STATISTICS

From Simple Studies, https://simplestudies.edublogs.org & @simplestudiesinc on Instagram

Units Covered:

KEY TERMS & DEFINITIONS

POPULATION - A collection of persons, things, or objects under study.

SAMPLE - To study the population, we select a **sample**. S**ampling** selects a portion, or subset, of the larger population and studies that portion—the sample—to gain information about the population.

STATISTIC - A number that represents a property of the sample. For example, if we consider one physics class as a sample of the population of *all* physics classes, then the average number of points earned by students in that one physics class at the end of the term is an example of a **statistic**.

PARAMETER - A numerical characteristic of the whole population that can be estimated by a statistic. (e.x. Since we considered all physics classes to be the population, then the average number of points earned per student over all the physics classes is an example of a parameter).

VARIABLE - Usually notated by letters such as *X* and *Y*, it's a characteristic or measurement that can be determined in relation to each member of a population. Variables may describe values like weight in kilograms or favorite academic subject.

NUMERICAL VARIABLES - Values with equal units such as weight in pounds and time in hours.

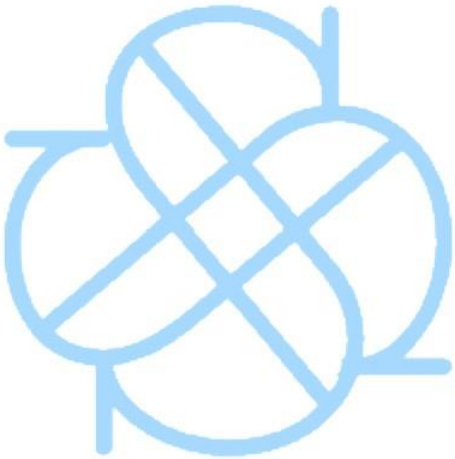
CATEGORICAL VARIABLES - Places the person/ thing into a category.

MEAN - The "average" you're used to, where you add up all the numbers and then divide by the number of numbers.

MEDIAN - The middle value in the list of numbers.

PROPORTION - A special type of ratio in which the denominator includes the numerator. For example, the **proportion** of deaths that occurred to males which would be deaths to males divided by deaths of both males and females.

DATUM - One item of information, one fact, one statistic on its own. More commonly known as "data"



INTRO TO STATISTICS

WHAT IS STATS?

Decisions or predictions are often based on data—numbers in context. These decisions or predictions would be easy if the data always sent a clear message, but the message is often obscured by variability. Statistics provides tools for **describing variability in data and for making informed decisions that take it into account.**

SAMPLE VS POPULATION

The population is the whole where the sample is a part of the whole. Think of it as a pizza, the population is the entire pizza and the sample is a slice of the pizza.

CLASSIFYING DATA

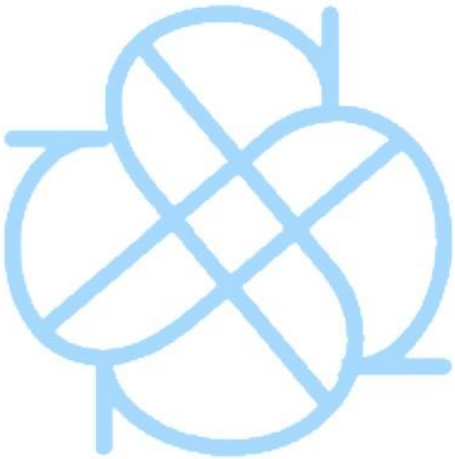
Classifying data is the **process of separating data into similar groups** based on their **characteristics**.

There are multiple types of data groups, there's:

- **Geographical**: Data that is classified with **reference to geographical locations** such as countries, states, cities, districts, etc. it is known as Geographical Classification.
- **Chronological:** Data that is **grouped according to time**, such a classification is known as a Chronological Classification.
- **Qualitative:** Under this classification, data is classified on the **basis of attributes or qualities** such as honesty, beauty, intelligence, literacy, marital status etc.
- **Quantitative:** This type of classification is made on the **basis of some measurable characteristics** like height, weight, age, income, marks of students, etc.

EXPERIMENTAL DESIGN

Experimental design is the branch of statistics that **deals with the design and analysis of experiments.** Participants are allocated to the different groups in an experiment with types of design including repeated measures, independent groups, and matched pairs designs. The researcher must decide how they will allocate the sample to the different experimental groups.



DESCRIPTIVE STATISTICS

FREQUENCY DISTRIBUTIONS

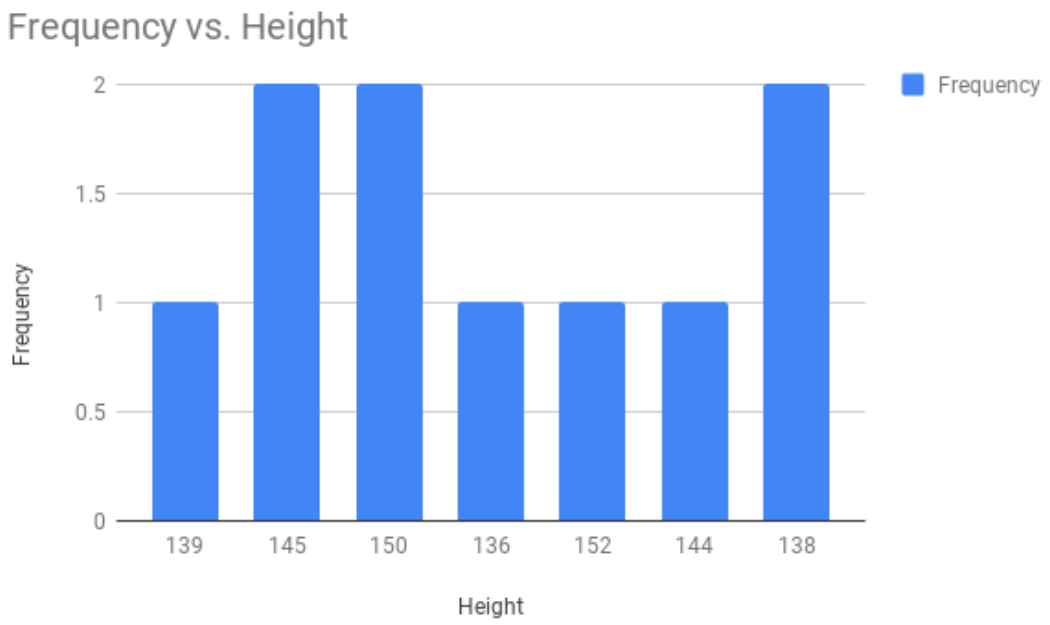
A representation, either in a graphical or tabular format, that **displays the number of observations within a given interval.** The intervals must be mutually exclusive and exhaustive.

An example of frequency distributions:

Height	Frequency
67in	3

62in	1
64in	5
70in	2

GRAPH OF FREQUENCY DISTRIBUTION



MEASURES OF CENTER

The four measures of center are mean, median, mode, and midrange.

- **Mean** - The average
- **Median** - Measure of center
- **Mode** - The value that appears most often.
- **Midrange** - The midway between the least and greatest value of the set.

WEIGHTED MEANS

A kind of average that instead of each data point contributing equally to the final mean, some data points **contribute more "weight" than others.**

MEASURES OF VARIATIONS

Summary measures to **describe the amount of variability** or spread in a set of data. The most common measures of variability are the range, the interquartile range (IQR), variance, and standard deviation.

- **Standard deviation:** A statistic that measures the **dispersion of a dataset** in relation to its mean and is calculated as the **square root of the variance.**
    - **Example -** The mean of the following two is the same: 15, 15, 15, 14, 16 and 2, 7, 14, 22, 30. However, the second is more spread out. If a set has a low standard deviation, the values are not as spread out.

- **IQR:** The **middle 50% of values** when ordered from lowest to highest.
    - To find the interquartile range (IQR), first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the **difference between Q3 and Q1.**
    - **Example -** Consider the following numbers: 1, 3, 4, 5, 5, 6, 7, 11. Q1 is the middle value in the first half of the data set. ... The interquartile range is Q3 minus Q1, so IQR = 6.5 - 3.5 = 3.

- **Range**: The **difference** between the **lowest and highest values.**
    - **Example -** `In (4, 6, 9, 3, 7) the lowest value is 3, and the highest is 9, so the range is 9 − 3 = 6.`

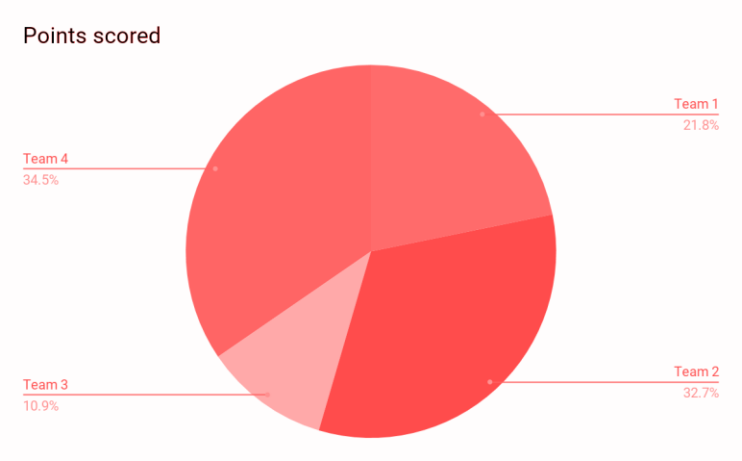
GRAPHS & STATISTICAL PRESENTATIONS

REPRESENTING DATA VISUALLY

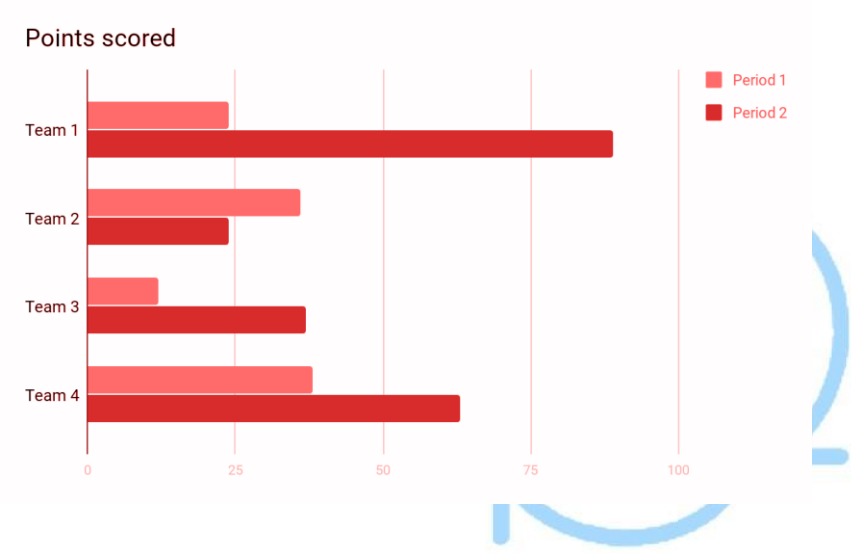
Visual representation of data includes multiple different types of charts, some of the most common are:

- **Pie charts**
- **Line charts**
- **Bar charts**
- **Bubble charts**

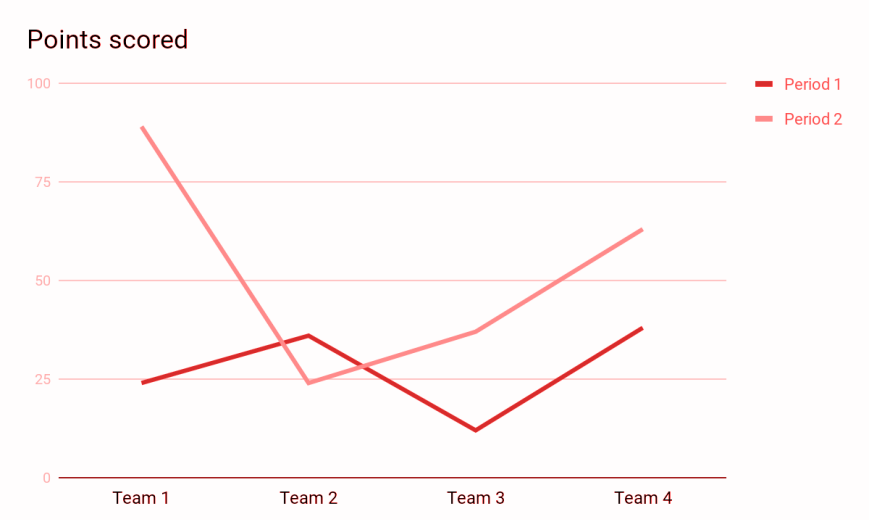
PIE CHART -



BAR CHART -



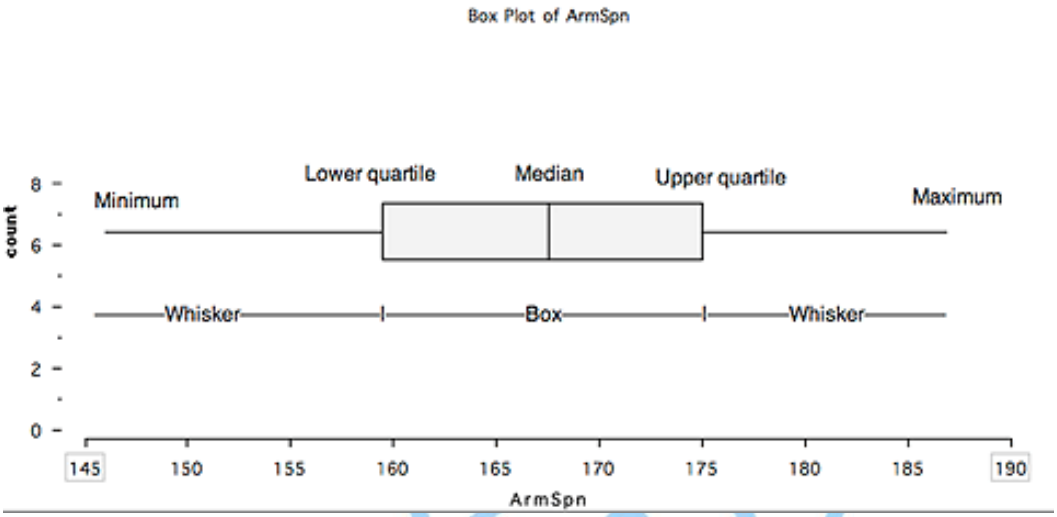
LINE CHART -



BOX AND WHISKER PLOTS

Also called a box plot—**displays the five-number summary of a set of data.** The five-number summary includes the **minimum, first quartile, median, third quartile, and maximum.** In a box plot, we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median.

Example -



STEM-AND-LEAF DIAGRAMS

A **table** used to display data. The **'stem'** on the left displays the **first digit** or digits. The **'leaf'** is on the right and displays the **last digit.**

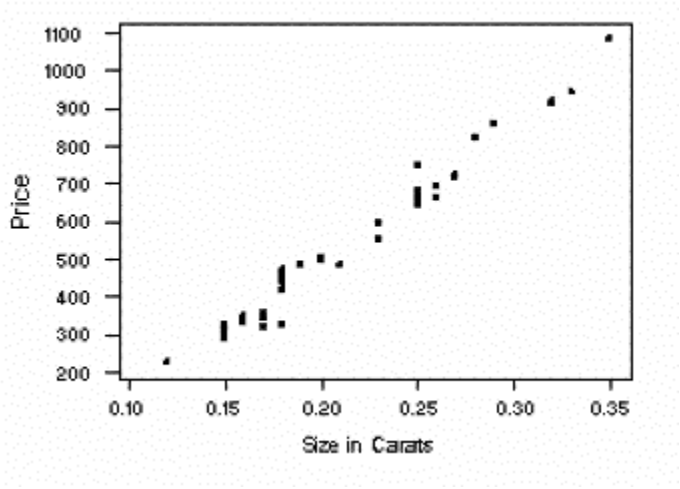
Example - 543 and 548 can be displayed together on a stem and leaf as 54 | 3,8.

Stem	Leaf
4	4 6 7 9
5	
6	3 4 6 8 8
7	2 2 5 6
8	1 4 8
9	
10	6

SCATTER PLOTS AND LINE GRAPHS

A scatter plot is a graph used to determine whether there is a **relationship between paired data.** In many real-life situations, scatter plots follow patterns that are approximately **linear**. If y tends to increase as x increases, then the paired data is a positive correlation.

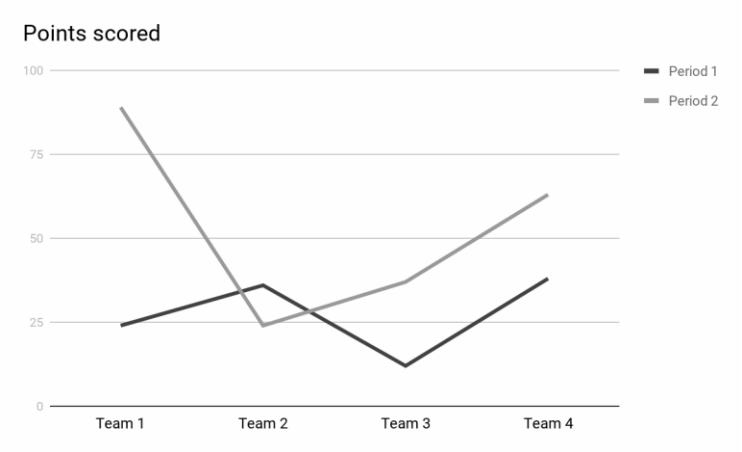
Example -



LINE GRAPHS

A **visual comparison of how two variables**—shown on both the x- and y-axes—are related or can vary with the other. It shows related information by drawing a **continuous line** between all the points on a grid.

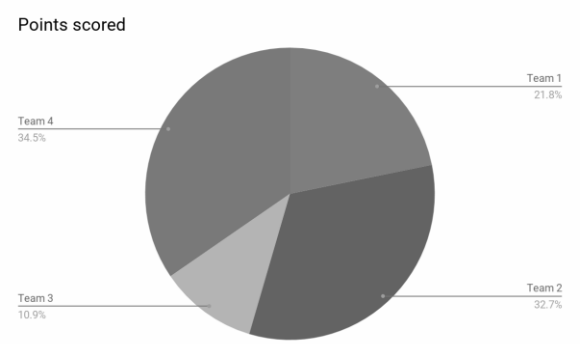
Example -



PIE CHARTS

A **circular statistical graphic,** which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice (and consequently its central angle and area), is proportional to the quantity it represents. Pie charts are generally used to **show percentage** or proportional data and usually the percentage represented by each category is provided next to the corresponding slice of pie. Pie charts are **good for displaying data for around 6 categories or fewer.**

Example -



PROBABILITY

BASIC PROBABILITY

The measure of the likelihood that a specific event will occur in a **random experiment.** Probability is quantified as a **number between 0 and 1**, where, loosely speaking, 0 indicates impossibility and 1 indicates certainty. The higher the probability of an event, the more likely it is that the event will occur.

THE MULTIPLICATION RULE

To find the probability of two events happening at the same time (this is also one of the AP Statistics formulas), you can use the general multiplication rule formula. The formula is = $P(A \cap B) = P(A)\,P(B|A)$ and the specific multiplication rule is P(A and B) = P(A) * P(B)

Example - If the probability of event A is 2/9 and the probability of event B is 3/9 then the probability of both events happening at the same time is (2/9)*(3/9) = 6/81 = 2/27.

THE ADDITION RULE

States the probability of two events is the **sum** of the probability that **either will happen** **minus** the probability that **both will happen.**

Example - When two events, A and B, are **non-mutually exclusive**, there is some overlap between these events. The probability that A or B will occur is the sum of the probability of each event, minus the probability of the overlap. P(A or B) = P(A) + P(B) - P(A and B)

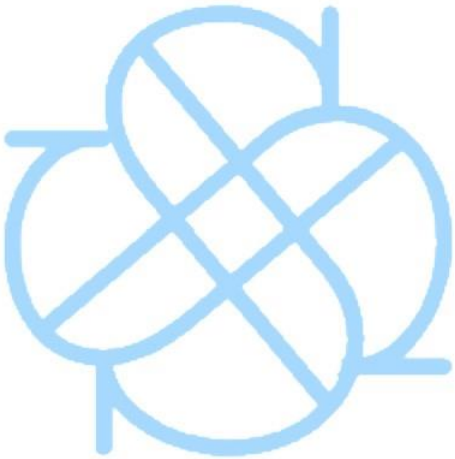
PERMUTATIONS & COMBINATIONS

A combination is a selection of all or part of a set of objects, with no regard to the order in which objects are selected. Suppose we have a set of three letters: A, B, and C. From this set, we may ask how many different ways we can select 2 letters from that set.

Example - Choosing 3 desserts from a menu of 10. C(10,3) = 120. Permutation: Listing your 3 favorite desserts, in order, from a menu of 10. P(10,3) = 720.

A **permutation** is an **arrangement of all or part of a set of objects**, with regard to the order of the arrangement. Suppose we have a set of three letters: A, B, and C. We may ask how many different ways we can arrange 2 letters from that set.

Example - suppose we have a set of three letters: A, B, and C. We might ask how many ways we can arrange 2 letters from that set. Each possible arrangement would be an example of a permutation. The complete list of possible permutations would be: AB, AC, BA, BC, CA, and CB.

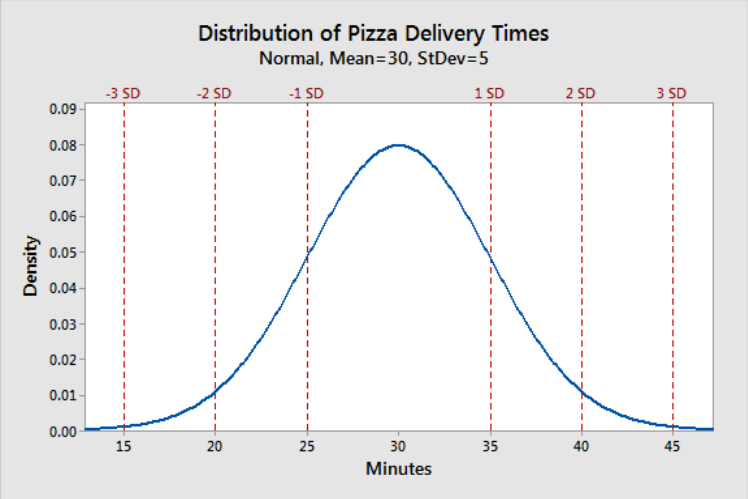


THE NORMAL DISTRIBUTION

WHAT IS A NORMAL DISTRIBUTION?

The normal distribution is a probability function that depicts the way values of a **variable are distributed.** It's a **symmetric distribution** where most of the observations cluster around the central peak and the probabilities for values further away from the mean and taper off equally in both directions.

Example -



FINDING PROBABILITIES

HOW TO FIND PROBABILITIES

1. Draw a picture of the **normal distribution.**

2. Translate the problem into one of the following: ***p*(*X* < *a*), *p*(*X* > *b*), or *p*(*a* < *X* < *b*).** Shade in the area on your picture.

3. Standardize *a* (and/or *b*) to a *z*-score using the ***z*-formula:**

4. Look up the *z*-score on the ***Z*-table** (see below) and find its corresponding probability.

a.Find the row of the table corresponding to the leading digit (ones digit) and first digit after the decimal point (the tenths digit).

b.Find the column corresponding to the second digit after the decimal point (the hundredths digit).

c.Intersect the row and column from Steps (a) and (b)

5. a. If you need a "less-than" probability — that is, ***p*(*X* < *a*)** — you're done.

b. If you want a "greater-than" probability — that is, ***p*(*X* > *b*)** — take one minus the result from Step 4.

c. If you need a "between-two-values" probability — that is, ***p*(*a* < *X* < *b*)** — do Steps 1–4 for *b* (the larger of the two values) and again for *a* (the smaller of the two values), and subtract the results.

FINDING VALUES

A **critical value** is a line on a graph that splits the graph into various sections. One or two of the sections is the **"rejection region"**; if your test value falls into that region, then you reject the null hypothesis.

Step 1: Subtract the confidence level from 100% to find the α level: 100% – 90% = 10%

Step 2: Convert Step 1 to a decimal: 10% = 0.10.

Step 3: Divide Step 2 by 2 (this is called "α/2"). 0.10 = 0.05. This is the area in each tail.

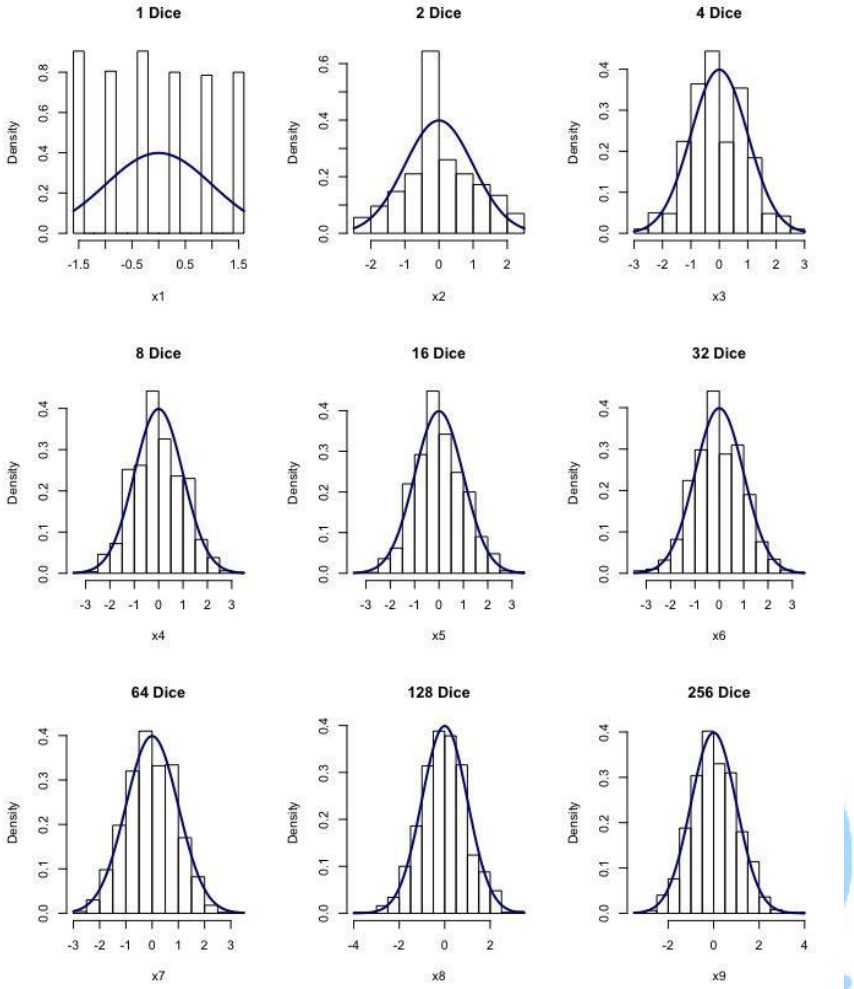
Step 4: Subtract Step 3 from 1 (because we want the area in the middle, not the area in the tail): 1 – 0.05 = .95.

Step 5: Look up the area from Step in the z-table. The area is at z=1.645. This is your critical value for a confidence level of 90%.

CENTRAL LIMIT THEOREM

The **central limit theorem** states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be **approximately normally distributed.**

Example -



LINEAR REGRESSION

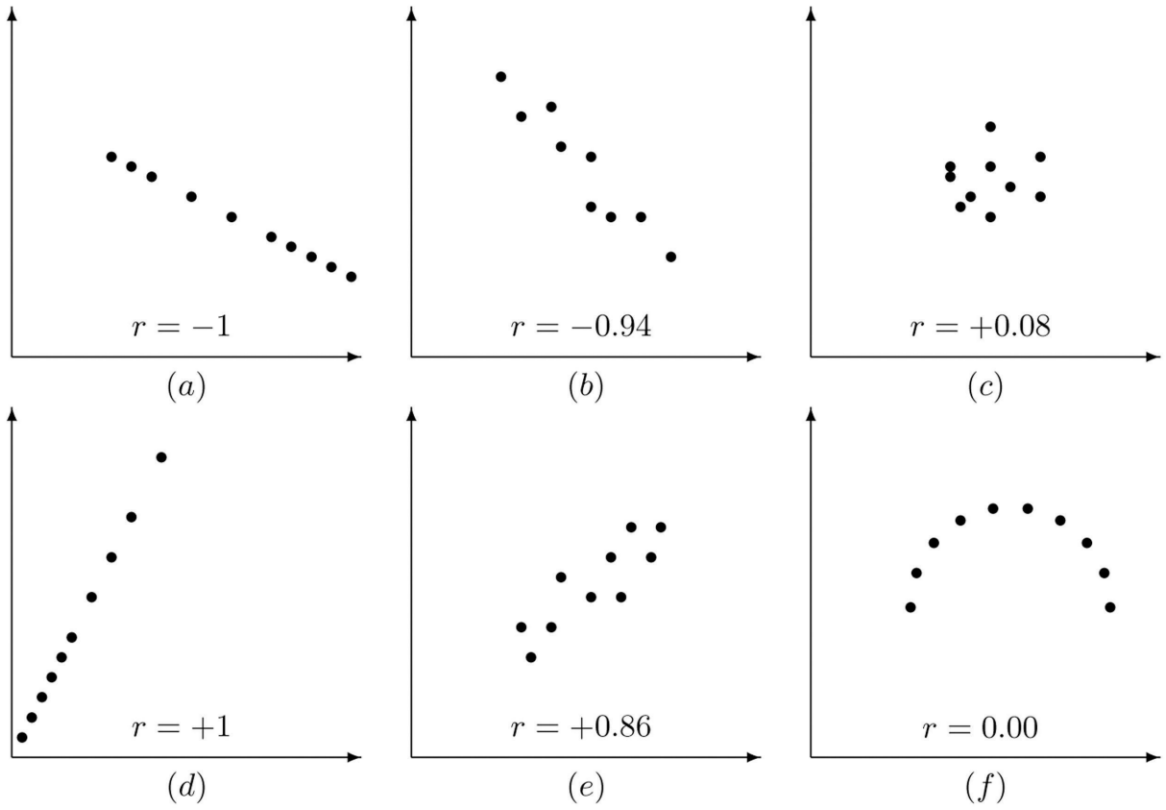
CORRELATION COEFFICIENT

The **correlation coefficient** is a statistical measure of the strength of the relationship between the relative movements of two variables. A calculated number greater than 1.0 or less than -1.0 means that there was an **error in the correlation measurement.**

HOW TO CALCULATE

Use the formula $(z_y)_i = (y_i - \bar{y}) / s_y$ and **calculate a standardized value** for each y_i. Add the products from the last step together. Divide the sum from the previous step by n – 1, where n is the total number of points in our set of paired data. The result of all of this is the **correlation coefficient r.**

Example -



LEAST SQUARES

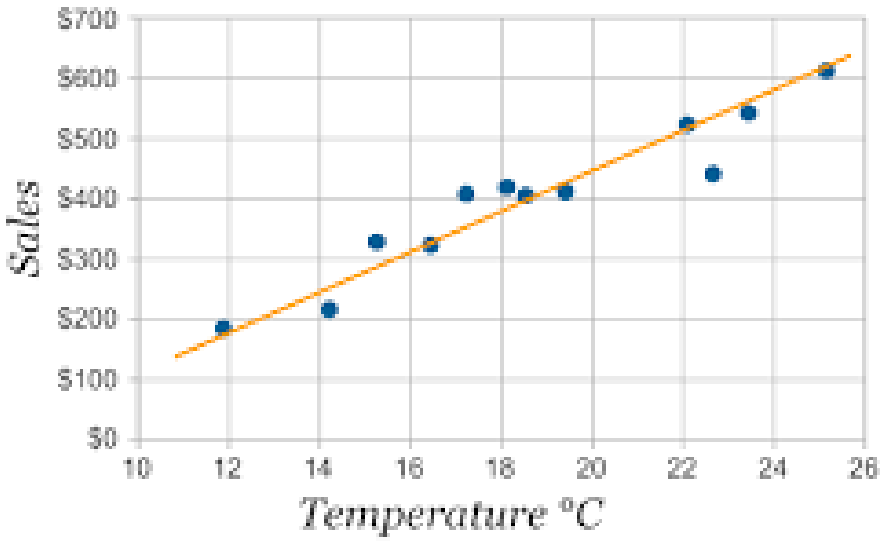
The **least squares method** is a statistical procedure to find the best fit for a set of data points by minimizing the sum of the offsets or residuals of points from the plotted curve. Least squares regression is **used to predict the behavior of dependent variables.**

HOW TO FIND

Steps

1. Step 1: For each (x,y) point calculate x^2 and xy.
2. Step 2: Sum all x, y, x^2 and xy, which gives us Σx, Σy, $Σx^2$ and Σxy (Σ means "sum up")
3. Step 3: Calculate Slope m:
4. m = N Σ(xy) − Σx Σy N Σ(x^2) − (Σx)^2
5. Step 4: Calculate Intercept b:
6. b = Σy − m Σx N.
7. Step 5: Assemble the equation of a line.

Example -



CHI SQUARES TEST

Pearson's **chi-square test** is a test used to determine whether or not there is a statistically significant difference between expected frequencies and observed frequencies in one or more categories of a contingency table.

WHERE IT CAN BE USED

The **Chi Square statistic** is commonly used for **testing relationships between categorical variables.** The null hypothesis of the **Chi-Square test** is that no relationship exists on the categorical variables in the population; they are independent.

Examples -

Color	Status	Frequency
Blue	Dead	7
Blue	Alive	129
Gold	Dead	9
Gold	Alive	46
Red	Dead	24
Red	Alive	215

	Like Scary Movies	
	Yes	No
Girls	32	38
Men	30	12
Total	62	50
Percentage	55.4%	44.6%

Bibliography

https://quizlet.com/522292658/psychology-unit-1-part-2-flash-cards/

https://stattrek.com/statistics/dictionary.aspx?definition=interquartile%20range

https://stattrek.com/descriptive-statistics/variability.aspx

https://www.michigan.gov/documents/mde/DLM_Math_Glossary_and_Examples_of_Mathematics_Terms_397694_7.pdf

https://www.corestandards.org/Math/Content/HSS

https://www.simplypsychology.org/experimental-designs.html

https://revisionmaths.com/gcse-maths-revision/statistics-handling-data/standard-deviation

https://www.dummies.com/education/math/statistics/how-to-find-statistical-probabilities-in-a-normal-distribution/

https://stattrek.com/online-calculator/combinations-permutations.aspx

https://www.statisticshowto.com/how-to-find-the-probability-of-two-events-occurring-together/

https://www.khanacademy.org/math/ap-statistics/quantitative-data-ap/histograms-stem-leaf/v/u08-l1-t2-we3-stem-and-leaf-plots

https://math.stackexchange.com/questions/1480904/given-a-95-confidence-interval-why-are-we-using-1-96-and-not-1-64

https://www.mathsisfun.com/data/least-squares-regression.html

https://www.investopedia.com/terms/c/correlationcoefficient.asp

https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/BS704_Probability12.html