QUIZ 1

- A **DATAFRAME** is an object that contains multiple variables. It is analogous to a spreadsheet.
- The collection of objects and things you have created in a session is known as your **WORKSPACE**.
- Variables that consist of numbers are called **NUMERIC** variables.
- A **NOMINAL VARIABLE** is a variable that uses numbers to represent different groups of data (e.g. biological sex)
- You only need to **INSTALL** a package one time. However, you will need to **REFERENCE** it every time you start a new session of R.
- Variables that consist of text are called **STRING** variables.
- Which symbols in R can be roughly translated as "is created from"? ←
- The **HEADER = TRUE** command tells R that a data file has variable names stored in the first row.
- When referring to a variable within a dataframe, what symbol goes between the name of the dataframe and the name of the variable? $
- When creating a grouping variable, you must tell R the **LEVELS**, or the numbers you have used to represent different groups, and the **LABELS**, or descriptions that correspond to each number.
- **STRING** values should always be placed inside quotes, but NUMERIC values are never placed inside quotes.
- What character separates multiple commands on the same line? ;
- The two most commonly used formats for importing data into R are TAB-**DELIMINATED** text and **COMMA-SEPARATED** values.
- A **FUNCTION** is something you do in R to create objects.
- In the wide format, each **ROW** represents data from one entity and each **COLUMN** represents one variable
- **NA** is used in R to denote data that are missing.
- An **OBJECT** is anything created in R, such as a variable, statistical model, etc.
- By setting the **WORKING DIRECTORY**, you can avoid typing the full file path every time you need to access a file.

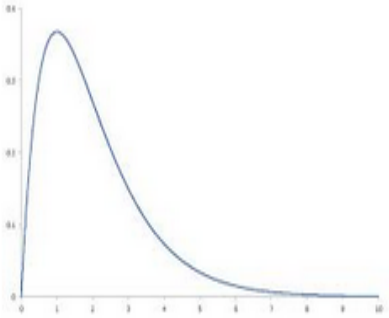
^ or **	Exponentiation
<=	Less than or equal to
>=	Greater than or equal to
==	Exactly equal to
!=	Not equal to
!x	Not X
X | Y	X OR Y
X & Y	X AND Y
isTRUE(x)	Test if x is true

-

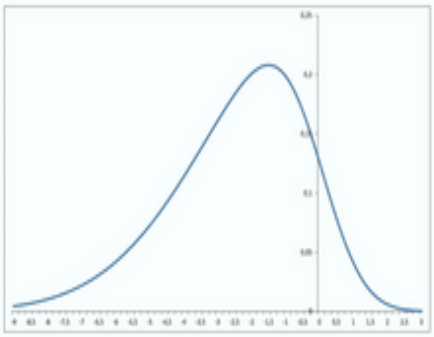
QUIZ 2

- **RANDOMIZATION** is a useful tool for minimizing sources of systematic variation other than the experimental manipulation by mixing up the order in which conditions are completed or the groups to which participants are assigned.
- Experimental research allows one to establish a **CAUSE AND EFFECT** relationships, whereas correlational research provides better **ECOLOGICAL** validity.
- Most statistical tests work by comparing **SYSTEMATIC** variation, or the effect of experimental manipulation, and **UNSYSTEMATIC** variation, or background noise.
- A frequency distribution with observations distributed symmetrically around the mean is called a **NORMAL** distribution
- A scientific hypothesis needs to be tested and rejected with empirical evidence.
- In a **BETWEEN SUBJECTS** design, each participant is exposed to only one experimental condition and their results are compared to other participants in other conditions.
- In a **WITHIN SUBJECTS** design, each participant completes all of the experimental conditions and their results are compared to themselves in previous conditions.
- Scientists never claim to prove the **ALTERNATIVE** hypothesis. Rather, they can only claim to reject the **NULL** hypothesis.
- **OUTCOME** VARIABALE is another name for dependent variable. **PREDICTOR VARIABLE** is another name for independent variable.
- **Categorical variable**: A variable made of distinct classes (biological sex, the city you were born in, type of car)
- **Ordinal variable:** A type of categorical variable with categories that have a natural order (class ranking, order of medals)
- **Interval variable:** A variable where the difference between two values is meaningful and consistent across the scale of measurement (1 vs. 2 is the same as 2 vs. 3)
- **Ratio variable:** Variable that has a meaningful zero point and values have meaningful multiplicative relationships (10 is twice as good as 5)
- **Theory: A possible explanation about an observed phenomenon**
- **Hypothesis:** A prediction that comes from a theory
- **Variable:** An entity that can take on different values
- **VALIDITY** describes whether an instrument measures the construct that it intends to measure. **RELIABILITY** describes the consistency witch which an instrument will measure its construct across situations.

This distribution is positively skewed:

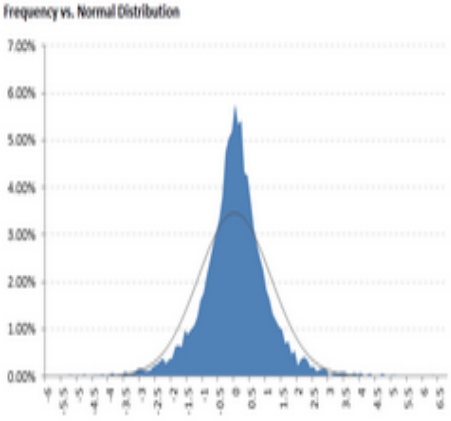


This distribution is negatively skewed:

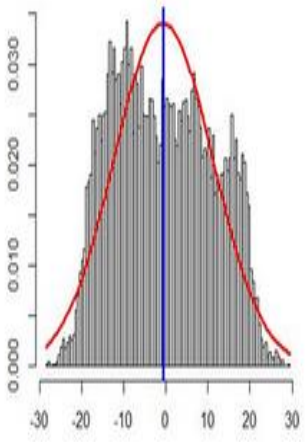


-

This distribution has positive kurtosis:



This distribution has negative kurtosis:



IQ scores have a standard deviation of 15 and a mean of 100. Calculate Z-scores for the following IQ scores (round your answers to one decimal place):

100: 0

85: -1

115: 1

109: .6

Isaac Newton was famously hit on the head by a falling apple while sitting under a tree (i).

He had the idea that the force acting upon the apple was the same as the force that keeps the moon in orbit (ii).

He predicted that if his idea was true, and the moon was to "fall" for one minute, the distance it fell could be calculated using the same equations as an apple on earth (iii).

He then determined the distance each object would fall, accounting for its size and its distance from the earth (iv).

He compared these distances and concluded that the force of gravity was indeed the same for the objects on earth, as well as in space (v).

i	i. Initial observations and rese
ii	ii. Generate theory
iii	iii. Generate hypotheses
iv	iv. Collect data to test theory
v	v. Analyze data

IQ scores have a standard deviation of 15 and a mean of 100. What is the probability that a randomly selected individual will have an IQ score that is equal to or greater than 115? Hint: you will need to use the appendix in your text book.

State the probability as a decimal rounded to 2 places, not a fraction or percentage:

.16

QUIZ 3

- What **can** you learn from a scatterplot?
    - Whether there seems to be a relation between two variables
    - What kind of relation exists between two variables
    - Whether any cases are substantially different from others
    - What **CAN'T** you learn from a scatterplot?
        - The significant of a relation between two variables
        - Amount of variance in one variable that is not explained by another
- A boxplot can give you information about skewness, range of scores, and variability of scores, but **NOT** kurtosis.
- Density plots are similar to **HISTOGRAMS,** except that they represent the distribution as a line rather than as bars.
- With ggplot2, you can build a quick plot using the qplot() function or you can build up a plot as a series of layers using the ggplot() function

The R code *new_graph <- ggplot(my_data, aes(number_of_drinks, BAC, colour=age))* produces a plot called [new_graph], with the variable [number_of_drinks] on the x axis and the variable [BAC] on the y axis. Data from each age group is assigned a different [color].

- Geoms and their descriptions: in book

QUIZ 4

- The **VARIANCE** is the average error between the mean and each individual's observed score
    - The square root variance is called the **STANDARD DEVIATION**
    - Because simply adding the deviances is not a valid estimate of error, it is necessary to utilize the sum of **SQUARED ERRORS**squared errors. Dividing the sum of squared errors **SQUARED ERRORS** by the number of observations minus 1 gives you the variance **VARIANCE**.
- The difference between an individual's observed score on some variable and the sample mean is called the **DEVIANCE**
    - Why does it not make sense to simply add up the sum of deviances to estimate total error? Adding the deviances results in a sum of zero
- **STANDARD ERROR/STANDARD ERROR OF THE MEAN** is the standard deviation of the sampling distribution.
    - The **standard error** represent how well a particular sample represents the population
- The process of calculating a parameter (such as the mean) from an infinite number of samples taken from a population, and plotting those parameters as a histogram produces a **SAMPLING DISTRIBUTION**
- **TYPE I ERROR** occurs when you conclude that an effect is present, when in fact there is no effect in the population.
    - The α-level is the acceptable probability of making a(n) Type I error
- **TYPE II ERROR** occurs when you assume that an effect is not present when it does actually exist.
- Statistical **POWER** refers to the ability to detect an effect of a desired size in a population of a given size.
    - Holding sample size constant, a larger effect will have greater statistical power.
    - Holding effect size constant, a larger sample will have greater statistical power.
- A **NULL HYPOTHESIS** states that a given effect is absent.
    - When conducting a null hypothesis significance test, a significant test statistic indicates that it is **NOT LIKELY** that the null hypothesis is true and allows you to **REJECT** the null hypothesis
- Because an infinite number of models can be specified for any given data, it is necessary to determine the **FIT** of a model to ensure that it accurately represents the data.
- It is useful to build **statistical models** because they allow you to make predictions about real-world processes and they allow you to make inferences about psychological processes that are otherwise inaccessible.
    - A model that explains more variation will produce a test statistic that is LARGER, and therefore LESS likely to occur by chance.

    - A model that explains less variation will produce a test statistic that is SMALLER, and therefore MORE likely to occur by chance.
    - A **TEST STATISTIC** is a ratio of systematic and unsystematic error
    - When there is an acceptably small probability that a result was not obtained by chance, we can say that the result is **STATISTICALLY SIGNIFICANT**
- **CENTRAL LIMIT THEOREM** states that in sufficiently large samples, the mean of the sampling distribution is equal to the population mean.
- **EFFECT SIZES** are objective and standardized measures of the magnitude of an observed effect.
- One-tailed hypothesis: Directional
    - A directional hypothesis requires a **SMALLER** test statistic to find a significant result because the range of significant results is larger.
- Two-tailed hypothesis: Non-directional (what we use!)
- Boundaries within which we believe the true value of the population mean will fall are called **CONFIDENCE INTERVALS**

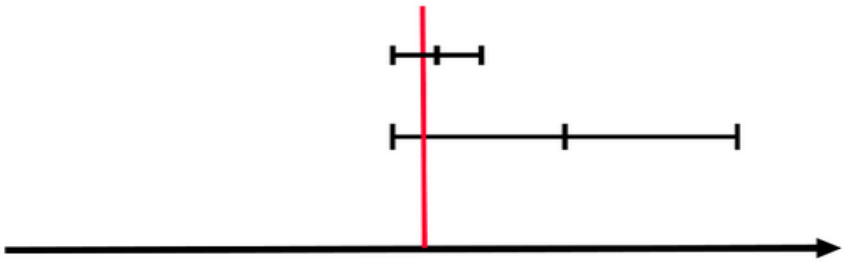
- **CALCULATING CONFIDENCE INTERVALS**

You measure social anxiety scores in a large sample of college students. You find that the mean score is 20 and the standard error is 4. Calculate a 95% confidence interval for the mean (round your answers to 2 decimal places).

Lower boundary: 12.16

Upper boundary: 27.84

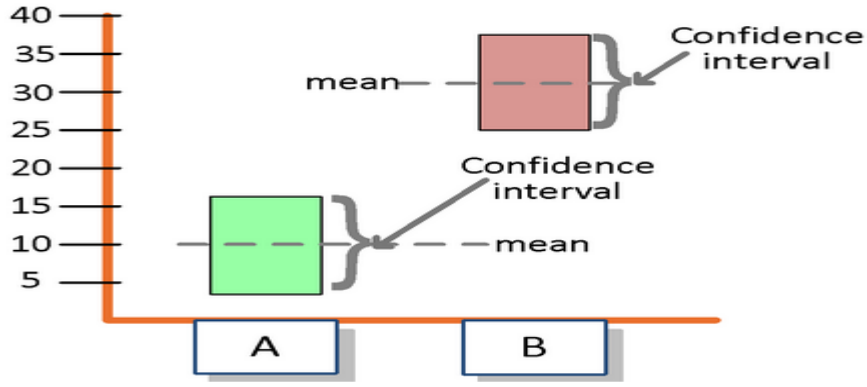
- You calculate 95% confidence intervals for the means of two samples and graph them, below. Is it plausible that these two means came from the same population?



- yes
- You collect 200 samples from a population and calculate a 95% confidence interval for the mean of each sample. Theoretically, how many of those

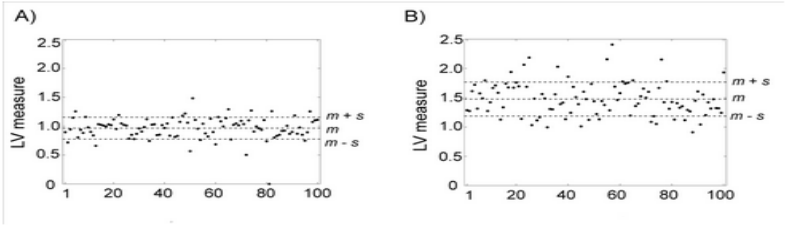
confidence intervals will NOT contain the true value of the population mean? 10
How many will contain the true value of the population mean? 190

You recruit 100 people to take part in an intervention designed to reduce anxiety. You randomly assign half to treatment A and the other half to treatment B. You measure participant's anxiety scores after treatment, and calculate a 95% confidence interval for the mean. The results are plotted below. Is it plausible that treatment A reduced anxiety scores?



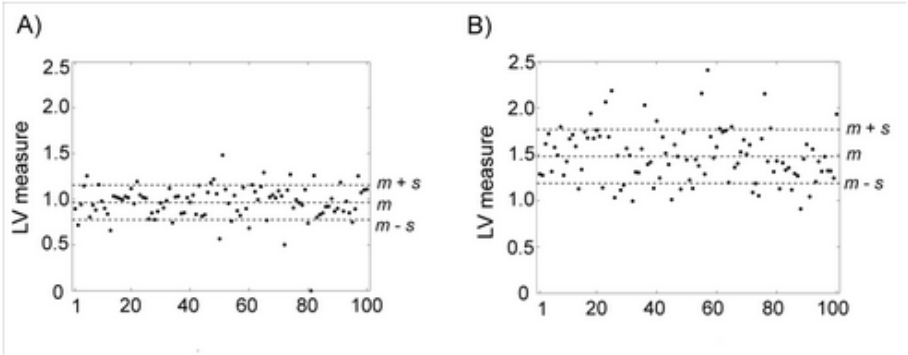
- Yes

Which of these graphs shows a larger standard deviation?



- A
- B

For which of these samples does the mean fit the data better as a model?



- A

QUIZ 5

- **POINT-BISERIAL** correlation and Biserial **BISERIAL** correlation are both used when one of the variables you are comparing is dichotomous.
    - Suppose that you are interested in the relation between height and intelligence. Rather than considering height as a continuous variable, you create a new variable height_di, in which you dichotomize height. You place everyone under 5'6" into a group called "short" and everyone over 5'6" into a group called "tall". To compare variability in height_di to variability in intelligence, you should calculate a **BISERIAL** Biserial correlation.
    - Suppose you are interested in the relation between degree program and intelligence. You have a variable called degree, where psychology students are coded 1 and medical students are coded 2. To compare variability in degree program to variability in intelligence you should calculate a Point-biserial **POINT-BISERIAL** correlation.
- A **PARTIAL** Partial correlation describes the relation between two variables while controlling for one or more other variables. A **BIVARIATE** Bivariate correlation describes the relation between just two variables
    - A **PARTIAL** correlation assesses the relation between two variables while controlling for the effect of a third variable on both of them. A **SEMI-PARTIAL** correlation assesses the relation between two variables while controlling for the effect of a third variable on just one of them.
        - Suppose you are interested in comparing the fuel efficiency of vehicles made by Toyota and Ford. You create two variables: "manufacturer" and "efficiency". However, you know that both manufacturers produce a wide range of vehicles (from small cars to large trucks). So, you create a third variable, "model". You want to find the shared variability between manufacturer and fuel efficiency, while keeping the effect of model on efficiency constant (but not the effect of model on manufacturer). You should calculate a Semi-partial **SEMI-PARTIAL** correlation for manufacturer and efficiency.
- Because a correlation is technically an **EFFECT SIZE** Effect size, it can be interpreted directly, without worrying about a p value
    - When the assumptions of Pearson's correlation coefficient are not met, **SPEARMAN'S** Spearman's **rho or KENDALL'S** Kendall's **tau** can be used instead.

    - Pearson's r assumes that the level of measurement of the data is Interval. In order to conduct a significance test, Pearson's r assumes that the sampling distribution is Normally distributed.
- If two samples are composed of the same participants, they are called Dependent **DEPENDENT**. If they are composed of different participants, they are called Independent **INDEPENDENT**.
- A **CORRELATION COEFFICIENT** is calculated by dividing the covariance of two variables by the product of each variable's standard deviation.
- Dividing the sum of cross-product deviations by degrees of freedom produces the covariance **COVARIANCE**, an indicator of the degree to which two variables are related.
    - Covariance is **NOT STANDARIDZED** Standardized, which means that it depends upon each variable's scale of measurement
- Squaring a correlation coefficient produces the **COEFFICIENT OF DETERMINATION** Coefficient of determination, an indicator of the amount of variability one variable shares with another.
- Multiplying the deviations of one variable by the corresponding deviations of another variable produces the **CROSS-PRODUCT DEVIATION**. If both variables vary in the same direction (either positively or negatively), this product will be POSITIVE Positive. If they vary in different directions, it will be NEGATIVE Negative.
- Assume that depression and ice cream eating frequency have a correlation coefficient of 0.25. What percentage of the variability in ice cream eating frequency is accounted for by variables other than depression? 93.75
- Excluding cases Listwise **LISTWISE** means that correlations are computed using only data from participants with no missing data for all variables. Excluding cases Pairwise **PAIRWISE** means that correlations are computed using data from participants with no missing data on the two variables being analyzed.

QUIZ 6

- The F-ratio utilizes mean squares, which are the sums of squares divided by their respective degrees of freedom **DEGREES OF FREEDOM**

- The residual **ANOVA** results in a regression output utilize the F statistic to test whether the regression model results in significantly better prediction than a model that uses the mean of the independent variable.
- In a regression model with only one predictor, R^2 is equivalent to the square of Pearson's R **PEARSON'S R**
- A t-test can be used to test the null hypothesis that the value of a regression coefficient is simple linear regression is equal to **ZERO**
- **F-ratio** is a statistical test of how much the model has improved the prediction of the outcome compared to the inaccuracy of the model
- Dividing the **MODEL** model sum of squares by the residual **TOTAL** sum of squares gives R^2 **R^2**, which represents the percentage of the variation in the outcome accounted for by the model.
    - The **TOTAL** total sum of squares represents the degree of inaccuracy when the most basic model (the mean) is applied to the data.
- The method of least squares is a way of finding the line of best fit by minimizing the sum of squared **RESIDUALS** residuals.
- **RESIDUALS** are the difference between a given subject's actual score on a variable and the score predicted by the regression model
- The **RESIDUAL SUM OF SQUARES** is the degree of inaccuracy when the regression model is applied to the data.
- **In R, the TILDE symbol indicates "predicted from" or "regressed on"**

$$Y_i = (b_0 + b_1X_1) + e_i$$

- Yi = outcome to
- be predicted
- b0 = the point where the linear model crosses the y-axis
- b1 = slope of the linear relationship
- X1 = predictor variable
- ei = difference between predicted and actual score for participant i

To test a hypotheses that number of western movies viewed per month is related to preference for drinking whiskey (on a scale from 1 to 10), I execute the following R command:

whiskey_regression <- lm(whiskey_preference ~ western_films, data = alcohol_use, na.*action* = na.*exclude)*

The results of the regression produce the equation $Y_i = (1.8) + (2.0)X_1 + e_i$

According to the model, what is the whiskey preference score of people who watch no western films? 1.8

What is the change in whiskey preference score for every additional western film that an individual views per month? 2.0

What whiskey preference score would the model predict for a person who watches 20 western films per month? 41.8

- To test a hypotheses that number of western movies viewed per month is related to preference for drinking whiskey (on a scale from 1 to 10), I execute the following R command:

whiskey_regression <- lm(whiskey_preference ~ western_films, data = alcohol_use, na.*action* = na.*exclude)*

The results of the regression produce the equation $Y_i = (1.8) + (2.0)X_1 + e_i$

According to the model, what is the whiskey preference score of people who watch no western films? 1.8

What is the change in whiskey preference score for every additional western film that an individual views per month? 2.0

What whiskey preference score would the model predict for a person who watches 20 western films per month? 41.8

QUIZ 7

- The Independent errors **INDEPENDENT ERRORS** assumption states that residual terms for observations must not be correlated with each other.
- In Hierarchical regression **HIERARCHICAL REGRESSION**, predictors are entered in a predetermined order. Known predictors from previous research are typically entered first, then new predictors are added to see if they improve the model.
- I run a regression to determine if number office plants (x_1) and number of windows (x_2) are related to employee satisfaction ratings. The results of the regression produce the equation $Y_i = (1.3) + (1.6)X_1 + (4.8)X_2 + e_i$.
    - What is predicted employee satisfaction rating for individuals who have 10 plants and 2 windows? 26.9
- I run a regression to determine if number office plants (x_1) and number of windows (x_2) are related to employee satisfaction ratings. The results of the regression produce the equation $Y_i = (1.3) + (1.6)X_1 + (4.8)X_2 + e_i$.
    - What is the predicted employee satisfaction rating for individuals with no plants and no windows? 1.3
- **STEPWISE REGRESSION** is an automated, atheoretical, process for selecting predictors in a model. Computer software adds or removes predictors sequentially, and uses a mathematical criterion of model fit to select the optimal combination of predictors.
- Within regression, it is assumed that residuals are NORMALLY Normally distributed and have a mean of Zero ZERO
- Cases with high **LEVERAGE** Leverage excessively influence the model as a whole. Including them in the model typically result in a decrease in **GENERALIZABILITY** Generalizability
- In simple regression, a regression line is a representation of the relationship between **ONE** One independent variable(s) and one dependent variable. However, in a regression with two independent variables, the relationship is modeled as a regression **PLANE** plane
- The Variable types **VARIABLE TYPES** assumption states that all independent variables must be quantitative or categorical, and the dependent variable must be quantitative, continuous, and unconstrained.
- The change in R^2 between two competing hierarchical models can be tested using the **ANOVA** ANOVA function.
- **MULTIPLE R^2** is the square of the correlation between the observed values of a dependent variable and the values of the variable predicted by a multiple regression model.

- **HETEROSCEDASCITY** refers to instances where residuals vary significantly at different levels of a predictor variable.
- **HOMOSCEDASCITY** refers to instances where residuals do not vary significantly at different levels of a predictor variable.
- Multicollinearity **MULTICOLLINEARITY** refers to correlation between two predictor variables. When it is too high, it leads to untrustworthy estimates of regression coefficients and limits the size of R.
    - The variance inflation factor is used to test for the presence of **MULTICOLLINEARITY** Multicollinearity .
- The process of assessing the accuracy of a model across different samples is called Cross-validation **CROSS-VALIDATION.**
- The Non-zero variance **NON-ZERO VARIANCE** assumption states that dependent variables do not have variances of 0.

QUIZ 8

- Because categorical predictors in regression must have values of 0 or 1, it is necessary to use **DUMMY** dummy coding to represent such variables.
- A variable with 5 categories will result in how many dummy variables? 4

I am interested in the relation between subjective intoxication and different methods of marijuana administration (flower, concentrate, edible, and tincture). I enter the following variables in a regression:

	(x_1)	(x_2)	(x_3)
Flower	0	0	0
Concentrate	1	0	0
Edible	0	1	0
Tincture	0	0	1

The results of the regression produce the equation $Y_i = (2) + (4.3)X_1 + (0.8)X_2 + (-1)X_3 + e_i$.

What is the difference in average level of subjective intoxication between concentrate and flower?

4.3000