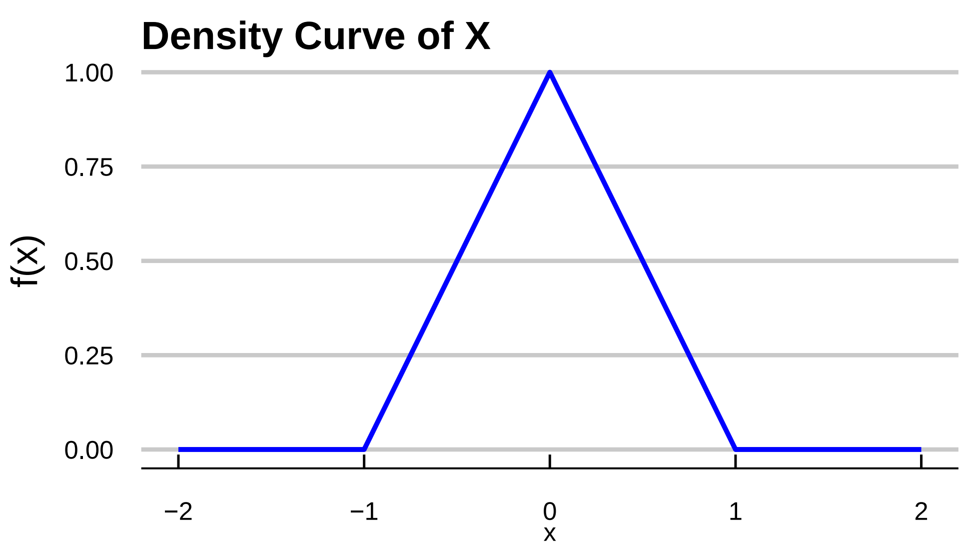
Solution: We seek $\mathbb{P}(162 \leq X \leq 180)$. Let's sketch a (not-to-scale) picture of the region of interest:



The area blue shaded region is precisely the probability we seek, which we can now see to be

$$\frac{1}{20}(170 - 162) = \frac{8}{20}$$

5. A random variable X has the following density curve:



(a) What is the state space of X?

Solution: We know that the state space S_X of a continuous random variable is the region over which the density is nonzero. From the density curve, we therefore see that

$$S_X = [-1,\ 1]$$

Name:______________________________ **Date:**_______________

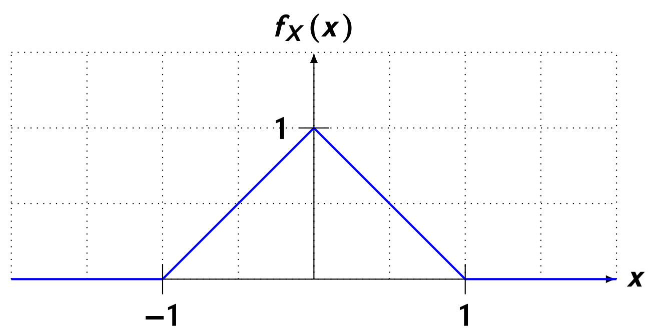
(b) Verify that $f_X(X)$ is a valid probability density function (p.d.f.) **Hint:** Recall that there are two properties a function $f_X(X)$ must satisfy in order to be a valid p.d.f..

Solution: We need only to verify that $f_X(x)$ is nonnegative everywhere, and that the area under the density curve is 1. The first condition is obviously met, just by looking at the density curve. To verify the second, we see that the area underneath the density curve is the sum of the areas of two triangles, each with base 1 and height 1: hence, the area under the density curve is

$$\frac{1}{2} \cdot (1) \cdot (1) + \frac{1}{2} \cdot (1) \cdot (1) = \frac{1}{2} + \frac{1}{2} = 1 \checkmark$$

(c) What is $\mathbb{P}(X \leq -1)$?

Solution:

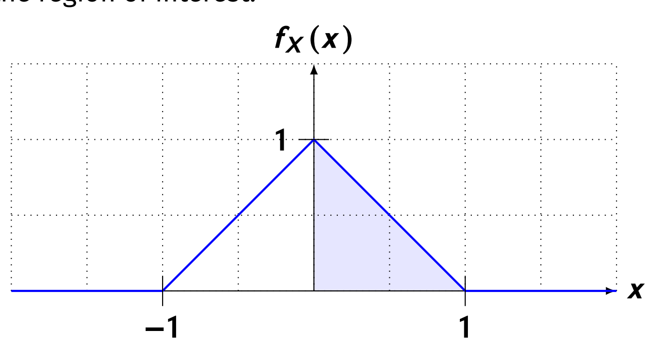


Indeed, this has area zero and hence $\boxed{\mathbb{P}(X \leq -1) = 0}$.

(d) What is $\mathbb{P}(X \geq 0)$?

Solution:

We again sketch the region of interest:



This is a triangle with base 1 and height 1, meaning its area (and, consequently, the prob-

Name:______________________ Date:______________

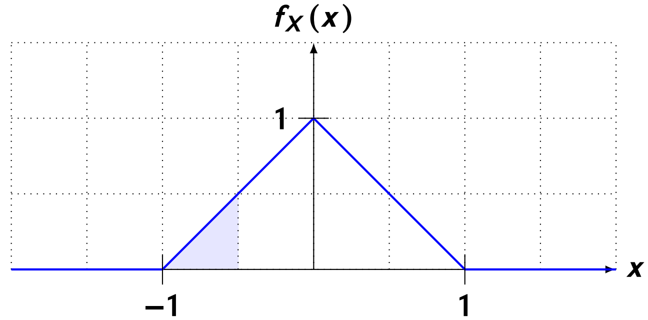
ability of interest) is

$$\frac{1}{2} \cdot (1) \cdot (1) = \boxed{\frac{1}{2}}$$

(e) What is $\mathbb{P}(-1 \leq X \leq -0.5)$?

Solution:

We again sketch the region of interest:



This is a triangle with base 0.5 and height 1, meaning its area (and, consequently, the probability of interest) is

$$\frac{1}{2} \cdot (0.5) \cdot (0.5) = \boxed{\frac{1}{8}}$$

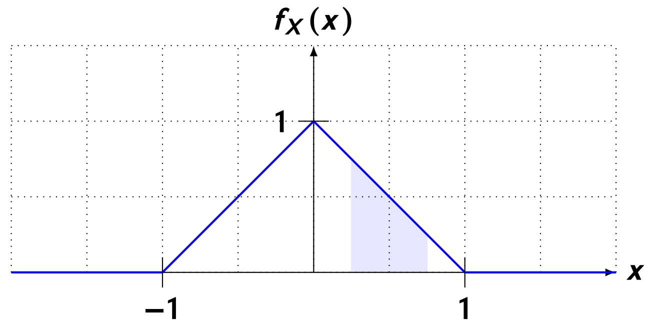
(f) What is $\mathbb{P}(-1 < X < -0.5)$?

Solution: For any continuous random variable X, we know we have $\mathbb{P}(a \leq X \leq b) = \mathbb{P}(a < X < b)$ meaning the desired probability is exactly the same as the one computed in part (e) above: $\boxed{1/4}$.

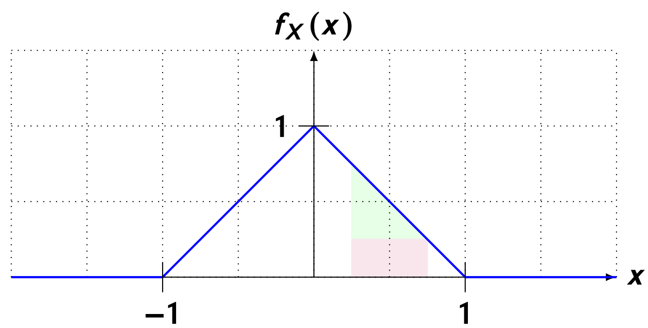
(g) What is $\mathbb{P}(0.25 \leq X \leq 0.75)$?

Solution:

We again sketch the region of interest:



There are a couple of ways we could find this area: we could try to look up the formula for the area of a trapezoid, or we could decompose the trapezoid into a rectangle and a triangle:



The area of the green triangle is

$$\frac{1}{2} \cdot (0.5) \cdot (0.5) = \frac{1}{8}$$

and the area of the purple rectangle is

$$(0.5) \cdot (0.25) = \frac{1}{8}$$

and hence, the total area is

$$\frac{1}{8} + \frac{1}{8} = \boxed{\frac{1}{4}}$$

If we instead wanted to use the formula for the area of a trapezoid, we would have obtained

$$\frac{0.25 + 0.75}{2} \cdot 0.5 = 0.25 = \frac{1}{4}$$

6. In each of the parts below, determine whether the provided quantity is a population parameter or a sample statistic. Use this to further determine whether the quantity is a deterministic (i.e. nonrandom) constant, or a random variable.

   (a) The proportion of trees in the Amazon rainforest that are mango trees.

Solution: This is a population parameter, and is therefore deterministic.

(b) The average AQI (Air Quality Index) of 25 randomly-selected cities on August 1, 2022.

Solution: This is a sample statistic, and is therefore a random variable.

(c) The longest amount of time a human can hold their breath underwater.

Solution: This is a population parameter, and is therefore deterministic.

(d) The median number of words of a sample of 100 randomly-selected books.

Solution: This is a sample statistic, and is therefore a random variable.

7. Suppose that 33% of a particular country's population has a college degree. A representative sample of 243 people is taken, and the proportion of these people who have a college degree is recorded.

(a) Define the parameter of interest, and use the notation discussed in Lecture.

Solution: Let p denote the proportion of the country's population that have a college degree.

(b) Define the random variable of interest; again use proper notation.

Solution: Let $\widehat{P}$ denote the proportion of people in a representative sample of 243 that have a college degree.

(c) Check whether the success-failure conditions are satisfied.

Solution: In this case, we know the value of p: $p = 0.33$. Additionally, $n = 243$, so we check:

1. $np = (243)(0.33) = 80.19 \geq 10$
2. $n(1-p) = 162.81 \geq 10$

We see that both conditions are satisfied.

(d) What is the probability that over 30% of the sample have college degrees?

Solution: We seek $\mathbb{P}(\widehat{P} \geq 0.3)$. By the Central Limit Theorem for Proportions (which we are able to invoke because the success-failure conditions are satisfied),

$$\widehat{P} \sim \mathcal{N}\left(0.33,\ \sqrt{\frac{(0.33)(1-0.33)}{243}}\right) \sim \mathcal{N}(0.33,\ 0.0302)$$

Therefore, we compute

$$\mathbb{P}(\widehat{P} \geq 0.3) = 1 - \mathbb{P}(\widehat{P} < 0.3) = 1 - \mathbb{P}\left(\frac{\widehat{P} - 0.33}{0.0302} < \frac{0.3 - 0.33}{0.0302}\right) = 1 - \mathbb{P}(Z \leq -0.99)$$

where $Z \sim \mathcal{N}(0,\ 1)$. From a normal table, we therefore see that the desired probability is

$$1 - 0.1611 = \boxed{0.8389 = 83.89\%}$$

(e) What is the probability that the proportion of people in the sample who have college degrees lies within 5% of the true proportion of 33%?

Solution: We now seek $\mathbb{P}(0.28 \leq \widehat{P} \leq 0.38)$, which we compute as

$$\begin{aligned}
\mathbb{P}(0.28 \leq \widehat{P} \leq 0.38) &= \mathbb{P}(\widehat{P} \leq 0.38) - \mathbb{P}(\widehat{P} \leq 0.28) \\
&= \mathbb{P}\left(\frac{\widehat{P} - 0.3}{0.0302} \leq \frac{0.38 - 0.3}{0.0302}\right) - \mathbb{P}\left(\frac{\widehat{P} - 0.3}{0.0302} \leq \frac{0.28 - 0.3}{0.0302}\right) \\
&= \mathbb{P}\left(Z \leq \frac{0.05}{0.0302}\right) - \mathbb{P}\left(Z \leq -\frac{0.05}{0.0302}\right) \\
&= \mathbb{P}\left(Z \leq 1.66\right) - \mathbb{P}\left(Z \leq -1.66\right) \\
&= 0.9515 - 0.0485 = \boxed{0.903 = 90.3\%}
\end{aligned}$$

8. The U.S. Department of Housing and Urban Development defines a person or household to be "rent-burdened" if 30% or more of the individual/household's income is spent on housing. A recent survey revealed that 42% of households in a representative sample of 150 households were rent-burdened.

(a) Define the parameter of interest.

Solution: Let p denote the proportion of households that are rent-burdened.

(b) Define the random variable of interest.

Solution: Let $\widehat{P}$ denote the proportion of housholds in a representative sample of 150 that are rent-burdened.

(c) Construct a 95% confidence interval for the true proportion of rent-burdened households, and interpret your interval in the context of the problem.

Solution: Our first task is to identify the sampling distribution of $\widehat{P}$, which entails checking the success-failure conditions. Since we don't know the value of p, we use the substitution approximation:

1) $n\widehat{p} = (150) \cdot (0.42) = 63 \geq 10$ ✓

2) $n(1 - \widehat{p}) = (150) \cdot (1 - 0.42) = 87 \geq 10$ ✓

Since both conditions are met, we can invoke the Central Limit Theorem for Proportions to conclude that $\widehat{P}$ will be approximately normally distributed. Hence, our Confidence Interval will take the form

$$\widehat{p} \pm z \cdot \sqrt{\frac{\widehat{p}(1 - \widehat{p})}{n}}$$

where z is the appropriately-selected quantile of the normal distribution. Since we want a 95% confidence interval, we take z to be negative one times the $(1 - 0.95)/2 \times 100\% = 2.5^{\text{th}}$ percentile of the standard normal distribution, which we know is around 1.96. Hence, our confidence interval is

$$\begin{aligned} (0.42) \pm (1.96) \cdot \sqrt{\frac{(0.42) \cdot (1 - 0.42)}{150}} &= (0.42) \pm (1.96) \cdot (0.0402) \\ &\approx [0.341208\,,\ 0.498792] \end{aligned}$$

The interpretation of this interval is:

> We are 95% certain that the true proportion of rent-burdened households is between 34.12% and 49.88%.

(d) Would you expect an 80% confidence interval for the true proportion of rent-burdened households to be wider or narrower than the 95% confidence interval you constructed in part (c)? Explain briefly.

Solution: We know that higher confidence levels lead to wider confidence intervals, meaning an 80% CI should be narrower than a 95% one.

Name:____________________________________ **Date:**__________________

(e) Construct an 80% confidence interval for the true proportion of rent-burdened households, andinterpret your interval in the context of the problem.

Solution: We can re-use a lot of our work from part (c) above: all we really need to change is our value of z. Now, we use negative one times the the $(1 - 0.8)/2 \times 100\% = 10^{\text{th}}$ percentile of the standard normal distribution, which from our normal table gives us a value of around 1.28. (Recall that this is equivalent to finding the 90$^{\text{th}}$ percentile, due to the symmetry of the normal distribution.) Hence, our confidence interval becomes

$$(0.42) \pm (1.28) \cdot (0.0402) \approx [0.368544\,,\ 0.471456]$$

and the interpretation of this interval is:

> We are 80% certain that the true proportion of rent-burdened households is between 36.9% and 47.1%.