STATS 2244

ACE Booklet Solutions

2025-26

STATS 2244 Final Exam Booklet Solutions (Winter 2025)

Midterm Material (Booklet Part 1)

B. Problem

Example:

1. Descriptive
2. Predictive
3. Causative
4. Descriptive
5. Causative
6. Predictive
7. Predictive

Real Exam Questions

Example 1.
The research question that has a predictive research goal is:

D) This question seeks to estimate or predict a specific outcome for an individual sunflower with a 5mg seed that is planted. Answer A) is descriptive, while C) is asking if a certain fertilizer causes higher yields of crops, so it is causative. B) is also descriptive.

Example 2.
The nature of the Research Question "Does the Skill Development Program lead to better skill improvement and player retention compared to traditional hockey practices?" is:
a. Causative

Explanation: This question is examining whether the Skill Development Program causes an improvement in skills and retention rates compared to traditional practices, indicating a focus on causal relationships between the type of training and its outcomes.

C. Plan: Sampling Design and Considerations

Example 1.

This is an example of stratified sampling because the apples are divided by type of apple. It is also probability sampling as it involves some element of chance, but it is NOT random. Random sampling is NOT correct because each apple doesn't have the same probability of being chosen since there are different numbers of apples in each group. Also, it isn't cluster as there is no census being performed where we randomly select a group(s) and test all apples in that group.

Example 2.

None of these are correct as they are NOT simple random samples where everyone has an equal chance to be chosen. a) is voluntary response b) is systematic, which is random, but not a SRS and c) is convenience sampling

Basic Knowledge

Example 1.
The answer is $\boxed{A}$, cluster since we're taking a census of a few randomly selected sections.
D is also correct as it is probability sampling.

Example 2.
The answer is $\boxed{C}$. Since we are taking a SRS of every one of the five pools.
Note: E. is not a sampling method, but rather a type of observational study. D. is not a sampling method either.

Example 3.
Both cluster and stratified require knowing about subjects in order to separate them into clusters or strata. SRS doesn't require any information. $\therefore$ *The answer is* $\boxed{A}$.
Note: D. and E. are not forms of sampling

Example 4.
The answer is $\boxed{C}$ since only 70 of 100 were handed back in.

Example 5.
The answer is $\boxed{A}$ as it is a stratified sample. It isn't SRS as each car doesn't have an equal chance to be chosen.

Real Exam Questions

Example 1.

The best description of the sampling strategy used by Chen is **B. Random sampling**.
He randomly selected four offices and then sent the questionnaire to every employee in those offices, which fits the definition of random sampling.

Example 2.

The statement that best describes the sampling frame the researcher used is **A. Students in the lecture halls of the five classes on the day the researcher visited**.
This option accurately reflects the specific group of individuals from which the researcher randomly selected participants for the study.

Example 3.

The sampling frame chosen by the city planners was:
B. Low-income families living in the identified neighborhood.
This is because the planners focused on a specific geographic area where they documented households before selecting a sample for their survey.

Example 4.

1. The research question is ***causative*** as it investigates whether the SDP directly affects skill improvement and player retention compared to traditional practices.

2. The population of interest is best described as **a) young children** since this is the whole group we want to find out about.

3. The answer is d). Since skill improvement is the response variable as it measures the outcome of the different practice methods, b) is incorrect. The practice method is an explanatory variable, so c) is incorrect. The location of the league is neither an explanatory or a response variable.

D. Plan: Study Design and Consideration

Example 1.

The response variable is the number of outbursts and the explanatory variable is the communication method: (levels: PECS or no PECS). We would hope to see a decrease in the number of outbursts when PECS are used.

Example 3.

It is an experiment since he is imposing a treatment, ie. assign them a type of mosquito deterrent with some receiving garlic and some a chemical.

The explanatory variable is the type of deterrent, ie. garlic, chemical
The response variable is the number of mosquitoes.

Example 4.

This is a measurement bias as everyone's weight is 2 pounds more.

If your doctor's office measures 10 patient weights in a day and all weights deviate due to the amount of clothing they wear, this is measurement error. I might have a thick sweater and shoes on and weigh in 5 pounds over my actual weight!

Example 5.

It is not a valid measurement because the price of your car isn't directly linked to your wealth or income so the value that we obtain for the price of the car won't validly reflect wealth/income.

Basic Knowledge

Example 1.
No treatment is imposed since they're just giving out questionnaires, so it is not A, D or E as they all relate to an experiment. B is not a study design at all, but rather a method of sampling.
$\therefore \textit{The answer is } \boxed{C}$ as case-control is a type of observational study when we select a group of cases that have a particular value for the response variable and a group of controls that don't have that value of the response variable.

Example 2.
They are comparing blood pressures to baseline levels from the beginning of the study. $\therefore$
$\textit{The answer is } \boxed{C}, \textit{matched pairs}$

Real Exam Questions

Example 1.
The study design that the management of Pasta Palace used is **B. Completely randomized design**. They randomly selected customers to receive either the discount email or the standard email, which aligns with the characteristics of a completely randomized design.

Example 2. The answer is A).
B) The study design did not incorporate a method of control.
This statement is incorrect; the third group was not pruned, therefore using a control group with no treatment.

C) Randomization was used in the study design.
This statement is incorrect as they did not randomly assign individuals from the sample to the treatment groups. The plants on the north side were all in the same treatment.

D) is incorrect because the study is experimental, with the students actively manipulating the pruning techniques.

Example 3.
A) One potential confounding variable that could influence the relationship between skipping breakfast and unhealthy snacking behaviors is irregular sleep patterns. Students with irregular sleep may be more likely to skip breakfast and also have increased cravings or poorer eating habits throughout the day.

B) To mitigate the confounding effect of irregular sleep patterns, the study design could include a question on sleep quality and duration in the questionnaire. By collecting data on participants' sleep habits, researchers can control for this variable in their analysis, allowing for a clearer understanding of the relationship between skipping breakfast and unhealthy snacking behaviors. This modification would effectively reduce confounding because it enables researchers to separate the effects of sleep patterns from the eating habits being studied, thus ensuring that the data reflects the true relationship between the two primary variables without the influence of another factor.

Alternative Answers:

A) One potential confounding variable that could influence the relationship between skipping breakfast and unhealthy snacking behaviors is the frequency of social activities. Students who are more socially active may have irregular eating patterns, including skipping breakfast and snacking unhealthily during gatherings.

B) To mitigate the confounding effect of social activity frequency, the study design could include questions about how often participants engage in social events and the types of foods typically consumed during these gatherings. By collecting this data, researchers can control for the impact of social activity on eating habits. This modification would effectively reduce confounding because it allows for a more accurate analysis of how breakfast consumption specifically relates to snacking behaviors, ensuring that any association found is not simply due to the influence of social eating contexts.

Example 4.

One concern with the sampling choice made by the editor is the potential for self-selection bias. The sampling frame is only "The Journal of Health Research" users who are logged on during a particular time when this "pop-up survey" appears. This is a form of voluntary response sampling that results in self-selection bias. For example, a certain portion of the population, such as the elderly may be less likely to click on the pop-up survey and therefore wouldn't be included in the sample. Depending on the time of day for the pop-up survey, more of certain age groups may be included as well. From my experience, most younger people prefer .pdf files so that they can annotate them more easily and this isn't generalizable to the population of interest which includes more elderly users who may prefer non-pdf files.

NOTE: You could just as easily argue that undercoverage is one concern with valid reasoning.

Example 5.

In this study, the **explanatory variable** is the type of dietary supplement (new supplement vs. placebo). The researchers will first **block** participants based on their age and baseline cholesterol levels to ensure that these factors do not confound the results. After creating these blocks, the researchers will randomly assign participants within each block into two groups, thereby using **randomization**. One group will receive the new dietary supplement (the **treatment** group), and the other will receive a placebo (the **control group**).

The **response variable** measured will be the change in cholesterol levels after a specified period. By using a **matched pairs design**, where each participant is paired with another in the same block based on similar characteristics, the study aims to control for variability and enhance the reliability of the findings. This design will provide clear insights into the supplement's effectiveness while accounting for the identified cofactors, like age and baseline cholesterol levels.

E. PLAN: Variables

Example 1.

15% is called a parameter as it describes the population (all of London). The statistic would be 12% since it describes the sample. The answer is D).

Basic Knowledge Questions

Example 1.

a) $interval$ – 0 doesn't mean absence of the variable
b) $ordinal$
c) $ratio$
d) $interval$
e) $ratio$
f) $ratio$
g) $ordinal$

Example 2.
interval b). has no meaningful zero...ratio does have a meaningful zero...quantitative is not unique to interval because both interval and ratio are quantitative.

Example 3.
The answer is A) because age, height and amount of loans are the only quantitative variables listed

Example 4. The answer is $\boxed{C}$.

Example 5. The answer is $\boxed{D}$.

Example 6. The answer is $\boxed{B}$ $\boxed{E}$ and $\boxed{G}$

Example 7. The answer is $\boxed{A}$ and $\boxed{E}$. Age is continuous if you look at exact age but discrete if you go up by the number of years.

Real Exam Questions

Example 1.

The type of variable for the data collected from this planned question will be **B. Ordinal**. This is because the response options reflect a ranking of satisfaction levels, where the responses can be ordered but the differences between the levels are not necessarily equal.

Example 2.

The answer is A). Discrete. The points accumulated is a quantitative variable, not categorical. It would also be ratio, since 0 is meaningful and both of these answers are absent from our options. In addition, it isn't continuous because there is only a countable number of options.

Example 3.

A) **Assessing the evidence for a claim**
The café owners aim to evaluate whether different types of dressings influence the amount used, which involves **testing a specific hypothesis** about consumer behavior.

B) How many comparison groups has the researcher created to address the research question? What are those comparison groups? (2 points)

The researcher has created **three comparison groups**:

1. Vinaigrette dressing (served with lunch salads)
2. Creamy ranch dressing (served with dinner salads)
3. Yogurt-based dressing (served with takeout salads)

C) Should the comparison groups you identified in part B be considered "independent" or "matched/paired"? Briefly justify/explain your answer with specific reference to relevant information from the scenario provided. (3 points)

The comparison groups should be considered matched pairs. There was a very similar exam question asked and most students got it marked incorrect as they put independent. The argument here is that these 3 comparison groups are matched/paired since there is a 1-1 matching amongst the value of the response variable across all groups. The response variable is the daily volume used. Since they collect the daily volume each day for each of the 3 groups, they are matching the first volume collected from vinaigrette, with the first one collected with creamy ranch as well as that for the yogurt-based dressing. They are measuring the amount or volume each day for a month in March.

F. DATA

Example 1.

The variables that would be classified as quantitative are:

1. Age in years
2. Height in centimeters
3. Weight in kilograms

Favorite color and blood type are categorical variables.

Example 5.

We can see that the median is approximately 73. We can find the range by using range=max – min = 93 – 63 = 30

We can see it is right skewed or positively skewed, so that means the mean is GREATER than the median. However, we cannot tell exactly what the mean of the data is from only a boxplot.

We can find the Q1 = lower quartile= 25^{th} percentile= 70 and the Q3=upper quartile = 75^{th} percentile=85 and use it to find IQR=Q3 – Q1 = 85 – 70 = 15

Example 6.

In this study, the response variable is the level of stress as reported by single mothers, which is measured on an ordinal scale (0, 1, 2, etc.). It is a categorical variable, not a quantitative one.

A. Relative frequencies – This involves percentages or proportions and it can be used for any type of data.

The median, range, and mean are only used to summarize quantitative data.

Example 7.

In a left skewed graph, the mean is LESS than the median, so we would choose histogram 1.

Basic Knowledge Questions

Example 1.

$\vert Min$	$\vert 1st\ Q$	$\vert Median$	$\vert 3rd\ Q$	$\vert Max$
5.03	3.77	3.6	11.1	

$\rightarrow$ $distances$ are not all equal $\therefore$ $not\ A$
Distance from Min to Median is 5.03+3.77=8.8
Distance from median to Max is 3.6+11.1=14.7 >8.8

$\therefore$ $right\ skewed$

$\therefore$ $The\ answer\ is\ \boxed{C}$.

Example 2.

a) true
b) false, it is about 28
c) true
d) false, it occurs in 4 years

$\therefore$ $The\ answer\ is\ \boxed{A}\ and\ \boxed{C}$

Example 3.

a) false
b) false, we can't tell mean from a box plot. The median is 0.4
c) true
d) true, 25% of values lie above Q3, which is approximately 0.5 for coal.

$\therefore$ $The\ answer\ is\ \boxed{C}\ and\ \boxed{D}$.

Example 4.

$\boxed{A}$ is the answer, since it is skewed left.

Real Exam Questions

Example 1.

Correct Answer: D, E and F are all correct.

Explanation: A boxplot is appropriate for summarizing the distribution of a continuous (quantitative) variable (BMI) across different categories (frequency of physical activity). It allows for easy comparison of the median, quartiles, and potential outliers for each group, effectively addressing the research question. A Means plot or a Stripchart could also be used as they can both display a quantitative variable, such as BMI, along with a categorical one, such as frequence of activity. Remember, this is categorical and ordinal since the choices are sedentary, moderately active, high active, etc.)

Recall, a scatterplot (G) is for 2 quantitative variables (and so is a dot plot), and a histogram (A) is for 1 quantitative variable. A bar graph (B) can be used for only categorical variables, although it will work for 1,2,3, etc. as you can use double, side by side or stacked bar graphs. A mosaic plot (C) is only for categorical variables.

Example 2.

Correct Answer: A or F, a Bar graph or a Mosaic plot

Explanation: A mosaic plot is ideal for visualizing the relationship between two categorical variables (dietary preference and exercise frequency). It allows for a comparison of the distribution of exercise frequencies across different dietary preferences, providing a clear view of how these two categorical variables interact. Another option is a bar graph.

Post-Midterm Material (Booklet Part 2)

A. Sampling Distributions

Statistical Inference

The number of days Tina spent working on this booklet is discrete while the mean amount of time first year students spend on social media is continuous.

Example 1.

0.64 is the sample proportion and it may differ from the true population proportion due to sampling variability.

Real Exam Questions

1. **B. The distribution of the population.**

Explanation: The histogram represents the total household income for all households that completed the U.S. Census survey, which includes data from the entire population being studied (all U.S. households). Therefore, it is the distribution of the population. If only a subset of households were surveyed, it would represent a sample, and the sampling distribution would describe the distribution of a statistic based on multiple samples.

2. **A. Sampling distribution**

Explanation:
The sampling distribution refers to the distribution of a statistic (in this case, the sample mean) across all possible samples of the same size (100 Canadians). This distribution shows how the statistic (mean) would vary if multiple samples were taken from the population.

3.
A. The value 0.80 is a parameter.
B. The sample size in the study is 500.
C. The value 0.75 (i.e., 375/500) should be considered a statistic.

Explanation:
- **A** is correct because the national average of 0.80 is a fixed value representing the population parameter.
- **B** is correct as the sample consists of 500 students who took the exam.
- **C** is correct since 0.75 is a sample statistic representing the proportion of students who passed the exam in this particular sample.
- **D** is not necessarily correct unless further details about the sampling procedure are provided.

4.
C. A histogram of the average test scores from simple random samples of 30 students each, from a group of 500 students in a Toronto school district.
D. A histogram of the medians of monthly sales figures for simple random samples of 12 stores out of 200 stores in a retail chain.

Explanation:

- **C** and **D** both refer to sampling distributions, where you are looking at a statistic (average test scores or medians) across multiple random samples. Sampling distributions describe the variability of a statistic (e.g., mean or median) from sample to sample.
- **A** is false since it is a distribution of weights (a variable) for a population (all 3^{rd} year university students in engineering in Toronto
- **B** is false because it is a distribution of daily high temperatures for a population, and again NOT from repeated samples, and therefore they do not represent sampling distributions.
- **Note: If they gave an answer that was only from one sample, that would also NOT be a sampling distribution**

B. Random Variables

Example 1.

This is a discrete random variable. X takes on only a finite, countable number of days, i.e. the possible values are integers from 0 to 7

Example 2.

This is a continuous random variable since the probability is represented by AREA, i.e. the area of the rectangle between 12 and 15.
Area = LxW= 3(1/10) = 0.3

Real Exam Questions

1. **D. The height (in centimeters) of a randomly selected student.**

Explanation: Height is a continuous random variable because it can take on any value within a given range (e.g., from 0 to a very tall value), and it can be measured with great precision. The other options are discrete variables, as they involve counting specific occurrences or categories (e.g., number of students or specific rooms).

Note: A is not a random variable at all. Since the exam has already happened, so there is no element of chance (they say most recent exam), there is no uncertainty. B is a discrete random variable, since it is a number of students and that is finite and countable and not continuous. It is a random variable since it is numeric (i.e. a count) and it is uncertain since the game is in the future, so unknown at this point in time. C can't be considered a random variable because there is no numerical outcome, as it is a room and not a number or numerical quantity.

2. **A. Number of cars passing through a toll booth during rush hour**

Explanation:

- **A** is a discrete random variable because it involves counting distinct items (cars), and the values are countable and finite.
- **B** and **D** are continuous variables because they involve measurements that can take any value within a range (height and time), and they can be infinitely precise depending on the measurement scale.
- **C is not a random variable since it has already happened, so there is no element of chance**

3.

A. The number of minutes a randomly selected commuter spends in traffic on their way to work
C. The number of home runs hit by a randomly selected player during a baseball season

Explanation:

- **A** and **C** are random variables because they can take on different values depending on the outcome of a random event (e.g., how long the commuter spends in traffic or how many home runs the player hits). Both are countable or measurable quantities that vary in different situations.
- **B** describes a categorical variable (color), not a random variable, as the color can only take on a specific set of categories (e.g., red, blue) and does not vary on a numerical scale.
- **D** is also a categorical variable, not a random variable, since it involves fixed categories (types of music) that don't involve numerical variation.

D. Analysis: Confidence Intervals

Example 1.

A. $\Pr\left(\theta - 1\sigma < \hat{\boldsymbol{\theta}}(\boldsymbol{x}) < \theta + 1\sigma\right)$

B. $\Pr\left(\theta - 2\sigma < \hat{\boldsymbol{\theta}}(\boldsymbol{x}) < \theta + 2\sigma\right)$

C. $\Pr\left(\theta - 3\sigma < \hat{\boldsymbol{\theta}}(\boldsymbol{x}) < \theta + 3\sigma\right)$

Example 2.

Since 21= 12 + 3(3) we know this value of 21 is 3 standard deviations away from the mean of 12.

Z=$\frac{Value-Mean}{SD} = \frac{21-12}{3} = \frac{9}{3} = 3$

E. Hypothesis Testing

Example 1.

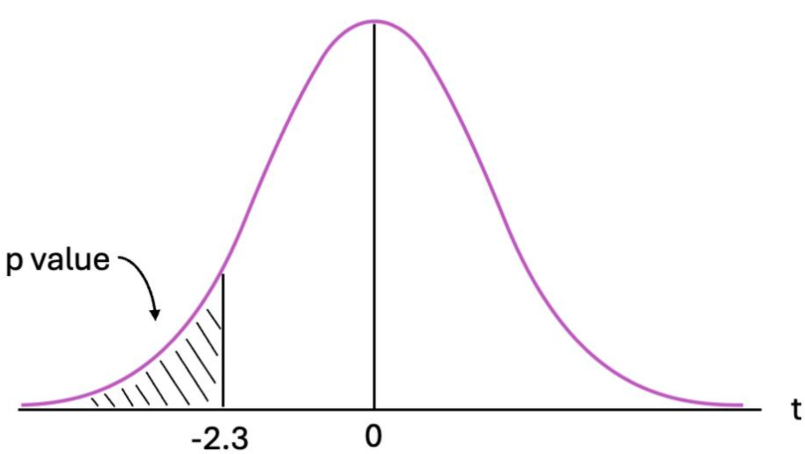
```
>summary(change)
Min. 1st Qu.  Median   Mean  3rd Qu.  Max.
-2     0        1.5     1.6     3       5

> sd (change)
[1] 2.22

> length (change)
[1]  10
```

In this case, the null hypothesis is H0: $\mu = 0$. This is the mean when we are looking at the case of having no impact on blood pressure, i.e. no change at all and here the mean change in blood pressure is zero.

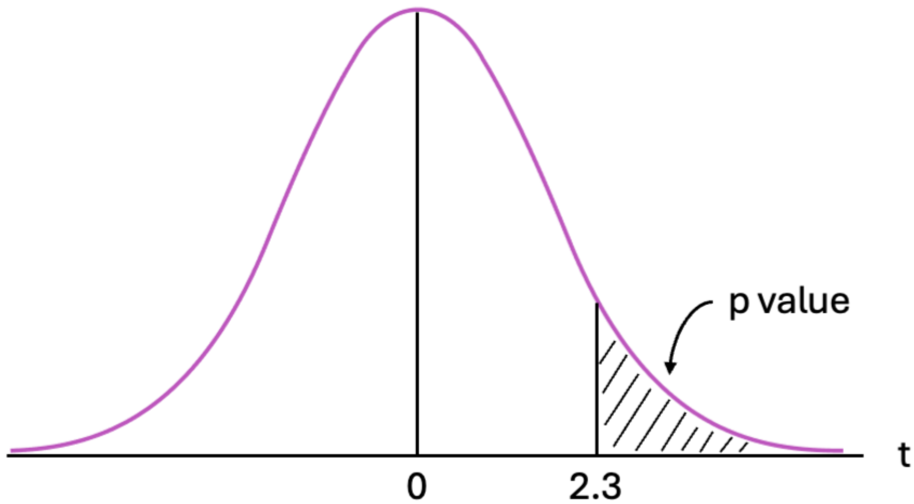
This is because we are looking for a mean difference of 0, i.e.) thinking there is no change in the blood pressure.

The alternative in this case is Ha: $\mu > 0$ since we are calculating the differences as "During – Before" and we are testing to see if their blood pressures ROSE during tax season, a positive result.

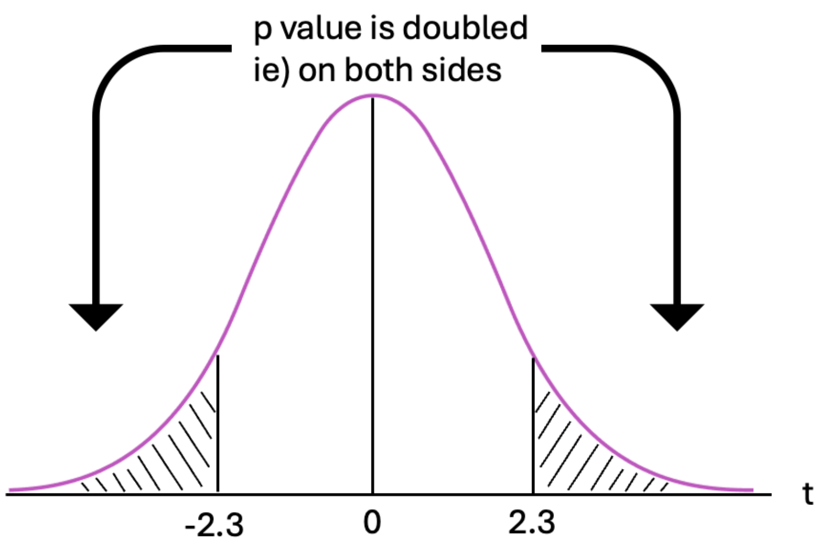
Example 2. A)



B)



C)



Real Exam Questions

1.

A. The research question is causative and aims to establish a relationship between the variables. The question is: How does the frequency of caffeine intake influence the exam performance of current Western students? The population of interest which is the complete collection of all individuals is current Western University Students. The explanatory variable which explains the differences and influences change is frequency of caffeine consumption. The response variable which measures the outcome of interest in this study is the exam results of the students.

B. The type of analysis goal of my research question is assessing evidence for a claim for a parameter. The suitable claim is "how does the frequency of caffeine intake influence exam performance of current Western students". The explanatory variable is the amount of caffeine and the response variable measures exam results.

2.

We use a confidence interval to estimate a population parameter, rather than a single value as the statistic will always have sampling error, and it will likely differ from the true population parameter. A confidence interval has the form:

Point estimate +/ critical value (standard deviation of the sampling distribution)

By using a confidence interval, we can address the problem that our sampling data only represents part of the population, since we are not using the whole population is our sample, but rather a portion of it. Since the statistic we calculate from our sample, such as the sample mean is subject to sampling error, it won't be exactly the same number as the population parameter, the population mean.

For example, perhaps our population mean is actually 40, but our sample mean we measure from our sample might be 35.

If we use a confidence interval of 35 +/ margin of error, then we are able to better account for this sampling error of the difference between our parameter and our statistic.

The confidence interval accounts for sampling error by incorporating the margin of error.

3.

- Change variable = eat breakfast – skip breakfast
- It is a paired test since each child is experiencing eating breakfast and skipping breakfast
- **Null Hypothesis (H_0):** There is no difference in the mean test scores between children who eat breakfast and children who skip breakfast. Mathematically, this is represented as:

$$\mu_{change=0}$$

- **Alternative Hypothesis (H_1):** Children who eat breakfast have higher mean test scores than children who skip breakfast. Mathematically, this is represented as:

$$\mu_{change>0}$$

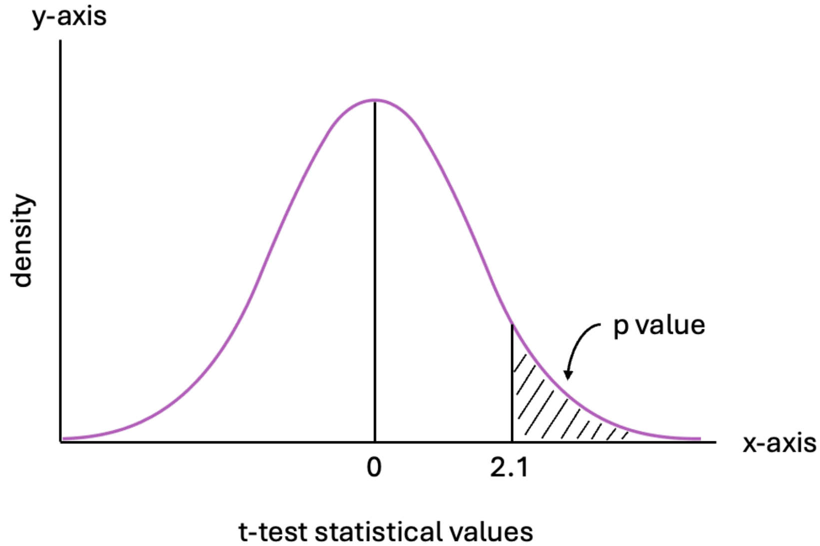
Note: if you did the "change" in the opposite order, you would have a less than test

Graph Description:

- **Distribution Type:** The distribution used is a **t-distribution** with **49 degrees of freedom** (since the sample size is 50, degrees of freedom = 50 - 1 = 49).
- **X-axis:** The axis represents the **test statistic values**, from negative to positive.
- **Y-axis:** This axis represents the **density** of the distribution.
- **Critical Region:** Since the alternative hypothesis suggests a one-tailed test (we are testing if children who eat breakfast perform better, i.e., higher test scores), the critical region is in the **right tail** of the distribution.
- **Test Statistic:** A value of **2.1** is plotted on the graph. This represents the calculated value of the test statistic from the study data.
- **P-value:** The **P-value** is the probability of obtaining a test statistic as extreme as 2.1, or more extreme, if the null hypothesis is true. This would be the area to the **right of the value 2.1** on the t-distribution graph.

t-distribution for mean change

eat breakfast – skip breakfast



- t-test statistic $= 2.1$
- t-distribution with $df = n - 1 = 49$
- since $\mu_{change} > 0$, we have a right tailed test
- the p-value is the area to the right or above 2.1
- the p-value is the probability of obtaining a test statistic that is as extreme or more extreme than your test statistic, assuming H_0 is true.

4.
A. The mean of the sampling distribution is 15 hours.
C. The standardized scores for the sample means will be Normally distributed.
Explanation:

- **A** is correct because the mean of the sampling distribution of sample means equals the mean of the population, which is 15 hours as long as we used random sampling
- **B** is incorrect because the standard deviation of the sampling distribution of sample means (also known as the standard error) is calculated as the population standard deviation divided by the square root of the sample size, but this can only be done if we used a SRS, but here our sample was a stratified random sample
- **C** is correct because we are told the sampling distribution of the sample means is Normally distributed, so the standardized scores (z-scores) will follow a Normal distribution with N(0,1).

5. Let's choose the **P-value** for this question.

The **P-value** is a critical component of the conclusion because it helps assess the strength of the evidence against the null hypothesis and whether your evidence is strong enough to show there is evidence against your claim. It represents the probability of obtaining the observed test statistic (or something more extreme) if the null hypothesis were true. A small p-value indicates stronger evidence against H0, so including that information helps readers to interpret the conclusion and understand why certain choices and conclusions were made.

Without the P-value, the conclusion would lack information about how convincing the evidence is for rejecting the null hypothesis. In this case, a P-value of 0.004 suggests that there is strong evidence that the new drug reduces cholesterol levels.

6. **Answer: C**

***Solution*:**

Answer: C

The interest here is in the mean difference in exam performance, comparing the number of correct answers for the two groups (S vs. L). Since the same students are tested under both conditions, this is a paired data situation. The data can be treated as a one-sample t-test for the mean difference. The null hypothesis is that the mean difference is zero (i.e., no difference in performance), and the alternative hypothesis is that the mean difference is greater than zero (i.e., better performance after a full night's sleep).

The P-value is the probability of obtaining a sample mean difference as extreme or more extreme than the observed value of 3.2 correct answers (i.e., if the true mean difference is zero). Therefore, the correct interpretation is: **the probability of taking a sample that has a mean difference in exam performance of 3.2 correct answers or greater, assuming there is really no difference in mean exam performance**.

Option **C** accurately reflects this definition of the P-value

Why the Other Options Are Incorrect:

- **A. "It is the probability that there is no difference in exam performance after a full night's sleep versus a short night's sleep, assuming there is really no difference in mean exam performance."**
  - This is **incorrect** because the P-value does not represent the probability of the null hypothesis itself. The P-value is the probability of observing a sample result given that the null hypothesis is true, but it is **not** the probability of the null

hypothesis itself. The P-value does not answer whether there is a difference, only how likely it is to observe the data assuming no difference exists.

- **B. "It is the probability that exam performance after a full night's sleep is typically better than after a short night's sleep, assuming there is really no difference in mean exam performance."**
  - This is **incorrect** because the P-value does not give the probability of performance being better after a full night of sleep (or after a short night of sleep). Instead, it tells us the probability of observing the **observed sample statistic** (a difference of 3.2 correct answers) or something more extreme, assuming that there is **no true difference** in performance. The P-value is not about the performance itself but about the test statistic under the assumption of no difference.
- **D. "It is the probability of taking a sample that has a mean exam performance of 3.2 correct answers, assuming there is really no difference in mean exam performance."**
  - This is **incorrect** because the statement confuses the **mean difference** with the actual performance on the exam. The P-value is related to the **mean difference** (3.2 correct answers) between the two groups (S and L), not the exam performance for each condition (full night vs. short night). The P-value reflects the probability of observing a mean difference as extreme as 3.2 correct answers, not the performance level itself.

Conclusion:

The P-value represents the probability of obtaining a result as extreme as 3.2 correct answers or more extreme, given that there is truly no difference in performance. **Option C** correctly captures this interpretation, making it the right choice.

7. **A. μ ≠ 1,200 mm**

Explanation:

The null hypothesis would state that the mean annual rainfall is exactly 1,200 millimeters ($\mu = 1{,}200$ mm). The alternative hypothesis tests whether the mean is different from 1,200 millimeters, which could be either greater than or less than 1,200 mm. This is a two-tailed test, and the appropriate alternative hypothesis is that the mean is not equal to 1,200 millimeters ($\mu \neq 1{,}200$ mm). Thus, **A** is the correct alternative hypothesis.

8. The **level of confidence** in the confidence interval would be negatively impacted. The critical value used in constructing the interval depends on the chosen confidence level and the assumption that the Normal model accurately fits the data. If this assumption is incorrect, and the sampling distribution is not Normal, the critical value will be inaccurate. As a result, the actual level of confidence could be either lower or higher than expected, depending on how the data deviate from the Normal model. This means the interval's reliability may be compromised, either overstating or understating the true confidence level.

9. ***B***

The confidence interval is estimating the **population parameter** (in this case, the average weight of all adult females in the city), based on a sample of 50 adult females. This interval tells us the range of plausible values for the true population mean weight, not just the sample mean or individual weights. Therefore, the correct interpretation is that the confidence interval gives the range of values for the **mean weight of all adult females in the city**, not just the sample.

10. A. There are **two samples** in this study:
   1. The **regular coffee beans** from each coffee plant
   2. The **decaffeinated coffee beans** from each coffee plant

B. The samples are **paired** because each coffee plant contributes two measurements: one for regular coffee beans and one for decaffeinated coffee beans. The measurements from the same plant are related (paired).

C. The **response variable** is the **weight of the coffee beans** (in grams).
   - This is a **quantitative** variable because it is measured on a continuous scale (weight in grams).

D. D. The null hypothesis (H_0) and alternative hypothesis (H_1) for the study are:
Null Hypothesis (H_0): There is no difference in the mean weight of regular and decaffeinated coffee beans.
H0: ud = 0 or u change = 0 where change =μregular - μdecaf

Alternative Hypothesis (Ha): There is a difference in the mean weight of regular and decaffeinated coffee beans.
Ha: u change ≠0

μ_regular represents the mean weight of regular coffee beans across all plants.
μ_decaf represents the mean weight of decaffeinated coffee beans across all plants.

11. The sampling distribution relevant to this hypothesis test is the sampling distribution of sample means. Since the sample is a simple random sample, we know that the sampling distribution of sample means will have a mean equal to the population mean; under the null hypothesis, this is assumed to be $\mu = 5$. The standard deviation of the sampling distribution of sample means (also known as the standard error) is given by the formula $\sigma/\sqrt{n}$, where σ is the population standard deviation and n is the sample size (in this case, $n = 300$). However, we don't know the exact value of the population standard deviation (σ), so we will use the sample standard deviation (s) to estimate it.

Given that the sample size is large ($n = 300$), the sampling distribution of sample means will be approximately Normal, according to the Central Limit Theorem. This means that even if the population distribution of exercise hours is not Normal, the sampling distribution of the sample means will be roughly Normal due to the large sample size. This Normality assumption allows us to conduct the t-test to evaluate the hypothesis about the population mean.

12. ***B***

In this case, 0.30 represents the proportion of households in the population that are known to have a pet dog. This is the population proportion, denoted as **p**, and is the value being used for comparison against the sample proportion (p-hat), which is based on the data collected from the 150 households

13. ***D***

The gym is interested in the percentage of all its members who attended at least one workout session per week, which is a proportion of the population. Therefore, the gym is interested in the population proportion **p** (the true proportion of all members who attend at least one session per week). The sample proportion, **p-hat**, will be calculated from the 80 sampled members, but the parameter of interest is the population proportion.

14. ***C***

The value 0.05 in this confidence interval represents the **margin of error**. It reflects the range of possible values within which the true population proportion is likely to fall, considering the sample data. The margin of error is determined by the confidence level and variability in the data, and it provides an indicator of the precision of the sample estimate. The sample estimate itself is 0.38, and the margin of error tells us how far the estimate might vary from the true population proportion.

All confidence intervals have the same general form and that is:
Estimate +/- margin of error
The estimate in this case is a proportion, but in other intervals it might be the sample mean, i.e. the statistic x-bar

15. ***C***

The confidence level (90%) in this context refers to the percentage of intervals that would correctly capture the true population mean if the sampling and interval construction were repeated many times. It does not imply that there is a 90% chance the true mean is within this specific interval, nor does it imply that 90% of players or games fall within this range. The correct interpretation is that 90% of confidence intervals from repeated sampling would contain the true population mean.

16. ***C***

The confidence level reflects the proportion of confidence intervals, generated through repeated sampling, that would contain the true population parameter. It is not a probability about any single interval but rather a statement about the long-term performance of the method used to construct the interval.

17. ***C***

The P-value is the probability, assuming the null hypothesis is true, of obtaining a test statistic at least as extreme as the one observed in the sample data. It helps assess the strength of the evidence against the null hypothesis. A smaller P-value indicates stronger evidence against the null hypothesis.

18. ***A***

In this case, the company is testing whether the mean battery life of the phones exceeds 10 hours. The objective is to compare the mean of the sample (from the random selection of phones) to a specific value (10 hours). This scenario requires a hypothesis test to evaluate evidence against the null hypothesis that the mean battery life is 10 hours. A confidence interval would provide an estimate for the mean, but since the company is testing if it is greater than 10 hours, a hypothesis test is the appropriate procedure.

19. ***A***

In this case, the bakery is interested in estimating the proportion (p) of customers who prefer the new recipe. A sample is taken to estimate this population proportion. A confidence interval will produce an estimate for p based on p-hat, reflecting the variability in sample proportions. The bakery is not testing a specific claim or comparing proportions, so a hypothesis test is not needed.

20. ***A***

In this case, the manufacturer is interested in estimating the mean (μ) fuel efficiency of the new car model. A sample is taken to estimate this population mean. A confidence interval will produce an estimate for μ based on x-bar, reflecting the sampling variability in sample means. Since the manufacturer is not testing a specific claim or comparing it to a target value, a hypothesis test is not needed.

21. ***B***

The hypotheses in a hypothesis test are framed in terms of population parameters (μ), not sample statistics ($\bar{x}$). The study is designed to test if there is evidence that the average income of high school graduates in Saskatchewan differs from the reported $45,000 in Alberta. Therefore, the alternative hypothesis (Ha) is that the population mean income for high school graduates in Saskatchewan (μ) is not equal to $45,000. Hence, the alternative hypothesis is **μ ≠ 45,000**.

22. ***D***

Solution:

The correct answer is d) It is the probability of obtaining a sample mean difference as extreme as 2.35 or greater, assuming there is really no difference in sleep quality between the groups.

Let's break down why this is the correct answer and why the other options are incorrect:

1. P-value and hypothesis testing: In hypothesis testing, the P-value represents the probability of obtaining a test statistic as extreme as the observed statistic (or more extreme) given that the null hypothesis is true. The null hypothesis typically assumes no effect, meaning there's no difference between the groups or conditions being studied.
2. Context of the question: The research team is testing whether physical activity affects sleep quality. The null hypothesis in this case is that physical activity does not impact sleep quality, and the alternative hypothesis is that more physical activity improves sleep quality. The test is one-sided, meaning it's testing if the effect of physical activity is greater than zero (i.e., improved sleep quality with more activity).
3. The test statistic of 2.35: The observed test statistic of 2.35 indicates how far the sample data is from the null hypothesis, in terms of standard errors. The P-value of 0.022 means that, assuming the null hypothesis (no difference in sleep quality) is true, there is a 2.2% probability of observing a sample mean difference as extreme as 2.35 or greater (in the direction suggested by the alternative hypothesis).

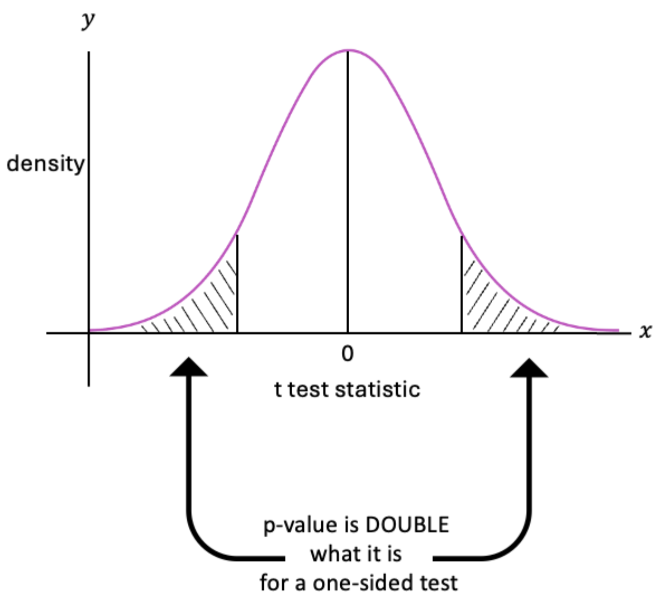
Why d) is correct:

- d) correctly describes the P-value as the probability of obtaining a sample result as extreme as the observed test statistic (2.35 or greater), assuming the null hypothesis is true. This is exactly what the P-value represents.
  Why the other options are incorrect:
- a) "It is the probability that there is no difference in sleep quality between adults with more physical activity and adults with less physical activity."
  This is incorrect because the P-value does not give the probability that the null hypothesis is true. The P-value tells us the probability of observing data as extreme as what we observed, assuming the null hypothesis is true.
- b) "It is the probability that the average sleep quality score of the sample is higher after increased physical activity than before."
  This is incorrect because the P-value is not about the probability of the sample mean being higher after physical activity; it's about the probability of observing a test statistic as extreme as the one calculated, assuming the null hypothesis is true.
- c) "It is the probability of observing a sample mean difference of 2.35 or greater, assuming there is no true difference in sleep quality."
  This is incorrect because the test statistic of 2.35 is not a sample mean difference; it's a test statistic derived from the data. The P-value measures the probability of obtaining a result as extreme as the observed test statistic, not the observed difference in means. Also, "sample mean difference" here is misleading, as the test statistic is what is actually being compared to the critical value.
  In conclusion, d) correctly defines what the P-value represents in the context of hypothesis testing, which is why it is the correct answer.

23. As sample size increases, the standard deviation of our sampling distribution decreases. In other words, our sampling distribution has less "spread". This is important in computing confidence intervals. The equation is CI=point estimate $\pm$ (critical value) (SD of sampling distribution). It can be seen that the SD of sampling distribution contributes to the margin of error. Thus, increasing the sampling size decreases the sampling distribution SD< and thus decrease the margin of error for a confidence interval computed for a random sample. Thus, smaller confidence intervals can be completed for higher confidence levels when a larger sample size is taken, and this provides a more precise estimate of the population parameter. This is intuitive as a larger sample reduces sampling error (difference between statistic and parameter).

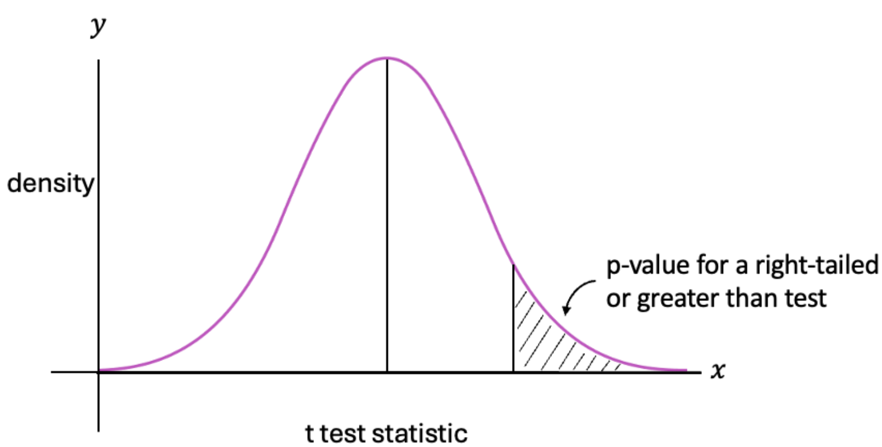
F. T Confidence Interval & Hypothesis Test for Differences in Means

p. 5 ***Solution:***
two-sided p-value



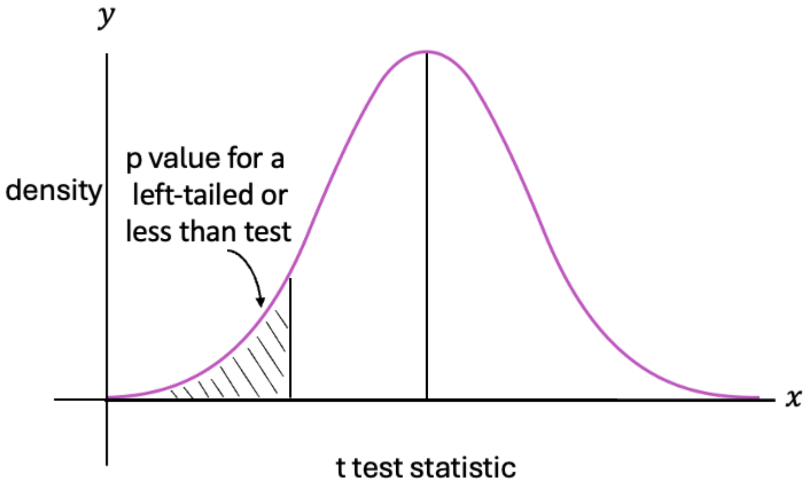
If we wanted to test to see if the first group mean is larger than the second mean, we would have:

Solution:



If we wanted to test to see if the first group mean is less than the second mean, we would have:

Solution:



Example 1. Look at the R output:

A. Data: women_ weight and men_weight are the two data vectors, so x=women's weight
And y=men's weight (for the first table)

B. H0: μ women – μ men=0
Vs. Ha: μ women – μ men$\neq$ 0

C. t test = - 2.7842

D. df=16 The formula is:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{s_1^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{s_2^2}{n_1}\right)^2}$$

E. The level of confidence used was 95% and the 95% confidence interval for μ women – μ men is
Is (-29.748019, -4.029759)

F. Sample mean (women)= 52.1 and sample mean (men)=68.98889

G. p-value=0.01327, which is fairly small, so we would conclude that women's average weight is most likely different than men's average weight. It is unlikely that the difference in mean weight

of women and men this large or larger than the observed difference of 2.78 has occurred due to random chance alone.
Note: The ability to obtain a small p-value can be influenced by both sample size (and sampling variability) as well as the distance between the hypothesized value and the actual value of the mean weight.

Real Exam Questions

1. ***Solution: A***

In this scenario, the manager is comparing the mean number of items purchased between two independent groups (two store locations). The appropriate statistical test to compare the means of two independent groups is the t-test for difference between means. Therefore, the correct answer is A.

2. ***Solution: B***

In this scenario, the researcher is comparing the mean test scores between two groups: male and female students. Since the comparison is between two independent groups, the most appropriate statistical test is a **t-test for the difference between means**, which is used to test if there is a significant difference between the means of two independent groups.

Therefore, the correct answer is **B**.

3.

A. H0: μ control – μ exercise=0
Vs. Ha μ control – μ exercise$\neq$ 0

B. t test = 3.776

C. df=19 The formula is:

$$df = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\left(\frac{1}{n_1-1}\right)\left(\frac{s_1^2}{n_1}\right)^2+\left(\frac{1}{n_2-1}\right)\left(\frac{s_2^2}{n_1}\right)^2}$$ which is also df=n1 +n2 - 2

D. The level of confidence used was 95% and the 95% confidence interval for μ control – μ exercise is (0.2274728, 0.7932545)

E. Sample mean (control)=5.064 and sample mean (exercise)=4.553636

F. p-value=0.001278, which is small, so we would conclude that we have enough evidence to conclude that the mean cholesterol between the control group and exercise group are different. It is unlikely that the difference in mean cholesterol this large or larger than the observed difference of 3.78 has occurred due to random chance alone.

Note: The ability to obtain a small p-value can be influenced by both sample size (and sampling variability) as well as the distance between the hypothesized value and the actual value of the mean weight.

G. Simple Linear Regression

Example 1. The slope will move closer to 0, as it will decrease, and it will become a stronger correlation without an outlier, so r^2 will increase. The answer is A).

Example 2.

The least squares regression equation is Height = b(Age) +a

Or $Y = \beta_1 x + \beta_0$

which would become Y=0.6350 x + 64.9283

0.6350 is the estimate of the slope parameter and 64.9283 is the estimate for the intercept parameter

The summary of the residuals is above the intercept and "age" and it shows the five-number summary for the residuals i.e. observed y-value – predicted y-value (for all n=12 children)

We can see that it says "10 degrees of freedom" which is df=n-2 for linear regression.

The estimated standard deviation of the residuals is called "residual standard error" and it's value is 0.256 (assumed constant for all x).

The standard error (SE) of the estimators, i.e. slope and y-intercept are given as 0.5084 and 0.0214 along with the t test statistics for each which are 127.71 and 29.66. The Pr(>$|t|$) are the p-values.

So, if we test:

Null hypothesis (H_0): $\beta_1 = 0$

Alternative hypothesis (H_A): $\beta_1 \neq 0$

Real Exam Questions

1. ***Solution: B and D***

Step 1: Assume the regression model

The regression equation is likely in the form:

$$\text{Disease Incidence} = b_1 \times \text{Pollution Index} + b_0$$

Where: y=$b_1 x + +b_0$

- b_0 is the intercept (disease incidence when the pollution index is 0).
- b is the slope (change in disease incidence for a one-unit increase in the pollution index).

Step 2: Create plausible coefficients and stats

- Slope coefficient $(b_1): -5.25$ (indicating that higher pollution is associated with a lower disease incidence, as the question suggests).
- Intercept (b_0): 55.30 (the disease incidence when the pollution index is 0).
- Standard error of b_1: 1.10.
- Standard error of b_0: 2.50.

Step 3: Check the options

We now check the statements:

A. If the pollution index increases by one unit, the incidence rate is expected to increase by 5.25 cases per 100,000.

- **Incorrect**. Since the slope is $b_1 = -5.25$. A one-unit increase in pollution results in a decrease of 5.25 cases of respiratory diseases per 100,000 residents.

B. The predicted disease incidence is 52.50 cases per 100,000 when the pollution index is 5 units.

- **Correct**. The prediction at a pollution index of 5 is given by:

$$\text{Disease Incidence} = 55.30 + (-5.25 \times 5) = 55.30 - 26.25 = 29.05$$

So the disease incidence would be 29.05 cases per 100,000.

C. If disease incidence increases by 1 case per 100,000, the pollution index is expected to increase by 5.25 units.

- **Incorrect**. This statement confuses the relationship. The slope tells us that for each increase of 1 unit in the pollution index, the disease incidence changes by -5.25 cases. If we want to find the change in the pollution index for a 1-unit change in disease incidence, we would need to rearrange the equation:

$$\Delta \text{Pollution Index} = \frac{\Delta \text{Disease Incidence}}{\beta_1} = \frac{1}{-5.25} \approx -0.19$$

This indicates that for a 1 case per 100,000 increase in disease incidence, the pollution index would increase by about 0.19 units, not 5.25 units.

D. The estimated disease incidence is about 55.3 cases per 100,000 when the pollution index is 0.

- **Correct**. The intercept So represents the disease incidence when the pollution index is 0, which is 55.30 cases per 100,000.

Conclusion
The correct answers are B and D. Statements A and C are incorrect.

2. ***Solution: B***

The points are tightly concentrated around a pattern (here a linear pattern), indicating a 'strong' association between lab and exam scores.

3. ***Solution: A***

Here, the linear association is negative and strong since the points fall close to the line. So, the correlation will be closest to -0.90. (strong is closer to -1)

4. ***Solution: B (False)***

While a correlation of 0.7 indicates a strong positive association between the number of hours of sleep and test performance, we cannot conclude that the increase in sleep directly causes better performance. Correlation does not imply causation, and other factors (e.g., study habits, prior knowledge) could also influence the test performance. Therefore, the correct answer is False.

5. ***Solution: C***

In this scenario, we are dealing with two quantitative variables: hours of exercise and anxiety scores, and the goal is to assess the relationship between them. The appropriate way to assess the relationship between two quantitative variables is to use correlation and regression analysis. Therefore, the correct hypothesis test would be a test for linearity, specifically testing the hypothesis for the slope of the regression line.

6. ***Solution:***

$\boxed{A}$ shows a larger correlation as the points are closer to the straight line.

7. ***Solution***:

a) Yes, since both variables are quantitative, we can use simple linear regression to determine the relationship between income and happiness.

b) Null hypothesis (H_0): $\beta_1 = 0$

   Alternative hypothesis (H_A): $\beta_1 \neq 0$

c) Slope = 0.71383 x 10000 and y-intercept = 0.20427 x10000

Equation y= 0.71383x + 0.20427 (in 10,000's)

d) T test statistic=38.505

e) p-value = 2×10^{-16} which is basically 0, which means the slope is NOT equal to 0

f) Since the p-value is so small, we would reject H0 and conclude that the there is evidence to support that there is a straight line relationship between x and y. i.e. that the slope is not equal to 0 and there is a linear relationship between income and happiness.

8. ***Solution: C***

In this scenario, the organization is exploring the relationship between two quantitative variables: the number of nurses on duty and the average number of patients treated. To assess the relationship between these two variables, the most appropriate statistical test is a **t-test for beta (slope)**, which is used to test whether there is a linear relationship between the independent variable (number of nurses) and the dependent variable (number of patients treated). Therefore, the correct answer is **C**.

9. ***Solution: D***

In this scenario, the gym owner is interested in the relationship between two quantitative variables: the price of the membership and the number of memberships sold. To assess whether there is a linear relationship between these two variables, the most appropriate statistical test would be a **hypothesis test for the slope of simple linear regression**, as it evaluates how changes in the price are associated with changes in the number of memberships sold. Therefore, the correct answer is **D**.

10. ***Solution:***

Linearity of Y vs X:

Be cautious here. The regression assumption of a "linear relationship between X and Y" does not evaluate whether a relationship exists; rather, it examines whether a linear equation can describe any existing relationship. The goal is to assess if the scatter in a scatterplot (or residual plot) shows a pattern that would indicate the relationship is not linear. This evaluation is independent of whether the slope is 0 or non-zero. If the scatterplot does not visually suggest a relationship, it doesn't mean that the linearity assumption is violated. The hypothesis test for the slope (testing if beta = 0) is the appropriate method to assess if a relationship exists.

Constant Variance of Y (or errors):

Simple linear regression assumes that the response variable (Y) or the error terms have constant variance (or standard deviation). This is evaluated by looking at the spread of the residuals in the residual plot. Constant variance would mean that the vertical spread of residuals remains roughly the same as age increases or decreases. If the residuals for lower values of time are much tighter compared to the larger spread at higher values of age, indicating a "fanning out" pattern. This suggests that the variability in the residuals increases with time, which clearly violates the constant variance condition. If there is a “fanning” in or out, the constant standard deviation or variance condition is violated.

Normality of Y (or errors):

Be mindful that invoking a large sample size in discussions of Normality typically refers to the Central Limit Theorem. This theorem states that the sampling distribution of sample means will approach a Normal distribution with large sample sizes. In regression, however, we need a sufficiently large sample size for each value of X, not just a large total sample size.

You can look at the QQ plot to determine if the normality assumption is met. If the points come close to the diagonal line, the normality assumption is met. We look to see if the values at the extremes differ from the regression line. However, if n is large enough, by the CLT, it can be considered a normal distribution.

Independence of Y (or errors) Assumption:

We need that observations of the response variable are independent. This requires that one value can't be matched to another value.

H. Analysis of Variance (ANOVA)

Example 1. From the ANOVA table, we have:

df numerator (Between groups) = 2, so there were k=3 groups

df denominator (Within groups=MSE=Mean Square Error)= 27, so N-k=27 and N-3=27 so there are N=30 total observations

The F test statistic is 4.467

The p-value is 0.021

The p-value is small, so we could reject H0 and we could conclude that there is enough evidence to say that there is some difference in the means. Recall, ANOVA doesn't allow us to conclude where the difference lies in the means, but rather just that there is some difference between them.

For the Tukey HSD, we can look at the p-values and decide which means there is enough evidence to say are different.

Example 2.

a) ***Hypotheses***: H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
H_a: at least one mean is different than the other means

b) k=4 groups; superman, spiderman, hulk, ninja turtles and N=30 total observations

c) MS (between) =1393.539 and MS (within)= MS (residual)= 167.561

d)F test statistic= 8.317

e) p-value = 0

f) We would conclude that there is some difference amongst the four means, and therefore there is some difference amongst the groups in terms of number of injuries.

Example 3.

a) ***Hypotheses***: $H_0: \mu_1 = \mu_2 = \mu_3$
H_a: at least two means differ significantly from one another

b) 1) The observations from populations are **independent samples**
*This is true if you take SRS's from each population

2) Each sample is obtained using a **SRS with replacement** from its population (or involves a completely randomized design)

3)Each sample is obtained from a population that is **normally distributed**

4) The **standard deviation/variances of each population are equal**

$$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$$

We check independence by ensuring the sampling was done with a SRS with replacement, and we check normality by examining the QQ plot of the residuals and constant variance by examining the stripchart of residuals.

c) The independent variable is the type of fertilizer and the response variable is the crop yield.

d) F test=7.863

e) p-value $=7 \times 10^{-4}$

f) df=2 means k=3 since this is k-1
df=93 means there are 96 total observations since N-k=93 and N-3=93, so N=96.

g) MS (between) groups= fertilizer= 3.0340

MS (within groups) = Residuals = 0.3859

h) The p-value we obtained is very small, so we would reject H0 that the means are all equal and conclude there is some difference amongst the means in terms of type of fertilizer and the crops yield. We don't know from ANOVA which means differ significantly.

i) We conduct a post-hoc test, such as Tukey's HSD test. Yes, from the Tukey multiple comparison table, we see that means 2 &1, have a confidence interval of $\mu_2 - \mu_1$ =(-0.1937, 0.546) with a p-value of 0.495, so we would say there is enough evidence to conclude that means 1&2 do NOT differ from one another. For means 1 &3, we have a p-value of 0.0006125, which is VERY small, so we would reject H0 that the means are equal and conclude those means ARE different. For means 2&3, the p-value is 0.0208, which is fairly small, so we would likely say there is a difference.

Real Exam Questions

1. ***Solution: A***

In this scenario, the principal is comparing the mean test scores across three different groups (9th, 10th, and 11th grades). Since we are dealing with more than two groups, the most appropriate statistical test is a **one-way ANOVA** (Analysis of Variance), which is used to test if there is enough evidence to conclude that there is some difference(s) in the mean scores across multiple groups.

Therefore, the correct answer is **A**.

2. ***Solution:***

a) **Answer:**
"number of points scored" is a **quantitative** and **discrete** variable. It is also ratio data.

b) The comparison groups should be considered **independent** because each game involved different teams, and the players wore different types of shoes. It is not paired/matched groups. The three groups are sneakers, basketball shoes, and casual shoes.

c) **Answer:**
An appropriate statistic to summarize the "number of points scored" would be the **mean**.

d) **Answer:**
The analysis goal is to **assess evidence for a claim** (i.e., hypothesis test) to determine if shoe type influences the amount of points scored.

e) Answer:
The most appropriate inference procedure to address this research question would be **one-factor ANOVA**, as we are comparing the mean number of points scored across three groups, i.e. more than two independent groups (sneakers, basketball shoes, casual shoes).

3. ***Solution:***

D. Conduct an ANOVA analysis to test whether the mean stress levels are different among the three groups.
Explanation:
Since the researcher is comparing the mean stress levels across three independent groups (no exercise, moderate exercise, high exercise), the most appropriate statistical procedure is a one-way ANOVA. This test allows the researcher to determine if there is enough evidence to conclude that there is some difference(s) in the mean stress levels among the three groups.

4. The number of pairwise comparisons is n(n-1)/2 where n is the number of groups
So, here we get 3(3-1)/2 = 3(2)/2=3 pairwise comparisons.

5. ***Solution: True***

The Type I error rate is influenced by multiple comparisons. Performing individual two-sample t-tests at an alpha level of 0.01 does not control the overall Type I error rate. This means that even if there are no true differences between the means, there is a greater than 1% chance of incorrectly rejecting at least one null hypothesis by chance (i.e., detecting false differences between means more than 1% of the time).

6. ***Solution: True***

After rejecting the null hypothesis in a one-way ANOVA, we know that at least one group mean is different from the others. To identify for which specific pairs of groups we have enough information to conclude they are actually different, we can conduct pairwise t-tests. However, since multiple t-tests increase the risk of Type I error (i.e., incorrectly rejecting the null hypothesis), it is important to apply a correction for multiple comparisons, such as the Bonferroni correction or Tukey's HSD (Honestly Significant Difference) test. These corrections adjust the significance level to control the overall Type I error rate across multiple comparisons. Therefore, it is appropriate to use pairwise t-tests after ANOVA, but with adjustments for multiple comparisons to ensure valid results.

7. ***Solution: C***

The most serious flaw in this research plan is that the three departments (marketing, sales, and finance) are not independent. Employees within each department may share common work-related factors or team dynamics that influence their physical activity levels, which would introduce dependence between the samples. Independence is a key assumption of one-way ANOVA, and violating this assumption would invalidate the results. While normality and equal variances are also important considerations, the lack of independence between the groups is the most significant issue.

8. ***Solution: B***

The **ANOVA table** (Answer B) provides information about the results of the ANOVA, such as the F-statistic and p-value, but it does not assess whether the assumptions of normality, equal variances, or independence are met. The other outputs—Normal Quantile (QQ) plots, Boxplots, and Residual plots—are used to evaluate the assumptions of normality and homogeneity of variances, which are critical when conducting a One-factor ANOVA. Therefore, the ANOVA table was probably not used to evaluate the model assumptions.

9.

a) ***Hypotheses***: H_0: $\mu_1 = \mu_2 = \mu_3\ etc$

H_a: at least two means differ significantly from one another

b)1) The observations from populations are **independent samples**

*This is true if you take SRS's from each population

2) Each sample is obtained using a **SRS with replacement** from its population (or involves a completely randomized design)

3)Each sample is obtained from a population that is **normally distributed**

5) The **standard deviation/variances of each population are equal**

$$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$$

We check independence by ensuring the sampling was done with a SRS with replacement, and we check normality by examining the QQ plot of the residuals and constant variance by examining the stripchart of residuals.

c) The explanatory variable is the experimental condition and the response variable is the average weights of the plants.

d) F test statistic = 4.846

e) P-value = 0.0159

f) There are 3 groups, since df=2=k-1 and k=3 groups
N-k=27 and N-3 = 27, so N=30 and there are 30 total observations amongst the 3 groups.

g) MS (between groups)= 1.8832
MS (within groups) = MS (residuals)=0.3886

h) The p-value we obtained is small at 0.0159, so we would reject H0 that the means are all equal and conclude there is some difference amongst the means in terms of type of experimental condition and the average weight of the plants. We don't know from ANOVA which means differ significantly.

i) Use Tukey's HSD test. We can see that the p-value is very small for the test of 1&2 treatments, so we can say that those two treatments DO differ significantly. The other two groups 1&control and 2& control do not.

I. Conclusions

Confidence Intervals

Example 1.

The **mean weight of a bag of chips** is 175g to 185 g (with 90% confidence, df=99)

OR

The **mean weight of a bag of chips** is 175g to 185g (with 90% confidence, n=100)

*Include the parameter, the interval estimate, the level of confidence, and either the sample size or degrees of freedom

Example 2.

The **mean difference in systolic blood pressure** is between -1.2 and 2.5 (with 95% confidence, n=50).

Example 3.

The **difference in proportion of students earning an A in organic chemistry vs. stats 2244** is:

5 ± 8.5 (99% confidence, n (organic chemistry)= 50, n (2244)=35)

*The point estimate and the margin or error is given

Hypothesis Testing

Example 4.

The **mean systolic blood pressure for male chartered accounts** was higher during tax season than before tax season ($\bar{x}(\text{after} - \text{before}) = 1.6$, s=2.2, n=10) (t=2.28, df=9, p=0.02)

Example 5.

The **proportion of undergrad students at Western with light brown hair** $\hat{p} = 0.25$ is lower than the typical Canadian average, i.e. 0.375. (Z= -2.58, p=0.0048, n=100)

*List the sample estimate, i.e. in this case the sample proportion, as well as the test statistic (if applicable), p-value and the sample size or degrees of freedom in the case of a t test)

Best of luck on the exam!!