AP Statistics: Study Guide

Key Exam Details

The AP® Statistics course is equivalent to a first-semester, college-level class in statistics. The 3-hour, end-of-course exam is comprised of 46 questions, including 40 multiple-choice questions (50% of the exam) and 6 free-response questions (50% of the exam).

The exam covers the following course content categories:

- Exploring One-Variable Data: 15%–23% of test questions
- Exploring Two-Variable Data: 5%–7% of test questions
- Collecting Data: 12%–15% of test questions
- Probability, Random Variables, and Probability Distributions: 10%–20% of test questions
- Sampling Distributions: 7%–12% of test questions
- Inference for Categorical Data: Proportions: 12%–15% of test questions
- Inference for Quantitative Data: Means: 10%–18% of test questions
- Inference for Categorical Data: Chi-Square: 2%–5% of test questions
- Inference for Quantitative Data: Slopes: 2%–5% of test questions

This guide will offer an overview of the main tested subjects, along with sample AP multiple-choice questions that look like the questions you'll see on test day.

Exploring One-Variable Data

On your AP exam, 15–23% of questions will fall under the topic of Exploring One-Variable Data.

Variables and Frequency Tables

A **variable** is a characteristic or quantity that potentially differs between individuals in a group. A **categorical** variable is one that that classifies an individual by group or category, while a **quantitative** variable takes on a numerical value that can be measured.

Examples of Variables	
Categorical variables	The country in which a product is manufactured The political party with which a person is affiliated The color of a car
Quantitative variables	The height, in inches, of a person The number of red cars that pass through an intersection in a day

It is important to recognize that it is possible for a categorical variable to look, superficially, like a number. For example, despite being composed of numbers, a zip code is categorical data. It does not represent any quantity or count; rather, it's simply a label for a location.

Quantitative variables can be further classified as discrete or continuous. A **discrete** variable can take on only countably many values. The number of possible values is either finite or countably infinite. In contrast, a **continuous** variable can take on uncountably many values. An important characteristic of a continuous variable is that between any two possible values another value can be found.

Graphs for Categorical Variables

A categorical variable can be represented in a **frequency table**, which shows how many individual items in a population fall into each category. For example, suppose a student was interested in which color of car is most popular. He collects data from the parking lot at school, and his results are shown in the following frequency table:

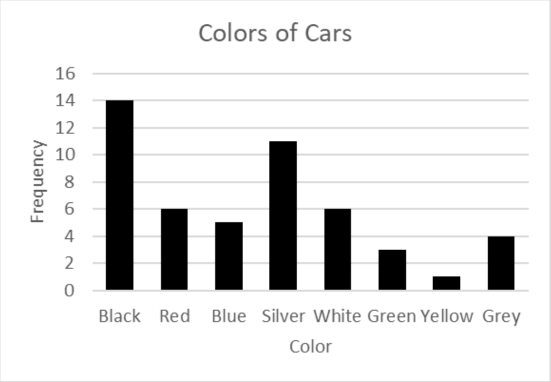
Color	Frequency
Black	14
Red	6
Blue	5
Silver	11
White	6
Green	3
Yellow	1
Grey	4

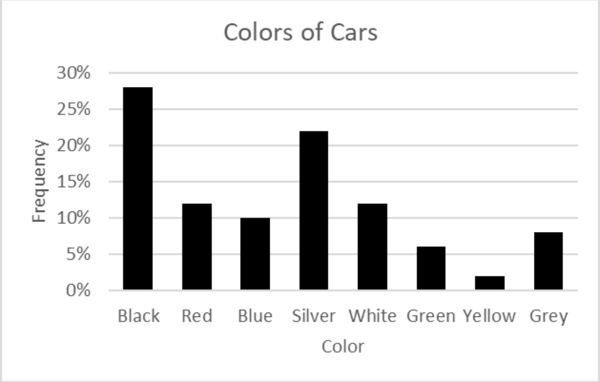
A **relative frequency table** gives the proportion of the total that is accounted for by each category. For example, in the previous data, 14 of the 50 cars, or 28%, were black. The full relative frequency table is as follows:

Color	Relative Frequency
Black	28%
Red	12%
Blue	10%
Silver	22%
White	12%
Green	6%
Yellow	2%
Grey	8%

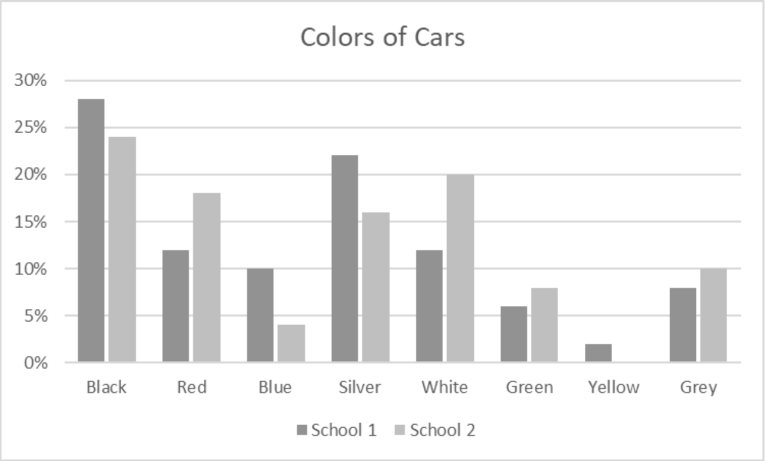
Note that the percentages add up to 100%, since all of the cars were of one of the colors represented in the table.

A **bar chart** is a graph that represents the frequencies, or relative frequencies, of a categorical variable. The categories are organized along a horizontal axis, with a bar rising above each category. The height of the bar corresponds to the number of observations of that category. The vertical axis may be labeled with frequencies or with relative frequencies, as in the following examples.



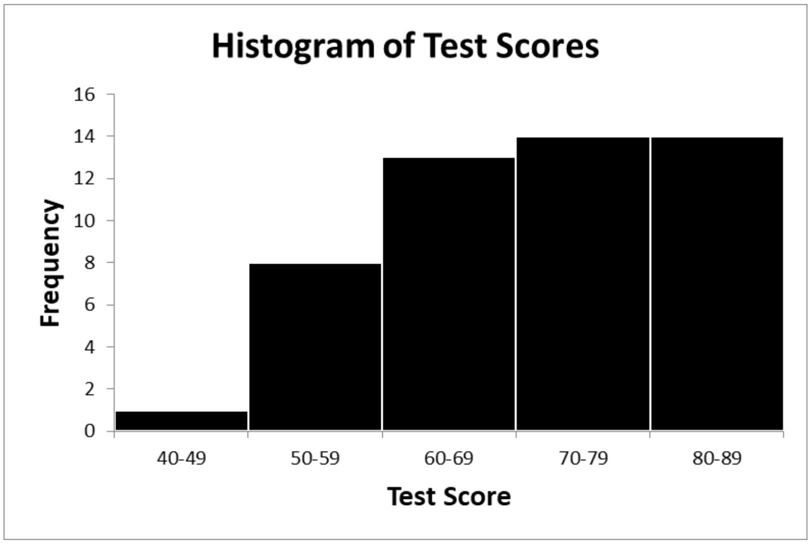


A bar chart representing data from more than one set is useful for comparing the frequencies across the sets. For example, suppose that the day after collecting the initial data on car colors, the student collected the same information from a parking lot at a nearby school. The results can be compared using the following bar chart, which shows the relative frequencies for each color, separated by school:



Graphs for Quantitative Variables

A **histogram** is related to a bar chart but is used for quantitative data. The data is split into intervals, or bins, and the number of data points in each interval is counted. The horizontal axis contains the different intervals, which are adjacent to each other, as they form a number line. The vertical axis shows the count for each interval. The following histogram represents the scores that 50 students received on a test:

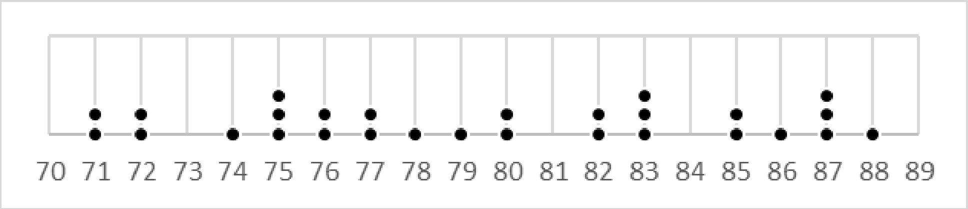


How the data is split into intervals can have a big impact on the appearance of the histogram. Two histograms that represent the same data can show different characteristics, depending on the choice of interval width.

A stem-and-leaf plot is another graphical representation of a quantitative variable. Each data value is split into a **stem** (one or more digits) and a **leaf** (the last digit). The stems are arranged in a column, and the leaves are listed alongside the stem to which they belong. The test score data is shown in the following stem-and-leaf plot:

4	9
5	1 3 5 5 6 9 9 0
6	0 1 3 3 3 4 4 5 6 8 8 8 9
7	1 1 2 2 4 5 5 5 6 6 7 7 8 9
8	0 0 2 2 3 3 3 5 5 6 7 7 7 8

In a **dotplot**, each data value is represented by a dot placed above a horizontal axis. The height of a column of dots shows how many repetitions there are of that value. The following is a subset of the test score data:



The Distribution of a Quantitative Variable

The distribution of quantitative data is described by reference to **shape**, **center**, **variability**, and unusual features such as outliers, clusters, and gaps.

When a distribution has a longer tail on either the right or left, the distribution is said to be **skewed** in that direction. If the right and left sides are approximately mirror images, the distribution is **symmetric**. A distribution with a single peak is **unimodal**; if it has two distinct peaks, it is **bimodal**. A distribution without any noticeable peaks is **uniform**.

An **outlier** is a value that is unusually large or small. A **gap** is a significant interval that contains no data points, and a **cluster** is an interval that contains a high concentration of data points. In many cases, a cluster will be surrounded by gaps.

Free Response Tip

If you are asked to compare two distributions, be sure to address both their similarities and differences. For example, perhaps they are both unimodal, but one is skewed while the other is symmetric. Perhaps one has an outlier while the other does not. In particular, be sure to note if one has greater variability than the other, even if you cannot quantify the difference.

Summary Statistics and Outliers

A **statistic** is a value that summarizes and is derived from a sample. Measures of center and position include the mean, median, quartiles, and percentiles. The commonly used measures of variability are variance, standard deviation, range, and IQR.

The **mean** of a sample is denoted $\bar{x}$, and is defined as the sum of the values divided by the number of values. That is, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$. The **median** is the value in the center when the data points are in order. In case the number of values is even, the median is usually taken to be the mean of the two middle values. The first quartile, Q_1, and the third quartile, Q_3, are the medians of the lower and upper halves of the data set.

The ideas behind the first and third quartiles can be generalized to the notion of percentiles. The ***p*th percentile** is the data point that has *p*% of the data less than or equal to it. With this terminology, the first and third quartiles are the 25th and 75th percentiles, respectively.

The **range** of a data set is the difference between the maximum and minimum values, and the **interquartile range**, or *IQR*, is the difference between the first and third quartiles. That is, $IQR = Q_3 - Q_1$.

Variance is defined in terms of the squares of the differences between the data points and the mean. More precisely, the variance s^2 is given by the formula $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$. The **standard deviation** is then simply the square root of the variance: $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$.

When units of measurement are changed, summary statistics behave in predictable ways that depend on the type of operation done.

Mean	$\overline{x}$	$c\overline{x}$	$\overline{x}+c$
Median/Quartile/Percentile	m	cm	$m+c$
Range/IQR	r	cr	r
Variance	s^2	c^2s^2	s^2
Standard deviation	s	cs	s

There are many possible ways to define an outlier. There are two methods commonly used in AP Statistics, depending on what statistic is being used to describe the spread of the distribution.

When the IQR is used to describe the spread, the *1.5IQR* rule is used to define outliers. Under this rule, a value is considered an outlier if it lies more than $1.5\times IQR$ away from one of the quartiles. Specifically, an outlier is a value that is either less than $Q_1 - 1.5\times IQR$ or greater than $Q_3 + 1.5\times IQR$.

On the other hand, if the standard deviation is being used to describe the variation of the distribution, then any value that is more than 2 standard deviations away from the mean is considered an outlier. In other words, a value is an outlier if it is less than $\overline{x}-2s$ or greater than $\overline{x}+2s$.

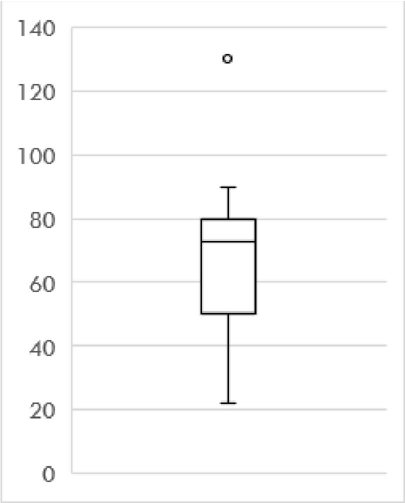
If the existence of an outlier does not have a significant effect on the value of a certain statistic, we say that statistic is **resistant** (or **robust**). The median and IQR are examples of resistant statistics. On the other hand, some statistics, including mean, standard deviation, and range, are changed significantly by an outlier. These statistics are called **nonresistant** (or **non-robust**).

Related to the idea of robustness is the relationship between mean and median in skewed distributions. If a distribution is close to symmetric, the mean and median will be approximately equal to each other. On the other hand, in a skewed distribution the mean will usually be pulled in the direction of the skew. That is, if the distribution is skewed right, the mean will usually be greater than the median, while if the distribution is skewed left, the mean will usually be less than the median.

Graphs of Summary Statistics

The **five-number summary** of a data set is composed of the following five values, in order: minimum, first quartile, median, third quartile, and maximum. A **boxplot** is a graphical representation of the five-number summary that can be drawn vertically or horizontally along a number line. In a boxplot, a box is constructed that spans the distance between the quartiles. A line, representing the median, cuts the box in two.

Lines, often called whiskers, connect the ends of the box with the maximum and minimum points. If the set contains one or more outliers, the whiskers end at the most extreme values that are *not* outliers, and the outliers themselves are indicated by stars or dots.



Note that the two sections of the box, along with the two whiskers, each represent a section of the number line that contains approximately 25% of the values.

Boxplots can be used to compare two or more distributions to each other. The relative positions and sizes of the sections of the box and the whiskers can demonstrate differences in the center and spread of the distributions.

The Normal Distribution

A **normal distribution** is unimodal and symmetric. It is often described as a bell curve. In fact, there are infinitely many normal distributions. Any single one is described by two parameters: the mean, μ, and the standard deviation, σ. The mean is the center of the distribution, and the standard deviation determines whether the peak is relatively tall and narrow or short and wide.

The **empirical rule** gives guidelines for how much of a normally distributed data set is located within certain distances from the center. In particular, approximately 68% of the data points are within 1 standard deviation of the mean, approximately 95% are within 2 standard deviations of the mean, and approximately 99.7% are within 3 standard deviations of the mean.

In practice, many sets of data that arise in statistics can be described as **approximately normal**: they are well modeled by a normal distribution, although it is rarely perfect.

The **standardized score**, or **z-score**, of a data point is the number of standard deviations above or below the mean at which it lies. The formula is $z = \frac{x-\mu}{\sigma}$. It is analogous to a percentile in the sense that it describes the relative position of a point within a data set. If the z-score is positive, the value is greater than the mean, while if it is negative, the value is less than the mean. In either case, the absolute value of the z-score describes how far away the value is from the center of the distribution.

Suggested Reading

- Starnes & Tabor. *The Practice of Statistics.* 6th edition. Chapters 1 and 2. New York, NY: Macmillan.
- Larson & Farber. *Elementary Statistics: Picturing the World*. 7th edition. Chapter 2. New York, NY: Pearson.
- Bock, Velleman, De Veaux, & Bullard. *Stats: Modeling the World*. 5th edition. Chapters 1–5. New York, NY: Pearson.
- Sullivan. *Statistics: Informed Decisions Using Data.* 5th edition. Chapters 2 and 3. New York, NY: Pearson.
- Peck, Short, & Olsen. *Introduction to Statistics and Data Analysis*. 6th edition. Chapters 3 and 4. Boston, MA: Cengage Learning.

Sample Exploring One-Variable Data Questions

Consider the following output obtained when analyzing the percent nitrogen composition of soil collected in neighborhoods near a water treatment facility in 2019.

NumCases = 55

Mean = 23.01

Median = 24.26

StdDev = 4.131

Min = 12.05

Max = 31.49

75th %ile = 30.12

A. The 25th percentile must be about 18.4.

B. Some outliers appear to be present.

C. The IQR is 19.44

D. About 10% of the values are in the range 30.12 to 31.49.

E. Soil levels at 11% exist in the sample, but are not prevalent.

Explanation:

The correct answer is B. An outlier is typically taken to be a data point that is more than two standard deviations from the mean. If you compute mean + 2(standard deviation), you get 31.272. Since the maximum is larger than this value, and 25% of the values are larger than 30.12, there must be *some* outliers in the data. Choice A is incorrect because the data need not be uniformly spaced and so, the manner in which the data is dispersed to the left of the median may not be the same as how it is dispersed to the right. Choice C is incorrect because 19.44 is the range, not the IQR. Choice D is incorrect because about 25% values are within this range. Choice E is incorrect because the minimum value in this data set is 12.05.

A researcher is interested in the age at which adolescents get their first paying job. She surveyed a simple random sample of 150 adolescents who have had at least one paying job before the age of 19. The distribution of the ages was found to be approximately normal with a mean of 15.2 years and a standard deviation of 1.6 years. According to the empirical rule, between which two ages do approximately 95% of the adolescents get their first paying job?

A. 13.2 years and 17.2 years

B. 15.2 years and 18.4 years

C. 12 years and 15.2 years

D. 12 years and 18.4 years

E. 13.6 years and 16.8 years

Explanation:

The correct answer is D. Let *X* be the age at which an adolescent gets his or her first paying job. Since *X* is assumed to be normal with mean 15.2 and standard deviation 1.6, the empirical rule states that about 95% of data will be within 2 st.dev of the mean and 15.2 − 2(1.6) = 12, 15.2 + 2(1.6) =18.4. So, 95% of adolescents get their first paying job between the ages of 12 years and 18.4 years. Choice A is incorrect because you used 2 instead of 2 times the standard deviation 1.6 when computing the margin of error. Choice B is incorrect because you forgot to subtract the margin of error 2(1.6) from the left endpoint. Choice C is incorrect because you forgot to add the margin of error 2(1.6) to the right endpoint. Choice E is incorrect because because you used 1(1.6) as the margin of error instead of 2(1.6). As such, this is the range for when approximately 68% of adolescents get their first paying job.

Thirty-six students completed an algebra exam consisting of 40 questions. The score distribution is described by the following stem-and-leaf plot:

0	0 1 1 6
1	2 2 3 5 7 8 8 8
2	1 8 8 8 8 9 9 9
3	1 2 2 2 4 6 6 6 6 6 8 8
4	0 0 0 0

The first quartile of the score distribution is equal to which of the following?

A. 17

B. 7

C. 36

D. 17.5

E. 29

Explanation:

The correct answer is A. Since there are 36 scores in the stem-and-leaf plot, the position of the 0.25(36) = 9th score, measured starting from the lowest score, is the 25th percentile, or first quartile. The score in the 9th position is 17. Choice B is incorrect because is incorrect because you likely forgot to include the stem "1" when reporting the score. Choice C is incorrect because this is the third quartile, not the first. Remember, the first quartile is the 25th percentile and the third quartile is the 75th percentile. Choice D is incorrect because you averaged the 9th and 10th scores. But, the position of the first quartile, or 25th percentile, is 0.25(36) = 9, an integer, so there is no need to average two scores. Choice E is incorrect because it is the median score, or second quartile.

Exploring Two-Variable Data

On your AP exam, 5–7% of questions will fall under the topic of Exploring Two-Variable Data.

Two Categorical Variables

When a data set involves two categorical variables, a **contingency table** can show how the data points are distributed categories. For example, suppose 600 high school students were asked whether or not they enjoy school. The students could be separated by grade level and by their answer to the question. The data might be organized as follows:

		Grade				
		9th	**10th**	**11th**	**12th**	***Total***
Do you enjoy school?	Yes	40	80	20	100	240
	No	50	30	100	180	360
	Total	90	110	120	280	600

Totals can be calculated for the rows and columns, along with a grand total for the entire table. The entries can be given as relative frequencies by representing the value in each cell as a percentage of either the row or column total. For example, the preceding data is shown below as relative frequencies based on the column totals:

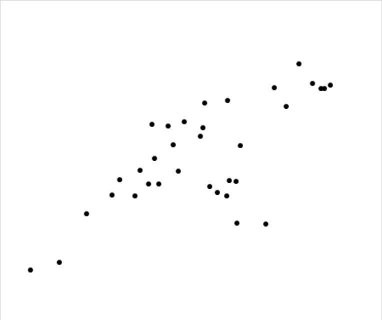
		Grade				
		9th	**10th**	**11th**	**12th**	***Total***
Do you enjoy school?	Yes	44%	73%	17%	36%	40%
	No	56%	27%	83%	64%	60%
	Total	100%	100%	100%	100%	100%

Note that since the percentages are relative to the row column totals, each column now has a total of 100%. The row totals are shown as a percentage of the table total and are referred to as a **marginal distribution**. If the entries are given as relative frequencies by dividing the total for the entire table, rather than by the row or column totals, the table is referred to as a **joint relative frequency**.

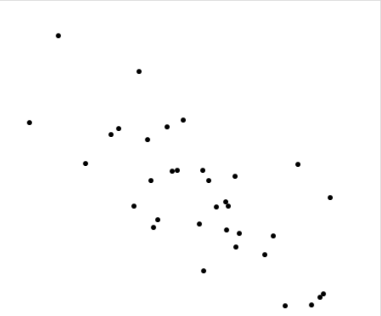
Two Quantitative Variables

When data consists of two quantitative variables, it can be represented as a **scatterplot**, which shows the relationship between the two variables. The variables are assigned to the *x*- and *y*-axes, and then each point can be represented by a point on the *xy*-plane. The variable that is chosen for the *x*-axis is often referred to as the **explanatory variable**, while the variable represented on the *y*-axis is the **response variable**.

A scatterplot shows what kind of **association**, if any, exists between the two variables. The **direction** of the association can be described as positive or negative; positive means that as one variable increases, the other increases as well, while negative means that as one variable increases, the other decreases.

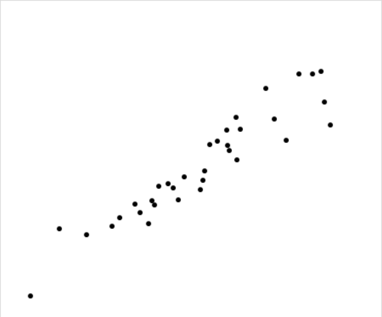


Positive Association

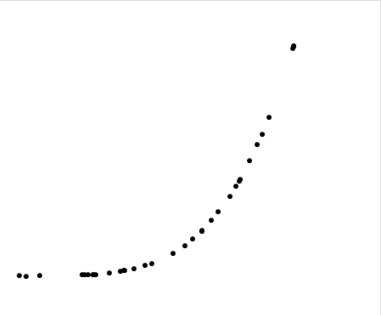


Negative Association

The **form** of an association describes the shape that the points make. In particular, we are generally most interested in whether or not the association is linear. When it is non-linear, it may also be described as having another form, such as exponential or quadratic.



Linear Association



Non-linear Association

The **strength** of an association is determined by how closely the points in the scatterplot follow a pattern (whether the pattern is linear or not). In the previous two examples, the non-linear plot shows a much stronger association than the linear plot, since the points more closely follow a particular curve.

Finally, a scatterplot might have some unusual features. Just as with data involving a single variable, these features include clusters and outliers.

Correlation

The **correlation** between two variables is a single number, *r*, that quantifies the direction and strength of a linear association:

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

In this formula, s_x and s_y denote the sample standard deviations of the *x* and *y* variables, respectively. Although it is possible to calculate by hand, it is implausible for all but the smallest data sets.

The correlation is always between –1 and 1. The sign of *r* indicates the direction of the association, and the absolute value is a measure of its strength: values close to 0 indicate a weak association, and the strength increases as the values move toward –1 or 1. If *r* is 0, there is absolutely no linear relationship between the variables, whereas an *r* of –1 or 1 indicates a perfect linear relationship.

It is important to note that a value close to –1 or 1 does not, by itself, imply that a linear model is appropriate for the data set. On the other hand, a value close to 0 does indicate that a linear model is probably not appropriate.

Regression and Residuals

A **linear regression model** is a linear equation that relates the explanatory and response variables of a data set. The model is given by $\hat{y} = a + bx$, where a is the y-intercept, b is the slope, x is the value of the explanatory variable, and $\hat{y}$ is the predicted value of the response variable.

The purpose of the linear regression model is to predict a *y* given an *x* that does not appear within the data set used to construct the model. If the *x* used is outside of the range of *x*-values of the original data set, using the model for prediction is called **extrapolation**. This tends to yield less reliable predictions than **interpolation**, which is the process of predicting *y*-values for *x*-values that are within the range of the original data set.

Since regression models are rarely perfect, we need methods to analyze the prediction errors that occur. The difference between an actual *y* and the predicted *y*, $y-\hat{y}$, is called a **residual**. When the residuals for every data point are calculated and plotted versus the explanatory variable, *x*, the resulting scatterplot is called a **residual plot**.

A residual plot gives useful information about the appropriateness of a linear model. In particular, any obvious pattern or trend in the residuals indicates that a linear model is probably inappropriate. When a linear model is appropriate, the points on the residual plot should appear random.

The most common method for creating a linear regression model is called **least-squares regression**. The least squares model is defined by two features: it minimizes the sum of the squares of the residuals, and it passes through the point $\left(\overline{x},\overline{y}\right)$.

The slope *b* of the least-squares regression line is given by the formula $b=r\cdot\frac{s_x}{s_y}$. The slope of the line is best interpreted as the predicted amount of change in *y* for every unit increase in *x*.

Once the slope is known, the *y*-intercept, *a*, can be determined by ensuring that the line contains the point $\left(\overline{x},\overline{y}\right)$: $a=\overline{y}-b\overline{x}$.

The *y*-intercept represents the predicted value of *y* when *x* is 0. Depending on the type of data under consideration, however, this may or may not have a reasonable interpretation. It always helps to define the line, but it does not necessarily have contextual significance.

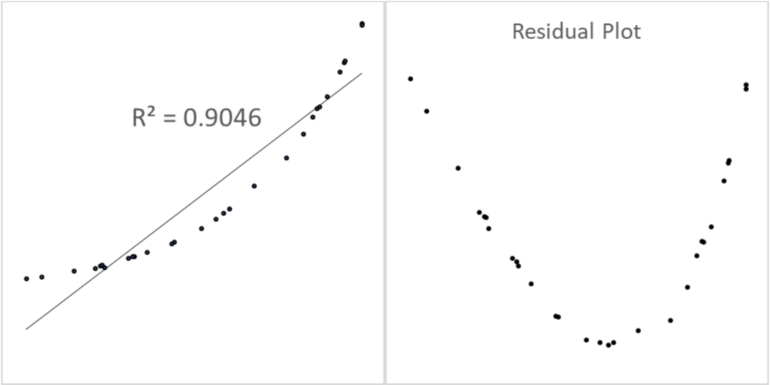
The square of the correlation *r*, or r^2, is also called the coefficient of determination. Its interpretation is difficult, but is usually explained as the proportion of the variation in *y* that is explained by its relationship to *x* as given in the linear model.

There are three ways to classify unusual points in the context of linear regression:

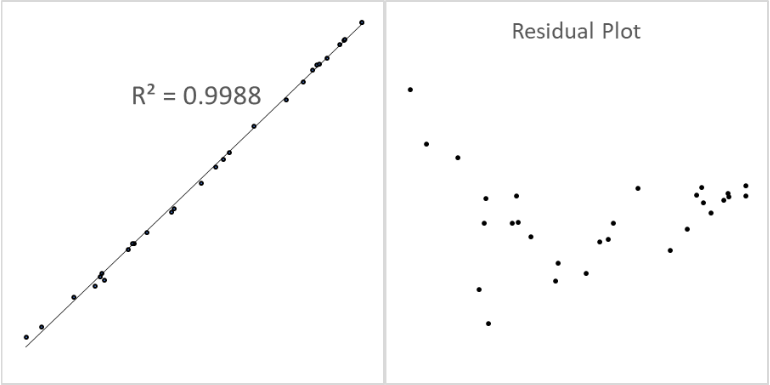
- A point that has a particularly large residual is called an **outlier**.
- A point that has a relatively large or small *x*-value than the other points is called a **high-leverage point**.
- An **influential point** is any point that, if removed, would cause a significant change in the regression model.

Outliers and high-leverage points are usually also influential.

There are situations in which transforming one of the variables results in a linear model of increased strength compared to the original data. For example, consider the following scatterplot, associated least-squares line, and residual plot:



Although the coefficient of determination is high, the residual plot shows a clear lack of randomness. This indicates that a linear model is not appropriate, despite the relatively high correlation. Here are the results of performing the same analysis on the data after taking the logarithm of all the *y*-values:



Not only is the correlation even higher now, the residual plot does not show any obvious patterns. This means that the data were successfully transformed for the pu rposes of fitting a linear model.

There are many other transformations that can be tried, including squaring or taking the square root of one of the variables.

Free Response Tip

If a free response question asks you to justify the use of a linear model for relating two variables, you can mention a correlation near -1 or 1. However, that is not a full justification on its own. You must also analyze the residuals as described in this section.

Suggested Reading

- Starnes & Tabor. *The Practice of Statistics*. 6th edition. Chapter 3. New York, NY: Macmillan.
- *Essentials of Statistics 6e*, Triola. Chapter 9.
- Bock, Velleman, De Veaux, & Bullard. *Stats: Modeling the World*. 5th edition. Chapters 6–9. New York, NY: Pearson.
- Sullivan. *Statistics: Informed Decisions Using Data.* 5th edition. Chapter 4. New York, NY: Pearson.
- Peck, Short, & Olsen. *Introduction to Statistics and Data Analysis*. 6th edition. Chapter 5. Boston, MA: Cengage Learning.

Sample Exploring Two-Variable Data Questions

For new trees of a certain variety between the ages of 6 months and 30 months, there is approximately a linear relationship between *height* and *age*. This relationship can be described by $y = 15.4 + 0.35x$, where y represents the height (in inches) and x represents the age (in months). The tree you planted in the front yard is 16.4 months old and is 23 inches tall. What is its residual according to this model?

A. 5.7400

B. 44.1435

C. −1.8565

D. 1.8565

E. 21.1435

Explanation:

The correct answer is D. The residual is the actual value minus the predicted value given by the linear model at the age of 16.4 months. This yields:

$$23 - (15.4 + 0.35(16.4)) = 1.8565$$

Choice A is the amount of growth experienced by the tree at an age of 16.4 months. Choice B is incorrect because you should have subtracted the actual height and the predicted height at an age of 16.4 months given by the linear model. Choice C is incorrect because this is the negative of the correct value, so you subtracted in the wrong order. Choice E is the predicted height for the age of 16.4 months provided by the linear model. You must subtract this from the actual height of the tree to get the residual.

The effects of a nutritional supplement on hamsters were examined by feeding hamsters various concentrations of the supplement in their daily water supply (measured in mg per liter). The time (in days) until the hamsters exhibited an increase in activity was recorded. A total of 21 different experiments were performed. A preliminary plot of the data showed that the relationship of time versus concentration was approximately linear. The output appears below:

Parameter	Estimate	Test Statistic T	Prob > |T|	Standard Error of Estimate
Intercept	3.415	4.932	0.0004	0.613
Concentration	0.36	0.84	0.041	0.028

Which of the following is the best fit regression line?

A. $y = 0.36 + 3.415x$

B. $y = 3.415 + 3.6x$

C. $y = 3.415 + 0.36x$

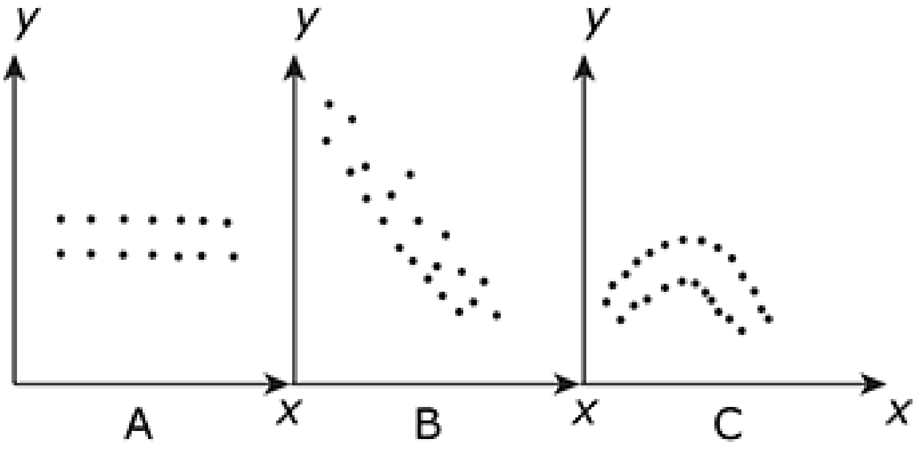
D. $y = 4.932 + 0.84x$

E. $y = 0.36x$

Explanation:

The correct answer is C. This choice is the result of correctly extracting the slope and intercept from the table, and inserting them in the model $y=\beta 0+\beta_1 x$. Choice A is the result of switching the slope and intercept. Choice B is incorrect because the slope is off by a factor of 10. Choice D is incorrect because you used the test statistics instead of the actual estimates of the slope and intercept provided. Choice E is incorrect because you neglected to include the intercept.

Consider the following three scatterplots:



Which of the following statements, if any, are true?

I. The intercept for the line of best fit for the data in scatterplot A will be positive.
II. The slope for the line of best fit for the data in scatterplot B will be negative.
III. There is no discernible relationship between the variables *x* and *y* in scatterplot C.

A. I only

B. II only

C. III only

D. II and III only

E. I and II only

Explanation:

The correct answer is E. Statement I is true because the best fit line is a horizontal line *above* the *x*-axis, so that its *y*-intercept will intersect the *y*-axis in a positive number. Statement II is true because the best fit line is a line whose slope is the same as the parallel lines along which the data in the scatterplot conform. Since these lines fall from left to right, the slope is negative. Statement III is false because there *is* a discernible relationship between *x* and *y* in scatterplot C, it is simply nonlinear.

Collecting Data

About 12–15% of the questions on your AP Statistics exam will cover the category of Collecting Data.

Planning a Study

The entire set of people, items, or subjects of interest to us is called a **population**. Because it is often not feasible to collect data from a population, a **sample**, or smaller subset, is selected from the population. One of the goals of statistics is to use sample data to make reliable inferences about populations.

Once a sample is selected, data collection must take place. In an **experiment**, the participants or subjects are explicitly assigned to two or more different **conditions**, or **treatments**. For example, a medical study investigating a new cold medication might assign half of the people in the study to a group that receives the medication, and the other half to a group that receives an older medication.

Experiments are the only way to determine causal relationships between variables. In the experiment just described, the manufacturer of the medication would like to be able to state that taking their medication causes a reduced duration of the cold.

When experiments are not possible to do for logistical or ethical reasons, **observational studies** often take their place. In an observational study, treatments are not assigned. Rather, data that already exists is collected and analyzed. As noted, an observational study can never be used to determine causality.

Whether a study is experimental or observational, it is important to keep in mind that the results can only be generalized to the population from which the sample was selected.

Data Collection

The methods used in collected data play a large role in determining what conclusions can be drawn from statistical analysis of the data. A **sampling method** is a technique, or plan, used in selecting a sample from a population.

When a sampling method allows for the possibility of an item being selected more than once, the sampling is said to be done **with replacement**. If that is not possible, so that each item can be selected at most once, the sampling is **without replacement**.

A **random sample** is one in which every item from the population has an equal chance of being chosen for the sample. A **simple random sample**, or **SRS**, is one in which every group of a given size has an equal chance of being chosen. Every simple random sample is also random, but the opposite is not true: some sampling techniques lead to random samples that are not simple random samples.

In a **stratified** sample, the population is first divided into groups, or strata, based on some shared characteristic. A sample is then selected from within each stratum, and these are combined into a single larger sample. A stratified sample may be random, but it will never be an SRS.

Another kind of sample is called a **cluster** sample. As with a stratified sample, the population is first divided into groups, called clusters. A sample of clusters is then chosen, and every item within each of the chosen clusters is used as part of the larger sample. Here again, a cluster sample may be random, but it will never be an SRS.

A **systematic** random sample consists of choosing a random starting point within a population and then selecting every item at a fixed periodic interval. For example, perhaps every 10th item in a list is chosen. Again, this kind of sample is not an SRS.

Each of these sampling methods has pros and cons that depend on the population from which they are drawn, as well as the kind of study being done.

Problems with Sampling

There are many potential problems with sampling that can lead to unreliable statistical conclusions. **Bias** occurs when certain values or responses are more likely to be obtained than others. Examples of bias include:

- **Voluntary response bias**, which occurs when a sample consists of people who choose to participate
- **Undercoverage bias**, which happens when some segment of the population has a smaller chance of being included in a sample
- **Nonresponse bias**, which happens when data cannot be obtained from some part of the chosen sample
- **Question wording bias**, which is the result of confusing or leading questions

A random sample, and specifically a simple random sample, is an important tool in helping to avoid bias, though it certainly does not guarantee that bias will not occur.

Experimental Design

A well-designed experiment is the only kind of statistical study that can lead to a claim of a causal relationship. A sample is broken into one or more groups, and each group is assigned a treatment. The results of the data collection that follows show the effect that the treatment had on the subjects.

In an experiment, the **experimental units** are the individuals that are assigned one of the treatments being investigated; these may or may not be people. When they are people, they are also called **participants** or **subjects**. The **explanatory variable** in an experiment is whatever variable is being manipulated by the experimenter, and the different values that it takes on are called **treatments**. The **response variable** is the outcome that is measured to determine what effects, if any, the treatments had. A potential problem in any experiment is the existence of **confounding variables**.

A confounding variable has an effect on the response variable, and may create the impression of a relationship between the explanatory and response variable even where none exists. When possible, confounding variables should be controlled for by careful design of treatments and data collection. Even when they cannot be entirely controlled for, they should be acknowledged as potentially having an effect on the results of the experiment.

A well-designed experiment should always consist of at least two treatment groups, so that the treatment under investigation can be compared to something else. Often, it is compared to a control group, whose sole purpose is to provide comparison data. The control group either receives no treatment, or treatment with an inactive substance called a placebo. It is important to realize, however, that there is a well document phenomenon called a **placebo effect**, in which subjects do respond to treatment with a placebo.

Blinding is a precaution taken to ensure that the subjects and/or the researcher do not know which treatment is being given to a particular individual. In a **single-blind** experiment, either the subject or the researcher does have this information, but the other does not. In a **double-blind** experiment, neither party has this information.

The experimental units should always be randomly assigned to the different treatment groups; if they are not, bias of the sort discussed in the previous section is likely to be an issue. In a **completely randomized design**, experimental units are assigned to treatment groups completely at random. This is usually done using random number generators, or some other technique for generating random choices. This design is most useful for controlling confounding variables.

In a **randomized block design**, the experimental units are first grouped, or **blocked**, based on a **blocking variable**. The members of each block are then randomly assigned to treatment groups. This means that all the values of the blocking variable are represented in each treatment group, which helps ensure that it does not act as a confounding variable in the experiment. A **matched pairs design** is a particular kind of block design in which the experimental units are first arranged into pairs based on factors relevant to the experiment. Each pair is then randomly split into the two treatment groups.

Free Response Tip

When a free response question asks you to describe an experimental design, be sure to explain *why* you are making the choices you are. For example, if you are blocking the experimental units, explicitly state why the variable you are blocking on might be confounding.

Suggested Reading

- Starnes & Tabor. *The Practice of Statistics.* 6th edition. Chapter 4. New York, NY: Macmillan.
- Larson & Farber. *Elementary Statistics: Picturing the World*. 7th edition. Chapter 1. New York, NY: Pearson.
- Bock, Velleman, De Veaux, & Bullard. *Stats: Modeling the World*. 5th edition. Chapters 10–12. New York, NY: Pearson.
- Sullivan. *Statistics: Informed Decisions Using Data.* 5th edition. Chapter 1. New York, NY: Pearson.
- Peck, Short, & Olsen. *Introduction to Statistics and Data Analysis*. 6th edition. Chapter 2. Boston, MA: Cengage Learning.

Sample Collecting Data Questions

A four-year liberal arts college is deciding whether or not to begin a new graduate degree program. They wish to assess the opinion of alumni of the college. The Alumni Affairs Department decides to mail a questionnaire to a random sample of 3500 alumni from the past 30 years. Of the 3500 mailed, 679 were returned, and of these, 218 supported the launching of a new graduate degree program.

Which of the following statements is true?

A. The population of this study consists of the 218 respondents who favor the graduate degree program.

B. The 3500 alumni who were randomly mailed a questionnaire is a representative sample of all alumni of the college for the past 30 years.

C. The population of this study consists of the 679 alumni who mailed back a response.

D. The 3500 alumni receiving the questionnaire constitute the population of the study.

E. Current students are part of the population of this study.

Explanation:

The correct answer is B. This is a very large sample of the graduates from the past 30 years of a small liberal arts college and so, is representative of that population. Choice A is incorrect because this is simply the number of respondents *in the sample* who had this opinion. The *population* is the broader group about which we are trying to infer an opinion on the matter. Choice C is incorrect because this is simply the number of alumni *in the sample* who responded to the questionnaire. The *population* is the broader group about which we are trying to infer an opinion on the matter. Choice D is incorrect because this is simply the size *of the sample*. The *population* is the broader group about which we are trying to infer an opinion on the matter. Choice E is incorrect because only the opinion of the alumni was of interest in this study.

Suppose a simple random sample of size 50 is selected from a population. Which of the following is true of such a sample?

I. It is selected so that every set of 50 subjects in the population has an equal chance of being the sample chosen.

II. It is drawn in such a manner so that every subject has the same chance of being selected.

III. Some members of the population have no chance of being selected, but those that can be selected have the same chance of being selected.

A. I only

B. II only

C. III only

D. I and II only

E. II and III only

Explanation:

The correct answer is D. If I were not true, then some subjects would necessarily have a different chance of being selected, which would render the sample as not being truly random. So, I must be true. Also, if different subjects had a different chance of being selected, the sample would not be truly random. So, II is true. Choice II is false; ALL members of the population must have the same chance of being selected in order for the sample to be random.

A college admissions officer wishes to compare the SAT scores for the incoming freshmen class to the current sophomore class. Which of the following is the most appropriate technique for gathering the data needed to make this comparison?

A. observational study

B. experiment

C. census

D. sample survey

E. a double-blind experiment

Explanation:

The correct answer is C. Making this comparison requires that you collect data for all members satisfying a certain characteristic (here, being an incoming freshman). This is precisely what is done in a census. Choice A s incorrect because you are not trying to make inferences about the effect of a treatment on a group of subjects. Rather, making this comparison requires that you collect data for all members satisfying a certain characteristic (here, being an incoming freshman). Choice B is incorrect because you are not conducting an experiment but making this comparison requires that you collect data for all members satisfying a certain characteristic (here, being an incoming freshman). Choice D is incorrect because there is no reason to take only a sample. All of the data is available for this set of people and so, a census study is more appropriate when trying to make the described comparison. Choice E is incorrect because you are not conducting an experiment but making this comparison requires that you collect data for all members satisfying a certain characteristic (here, being an incoming freshman).

Probability, Random Variables, and Probability Distributions

On your AP Statistics exam, 10–20% of questions will cover the topic of Probability, Random Variables, and Probability Distributions.

Basic Probability

The field of probability involves random processes. That is, processes whose results are determined by chance. The set of all possible outcomes is called the **sample space**, and an **event** is any subset of the sample space.

The probability of an event is the likelihood of it occurring and is represented as a number between 0 and 1, inclusive. If the chance process is repeatable, the probability can be interpreted as the relative frequency with which the event will occur if the process is repeated many times.

If all of the outcomes in the sample space are equally like to occur, then the probability of an event E is the ratio of the number of outcomes in E to the number of outcomes in the sample space.

The completement of an event E, denoted E' or E^c, is the event that consists of all outcomes that are not in E. The probability of an event and its complement always sum to one: P€ + P(E') = 1. Rearranging the terms, this is equivalent to P(E') = 1 – P€.

In many real-world situations, probabilities can be very difficult to calculate. When this happens, **simulation** can be used. Simulation is a technique in which random events are simulated in a way that matches as closely as possible the random process that gives rise to the probability. This is usually done by generating random numbers. The simulation can be repeated many times, and the simulated outcome examined for each repetition. The relative frequency of an event in this sequence of simulated outcomes is an estimate of the probability of the event.

Joint and Conditional Probability

When a probability involves two events both occurring, it is referred to as a **joint probability**. The joint event is denoted using a $\cap$, as in $A \cap B$.

Sometimes we are interested in a probability that depends on knowledge about whether or not another event occurred. This is called a **conditional probability**. The probability that an A will occur given that another event B is known to have occurred is denoted P(A|B), and its value is given by $P(A|B)=\dfrac{P(A\cap B)}{P(B)}$.

Rearranging the terms in this formula, we get the **multiplication rule** for joint probabilities: $P(A\cap B)=P(A)\cdot P(B|A)$.

If *P*(*A*|*B*) = *P*(*A*), then events *A* and *B* are said to be **independent**. The significance of independence is that whether or not one of the events occur has no influence on the probability of the other event. The roles of *A* and *B* can always be switched, so that *P*(*B*|*A*) = *P*(*B*) will also be true if *A* and *B* are independent. Another important consequence of independence is that the multiplication rule simplifies to $P(A\cap B)=P(A)\cdot P(B)$. This last equation can also be used to check for independence.

Unions and Mutually Exclusive Events

The event consisting of either *A* or *B* occurring is called a union, and is denoted by $A\cup B$. Its probability is given by the addition rule: $P(A\cup B)=P(A)+P(B)-P(A\cap B)$. Note that this is inclusive, so that any outcomes that are in both A and B are included in $A\cup B$.

Two events are called **mutually exclusive** if they cannot both occur, so that their joint probability is 0. In other words, *A* and *B* are mutually exclusive if $P(A\cap B)=0$. When this occurs, the last term in the addition rule given previously is 0. Therefore, if A and B are mutually exclusive, the addition rule simplifies to $P(A\cup B)=P(A)+P(B)$.

Free Response Tip

Do not assume events are mutually exclusive unless you are sure they really are! There is no downside to using the full addition rule. If it happens that they are mutually exclusive, the last term will simply not contribute to the probability.

Random Variables and Probability Distributions

A **random variable** is a variable whose numerical value depends on the outcome of a random experiment, so that it takes on different values with certain probabilities. A random variable is called **discrete** if it can take on finitely or countably many values. The sum of the probabilities of the possible values is always equal to 1, since they represent all possible outcomes of the experiment.

A **probability distribution** represents the possible values of a random variable along with their respective probabilities. It is often represented as a table or graph, as in the following example:

X	**1**	**2**	**3**	**4**	**5**
P(*X* = *x*)	0.2	0.3	0.1	0.25	0.15

The table shows a random variable *X* that can take on each of the values 1, 2, 3, 4, and 5. It takes on the value 1 with probability 0.2, the value 2 with probability 0.3, and so on. Note that the sum of the probabilities is 0.2 + 0.3 + 0.1 + 0.25 + 0.15 = 1, as expected. The notation *P*(*X* = *x*) in the second row represents the probability of the random variable (*X*) taking on one of its possible values (*x*).

Sometimes it is beneficial to have a **cumulative probability distribution**, which shows the probabilities of all values of a random variable less than or equal to a given value.

The cumulative distribution for the example in the previous table is as follows:

X	**1**	**2**	**3**	**4**	**5**
$P(X \le x)$	0.2	0.5	0.6	0.85	1

A probability distribution has a mean and a standard deviation, just like a population. The **mean**, or **expected value**, of a discrete random variable X is $\mu_X = \sum X_i \cdot P(x_i)$. Its standard deviation is $\sigma_X = \sqrt{\sum (x_i - \mu_X)^2 \cdot P(x_i)}$.

Combining Random Variables

If *X* and *Y* are two discrete random variables, a new random variable can be constructed by combining *X* and *Y* in a linear combination $aX + bY$, where *a* and *b* are any real numbers. The mean of this new random variable is $\mu_{aX+bY} = a\mu_X + b\mu_Y$. If the two variables are independent, so that information obtained about one of them does not affect the distribution of the other, then the standard deviation of the linear combination is $\sigma_{aX+bY} = \sqrt{a^2\sigma_Y^2 + b^2\sigma_Y^2}$. If the variables are not independent, the computation of the standard deviation of the linear combination is well beyond the scope of AP Statistics.

A single random variable can also be transformed into a new one by means of the linear equation $Y = a + bX$. The mean of the transformed variable is $\mu_Y = a + b\mu_X$, and its standard deviation is $\sigma_Y = |b|\sigma_X$. In addition, if *a* and *b* are both positive, then the distribution of *Y* has the same shape as the distribution of *X*.

Binomial and Geometric Distributions

A **Bernoulli trial** is an experiment that satisfies the following conditions:

- There are only two possible outcomes, called **success** and **failure**
- The probability of success is the same every time the experiment is conducted

We will let *p* denote the probability of success. Because failure is the complement of success, the probability of failure is then 1 – *p*.

Consider repeating a Bernoulli trial *n* times and counting the number of successes that occur in these repetitions. If we call the number of successes *X*, then *X* is called a **binomial random variable**. The probability of exactly *x* successes in *n* trials is given by $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ Here $\binom{n}{x}$ is the binomial coefficient often referred to as a combination. Its value is $\binom{n}{x} = \dfrac{n!}{x!(n-x)!}$.

The mean of a binomial random variable is $\mu_X = np$, and its standard deviation is $\sigma_X = \sqrt{np(1-p)}$.

A **geometric random variable** is also related to Bernoulli trials. Unlike a binomial random variable, a geometric random variable *X* is the number of the trial on which a success first occurs. The value is given by $P(X = x) = (1-p)^{1-x} p$. Its mean is $\mu_X = \dfrac{1}{p}$ and its standard deviation is $\sigma_X = \dfrac{\sqrt{1-p}}{p}$.

Free Response Tip

Be careful to not get confused by the terms **success** and **failure** in the description of binomial and geometric distribution. They do not necessarily have any bearing on success and failure as the words might generally be applied in any given situation. For example, if a problem involves counting the number of phones in a case of 20 produced in a factory, it would be advantageous to refer to a phone being defective as a success, even though it is certainly not that from the perspective of the manufacturer!

Suggested Reading

- Starnes & Tabor. *The Practice of Statistics*. 6th edition. Chapters 5 and 6. New York, NY: Macmillan.
- Larson & Farber. *Elementary Statistics: Picturing the World*. 7th edition. Chapters 3 and 4. New York, NY: Pearson.
- Bock, Velleman, De Veaux, & Bullard. *Stats: Modeling the World*. 5th edition. Chapters 13–16. New York, NY: Pearson.
- Sullivan. *Statistics: Informed Decisions Using Data.* 5th edition. Chapters 5 and 6. New York, NY: Pearson.
- Peck, Short, & Olsen. *Introduction to Statistics and Data Analysis*. 6th edition. Chapters 6 and 7. Boston, MA: Cengage Learning.

Sample Probability, Random Variables, and Probability Distributions Questions

The probability that Valley Creek will flood in any given year has been estimated from 150 years of historical data to be 0.20. Which of the following is an accurate interpretation of this statement?

A. Valley Creek will flood once every five years.

B. In the next 50 years, Valley Creek will flood about in about 10 of those years.

C. In the next 100 years, Valley Creek cannot flood fewer than 20 times.

D. In the last 50 years, Valley Creek flooded exactly 10 times.

E. In the next 50 years, Valley Creek will flood exactly 10 times.

Explanation:

The correct answer is B. In the long run, this statement means that Valley Creek floods about 20% of the time. Since 20% of 50 is 10, we expect it to flood about 10 times. The statement in choice A is probabilistic; it does not literally imply that the creek will *necessarily* flood every fifth year, but rather of the past 150 years, it has flood 30 times. Choices C and E are also incorrect because it does not literally imply that the creek will *necessarily* flood every fifth year; in the long run, Valley Creek floods 20% of the time, not necessarily exactly 20% of the time. Choice D is incorrect because the past 150 years were used to formulate this probabilistic statement. It could be the case that Valley Creek flooded 30 times in the first 100 years and never thereafter.

The probability that a visitor of the local botanical gardens walks through the rose garden is 0.65, and the probability that a visitor meanders through the new meadow is 0.45. The probability that a visitor does both activities on the same day is 0.32. What is the probability that a visitor does at least one of the activities on a given day?

A. 0

B. 0.2925

C. 0.78

D. 0.22

E. 0.50

Explanation:

The correct answer is C. Let A be the event "walks through the rose garden" and B the event "meanders through the new meadow." We must compute $P(A \cup B)$. To do so, use the addition formula, as follows:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$= 0.65 + 0.45 - 0.32 = 0.78$$

Choice A is incorrect because this event is far from impossible. Use the addition formula to compute the probability of the event "walks through rose garden OR meanders through new meadow." Choice B is incorrect because when computing $P(A \cup B)$, you multiplied the probabilities $P(A)$ and $P(B)$, which is incorrect; you must use the addition formula. Choice D is incorrect because this is the probability that a visitor does neither of these two activities on a given day. Choice E is incorrect because there is not a 50-50 chance of this event occurring. You must use the addition formula to compute the probability of the event "walks through rose garden OR meanders through new meadow."

To study the relationship between township and support for a certain amendment concerning property tax, 200 registered voters were surveyed with the following results:

	Against amendment	For amendment	Neutral
Hawk Township	35	62	3
Caln Township	2	40	8
Front Township	39	6	5

What percentage of those surveyed were against the amendment and were residents of Front Township?

A. 80.5%

B. 51.3%

C. 78%

D. 19.5%

E. 39%

Explanation:

The correct answer is D. The event of interest is “against amendment AND lives in Front Township.” The number of respondents satisfying this criterion is in the lower left cell of the table. Hence, the percentage satisfying this criterion is 39/200 = 19.5%. Choice A is the percentage of those sampled that satisfies *neither* condition. Choice B is incorrect because you computed a *conditional* probability assuming “against amendment” as given information. As the problem is stated, you are looking for the probability of an “AND” event. Choice C is incorrect because you computed a *conditional* probability assuming “lived in Front Township” as given information. As the problem is stated, you are looking for the probability of an “AND” event. Choice E is incorrect because this is the *number* of respondents satisfying the criterion, not the percentage. You must divide this by the total sample size, 200.

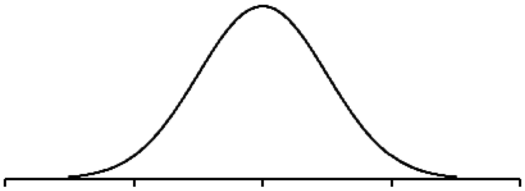
Sampling Distributions

About 7–12% of the questions on your exam will cover the topic of Sampling Distributions.

The Normal Distribution

Unlike a discrete random variable, a **continuous random variable** can take on *any* value within some set. Instead of probabilities being associated with individual values of the variable, probabilities are assigned to every possible interval of values. The most important type of continuous random variable is a **normal** random variable.

A normal distribution has a shape that is often described as a bell-shaped curve. It is unimodal and symmetric:



The probability associated with any given interval is given by the area under the normal curve over that interval. The intervals are generally one of four types:

- $X < a$. This is called a *left-tailed* interval. If the probability associated with it is $P(X < a) = \frac{p}{100}$, this means that the smallest *p*% of values in the population are less than *a*.
- $X > a$. This is called a *right-tailed* interval. If the probability associated with it is $P(X > a) = \frac{p}{100}$, this means that the largest *p*% of values in the population are greater than *a*.
- $|X| > a$, where $a > 0$. This is a *two-tailed* interval. If the probability associated with it is $P\left(|X| > a\right) = \frac{p}{100}$, this means that the largest $\frac{p}{2}$% of values are greater than *a* , and the smallest $\frac{p}{2}$% of values are less than $-a$.

- $a < X < b$. If the probability associated with this interval is $P(a < X < B) = \frac{p}{100}$, this means that *p*% of the values are between *a* and *b*.

As mentioned previously, a normal distribution is defined by two parameters: its mean, n, and its standard deviation, σ. Every combination of μ and σ determine a different normal distribution. The **standard normal distribution** is the normal distribution with $m = 0$ and $s = 1$.

Areas under any normal curve can be found using a calculator or computer software. There are also standard normal tables in many textbooks that can be used to find the probabilities of the standard normal distribution. Using *z*-scores, however, the probabilities in *any* normal distribution can be made equivalent to the probabilities in a standard normal distribution. First, find the *z*-score(s) of the endpoint(s) of the interval of interest. Then simply use the standard normal distribution to find the probability of the new interval. This probability is also the correct value for the original interval.

Central Limit Theorem

Consider a statistic of interest, such as a mean, median, standard deviation, or proportion of a large population. Now take all possible samples of a given size from within this population and calculate the statistic for each sample. The resulting values would themselves form a distribution. This distribution is called the **sampling distribution** of the statistic.

The **Central Limit Theorem** is the most important tool needed in doing inferential statistics. It states that under certain assumptions, the sampling distribution of the mean will be approximately normally distributed. One requirement is that either the original population itself be normally distributed, or that the sample size be at least 30. In addition, the samples must be independent of each other.

Since we are going to use sampling distributions to infer parameters of the population, it is important to know when this will give accurate values. A statistic is an **unbiased estimator** of its corresponding parameter if the mean of its sampling distribution is equal to the parameter. For example, sample mean is an unbiased estimator, while sample standard deviation is not.

Sampling Distribution for Proportions

Consider a population for which we are interested in the proportion that satisfy some condition. This means there is a categorical variable, and that we want to know the proportion *p* of values that are in a certain category. The sample proportions $\hat{p}$ from all independent samples of size *n* form a sampling distribution with mean $\mu_{\hat{p}} = p$ and standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$. If the samples are not independent, then the standard deviation is not as given; however, if the sample size *n* is less than 10% of the population size, it is very close to accurate. If $np \geq 10$ and $n(1-p) \geq 10$, the sampling distribution is approximately normal.

If we are not interested in a single population proportion, but rather in the difference between two proportions p_1 and p_2, from populations with samples of sizes n_1 and n_2, the distribution is given by $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$ and $\mu_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$. Here again, if the samples are not independent the standard deviation is not quite correct, but it is very close if the sample sizes are less than 10% of the population sizes. If $n_1 p_1$, $n_1(1-p_1)$, $n_2 p_2$, and $n_2(1-p_2)$ are all at least 10, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal.

Sampling Distribution for Means

The sampling distribution of sample means is also easy to describe. If independent samples of size n are taken from a population with mean μ and standard deviation σ, then the sampling distribution has $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Even in dependent samples, the standard deviation is accurate if the sample size is less than 10% of the population. The sampling distribution is approximately normal if either the population itself is approximately normal or if $n \geq 30$.

If samples of sizes n_1 and n_2 are taken from two independent populations with means μ_1 and μ_2 and standard deviations σ_1 and σ_2, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ has mean $\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2$ and standard deviation $\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. There are similar conditions here as well: the standard deviation given is accurate even for dependent samples if the sample sizes are less than 10% of their respective populations, and the sampling distribution is approximately normal if either the two populations are approximately normal or if n_1 and n_2 are both at least 30.

Summary of Sampling Distributions

Sampling Distribution	Parameter	Statistic	Mean of Sampling Distribution	Standard Deviation of Sampling Distribution	Conditions
Proportion	p	$\hat{p}$	p	$\sqrt{\frac{p(1-p)}{n}}$	$np \geq 10$ $n(1-p) \geq 10$ n less than 10% of population
Difference of proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$	$n_1 p_1 \geq 10$ $n_1(1-p_1) \geq 10$ $n_2 p_2 \geq 10$ $n_2(1-p_2) \geq 10$ n_1 and n_2 both less than 10% of populations
Mean	μ	$\overline{x}$	μ	$\frac{\sigma}{\sqrt{n}}$	$n \geq 30$ n less than 10% of population
Difference of means	$\mu_1 - \mu_2$	$\overline{x}_1 - \overline{x}_2$	$\mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$n_1 \geq 30$ $n_2 \geq 30$ n_1 and n_2 both less than 10% of population

Suggested Reading

- Starnes & Tabor. *The Practice of Statistics.* 6th edition. Chapter 7. New York, NY: Macmillan.
- Larson & Farber. *Elementary Statistics: Picturing the World*. 7th edition. Chapter 5. New York, NY: Pearson.
- Bock, Velleman, De Veaux, & Bullard. *Stats: Modeling the World*. 5th edition. Chapter 17. New York, NY: Pearson.
- Sullivan. *Statistics: Informed Decisions Using Data.* 5th edition. Chapters 7 and 8. New York, NY: Pearson.
- Peck, Short, & Olsen. *Introduction to Statistics and Data Analysis*. 6th edition. Chapter 8. Boston, MA: Cengage Learning.

Sample Sampling Distributions Questions

Which of the following statements is (are) true?

I. The larger the sample, the smaller the variance of the sampling distribution.

II. Sampling distributions from non-normal populations are approximately normal when the sample size is large.

III. If the population size is much larger than the sample size, then the variance of the sampling distribution remains unchanged, no matter what the sample size is.

A.I only

B. III only

C. I and II only

D. I and III only

E. I, II, and III

Explanation:

The correct answer is C. Statement I is true because the variance of the sampling distribution is $\frac{s^{2}}{n}$, where n is the sample size. So, as n increases, this fraction decreases. Statement II is true because it is a direct consequence of the Central Limit Theorem.

Which of the following is a consequence of the Central Limit Theorem for a sample size n?

A. A standard deviation of the sample mean random variable is greater than the population standard deviation.

B. The expectation of the sample mean random variable is equal to the population mean μ when n is large.

C. The sampling distribution of the sample mean is always normal for any sample size n.

D. The standard deviation of the set of sample mean random variable is equal to the population standard deviation σ.

E. mean of the sample mean random variable is always less than the mean of the population mean for any sample size n.

Explanation:

The correct answer is B. This is exactly what the Central Limit Theorem guarantees for large n. Think of it as saying for large enough sample size, we expect the average of the data values to estimate well the target, which is the population mean μ.

Does playing the television during an 8-hour workday reduce a pet Siberian Husky dog's activity level during the day? An experiment was conducted where a group of Siberian Huskies was divided into two groups. The television was played in the household for one group, and it was not played for the control group. Activity level was assessed as being the number of hours the dog was engaged in activities other than lying down or eating. The average decrease in activity level for the groups measured is 3.6 hours.

A 95% confidence interval for the difference (treatment – control) in the mean activity levels was computed to be (2.5, 4.7). Which of the following is an accurate interpretation of this interval?

A. We do not know the true decrease in activity level in Siberian Huskies due to television exposure, but we are 95% confident that the increase in the mean decrease lies in this interval.

B. Because the confidence interval does not include zero, we are 95% confident that the true decrease in activity level in Siberian Huskies is 3.6 hours.

C. We are 95% confident that the average decrease in activity level in the sample is 3.6 hours.

D. Because the confidence interval does not contain zero, we are 95% confident that there was no effect of playing the television on decreasing activity level in Siberian Huskies.

E. The activity level of 95% of the Siberian Huskies decreased by between 2.5 and 4.7 hours.

Explanation:

The correct answer is A. There are various ways to interpret "confidence," one of which is listed here. What we *can* infer that since the left endpoint of the confidence interval is greater than zero, we can be 95% confident that playing the television had an effect on decreasing activity level. We do not have the raw data to confirm the statement in choice E, and this conclusion need not be true in general. There could be several extreme outliers that prevent this conclusion from holding true.

Inference for Categorical Data: Proportions

On your AP Statistics exam, about 12–15% of questions will fall under the category of Inference for Categorical Data: Proportions.

Overview of Confidence Intervals

A **point estimate** of a parameter is a single value that is used as an estimate of a parameter value. A **confidence interval** for a parameter is an interval in which the parameter is likely to lie.

Construction of a confidence interval, regardless of the parameter it is being used to estimate, follows several common steps:

1. **Check the relevant conditions**. These will vary depending on the parameter being estimated, but generally include a condition involving independence of samples, and a condition that assures normality of the relevant sampling distribution.
2. **Choose a confidence level for the interval**. This is a percentage that gives the confidence with which the interval constructed will contain the parameter. Values of 90%, 95%, and 99% are most common.
3. **Calculate a point estimate for the parameter**. This will be the corresponding statistic calculated from the sample data.
4. **Calculate the standard error of the sampling distribution**. This is used as an estimate for the standard deviation of the sampling distribution, which is usually unknown, and is abbreviated *SE*.
5. **Find the critical value associated with the chosen confidence level**. This will be based on a standard distribution that varies depending on the parameter of interest. It is denoted with an asterisk, as in z^* or t^*.
6. **Calculate the margin of error**. The margin of error is half the width of the confidence interval and is equal to the product of the critical value and the standard error.
7. **Compute the endpoints of the confidence interval**. These are given by subtracting the margin of error from, and adding the margin of error to, the point estimate.

It is important to realize that not every confidence interval for a parameter will contain the true value of the parameter. Since each interval is constructed using a random sample, some of them will not, in fact, contain the population parameter. However, if we were to repeatedly obtain a new random sample, and construct a confidence interval, with a confidence level of C%, based on that sample, we would find over time that approximately C% of the intervals thus constructed would contain the population parameter.

Free Response Tip

Always provide a concluding, interpretive statement when constructing a confidence interval. It should include the confidence level, a descriptive context of what the data represents, the population from which the data was gathered, and the interval itself. The examples in each of the following sections give examples of such a statement.

As sample size increases, the width of a confidence interval decreases (assuming all other values remain constant). On the other hand, the width of a confidence interval increases as the confidence level increases (again, assuming all other values are unchanged). This means that if you want to have a smaller confidence interval, you have two choices: decrease the confidence level, or increase the sample size.

Confidence Intervals for One Proportion

There are two conditions that need to be checked when constructing a confidence interval for a proportion *p* based on a sample proportion $\hat{p}$.

1. Independence of samples: this is verified by the data being collected randomly or by the experiment being done with random assignment. If sampling is done without replacement, it is necessary that the sample size be less than 10% of the population size.
2. Normality of the sampling distribution of $\hat{p}$: both $n\hat{p}$ and $n(1-\hat{p})$ need to be at least 10.

The standard error of $\hat{p}$ is $SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Since the sampling distribution of $\hat{p}$ is normal, the critical value it is a *z*-value, notated $z*$, for which the desired percentage of a normal distribution lies between $-z^*$ and z^*. The margin of error is then given by $ME = z*\cdot SE_{\hat{p}} = z*\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Finally, the confidence interval is as follows:

$$(\hat{p}-ME, \hat{p}+ME) = \left(\hat{p}-z*\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p}+z*\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right).$$

For example, consider a college student who wants to estimate, with 95% confidence, the proportion of the 12,000 students at her school who have stayed awake all night studying for an exam. She surveys a random sample of 96 students and finds that 24 of them say that they have done so.

First, we must check the conditions. The student population was randomly sampled, and 96 is certainly less than 10% of 12,000. Additionally, there were 24 students who answered yes to the survey, and 72 who answered no, so both $\hat{p}$ and $1-\hat{p}$ are greater than 10.

Based on the given numbers, $\hat{p}=\frac{24}{96}=0.25$ and $SE_{\hat{p}}=\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}=\sqrt{\frac{0.25(1-0.25)}{96}}\approx 0.044$ By examining a standard normal table, or consulting a calculator, we find that the interval $(-1.96,1.96)$ contains 95% of a standard normal distribution. Thus, we have $z^*=1.96$. We can therefore calculate the margin of error as $ME=z^*\cdot SE_{\hat{p}}=1.96*0.044\approx 0.086$, so the 95% confidence interval is $(\hat{p}-ME,\hat{p}+ME)=(0.25-0.086,0.25+0.086)=(0.164,0.336)$.

We are 95% confident that the proportion of students at this college who have stayed awake all night studying for an exam is between 16.4% and 33.6%.

Confidence Intervals for Two Proportions

There are two conditions that need to be checked when constructing a confidence interval for a difference of two proportions, p_1-p_2, based on the difference of sample proportions, $\hat{p}_1-\hat{p}_2$.

1. **Independence of samples**: this is verified by the data being collected randomly or by the experiment being done with random assignment. If sampling is done without replacement, it is necessary that both sample sizes be less than 10% of their respective population sizes.
2. **Normality of the sampling distribution of $\hat{p}$**: all of $n_1\hat{p}_1$, $n_1(1-\hat{p}_1)$, $n_2\hat{p}_2$, and $n_2(1-\hat{p}_2)$ need to be at least 10.

The standard error of $\hat{p}$ is $SE_{\hat{p}_1-\hat{p}_2}=\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$. Since the sampling distribution of $\hat{p}_1-\hat{p}_2$ is normal, the critical value is a *z*-value, notated z^*, for which the desired percentage of a normal distribution lies between $-z^*$ and z^*. The margin of error is then given by $ME = z^* \cdot SE_{\hat{p}_1 - \hat{p}_2} = z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$. Finally, the confidence interval is as follows: $\left(\hat{p}_1 - \hat{p}_2 - z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \hat{p}_1 - \hat{p}_2 + z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right)$.

For example, consider the question of how more or less likely men are to drive a red car compared to women at the 90% confidence level. Suppose that from a random sample of 200 men and 240 women, 12 men and 18 women drive red cars. Let p_1 be the proportion of men who drive red cars, and p_2 be the proportion of women who drive red cars. Then we have $\hat{p}_1 = \frac{12}{200} = 0.06$ and $\hat{p}_2 = \frac{18}{240} = 0.075$, so $\hat{p}_1 - \hat{p}_2 = -0.015$.

The sample was stated to be random, and both 200 and 240 are certainly less than 10% of the population of men and women. Additionally, there were at least each of the possible categories (men who drive red cars, men who don't drive red cars, women who drive red cars, and women who don't drive red cars), so all of the conditions are met.

We can now calculate $SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{0.06(1-0.06)}{200} + \frac{0.075(1-0.075)}{240}} \approx 0.024$. By examining a standard normal table, or consulting a calculator, we find that the interval $(-1.645, 1.645)$ contains 90% of a standard normal distribution, so we have $z^* = 1.645$. We can now calculate the margin of error as $ME = z^* \cdot SE_{\hat{p}_1 - \hat{p}_2} = 1.645 * 0.024 \approx 0.0395$, so the 95% confidence interval is $(-0.015 - 0.0395, -0.015 + 0.0395) = (-0.0545, 0.0245)$.

We are 90% confident that the difference in the proportions of men and women who drive red cars is between -5.45% and 2.45%.

Note that this interval contains both positive and negative values. This means that based on the data, we cannot be 90% confident that women are more likely than men to drive red cars.

Overview of Hypothesis Testing

A **hypothesis test** is a procedure for testing a claim about a population based on a sample. The **null hypothesis***,* denoted H_0, is the assumption about that population that is assumed to be correct, and can only be rejected if statistical evidence is sufficiently strong. The **alternative hypothesis**, denoted H_a, is the alternative possibility about the population for which statistical evidence is gathered.

The null hypothesis consists of an equality statement about a population parameter, while the alternative hypothesis contains an inequality. If the alternative hypothesis involves the < or > inequalities, it is referred to as **one-sided**. If it involves the $\neq$ inequality, it is **two-sided**.

The following steps outline the procedure for performing a hypothesis test:

1. Check conditions for independence and a normal sampling distribution. These conditions are usually completely analogous to those for constructing confidence intervals.
2. State the null and alternative hypotheses.
3. Choose a **significance level** for test. This significance level is referred to as α . Common values are 0.1, 0.05, and 0.01. These are sometimes given as percentages: 10%, 5%, and 1%.
4. Compute a **test statistic**. This is usually a standardized score (*z* or *t*) from a probability distribution.
5. Using the test statistic, determine a *p*-value. A *p*-value is the probability of obtaining a test statistic at least as extreme as the one you got, under the assumption that the null hypothesis is true. This will depend on whether the alternative hypothesis is one-sided or two-sided.
6. Decide whether or not to reject the null hypothesis. If p £ a, reject the null hypothesis. If $p > \alpha$, fail to reject the null hypothesis.
7. Write a conclusion in context of the problem as stated.

The null hypothesis can never be confirmed by a hypothesis test. Rather, there is either sufficient evidence to reject, or there is not. Failure to reject it should not be mistaken for evidence in favor of it.

Hypothesis Test Errors

There are two types of errors that can occur when performing a hypothesis test. A **Type I error** occurs when the null hypothesis is rejected even though it is, in fact, true. This is sometimes referred to as a **false positive**. The probability of a Type I error occurring is precisely α , the significance level chosen for the test.

A Type II error, also called a **false negative**, occurs when the null hypothesis is not rejected, even though it is actually false. The **power** of a hypothesis test is the probability that the null hypothesis will be correctly rejected if it is actually false. The probability of a Type II error occurring, denoted β , is then related to the power of a test: $\beta + power = 1$.

α and β have an inverse relationship with each other: if other values remain the same, then one of these increases while the other decreases. This means that reducing the chances of committing one type of error increases the chances of making the other type of error. A decision often has to be made, in context, as to whether it is more necessary to avoid a Type I or Type II error.

There are other factors that influence β. It decreases when the sample size increases, when the standard error of the statistic decreases, and when the null hypothesis is farther from the true value of the parameter. If one of these occur, then the probability of a Type II error is decreased without an increase in the probability of a Type I error.

Hypothesis Tests for One Proportion

The null hypothesis for a population proportion is $H_0 : p = p_0$ (for some hypothesized population proportion p_0 determined in context), and the alternative hypothesis is one of $H_a : p < p_0$, $H_a : p > p_0$, or $H_a : p \neq p_0$.

The necessary conditions for independence are that the sample is random, and if the sampling is done without replacement, that the sample size is less than 10% of the population. Normality of the sampling distribution is confirmed by checking that $np_0 \geq 10$ and $n(1-p_0) \geq 10$. Note that the proportion used here is the one assumed in the null hypothesis.

Since the sampling distribution of $\hat{p}$ is normal, the test statistic is a *z*-statistic; that is, it comes from a standard normal distribution. Its value is $z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$.

For example, a major car manufacturer claims that 97% of its customers are happy with their cars. An industry analyst thinks the true number is lower. In a survey of 600 randomly chosen customers, he finds that 573 are happy with their cars. Is there sufficient evidence to reject the manufacturer's claim at the 5% significance level?

First note that the necessary conditions are satisfied: the sample is random, it is (presumably) less than 10% of the population, and both 573 and 600 – 573 = 27 are at least 10. Also, we are given $\alpha = 0.05$.

The null hypothesis is $H_0 : p = 0.97$, and the alternative is $H_a : p < 0.97$. The other relevant values are $\hat{p} = \dfrac{573}{600} = 0.955$, and *n* = 600. Now we can calculate the test statistic: $z = \dfrac{0.955 - 0.97}{\sqrt{\dfrac{0.97(1-0.97)}{600}}} \approx -2.15$. Since this is a one-sided test, with $H_a : p < 0.97$, we need the left-tailed probability of a standard normal value of -2.15. This value is *p* = 0.0158.

Since $\alpha = 0.05$, we have $p < \alpha$, so we reject the null hypothesis. At the 5% significance level, there is sufficient evidence to reject the manufacturers claim that the proportion of its customers who are happy with their cars is 97%.

Note that if the test were conducted again with $\alpha = 0.01$, we would fail to reject the null hypothesis.

Hypothesis Tests for Two Proportions

When conducting a hypothesis test for the difference of two proportions, the null hypothesis is $H_0 : p_1 = p_2$. The one-sided alternative hypotheses are $H_a : p_1 < p_2$ or $H_a : p_1 > p_2$, and the two-sided alternative is $H_a : p_1 \neq p_2$.

The samples need to be random, and both should represent less than 10% of their respective populations. To check that the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is normal, we use a *pooled*, or *combined* proportion: $\hat{p}_c = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$. We then require that each of $n_1\hat{p}_c$, $n_1(1-\hat{p}_c)$, $n_2\hat{p}_c$, and $n_2(1-\hat{p}_c)$ be at least 10.

If these conditions are met, the sampling distribution is normal, so the test statistic is again a *z*-statistic. It is given by $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c(1-\hat{p}_c)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$, where $\hat{p}_c$ is the pooled proportion as described.

For example, a researcher believes that high school students in Pennsylvania and Ohio are equally likely to have taken a statistics course. In a random sample of 255 Pennsylvania high school students, 22% indicated that they have taken a statistics course. In a random sample of 340 Ohio high school students, 20% indicated the same. Let us test this researcher's claim at the 0.01 significance level.

The null hypothesis is $H_0 : p_1 = p_2$, and the alternative hypothesis is $H_a : p_1 \neq p_2$. The necessary conditions are all easily checked. The samples are random, and clearly represent less than 10% of all high school students in their respective states. In both states, the number of respondents who did take a statistics course and the number who did not are both at least 10.

The pooled proportion is $\hat{p}_c = \dfrac{255 \cdot 0.22 + 340 \cdot 0.2}{255+340} \approx 0.2086$. The test statistic is $z = \dfrac{0.22-0.2}{\sqrt{0.2086(0.7914)\left(\dfrac{1}{255}+\dfrac{1}{340}\right)}} \approx 0.59$. Since the alternative hypothesis is two-sided, we need to find the probability associated with a *z*-score of less than -0.59 or greater than 0.59. This probability is $p = 0.5552$.

Clearly $p > \alpha$, so we fail to reject the null hypothesis. There is not sufficient evidence to reject the researcher's claim that the proportions of high school students in Pennsylvania and Ohio who have taken a statistics course are equal.

Suggested Reading

- Starnes & Tabor. *The Practice of Statistics.* 6th edition. Chapters 8–10. New York, NY: Macmillan.
- Larson & Farber. *Elementary Statistics: Picturing the World*. 7th edition. Chapters 6–8. New York, NY: Pearson.
- Bock, Velleman, De Veaux, & Bullard. *Stats: Modeling the World*. 5th edition. Chapters 18–21. New York, NY: Pearson.
- Sullivan. *Statistics: Informed Decisions Using Data.* 5th edition. Chapters 9–11. New York, NY: Pearson.
- Peck, Short, & Olsen. *Introduction to Statistics and Data Analysis*. 6th edition. Chapters 9–11. Boston, MA: Cengage Learning.

Sample Inference for Categorical Data: Proportions Questions

During the pre-deployment check of a city snowplow, the driver discovers a warning light that indicates that the coolant level may be low. If he decides to check the coolant level, it will delay the deployment by 30 minutes. If he decides to ignore the warning light, the snowplow might overheat before completing its route. If a hypothesis were to be performed, which of the following would be the appropriate null hypothesis and Type I error?

A.

Null hypothesis: assume the coolant level should be checked

Type I error: decide to ignore the warning light when, in fact, the coolant level was low

B.

Null hypothesis: assume the coolant level should be checked

Type I error: decide to check the coolant level when, in fact, the level is sufficiently high

C.

Null hypothesis: assume that the snowplow already has a delayed deployment

Type I error: decide to ground the snowplow until repairs can be made even if it is not broken

D.

Null hypothesis: assume the warning light can be ignored

Type I error: decide to ignore the warning light when, in fact, the coolant level was low

E.

Null hypothesis: assume the warning light cannot be ignored

Type I error: decide to check the coolant level when, in fact, the level is low

Explanation:

The correct answer is A. The status quo is that the warning light is correct unless there is sufficient evidence suggesting otherwise. A Type I error is committed when rejecting a true null hypothesis, which is what is listed. Choice B is incorrect because while the null hypothesis listed is correct, the Type I error is not; a Type I error is committed when rejecting a true null hypothesis. Choice C is incorrect because there is no indication in the described scenario that the snowplow is malfunctioning; the null hypothesis should be the status quo (meaning the driver should follow the recommendation of the warning light), and a Type I error is committed when rejecting a true null hypothesis. Choice D is incorrect because if this is taken to be the null hypothesis, the type of error described is a Type II error, not Type I. Choice E is incorrect because if this is taken to be the null hypothesis, what is listed as a Type I error is not an error at all.

You have measured the daily water intake of a random sample of students that regularly visit the school recreation center. A 99% confidence interval for the mean daily water intake (in fluid ounces) for these students is computed to be (78, 90). Which of the following statements is a valid interpretation of this interval?

A. If the sampling procedure were repeated many times, then approximately 99% of the sample means would be between 78 and 90.

B. The probability that the sample mean of the students sampled falls between 78 and 90 is equal to 0.99.

C. If the sampling procedure were repeated many times, then approximately 99% of the resulting confidence intervals would contain the mean daily water intake for all students who regularly visit the recreation center.

D. About 99% of the sample of students has a daily water intake between 78 and 90 fluid ounces.

E. About 99% of the students who regularly visit the recreation center have a daily water intake between 78 and 90 fluid ounces.

Explanation:

The correct answer is C. There are various ways to interpret “confidence,” and the one presented in this choice is one of them. Choice A is incorrect because you cannot draw such a conclusion about the means of other samples. The interpretation of a confidence interval concerns how many such intervals we would expect to contain the *true* mean daily water intake. Choice B is incorrect because the sample mean *is* definitely in this interval; in fact, it is the midpoint since the margin of error is added and subtracted from it to generate the interval. So, this probability is 1.00. Choice D is incorrect because without having the actual raw data, we cannot determine the number of students for which this is true. There could be a couple of extreme outliers in the data set that could throw off the percentage of those with daily water intake amounts in the interval. Choice E is incorrect because you cannot use a single confidence interval to make such a conclusion about the whole population.

The owner of an electronics store believes that the latest shipment of batteries contains more “dead on arrival” batteries than the usual 2%. He tests the null hypothesis that the proportion of dead batteries equals 2% against the alternative hypothesis that the proportion is greater than 2%. The results of a simple random sample of 53 randomly-chosen packages in this shipment are given by:

Test of H_0: $p = 0.02$ versus H_A: $p > 0.02$

Success = Battery is dead on arrival

Variable	*X*	*n*	Sample *p*	*p* value
Dead on arrival	12	53	0.2264	0.0006

Which of the following conclusions can be reached?

I. The *p*-value of 0.0006 indicates that it is not very likely to get an observed value of 0.2264 if the null hypothesis is true.

II. The owner can be quite confident that this batch has significantly more “dead on arrival” batteries than the typical shipment.

III. The *p*-value of 0.0006 tells us that we cannot reject the null hypothesis and that the shipment has 2% or less “dead on arrival” batteries.

A. I only

B. II only

C. III only

D. I and II only

E. II and III only

Explanation:

The correct answer is D. Statement I is true because the *p*-value is the probability of rejecting a true null hypothesis. As such, it is very unlikely to have gotten a test statistic value of 0.2264 if the null hypothesis is, in fact, true. Statement II is true because the *p*-value is so low that we reject the null hypothesis in favor of the alternative hypothesis that $p > 0.02$. So, the owner can be quite confident that this batch has significantly more “dead on arrival” batteries than the typical shipment. Statement III is false; this low of a *p*-value means that you *do* reject the null hypothesis in favor of the alternative hypothesis.

Inference for Quantitative Data: Means

About 10–18% of the questions on your AP Statistics exam will cover the topic Inference for Quantitative Data: Means.

The *t*-distribution

Constructing confidence intervals and carrying out hypothesis tests involving means follows the same steps as for proportions. However, the formulas for the standard errors and test statistics all involve the standard deviation, σ, of the population in question. Since this is usually unknown, the best we can do is estimate it using the sample standard deviation, *s*. When this is done, the resulting distribution is no longer normal. Instead, it follows a distribution, called a ***t*-distribution**, that resembles a normal distribution but has more of its probability in its tails.

There is not a single *t*-distribution. Rather, there is an infinite family of them that depend on a parameter called the **degrees of freedom**, abbreviated *df*. The value of *df* used in a particular situation usually depends on the sizes of the samples involved. As *df* increases, the *t*-distribution resembles a normal distribution more closely.

Confidence Intervals for One Mean

The conditions that need to be tested when constructing a confidence interval, as usual, are related to independence and normality of the sampling distribution. Specifically, the sample should be random, and when sampling without replacement, should represent less than 10% of the population. In addition, the sample size should be at least 30, or the sample data should be symmetric and free of outliers.

If these conditions are met, the standard error is given by $SE_{\overline{x}} = \frac{s}{\sqrt{n}}$, where *s* is the sample standard deviation. Since *s* is being used as an estimate for σ, the relevant distribution here is a *t*-distribution with *df* = *n* – 1, where as usual, *n* is the sample size. Therefore, the critical value is a *t*-value, $t*$. The margin of error, in turn, is $ME = t*\cdot\frac{s}{\sqrt{n}}$. The confidence interval is then constructed as before. Since the point estimate of μ is $\overline{x}$, the interval is $\left(\overline{x} - ME, \overline{x} + ME\right) = \left(\overline{x} - t*\cdot\frac{s}{\sqrt{n}}, \overline{x} + t*\cdot\frac{s}{\sqrt{n}}\right)$.

For example, a sociologist is interested in the number of hours per week that adults in the United States spend on social media. He collects a random sample of 78 adults, and surveys them on their social media habits. He finds a sample mean of 4.7 hours per week, with a standard deviation of 1.4.

To construct a 90% confidence interval for the mean number of hours per week spent on social media by all adults in the United States, we begin by checking the conditions. The sample size is greater than 30, and the data is from a random sample that is certainly less than 10% of the population.

The point estimate is $\bar{x} = 4.7$ and the sample standard deviation is $s = 1.4$. Calculating the standard error, we get $SE_{\bar{x}} = \frac{1.4}{\sqrt{78}} \approx 0.16$. The critical *t*-value with $n-1=77$ degrees of freedom for which 90% of the distribution lies between -*t* and *t* is 1.665. Therefore, the margin of error is $ME = 1.665 \cdot 0.16 = 0.2664$. Finally, the 90% confidence interval is $(4.7-0.2664, 4.7+0.2664) = (4.4336, 4.9664)$.

We are 90% confident that the mean number of hours per week spent on social media by adults in the United States is between 4.4336 and 4.9664.

Confidence Intervals for Two Means

Constructing a confidence interval for the difference between two independent means are like those for a single mean. The sampling should be random, consist of less than 10% of each population, and both samples should either be at least 30 or symmetric and free of outliers.

The point estimate is $\bar{x}_1 - \bar{x}_2$ and the standard error is $SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$. The critical value comes from a *t*-distribution. The correct value for df depends on the sample sizes, but is not usually practical to calculate by hand. It can be calculated using technology or will be provided. As usual, the margin of error is given by $ME = t^* \cdot SE_{\bar{x}_1-\bar{x}_2}$, and the confidence interval is $(\bar{x}_1 - \bar{x}_2 - ME, \bar{x}_1 - \bar{x}_2 + ME)$.

Hypothesis Tests for Means

When conducting a hypothesis test for a mean, the null hypothesis is $H_0 : \mu = \mu_0$, where μ_0 is the hypothesized value of the mean. The alternative hypothesis is either $H_a : \mu < \mu_0$, $H_a : \mu > \mu_0$, or $H_a : \mu \neq \mu_0$.

The sample should be random, and if collected without replacement, should be less than 10% of the population. If the sample size is not at least 30, the distribution of the sample data should be symmetric.

The test statistic comes from a *t*-distribution with $df = n-1$, and is $t = \dfrac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$. The *p*-value is computed from this distribution, and the interpretation remains the same as usual.

When comparing two means, μ_1 and μ_2, the null hypothesis is $H_0 : \mu_1 = \mu_2$. The possible alternative hypotheses are $H_a : \mu_1 < \mu_2$, $H_a : \mu_1 > \mu_2$, and $H_a : \mu_1 \neq \mu_2$. Both samples should meet the conditions given for the case of a single mean.

The test statistic is given by $t = \dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$. As mentioned, the degrees of freedom for this *t*-distribution are impractical to compute manually, but can be found using technology or will be given. As usual, a *p*-value is found using this distribution, and the conclusion then depends on whether $p \leq \alpha$ or $p > \alpha$.

Inference with Matched Pairs

When the two means to be compared are matched pairs, they should be treated as a single mean. First, find the difference of each pair, and then follow the inference procedures for single mean confidence intervals and hypothesis tests, using the differences as the sample.

For example, consider the following table, which shows the heart rates, in beats per minute, of seven people both after and before a 100-meter sprint.

	Person #1	Person #2	Person #3	Person #4	Person #5	Person #6	Person #7
After	82	103	99	84	110	91	125
Before	63	65	70	59	72	72	79

Since these are matched pairs, any inference done with them should first begin with finding differences. Here are the seven values obtained by subtracting the "before" heart rate from the "after" heart rate.

	Person #1	Person #2	Person #3	Person #4	Person #5	Person #6	Person #7
Difference	19	38	29	25	38	19	46

To construct a confidence interval for $\mu_1 - \mu_2$, follow the procedures for constructing a confidence interval for a single mean, and use the sample values given by the differences.

Free Response Tip

Be careful when presented with two samples. The fact that two samples have the same size does not mean they are necessarily matched pairs. There must be a meaningful and direct connection between the samples (such as coming from the same person, as in the heart rate example) for them to be matched pairs. In addition, be certain that the order of subtraction is well-defined and applied consistently for every pair.

Suggested Reading

- Starnes & Tabor. *The Practice of Statistics.* 6th edition. Chapters 8–10. New York, NY: Macmillan.
- Larson & Farber. *Elementary Statistics: Picturing the World*. 7th edition. Chapter 6–8. New York, NY: Pearson.
- Bock, Velleman, De Veaux, & Bullard. *Stats: Modeling the World*. 5th edition. Chapters 22–24. New York, NY: Pearson.
- Sullivan. *Statistics: Informed Decisions Using Data.* 5th edition. Chapters 9–11. New York, NY: Pearson.
- Peck, Short, & Olsen. *Introduction to Statistics and Data Analysis*. 6th edition. Chapters 9–11. Boston, MA: Cengage Learning.

Sample Inference for Quantitative Data: Means Questions

You have measured the daily water intake of a random sample of students that regularly visit the school recreation center. A 99% confidence interval for the mean daily water intake (in fluid ounces) for these students is computed to be (78, 90). Which of the following statements is a valid interpretation of this interval?

A. If the sampling procedure were repeated many times, then approximately 99% of the sample means would be between 78 and 90.

B. The probability that the sample mean of the students sampled falls between 78 and 90 is equal to 0.99.

C. If the sampling procedure were repeated many times, then approximately 99% of the resulting confidence intervals would contain the mean daily water intake for all students who regularly visit the recreation center.

D. About 99% of the sample of students has a daily water intake between 78 and 90 fluid ounces.

E. About 99% of the students who regularly visit the recreation center have a daily water intake between 78 and 90 fluid ounces.

Explanation:

The correct answer is C. There are various ways to interpret "confidence," and the one presented in this choice is one of them. Choice A is incorrect because you cannot draw such a conclusion about the means of other samples. The interpretation of a confidence interval concerns how many such intervals we would expect to contain the *true* mean daily water intake. Choice B is incorrect because the sample mean *is* definitely in this interval; in fact, it is the midpoint since the margin of error is added and subtracted from it to generate the interval. So, this probability is 1.00. Choice D is incorrect because without having the actual raw data, we cannot determine the number of students for which this is true. There could be a couple of extreme outliers in the data set that could throw off the percentage of those with daily water intake amounts in the interval. Choice E is incorrect because you cannot use a single confidence interval to make such a conclusion about the whole population.

A rock gym manager used a random sample of 300 rock climbers to obtain a 95% confidence interval for the mean time (in minutes) it takes to complete a difficult climbing route. This interval was (12.2, 14.3). If he had used a 90% confidence interval instead, the confidence interval would have been

A. wider and would have involved a smaller risk of being incorrect.

B. narrower and would have involved a smaller risk of being incorrect.

C. wider, but it cannot be determined whether the risk of being incorrect would be greater or smaller.

D. wider and would have involved a larger risk of being incorrect.

E. narrower and would have involved a larger risk of being incorrect.

Explanation:

The correct answer is E. As the confidence interval decreases, we are less "confident" that the interval we produce actually contains the true mean. The only way for this to occur is to shrink the interval (or make the target smaller, so to speak). In so doing, there is a greater chance that the interval does not contain the true mean, so the risk of being incorrect is larger.

You have sampled 30 amateur tennis players in New York to determine the mean first-serve ball speed. A 95% confidence interval for the mean first-serve ball speed is 85 to 94 miles per hour. Which of the following statements gives a valid interpretation of this interval?

A. If this sampling procedure were repeated several times, 95% of the resulting confidence intervals would contain the true mean first-serve ball speed of amateur tennis players in New York.

B. If this sampling procedure were repeated several times, 95% of the sample means would be between 85 and 94.

C. If 100 samples were taken and a 95% confidence interval was computed, 5 of the intervals would be included within the confidence interval (85, 94).

D. 95% of the population of all New York amateur tennis players have a first-serve ball speed between 85 and 94 miles per hour.

E. 95% of the 30 amateur tennis players sampled have a mean first-serve ball speed between 85 and 94 miles per hour.

Explanation:

The correct answer is A. There are various ways to interpret “confidence,” one of which is described here as the result of repeated sampling. Choice B is incorrect because 85 and 94 were the endpoints of a specific confidence interval that was linked to a particular sample, and hence a particular sample mean. Different samples will generate different sample means, and the one that was used to produce the given confidence interval could have contained extreme values not characteristic of the overall population. Choice C is incorrect because there is no reason to believe that a confidence interval formed from a single sample of 30 players should have this strong of a relationship to confidence intervals formed using different samples. The 100 confidence intervals formed may very well overlap in various ways, but they will not, in general, be included within any given confidence interval. Choice D is incorrect because different samples will generate different sample means, and the one that was used to produce the given confidence interval could have contained extreme values not characteristic of the overall population. So, there is no way to tell if 95% of all such tennis players have a first-serve ball speed in this interval. Choice E is incorrect because there is no way to determine this from the information provided. The tennis players could have put forth much more of an effort to produce their fastest serves when the data was being collected. So, their individual average first-serve ball speeds may very well be much lower.

Inference for Categorical Data: Chi-Square

On your AP exam, 2–5% of the questions you encounter will cover the topic of Inference for Categorical Data: Chi-Square.

Expected Counts

With categorical data, one can propose a variety of hypotheses that pertain to absolute or relative counts of the different categories. Usually, the null hypothesis is presented in the form of proportions of the population that fall into different categories. An **expected count** is a count that is consistent with the null hypothesis.

The value used to measure the difference between the expected counts and the observed counts comes from a χ^2, or chi-square distribution. Like the *t*-distributions, this actually refers to a family of distributions that depend on the degrees of freedom associated with a sample. Chi-square distributions are positive and right skewed.

Goodness-of-Fit Test

The chi-square goodness-of-fit test is used to determine whether a population fits a certain distribution. The null hypothesis is given in the form of a proportion for every possible category. The alternative hypothesis simply states that at least one of the given proportions is not correct.

The usual condition for independence applies: the sample should be random, and when done without replacement, should be less than 10% of the population. The other condition for this test is that all expected counts (as described) should be greater than 5.

To carry out the test, first find all expected counts. The expected count for each category is calculated by multiplying the hypothesized proportion for that category by the sample size. The test statistic is then given by $\chi^2 = \sum \frac{(O - E)^2}{E}$, where *O* is the observed count, *E* is the expected count, and the sum is taken over all categories. This statistic comes from a distribution with degrees of freedom equal to one less than the number of categories. The *p*-value is the probability of obtaining a test statistic at least as large as the one found. Since the distribution is positive, this is a right-tail area, which can be found using a table or calculator. If $p \le \alpha$, the null hypothesis is rejected, and if $p > \alpha$, it is not rejected.

Homogeneity Test

A chi-square homogeneity test determines whether two or more populations follow the same categorical distribution. The null hypothesis is that the distribution of the categorical variable across the populations is the same, whereas the alternative hypothesis is that the distributions are not all the same. Note that this alternative does not preclude one or more of the values having the same proportions across populations, only that the overall distributions are not identical. The necessary conditions for this test are that the data come from a simple random sample, and that all expected counts be at least 5.

Since the sample data come from two populations, it is represented in tabular form. The expected count for each cell in the table is given by $E = \frac{(row\ total)\cdot(column\ total)}{table\ total}$. The test statistic for the homogeneity test is $\chi^2 = \sum \frac{(O-E)^2}{E}$, just as with the goodness-of-fit test. However, the degrees of freedom is now equal to (*number of rows* – 1)(*number of columns* – 1). As with the goodness-of-fit test, the *p*-value is given by the right tail of the appropriate chi-square distribution.

For example, the following table shows the number of men and women who responded to a question asking which of four factors they consider most important in the workplace. We will conduct a chi-square homogeneity test at the $\alpha = 0.1$ significance level.

	Compensation	Convenience	Respect	Engagement	Total
Men	13	5	7	5	30
Women	20	10	12	18	60
Total	33	15	19	23	90

The expected counts, calculated using the formula given, are as follows. For example, the first cell is calculated as $\frac{30\cdot 33}{90} = 11$

	Compensation	Convenience	Respect	Engagement
Men	11	5	6.33	7.67
Women	22	10	12.67	15.33

We can now calculate the test statistic. The first cell contributes $\frac{(11-13)^2}{11} \approx 0.36$, the second cell contributes $\frac{(5-5)^2}{5} = 0$, and so on for the other six cells. The total for all of them is $\chi^2 \approx 2.042$, and there are $(2-1)(4-1) = 3$ degrees of freedom. The probability of obtaining a χ^2 of 2.042 or greater in this distribution is 0.5637. Since $p > \alpha$, we fail to reject the null hypothesis. There is not sufficient evidence to conclude that the distribution of this variable is different between men and women.

Independence Test

The chi-square independence test is nearly identical in execution to the homogeneity test. However, it is relevant in a different situation, and this is reflected in the null and alternative hypotheses. Unlike the homogeneity test, which is concerned with a categorical variable across different populations, the independence test deals with two different categorical variables from a single population.

The question it helps answer is whether the two variables are independent of each other, or whether there is a dependency relationship between them. The null hypothesis is that the two variables are independent, and the alternative hypothesis is that they are dependent.

Whereas the homogeneity test required simple random samples, the independence test requires stratified random samples. The expected counts, test statistic, degrees of freedom, and interpretation of the *p*-value carry over from the homogeneity test without any changes.

Suggested Reading

- Starnes & Tabor. *The Practice of Statistics.* 6th edition. Chapter 11. New York, NY: Macmillan.
- Larson & Farber. *Elementary Statistics: Picturing the World*. 7th edition. Chapter 10. New York, NY: Pearson.
- Bock, Velleman, De Veaux, & Bullard. *Stats: Modeling the World*. 5th edition. Chapter 25. New York, NY: Pearson.
- Sullivan. *Statistics: Informed Decisions Using Data.* 5th edition. Chapter 12. New York, NY: Pearson.
- Peck, Short, & Olsen. *Introduction to Statistics and Data Analysis*. 6th edition. Chapter 12. Boston, MA: Cengage Learning.

Sample Inference for Categorical Data: Chi-Square Questions

A survey was conducted to study people's attitudes toward music containing explicit lyrics. A random sample of 1000 adults in the age categories 18 − 25, 26 − 35, and 36 − 45 were selected to participate. They were classified according to age group and response to the question, "Do you think there is a link between listening to music containing explicit lyrics and bullying?"

The data are:

	YES	NO
18 – 25	20	350
26 – 35	180	200
36 – 45	250	0

Which of the following statements is NOT correct?

A. The null hypothesis is that age category and opinion about bullying are independent.

B. The alternate hypothesis is that the proportion of people in various age groups who say YES is different from the proportion in various age groups who say NO.

C. A Type I error would be to conclude that opinions differed across age groups when, in fact, they do not.

D. A Type II error would be to conclude that opinions are the same across age groups when they are actually different.

E. The Chi-square test for independence cannot be performed because all cells must have nonzero entries in order to do so.

Explanation:

The correct answer is E. There is no such requirement to perform a Chi-square test. You might be mistaking observed frequencies for expected frequencies. Often, a minimum *expected* frequency of 5 is imposed to run this test, but this is not true of the *observed* frequencies. The other statements are all true.

Are all members of the kitchen staff equally prone to the same types of injuries? To investigate this question, a restaurant owner asks a consultant to classify accidents reported this month for all the restaurants she owns by type and the job performed in the kitchen. The results are tabulated below:

	Accident Type		
Role in Kitchen	**Sprain**	**Burn**	**Cut**
Cook	3	20	7
Prep Station	8	2	10
Dishwasher	7	8	11

A Chi-square test for independence was performed and gave a test statistic of 11.24. If we test at the significance level α=0.05, which of the following is true?

A. There appears to be no association between accident type and kitchen role.
B. Role seems to be independent of accident type.
C. Accident type does not seem to be independent of kitchen role.
D. There appears to be an 11.24% correlation between accident type and kitchen role.
E. The proportion of sprains, burns, and cuts appear to be similar across all kitchen roles.

Explanation:

The correct answer is C. Note that df = (3 – 1)(3 – 1) = 4. At the α = 0.05 significance level, the cut-off for the critical region is 9.49. Since the test statistic is larger than this value, we would reject the null hypothesis of independence of accident type and role in kitchen in favor of the existence of some association between them.

A random sample of 100 people was asked to state whether he or she graduated from college and to state his or her TV show preference (Reality TV, Sit-Coms, or Crime Drama). The results are provided below:

	TV Genre Preference		
College Graduate?	**Reality TV**	**Sit-Coms**	**Crime Drama**
Yes	5	18	21
No	11	30	15

A Chi-square test is used to test the null hypothesis that college graduate status and TV show genre preference are independent. Which of the following statements is correct?

A. Accept H_A at the 0.05 significance level

B. Reject H_0 at the 0.10 significance level

C. Reject H_0 at the 0.01 significance level

D. Reject H_0 at the 0.05 significance level

E. Accept H_A at the 0.01 significance level

Explanation:

The correct answer is B. Define the null and alternative hypotheses as follows:

H_0: The probability of being in each cell is the product of the corresponding pair of marginal probabilities (that is, college graduate status and TV show genre preference are independent)

H_A: For at least one of the cells, the probability of belonging to that cell is not equal to the product of the corresponding pair of marginal probabilities (that is, college graduate status and TV show genre preference are not independent)

Below are the observed and expected frequencies. In the cells, the observed O is the number outside the parentheses and the expected E is enclosed in parentheses.

	Reality TV	Sit Com	Crime Drama	Row Marginal Probability
College Graduate	5 (7.04)	18 (21.12)	21 (15.84)	0.44
Not a College Graduate	11 (8.96)	30 (26.88)	15 (20.16)	0.56
Column Marginal Probability	0.16	0.48	0.36	1.00

Next, we compute $\frac{(O - E)^2}{E}$ for each cell:

	Reality TV	Sit Com	Crime Drama
College Graduate	0.591	0.461	1.681
Not a College Graduate	0.464	0.362	1.321

The test statistic is $c^2 = \text{å} \frac{(O - E)^2}{E}$ = 4.88. Observe that df = number of cells – 1 – number of parameters being estimated = 6 – 1 – 3 = 2.

The cut-off for the critical region with α = 0.01 is 9.21, with α = 0.05 is 5.99, and with α = 0.10 is 4.61.

Since the test statistic value is 4.88, we do not reject the null hypothesis at the significance levels α=0.05 or 0.01, but we *do* reject the null hypothesis at the α=0.10 level.

Inference for Quantitative Data: Slopes

Finally, 2–5% of the questions on your AP Statistics exam will cover the topic Inference for Quantitative Data: Slopes.

Recall that the model $\hat{y} = a + bx$ describes a linear relationship between the two variables *x* and *y*. The values of *a* and *b* are calculated based on a sample, so they are statistics. The parameter that *b* estimates is the true value of the slope when the entire population is considered and is denoted β. Just like we construct confidence intervals and carry out hypothesis tests for other parameters (such as proportions and means), we can also do this for β.

There are several conditions that must be checked in these situations:

- *x* and *y* should have a linear relationship. This can be checked using residual plots, as shown in an earlier section.
- σ_y should be the same for all *x*. That is, the standard deviation of all the *y*-values associated with any given *x*-value should be the same regardless of the *x*-value. This is difficult to check in practice, but again, a residual plot can be useful.
- Data should come from a random sample.
- The *y*-values associated with any given *x* should be approximately normal.

Confidence Intervals for Slopes

If the response variable, *y*, has standard deviation σ_y, then σ_y can be estimated using the standard deviation of the residuals. This is calculated using the formula $s = \sqrt{\frac{1}{n-2}\sum(y_i - \hat{y}_i)^2}$.

The sampling distribution of the slope b is then $SE_b = \frac{s}{s_x\sqrt{n-1}}$, where s_x is the sample standard deviation of the *x*-values. This comes from a *t*-distribution with $df = n-2$. The point estimate, of course, is the slope *b* calculated from the sample.

As usual, the margin of error is $ME = t^* \cdot SE_b$, and the confidence interval is $(b - ME, b + ME)$.

Free Response Tip

If a confidence interval for the slope contains both positive and negative values, then in particular it contains 0. This means that at the confidence level specified, you are not even sure that there is *any* linear relationship between *x* and *y*. If this happens, be sure to mention it in your conclusion.

Hypothesis Tests for Slopes

The null hypothesis for a test of slope is $H_0 : \beta = \beta_0$, where β_0 is the hypothesized slope. The alternative hypothesis is one of $H_a : \beta < \beta_0$, $H_a : \beta > \beta_0$, or $H_a : \beta \neq \beta_0$. Commonly, the null hypothesis will be $H_0 : \beta = 0$. If this is rejected, it establishes the fact that there is some relationship between *x* and *y*.

The test statistic comes from a *t*-distribution with $df = n-2$, and is given by the formula $t = \dfrac{b-\beta_0}{SE_b}$, where $SE_b = \dfrac{s}{s_x\sqrt{n-1}}$ as described previously. A *p*-value is obtained from the *t*-distribution as usual, and is interpreted in relation to the chosen α.

Suggested Reading

- Starnes & Tabor. *The Practice of Statistics.* 6th edition. Chapter 12. New York, NY: Macmillan.
- Larson & Farber. *Elementary Statistics: Picturing the World*. 7th edition. Chapter 9. New York, NY: Pearson.
- Bock, Velleman, De Veaux, & Bullard. *Stats: Modeling the World*. 5th edition. Chapter 26. New York, NY: Pearson.
- Sullivan. *Statistics: Informed Decisions Using Data.* 5th edition. Chapter 14. New York, NY: Pearson.
- Peck, Short, & Olsen. *Introduction to Statistics and Data Analysis*. 6th edition. Chapter 13. Boston, MA: Cengage Learning.

	Parameter	Point Estimate	Standard Error	Distribution
One Proportion	p	$\hat{p}$	$\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$	Standard normal *z*-distribution
Two Proportions	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	Standard normal *z*-distribution
One Mean	μ	$\overline{x}$	$\dfrac{s}{\sqrt{n}}$	*t*-distribution with $df = n-1$
Two Means	$\mu_1 - \mu_2$	$\overline{x}_1 - \overline{x}_2$	$\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$	*t*-distribution with df found using technology
Slope	β	b	$\dfrac{s}{s_x\sqrt{n-1}}$, where $s = \sqrt{\dfrac{1}{n-2}\sum(y_i - \hat{y}_i)^2}$	*t*-distribution with $df = n-2$

	Null Hypothesis	Alternative Hypotheses	Test Statistic	Distribution
One Proportion	$H_0 : p = p_0$	$H_a : p < p_0$ $H_a : p > p_0$ $H_a : p \neq p_0$	$z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$	Standard normal *z*-distribution
Two Proportions	$H_0 : p_1 = p_2$	$H_a : p_1 < p_2$ $H_a : p_1 > p_2$ $H_a : p_1 \neq p_2$	$z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c\left(1-\hat{p}_c\right)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$	Standard normal *z*-distribution
One Mean	$H_0 : \mu = \mu_0$	$H_a : \mu < \mu_0$ $H_a : \mu > \mu_0$ $H_a : \mu \neq \mu_0$	$t = \dfrac{\overline{x} - \mu_0}{\dfrac{s}{\sqrt{n}}}$	*t*-distribution with $df = n-1$
Two Means	$H_0 : \mu_1 = \mu_2$	$H_a : \mu_1 < \mu_2$ $H_a : \mu_1 > \mu_2$ $H_a : \mu_1 \neq \mu_2$	$t = \dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$	*t*-distribution with df found using technology
Goodness-of-Fit	$H_0 :$ hypothesized proportions for every category	$H_a :$ at least one of the hypothesized proportions is wrong	$\chi^2 = \sum \dfrac{(O-E)^2}{E}$	Chi-square distribution with $df = (\#categories) - 1$

Homogeneity	H_0 : all populations have the same distribution	H_a : at least one population has a distribution that is different	$\chi^2 = \sum \frac{(O-E)^2}{E}$	Chi-square distribution with $df =$ $(\#rows - 1)(\#columns - 1)$
Homogeneity	H_0 : the categorical variables are independent	H_a : the categorical variables are dependent	$\chi^2 = \sum \frac{(O-E)^2}{E}$	Chi-square distribution with $df =$ $(\#rows - 1)(\#columns - 1)$
Slope	$H_0 : \beta = \beta_0$	$H_a : \beta < \beta_0$ $H_a : \beta > \beta_0$ $H_a : \beta \neq \beta_0$	$t = \frac{b - \beta_0}{\left(\frac{s}{s_x \sqrt{n-1}}\right)}$	*t*-distribution with $df = n - 2$

Sample Inference for Quantitative Data: Slopes Questions

Use the information below to answer the two questions that follow.

Growth hormones are often used to increase the weight gain of turkeys. In an experiment involving 25 turkeys, five different doses of growth hormone (0, 0.25, 0.5, 0.75, and 1.0 mg/kg) were injected into turkeys (five for each dose) and the subsequent weight gain recorded. A linear relationship appears to hold for the data. The following output regarding the regression line was obtained:

Parameter	Estimate	Test Statistic T	Prob > |T|	Standard Error of Estimate
Intercept	4.2153	3.77	0.0071	1.2210
Dose	5.3180	3.18	0.0215	1.6340

The best fit regression line and 95% confidence interval for its slope are given by which of the following?

A. Regression Line: $y = 5.3180 + 4.2153x$
   95% CI for slope: 4.2153 ± 1.714(1.2210)
B. Regression Line: $y = 5.3180x + 4.2153$
   95% CI for slope: 5.3180 ± 2.069(1.634)
C. C.Regression Line: $y = 5.3180x$
   95% CI for slope: 5.3180± 1.96(1.634)
D. Regression Line: $y = 5.3180 + 4.2153x$
   95% CI for slope: 4.2153 ± 2.069(1.2210)
E. Regression Line: $y = 5.3180x + 4.2153$
   95% CI for slope: 5.3180 ± 1.714(1.6340)

Explanation:

The correct answer is B. You correctly identified the slope and intercept for the regression line from the output, and when constructing the 95% CI for the slope, the formula is slope ± $t0.05/_2$. Here, the cut-off $t0.05/_2$ with df =23 is 2.069 (from the *t*-table). So, choice B is the correct answer. Choice A is incorrect because you interchanged the slope and intercept in the equation of the regression line, and you used the cut-off $t0.05$ with df = 23 instead of $t0.05/_2$ with df =23 when constructing the CI. Choice C is incorrect because you forgot to include the intercept in the equation of the regression line. Also, a less critical error, but one nonetheless, was that you used the cut-off $z0.05 = 1.96$ when constructing the CI instead of $t0.05/_2$ with df =23, which is 2.069. Choice D is incorrect because you interchanged the slope and intercept in the equation of the regression line. Choice E is incorrect because you used the cut-off $t0.05$ with df = 23 instead of $t0.05/_2$ with df =23 (which is 2.069 from the t-table) when constructing the CI.

Which of the following lists appropriate null and alternative hypotheses to test the slope, the test statistic, and *p*-value?

A. Hypotheses: H_0: $\beta_1 = 0$ versus H_A: $\beta_1 \neq 0$
   Test Statistic: T = 3.18
   p-value: $0.20 < p < 0.50$
B. Hypotheses: H_0: $\beta_0 = 0$ versus H_A: $\beta_0 \neq 0$
   Test Statistic: T = 3.18
   p-value: $0.20 < p < 0.50$

C. Hypotheses: H_0: $\beta_1 \neq 0$ versus H_A: $\beta_1 = 0$
Test Statistic: T = 3.18
p-value: $0.20 < p < 0.50$

D. Hypotheses: H_0: $\beta_0 = 0$ versus H_A: $\beta_0 > 0$
Test Statistic: T = 3.77
p-value: $p < 0.01$

E. Hypotheses: H_0: $\beta_1 = 0$ versus H_A: $\beta_1 > 0$
Test Statistic: T = 3.18
p-value: $0.20 < p < 0.50$

Explanation:

The correct answer is A. The null and alternative hypotheses for the slope are standard for a linear regression model. The test statistic is the one associated with *dose*, and the *p*-value is 0.0215, which satisfies this inequality. Choice B is incorrect because the hypotheses are listed for the *intercept*, not the slope. Choice C is incorrect because the null and alternative hypotheses are reversed. Choice D is incorrect because the stated hypotheses concern the *intercept*, not the slope. And, the alternative hypotheses should not be an inequality here but rather "not equal to" since we are trying to assess the existence of *any* trend, positive or negative. Choice E is incorrect because the alternative hypotheses should not be an inequality here but rather "not equal to" since we are trying to assess the existence of *any* trend, positive or negative.