Practice Exam #1

Name: _____

Directions

- Provide numeric information and show the supporting calculations whenever applicable
- Answer each question concisely, writing no more than the indicated amount
- Avoid adding superfluous or unrelated statements to your answers (if you make an incorrect statement you'll be penalized, even if you get the main part of the question correct)
- I tend to write long exams, so work quickly and don't agonize over any particular questions

Formulas

Common Distributions:

Distribution	Parameters	Expected Value	Variance
Bernoulli	p	p	$p * (1-p)$
Binomial	n, p	$n * p$	$n * p * (1-p)$
Normal	μ, σ	μ	σ^2

Summary Measures:

Measure	Formula
Mean	$\bar{x} = \frac{\sum_i x_i}{n}$
Standard Deviation	$s_x = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}$
Correlation	$r_{xy} = \frac{1}{n-1} \sum_i \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$

Conceptual Questions (15 pts)

A: (5 pts)

Sampling bias and *sampling variability* are two reasons why an estimate from a sample might not accurately reflect a characteristic of a population. Briefly explain the difference between these two concepts (1-3 sentences).

> Sampling bias describes a systematic tendenacy for samples chosen using a particular sampling procedure to produce estimates that are consistently above or below the truth. Sampling variability refers to variation in samples, that is a given sample may overestimate or underestimate the truth, but on average they will be on-target.

B: (5 pts)

In your own words, conceptually describe the main result of *Central Limit Theorem*? Try to avoid introducing any symbols or notation in your answer. (1-3 sentences)

> Central Limit Theorem says that the distribution of sample averages will be a Normal curve with a known mean and known standard deviation (called Standard Error).

C: (5 pts)

In your own words, why is *random assignment* an important aspect of a well-designed experiment? That is, what how does random assignment help reduce the number of possible explanations for an observed result. (1-3 sentences)

> Random assignment balances any potentionally confounding characteristics in the treatment and control groups. This means that such variables cannot fulfil the definition of confounding (since they're not associated with the explanatory variable), and therefore cannot impact the observed results.

Application #1 - Do Dogs Recognize Human Directives? (34 pts)

Researchers conducted an experiment involving 6 different dogs repeatedly making choices. At the start of each repetition, a dog was positioned 2.5 meters from the experimenter. The experimenter had two cups, one on each side of them. Before each trial, experimenter flipped two coins. The first determined whether they would point using either own arm, or a mechanical arm; and the second determined if they should point the cup to the left or right of them. After the experimenter pointed, the dog was then allowed to go to drink from one of the two cups. This process was repeated 24 times for each dog.

We'll focus on the results for one of the dogs, Harley. When the experimenter pointed using their own arm, Harley chose the cup that was pointed at 10 of 12 times. When the experimenter pointed using the mechanical robotic arm, Harley chose the cup that was pointed at 8 of 12 times.

A (2 pts)

Consider a data spreadsheet documenting Harley's experimental results. How many rows are in this spreadsheet? Briefly explain what each represents (short phrase).

Harley made 24 choices throughout the experiment, each cup choice is a row in this spreadsheet.

B (2 pts)

Consider a data spreadsheet documenting Harley's experimental results. How many variables are in this spreadsheet? Briefly name each as well as whether it's a categorical or numeric variable (a few short phrases)

At minimum there are two variables - mech/human (categorical) and correct/not (Categorical).
It's okay if you also recorded things like direction (left/right - categorical), etc.

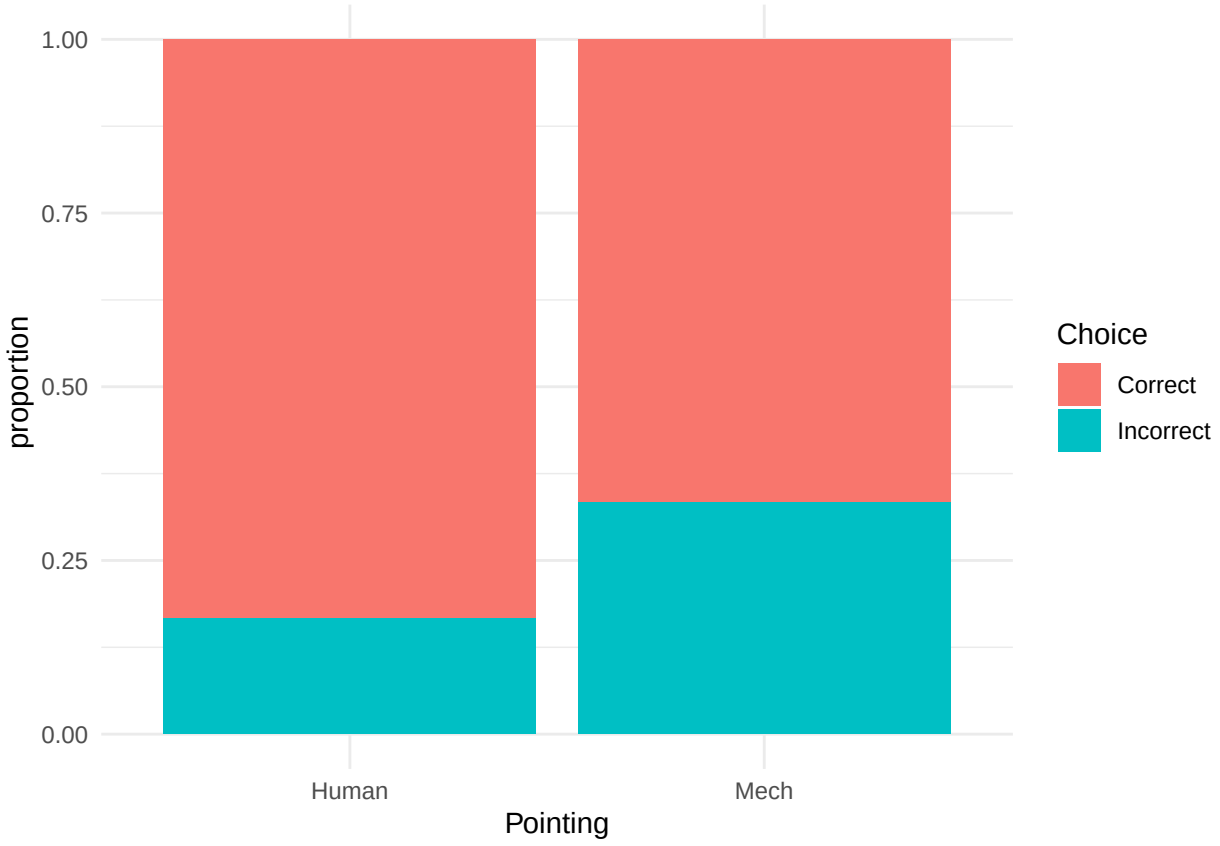
C: (2 pts)

Create a two-way frequency table (contingency table) summarizing the results of this experiment. (No explanation needed)

	Correct	Incorrect
Human Arm	10	2
Mechanical Arm	8	4

D: (3 pts)

Sketch an appropriate graph summarizing the results of this experiment (don't spend more than a minute or so on your sketch, I'm looking for the general type of graph and not the precise values in it).



E: (4 pts)

In these data, Harley was more likely to select the correct cup when directed by a human arm. Are you concerned that this outcome be explained by the presence of a confounding variable? Briefly explain. (1-2 sentences)

> I am not particularly concerned, the explanatory variable was randomized so confounding variable is unlikely (though with a smallish sample like this, it's not impossible).

F: (3 pts)

Consider *only* the instances where the experimenter pointed using their own (human) arm. Let the random variable X_i denote the i^{th} selection by Harley among these 12 instances. What is the sample space of X_i? (no explanation necessary)

> Note that X_i is a Bernoulli random variable, so $S = \{0, 1\}$; it's okay if you say something like $\{Right, Wrong\}$

G: (4 pts)

Consider a *null model* where the experimenter pointing with their own arm had no impact on Harley's choice. Under this model, what is the probability distribution of X_i:

x_i	0	1
$P(X_i = x_i)$	.5	.5

H: (3 pts)

Let the random variable $Y = \sum_{i=1}^{12} X_i$ denote the number of "correct" cup choices when the experimenter pointed using their own arm. What is the expected value of Y under the null model described in part F? (Show your calculation, no explanation necessary)

Note that Y is a Binomial random variable, so under the null model, $E(Y) = 12 * .5 = 6$

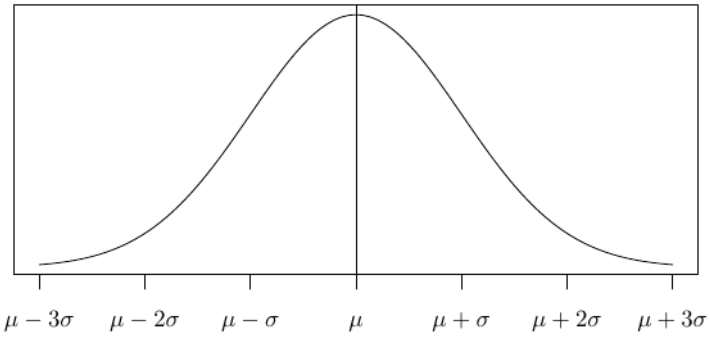
I: (3 pts)

For the random variable Y described in Part G, what are the mean and standard deviation you'd use to approximate probability distribution of Y with a normal curve? (Report your final answer as $N(?, ?)$ where the question marks are replaced by the proper numeric values).

Noting again that Y is a Binomial random variable, $Var(Y) = 12 * .5 * .5 = 3$, so $StdDev(Y) = \sqrt{3} = 1.73$

J: (4 pts)

On the normal curve below, label the mean and standard deviation with the values you found in Part I. Then shade the area representing the one-sided p-value for this scenario (recall that Harley chose the correct cup 10 of 12 times). Are you convinced that Harley "understands" when the experimenter points to a cup with their own arm? Briefly explain. (1-3 sentences)



The mean of this curve is 6, and the standard deviation is 1.73; which means the observed 10 of 12 successes correspond to a little more than 2 standard deviations above average. Thus the area representing outcomes at least as extreme as 10 of 12 is very small, perhaps 1-2% of the curve (ie: p-value is approximately 0.01). This is very convincing evidence that Harley "understands" since it's unlikely for 10 of 12 correct cups to be chosen by chance, and it's unlikely that any other factor might explain these choices (ie: design rules out confounding)

K: (4 pts)

Now consider the instances where the experimenter pointed using the mechanical arm. Are you convinced that Harley "understands" when the experimenter points to a cup with a mechanical arm? Briefly explain. (2-3 sentences)

In this scenario, 8 of 12 correct choices is only a little more than 1 standard deviation above average. Thus the p-value here will be pretty large (approximately 0.2). So I am not convinced because Harley would be expected to make at least 8 of 12 correct choices roughly 1/5 of the time just by random chance if the null model were correct.

Application #2 - Describing Diamonds (23 pts)

This application involves data on 53,940 diamonds sold in the past year by a large online retailer. Because diamond sales are pretty consistent from year-to-year, the retailer believes that these sales accurately reflect the all diamonds sold by the retailer.

The variables used in this application include:

- **price**: how much the diamond sold for (in US dollars)
- **carat**: the diamond's size (in carats)
- **cut**: the quality of the diamond's cut, ranging from "Fair" (poor cuts) to "Ideal" (flawless cuts)
- **color**: the quality of the diamond's color, ranging from "D" (completely colorless) to "J" (slightly yellow)

A: (2 pts)

Suppose this retailer is interested in learning about all diamonds they've sold in the last three years. Based upon the information given, what is the *target population*? (Short phrase)

> The target population are all of the diamonds sold by the company (presumably several years)

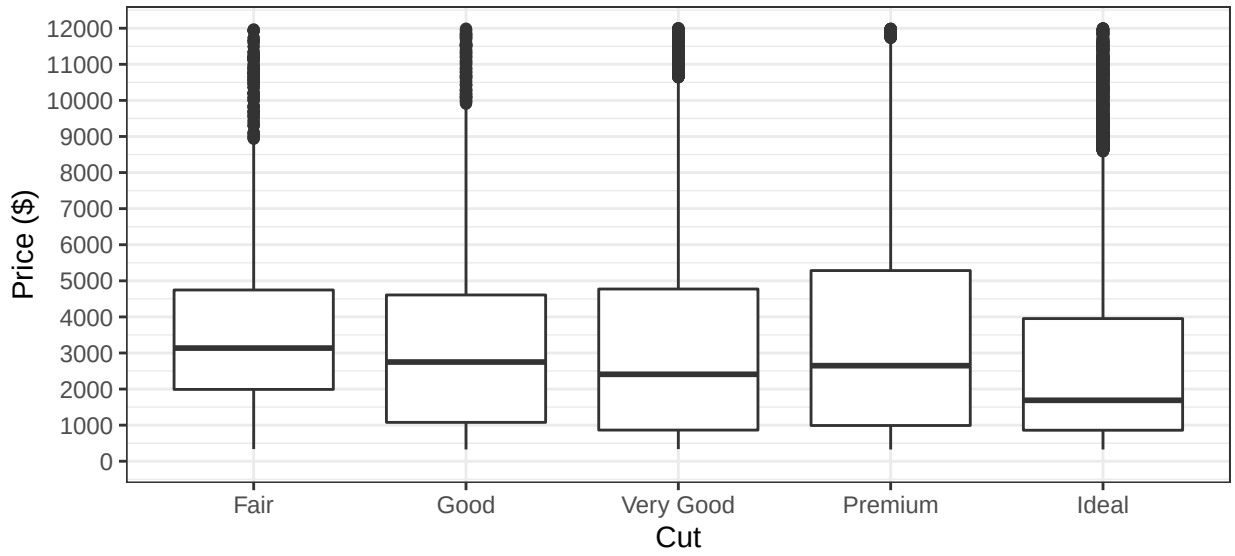
B (4 pts)

What is the *sample*? Are you concerned about there being sampling bias in how this sample was collected? (1-2 sentences)

> While this is not a random sample, because company says that sales are consistent on a yearly basis I wouldn't be too concerned about sampling bias.

C: (4 pts)

The graph below displays the relationship between "Cut" and "Price":



Compare the first quartile (Q1) of sale price for "Fair" diamonds with the first quartile (Q1) of sale price for "Ideal" diamonds and report the observed difference. Based *solely upon this comparison of first quartiles*, do the variables "Cut" and "Price" appear to be associated? Briefly explain. (1-2 sentences)

> Based upon first quartiles, the lower 25% of Ideal diamonds sell for about 1000 dollars less than the lower 25% of Fair diamonds. So yes, Cut and Price are associated because the distribution of Price depends upon the Cut.

D: (4 pts)

The table below describes the relationship between "Cut" and "Carat":

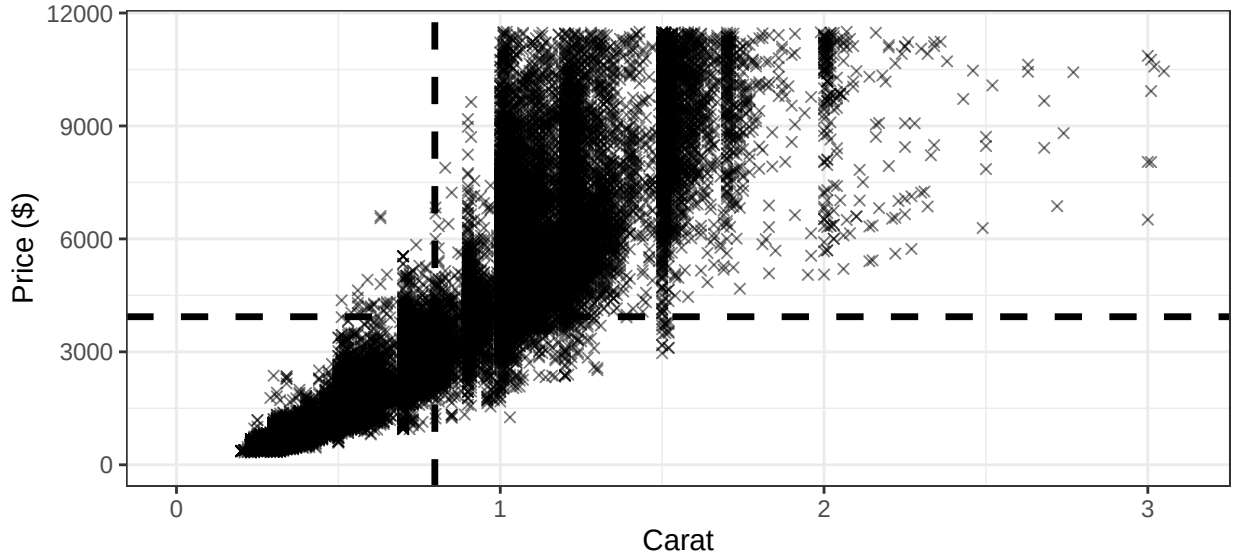
cut	mean	median	sd	n
Fair	1.0461366	1.00	0.5164043	1610
Good	0.8491847	0.82	0.4540544	4906
Very Good	0.8063814	0.71	0.4594354	12082
Premium	0.8919549	0.86	0.5152616	13791
Ideal	0.7028370	0.54	0.4328763	21551

Use the table above to determine whether the variables "Cut" and "Carat" appear associated. You do not need to show any calculations for this question, but you should write 1-2 sentences justifying your answer.

> The two variables are very clearly associated. Notice the different means, medians, etc. when looking different rows (different Cuts)

E: (3 pts)

The visual below displays the relationship between "Carat" and "Price". Note that the dashed lines indicate the mean values of each variable.



Based upon the plot above, *estimate* the *correlation coefficient* between the variables "Carat" and "Price". (No writing necessary)

> Any moderate to strong positive correlation will suffice here. Note that the actual correlation is 0.92, but it'd be tough to judge this from the scatterplot.

F: (6 pts)

Using all of the information available to you (including your previous answers to parts of this application), which of the following do you believe is the *most likely* explanation for the relationship you saw between the variables "Cut" and "Price"?

A) Random chance likely explains why "Fair" diamonds appear to be worth more than "Ideal" diamonds
B) Sampling bias is making "Fair" diamonds appear to be worth more than "Ideal" diamonds
C) A confounding variable is making "Fair" diamonds appear to be worth more than "Ideal" diamonds
D) "Fair" diamonds are actually worth more than "Ideal" diamonds

State your answer (A, B, C, or D), and provide a 1-3 sentence explanation, you may reiterate things you mentioned in earlier questions as part of your response.

> The most likely explanation is that "Carat" is a confounding variable that makes "Fair" diamonds appear to be more valuable than "Ideal" diamonds, but it's actually the case that "Fair" diamonds tend to be much larger in size (which is the biggest determinant of price).