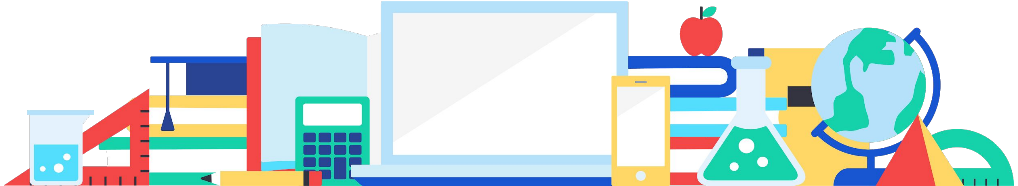
UPchieve

Statistics Concepts & Vocab

Updated September 2020



Introduction to Statistics

Types of Data

Data can be qualitative or quantitative.

- **Qualitative data** is descriptive information (it describes something)
- **Quantitative data** is numerical information (numbers)
  a. Quantitative data can be discrete or continuous. Discrete data is counted, continuous data is measured.

Example: What do we know about this cat?

Qualitative:

- He is grey and white

Quantitative:

- Discrete:
  - He has 4 legs
  - He has 2 eyes
- Continuous:
  - He weighs 2.2 lb
  - He is 8 inches tall

Levels of Measurement

Data can be classified into four levels of measurement:

- Nominal - Qualitative data in categories, not ordered
- Ordinal - Qualitative data that can be ordered and ranked
- Interval - Has a definite ordering, and differences can be measured, but no true zero value
- Ratio - Similar to interval, differences can be measures, and can have a zero value

Nominal	Ordinal	Interval	Ratio
Eye color	Place in a race (1st, 2nd, etc)	Temperature in Celsius	Weight
Gender	Socioeconomic status (low income, middle income)	Credit score	Pulse
Blood type			Flow rate

Sampling

A **population** includes all of the elements from a set of data. A **sample** consists one or more observations drawn from the population. There are four main types of sampling:

- **Simple random sampling** - All individuals are equally likely to be chosen for the sample
- **Stratified sampling** - The population is divided into groups, then a simple random sample is chosen from *each* group
- **Cluster sampling** - The population is divided into groups, and a number of clusters are chosen at random. *All* individuals in those clusters are chosen for the sample
- **Systematic sampling** - A list is created of every member of the population. From the list, we randomly select every nth element from the list

Descriptive Statistics: Representing data numerically

Measures of center

Measures of center use data points to approximate and understand a “middle value” or “average” of a given data set. The three most commonly used measures of center are the **mean**, **median**, and **mode**.

The **mean** is also known as the average. To find the mean, we add all values, and divide by the number of values.

To find the **median**, place the numbers in value order and find the middle.

The **mode** is the number that occurs the most often.

Example data set: 3, 3, 4, 5, 6, 7, 8

Mean: $$\frac{3+3+4+5+6+7+8}{7} = \frac{36}{7} = 5.1$$

Median: 3, 3, 4, 5, 6, 7, 8

Mode: 3

Measures of spread

Measures of spread of are used to describe the variability of a data set. The most commonly used measures of spread are the **range**, **quartiles**, **variance**, and **standard deviation**.

The **range** is the difference between the highest and the lowest values in a data set.

Quartiles divide the data into four equal groups. The **first quartile (Q1)** is the median of the lower half of the dataset. The **third quartile (Q3)** is the median of the upper half of the data set. The median is also known as the second quartile: it perfectly splits the data in half.

The **interquartile range (IQR)** is the difference between Q3 and Q1.

Example data set: 3, 3, 4, 5, 6, 7, 8

Range: 8 - 3 = 5

Q1: 3, 3, 4, 5, 6, 7, 8 (Remember, the median is 5, and Q1 is the median of the lower half. The lower half here is 3, 3, 4)

Q3: 3, 3, 4, 5, 6, 7, 8 (Remember, the median is 5, and Q3 is the median of the upper half. The upper half here is 6, 7, 8)

Interquartile range = Q3 - Q1 = 7 - 3 = 4

Measures of spread

The **variance** is the average of the squared differences from the mean. The steps to find the variance are as follows:

- Find the mean
- For each number, subtract the mean and square the result
- Find the average of the squared differences

The **standard deviation** is the square root of the variance

Example data set: 3, 3, 4, 5

Mean = $\dfrac{3+3+4+5}{4} = 3.75$

Variance = $\dfrac{(3-3.75)^2+(3-3.75)^2+(4-3.75)^2+(5-3.75)^2}{4} = 0.688$

Standard deviation = $\sqrt{.688} = 0.83$

Finding outliers

An **outlier** is a data point that is an abnormal distance from other points. In other words, it is data that lies **outside the other values** in the set.

An outlier is any data point that is over 1.5 IQRs below the first quartile (Q1) or above the third quartile (Q3).

Example data set: 3, 10, 14, 19, 22, 29, 32, 36, 49, 70. Are there any outliers?

An outlier would be any number *lower* than Q1 - 1.5*IQR or *higher* than Q3 + 1.5*IQR.

$$Q1 = 14 \qquad Q3 = 36 \qquad IQR = Q3 - Q1 = 36 - 14 = 22$$

$$Q1 - 1.5 * IQR = 14 - 1.5 * 22 = -19$$

$$Q3 + 1.5 * IQR = 36 + 1.5 * 22 = 69$$

Any number in our dataset that is below -19 or above 69 is an outlier.

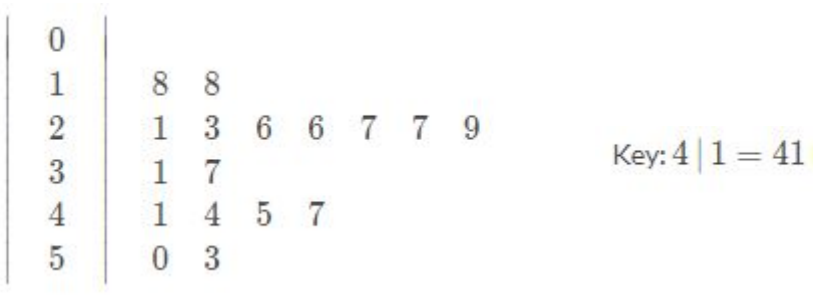
Therefore, 70 is an outlier.

Descriptive Statistics: Representing data graphically

Plots

Stem and Leaf plot

A stem and leaf plot is a table. The "stem" on the left displays the first digit, and the "leaf" on the right displays the last digit. The following stem and leaf plot displays the following data: 18, 18, 21, 23, 26, 26, 27, 27, 29, 31, 37, 41, 44, 45, 47, 50, 53

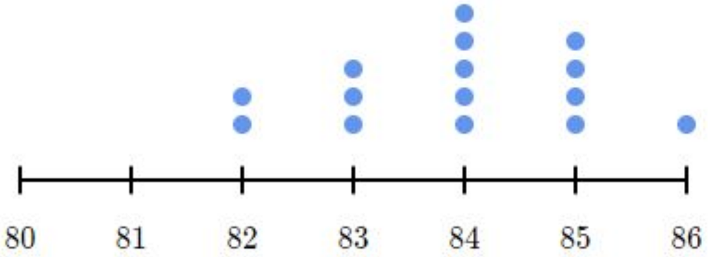


0	
1	8 8
2	1 3 6 6 7 7 9
3	1 7
4	1 4 5 7
5	0 3

Key: $4 \mid 1 = 41$

Dotplot

A dotplot is a graphical display of data using a dot for each data point. The following dotplot displays the following data: 82, 82, 83, 83, 83, 84, 84, 84, 84, 84, 85, 85, 85, 85, 86

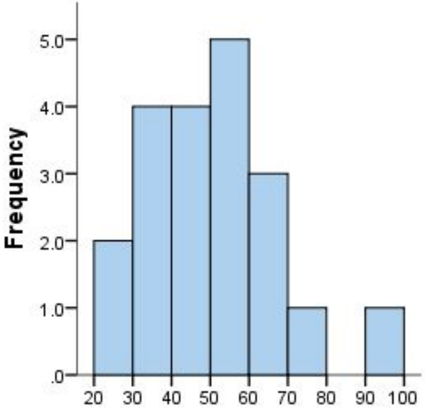


Plots

Histogram

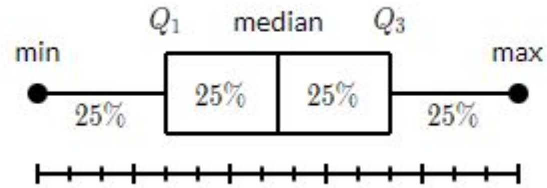
A histogram is a graphical display of data using bars of different heights. The height of each bar shows how many fall into each range. The following histogram displays the data from the following dataset:

36	25	38	46	55	68	72	55	36	38
67	45	22	48	91	46	52	61	58	55



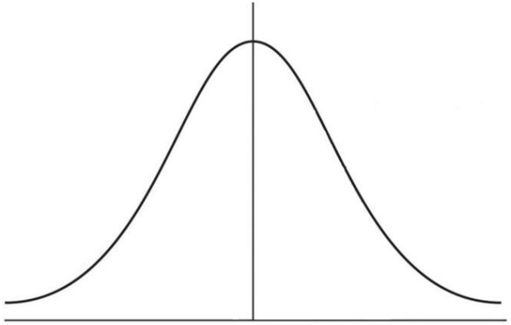
Box and Whisker Plot

A box and whisker plot is graphical display that shows quartiles 1, 2, and 3 in a box, with lines extending to the maximum and minimum value.

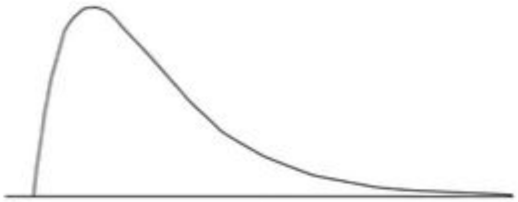


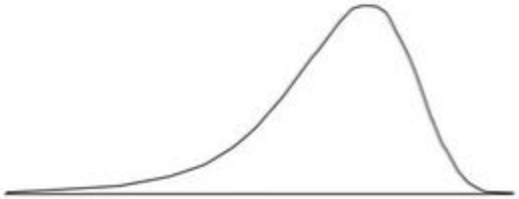
Skew

Data can be skewed, meaning it tends to have a long tail on one side or the other. Skewed data is not equally distributed on both sides of the distribution, like normal data is.



The normal curve





A **right-skewed distribution** has a long right tail. Right-skewed distributions are also called *positive-skew* distributions.

A **left-skewed distribution** has a long left tail. Left-skewed distributions are also called *negatively-skewed* distributions.

Relationships between Variables

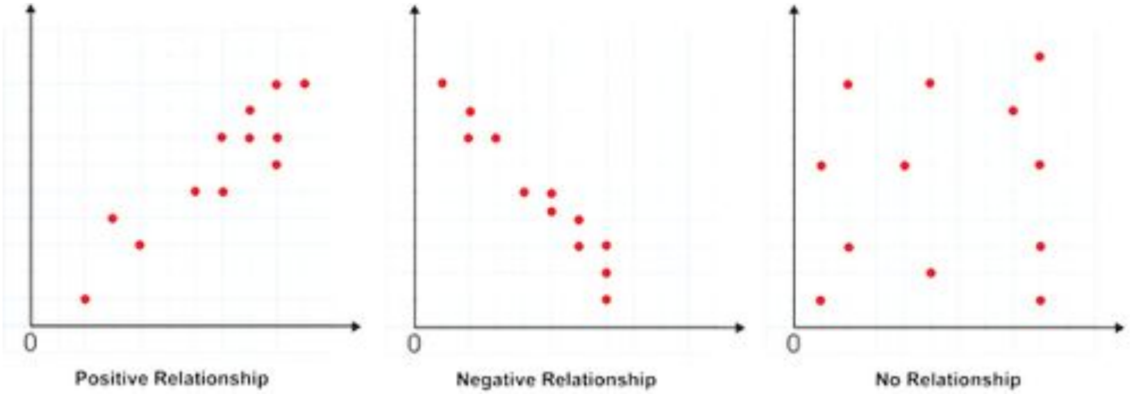
Scatterplots

An explanatory variable (X) is a variable whose values are used to explain or predict corresponding values for the response variable (Y).

Scatterplot is a great way to visualize the relationship of two variables.

Positive relationship – when X increases, Y increases

Negative relationship – when X increases, Y decreases



Positive Relationship

Negative Relationship

No Relationship

Correlation Coefficient

Correlation coefficient (r) – a measure of the strength and direction of a linear relationship.

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

Properties:

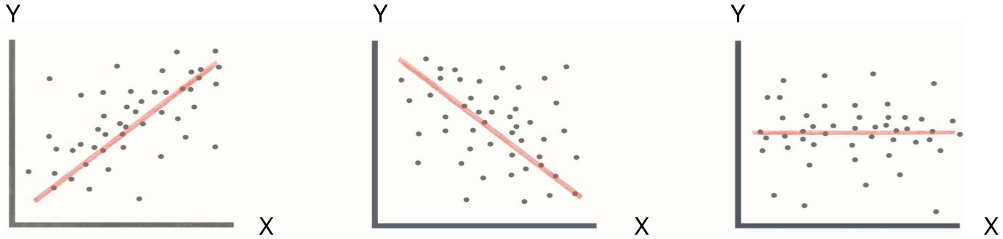
- Possible r values are between -1 and 1, inclusive
  - **r > 0** – positive relationship
  - **r < 0** – negative relationship
  - **r = 0** – no relationship
- The closer | r | is to 1, the more linear the relationship is.
- r has no units

Coefficient of determination (r^2) – measure of how well the regression line fits the data; more specifically, the proportion of variation in the response variable that is explained by the explanatory variable in the model.

- Possible r^2 values are between 0 and 1, inclusive.
- The higher the r^2, the better the regression line is in representing the data.

Linear regression model

Finding the line of best fit



The line of best fit – a line through a scatter plot of data points that best expresses the relationship between those points

Review: the equation of a line is **y = a + bx**, where a is the intercept and b is the slope

Intercept and slope for the line of best fit:

$$a = \bar{Y} - b\bar{X} \qquad b = r\frac{S_Y}{S_X}$$

- r is the correlation coefficient
- S_y is the standard deviation of the Y variable
- S_x is the standard deviation of the X variable
- $\bar{Y}$ Is the mean of the Y variable
- $\bar{X}$ Is the mean of the X variable

Linear regression model (cont.)

Interpreting a regression model

How to interpret regression coefficients:

- **Slope** – how much variable Y increases with 1 unit increase in X
- **Intercept** – the value of Y when X is 0. Sometimes, the y-intercept of the line does not have a logical interpretation in context.

Example:

We want to examine the relationship between students' study time (hours - h) and test score (s). After performing linear regression, we find the following relationship:

$$s = 10t + 40$$

Here, we see that the slope is 10 and the intercept is 40. Therefore, we can say that:

- For every 1 extra hour of study time, a student's test score, on average, increases by 10 points.
- If students does not study (study time = 0 hours), then they, on average, will receive 40 points on their test.
- There is a positive relationship between students' study time and test score.

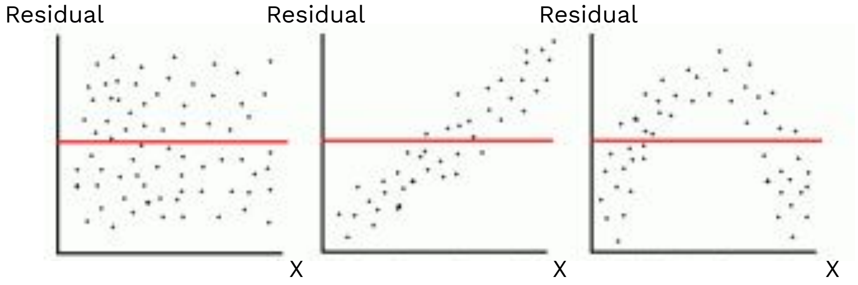
Residuals

Residuals (e) – the difference between the observed Y value (from data points) and the predicted Y value (based on the linear model).

$$e = \bar{Y} - \hat{Y}$$

A scatterplot of the residual against the explanatory (X) variable is useful to determine whether the model is a good fit for the data. This is called a **residual plot**.

- A residual plot with a **random pattern** – the variables have a linear relationship; the linear model is a good fit.
- A residual plot with a **non-random pattern** (e.g. parabolic / linear shape) – the variable have a non-linear relationship; the linear model is not a good fit.



Residual plot with random pattern

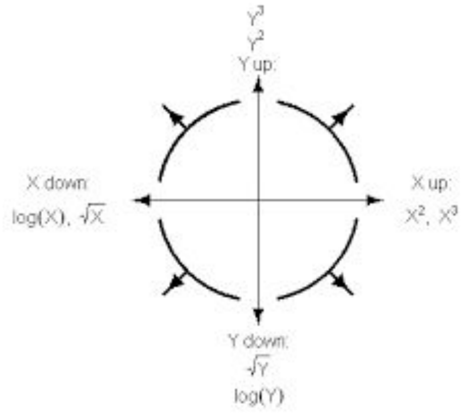
Residual plot with non-random pattern

Departure from Linearity

Vocabulary:

- **Outlier** – a point that does not follow the general trend shown in the rest of the data and has a large residual
- **High leverage point** – A point with a substantially larger or smaller x-value than the other observations have
- **Influential point** – a point that, if removed, changes the relationship substantially. Outliers and high leverage points are often influential.

Transformations – when Y and X have a non-linear relationship, transformations of y or x variables (or both) are often useful to make the relationship more linear. Common transformations: natural logarithm, powers



Choosing a transformation

For power transformations, the graph on the left shows the appropriate transformation for different non-linear graph shapes.

Random Variables

Discrete vs Continuous

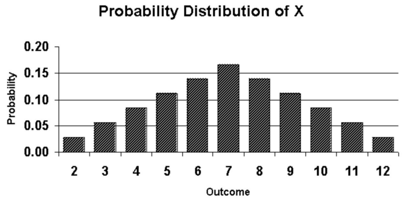
Random variable – a variable whose value is a numerical outcome of a random phenomenon

Discrete random variable – a random variable with a countable number of possible values

Its probability distribution is represented by a **probability histogram**

Example:

- Number of students attending a class
- Students' grade level (A, B, C, etc)
- Number of cars in a parking lot

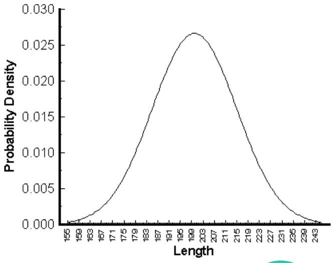


Continuous random variable – a random variable that takes all values in a given interval of numbers

Its probability distribution is represented by a **density curve**

Example:

- Height / weight of students in a class
- Distance between students' house and the school

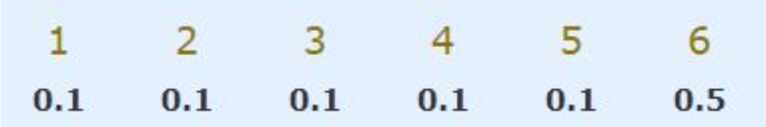


Expected Value

If we know the probability p of every value x, we can calculate the expected value (mean) of X.

$$\mu = \Sigma xp$$

Consider this probability distribution of a weighted die:



1	2	3	4	5	6
0.1	0.1	0.1	0.1	0.1	0.5

The expected value can be calculated as follows:

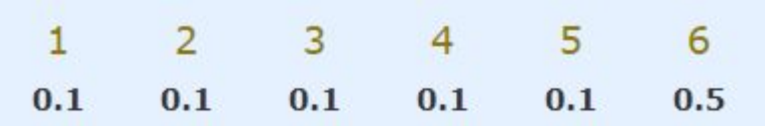
$$(1*0.1)+(2*0.1)+(3*0.1)+(4*0.1)+(5*0.1)+(6*0.5)=4.5$$

Variance

We can also find the variance from a probability distribution. The formula is:

$$\mathrm{Var}(X) = \Sigma x^2 p - \mu^2$$

Consider the same probability distribution:



1	2	3	4	5	6
0.1	0.1	0.1	0.1	0.1	0.5

To find the variance:

- Square each value and multiply by its probability
- Sum them up
- Then subtract the square of the expected value, μ^2

$$((1^2 * 0.1) + (2^2 * 0.1) + (3^2 * 0.1) + (4^2 * 0.1) + (5^2 * 0.1) + (6^2 * 0.5)) - (4.5^2)$$

$$23.5 - 20.25 = 3.25$$

The standard deviation is the square root of the variance.

Normal Distribution

Normal distribution

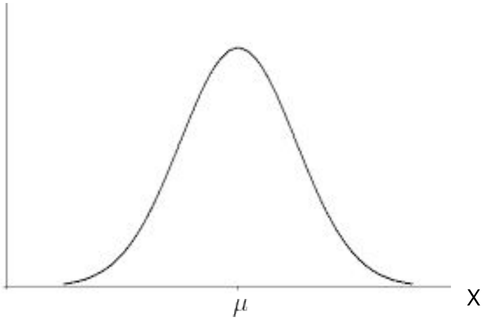
Review on distribution curves:

- Total area under the curve is 1
- Height of the curve at a given x value is the proportion of data with the given x value

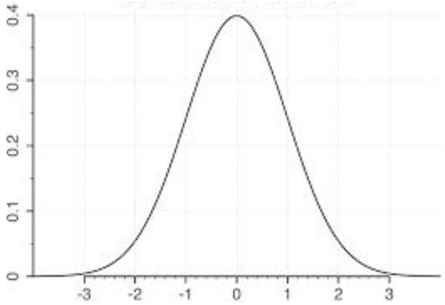
Properties of a normal curve:

- Symmetrical
- The mean, median and mode are all equal: the center of the curve / distribution (due to symmetry)
- Shape is determined by the standard deviation of the distribution

Standard Normal Distribution – A special case of normal distribution where the mean is 0 and the standard deviation is 1.



Normal Distribution



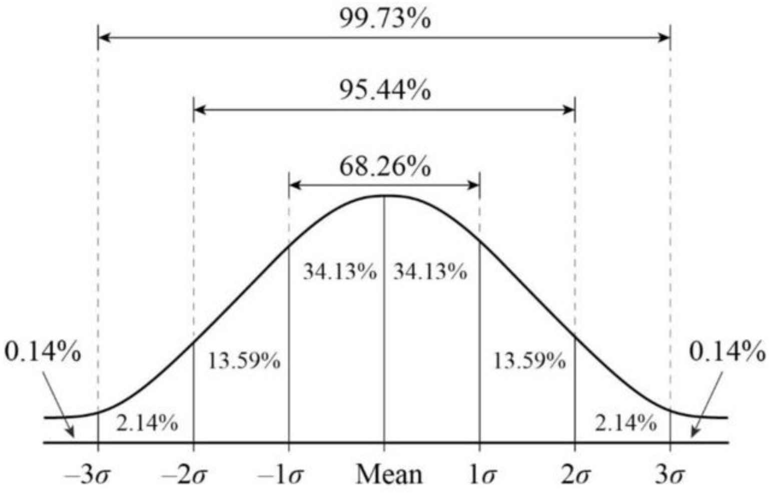
Standard Normal Distribution

Area under Normal Curve

Empirical Rule

By the **Empirical Rule**, almost all data lies within 3 standard deviations of the mean for a normal distribution.

- ~68% of data lies falls within 1 standard deviation from the mean
- ~95% of data lies falls within 2 standard deviation from the mean
- ~99.7% of data lies within 3 standard deviation from the mean



Probabilities under normal distribution

Z-score – the number of standard deviation a point is from the mean

$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean
$\sigma =$ Standard Deviation

How to find probabilities under normal distribution?

1. Convert x value into Z-score
2. Use formulas below to solve
   a. Proportion of x values less than a:
   $$P(x < a) = \Phi(z_a)$$
   b. Proportion of x values more than a:
   $$P(x > a) = 1 - P(x < a) = 1 - \Phi(z_a)$$
   c. Proportion of x values between a and b:
   $$P(a < x < b) = P(x < b) - P(x < a) = \Phi(z_b) - \Phi(z_a)$$

Note:

- z_a represents the z-score of a
- $\Phi(z_a) = P(Z < z_a)$
- $\Phi()$ values can be found in the standard normal / Z-score table (found in next slide)
- $\Phi(-z) = 1 - \Phi(z)$

Probabilities under normal distribution (cont.)

Sample Questions

Find the probability of x being less than a certain x value

An exam score is normally distributed with a mean of 67 out of 100 and a standard deviation of 9. The passing grade is 50. What is the proportion of students who did not pass?

Goal: find the proportion of students with a score of less than 50

1. Convert into z score

$$z_{50} = \frac{50-\mu}{\sigma} = \frac{50-67}{9} = -1.89$$

2. Apply the probability formula

$$P(x < 50) = P(Z < z_{50}) = P(Z < -1.89) = \Phi(-1.89)$$

3. Use the standard normal table to find the probability

$$\Phi(-1.89) = 1 - \Phi(1.89) = 1 - 0.9706 = 0.0294$$

Answer: 2.94% of students did not pass the exam.

Probabilities under normal distribution (cont.)

Sample Questions

Find data value given probability:

An exam score is normally distributed with a mean of 67 out of 100 and a standard deviation of 9. From the data, 30% of students did better than Josh. What score did Josh get?

1. **Find z-score given probability**
   a. Apply the probability formula

$$P(x > J) = P(Z > z_j) = 0.3$$

$$\Phi(z_j) = P(Z < z_j) = 1 - P(Z > z_j) = 1 - 0.3 = 0.7$$

   b. Use the standard normal table to find Josh’s z-score

$$z_j = \Phi^{-1}(0.7) = 0.758$$

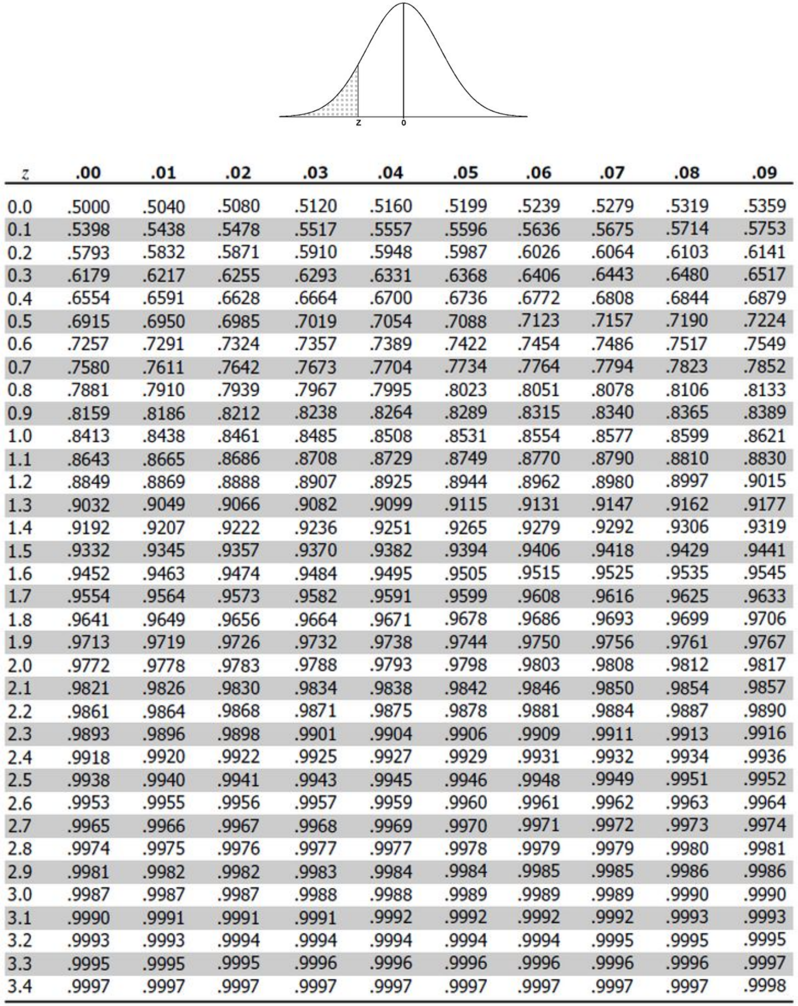
2. **Transform z-score into a data value (x-value)**
   a. Use the z-score formula to solve for Josh’s test score (J)

$$z_j = \frac{J - \mu}{\sigma} = \frac{J - 67}{9} = 0.758$$

$$J = (0.758)(9) + 67 = 73.82$$

Answer: Josh got 73.82 on the exam.

Standard Normal / Z-score Table



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Central Limit Theorem

Central Limit Theorem – the sampling distribution of the sample means approaches a normal distribution as the sample size gets larger — no matter what the shape of the population distribution.

- If a sample is large enough (n > 30), then it can be approximated as normally distributed.
- As such, the probability formula for normal distribution and the standard normal table can be used to calculate the probability for that sample.

If the population has mean μ and standard deviation σ, then the sample mean is

And the sample standard deviation is $\bar{X} = \mu$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$$

Use the sample mean and sample standard deviation for the probability and z-score calculations.

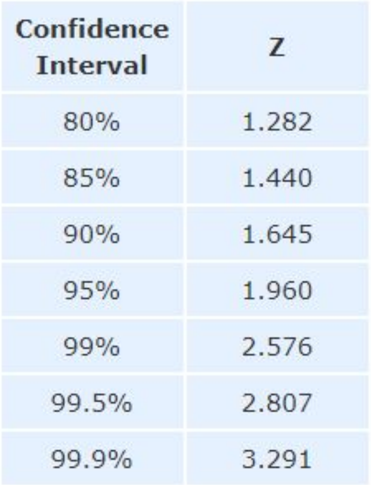
Statistical Inference: Confidence Intervals

Confidence interval for a mean

A **confidence interval** is a range of values we are fairly sure our true value lies in.

The following formula can be used to find the confidence interval for the mean X̄:

$$\overline{x} \pm z\frac{s}{\sqrt{n}}$$



Confidence Interval	Z
80%	1.282
85%	1.440
90%	1.645
95%	1.960
99%	2.576
99.5%	2.807
99.9%	3.291

Where X̄ is the sample mean, z is the z-value from the table, s is the standard deviation, and n is the number of observations.

$\frac{s}{\sqrt{n}}$ is known as the standard error.

How to interpret a confidence interval:

A researcher wanted to estimate the mean age of students at her university, so she took a random sample of 30 students. Their mean age was X̄ = 31.8 and the standard deviation was 4.3 years. A 95% confidence interval based on this data is (30.2, 33.4)

This means that we are 95% confident that the true mean age is between 30.2 and 33.4 years old.

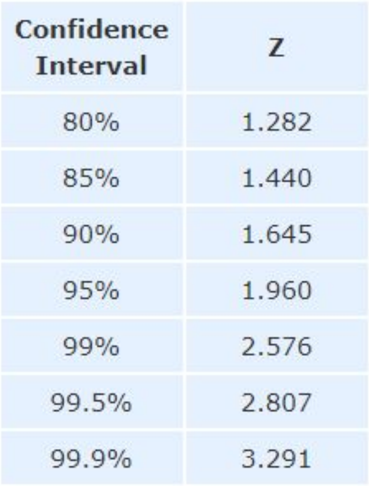
Confidence interval for a proportion

We can also create confidence intervals to estimate a proportion. The following formula can be used:

$$\hat{p} \pm z^{*}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Where p̂ is the sample proportion, z is the z-value from the table, and n is the number of observations.

$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is known as the standard error.



Confidence Interval	z
80%	1.282
85%	1.440
90%	1.645
95%	1.960
99%	2.576
99.5%	2.807
99.9%	3.291

Confidence interval for the difference between two means

We can also use confidence intervals to estimate the difference between two means. The following formula is used:

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Where $\bar{X}_1$ and $\bar{X}_2$ are the sample means, t is the critical value from the t table, s_1 and s_2 are the sample standard deviations, and n_1 and n_2 are the sample sizes.

There is a different formula that is used when the population standard deviations are unknown, but assumed to be equal. S_p is the pooled estimate of the standard deviation.

$$(\bar{x}_1 - \bar{x}_2) \pm z\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where

$$S_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

Confidence interval for the difference between two proportions

We can also use confidence intervals to estimate the difference between two proportions. The following formula is used:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Where $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions, z is the critical value from the z table, and n_1 and n_2 are the sample sizes.

Statistical Inference: Hypothesis tests

Hypothesis test

Hypothesis testing steps:

1. State the null and alternative hypothesis
2. Find the value of the test statistic
3. Find the p-value
4. Interpret the results and conclude whether to reject the null hypothesis
   a. If the p-value is less than alpha, we reject the null hypothesis
   b. If the p-value is greater than alpha, we fail to reject the null hypothesis

Hypothesis test example

A principal at a certain school claims that the students in his school are above average intelligence. A random sample of thirty students IQ scores have a mean score of 112. Is there sufficient evidence to support the principal's claim at? The mean population IQ is 100 with a standard deviation of 15. Use α = 0.05. The hypotheses can be written as follows:

H_0: μ=100 (The mean is 10)

H_A: μ > 100 (The mean is greater than 100)

We use the following formula to find the test statistic:

$$Z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{112.5 - 100}{15/\sqrt{30}} = 4.56$$

We use the test statistic to find the p-value. You can use an online calculator, like this one:

https://goodcalculators.com/p-value-calculator/

The p-value type is right tailed when the alternative hypothesis uses > or ≥, left tailed when the alternative hypothesis uses < or ≤, and two tailed when the alternative hypothesis uses ≠. Our p-value using a z-score of 4.56 and a right tailed test is **0.0000026**.

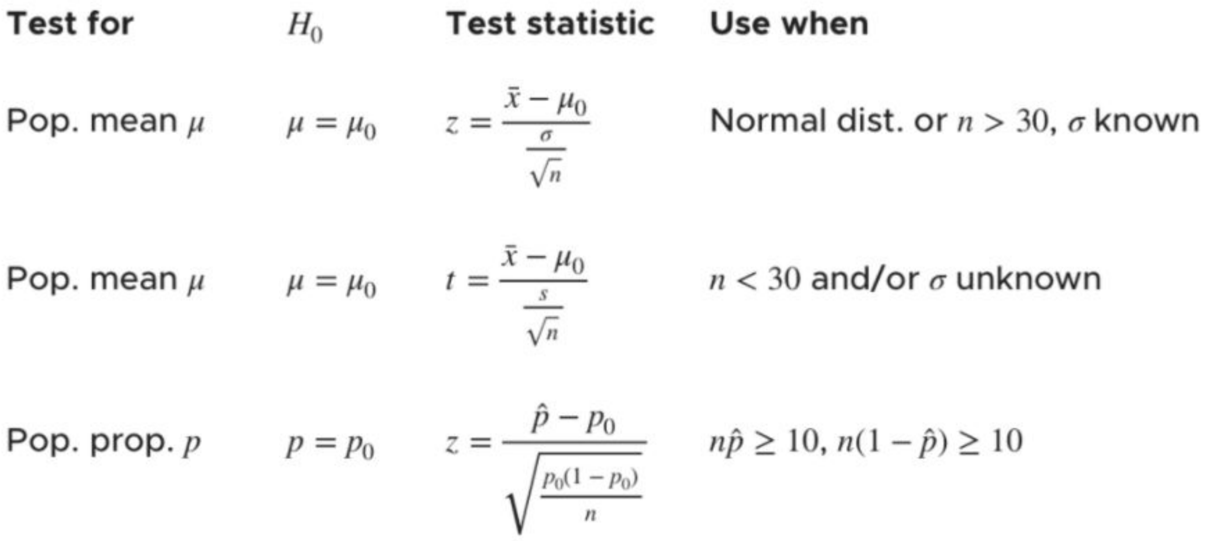
Hypothesis test example continued

We come to a conclusion by comparing our p-value to our α value. In this case,
the p-value < α, since $0.0000026 < 0.05$.

When the p-value is less than alpha, we reject the null hypothesis. We have enough evidence to conclude that the mean IQ is greater than 100.

Test statistic formulas

We use different formulas for the test statistic, depending on if we are testing for a mean or a proportion, and depending on whether the standard deviation is known. Refer to the following table.



Test for	H_0	Test statistic	Use when
Pop. mean μ	$\mu = \mu_0$	$z = \dfrac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$	Normal dist. or $n > 30$, σ known
Pop. mean μ	$\mu = \mu_0$	$t = \dfrac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$	$n < 30$ and/or σ unknown
Pop. prop. p	$p = p_0$	$z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$	$n\hat{p} \geq 10$, $n(1 - \hat{p}) \geq 10$

Type I and Type II Errors

Two types of errors can result from a hypothesis test:

- Type I error: A Type I error occurs when the researcher rejects a null hypothesis when it is true.
- Type II error: A Type II error occurs when the researcher fails to reject a null hypothesis that is false.

Probability

Permutation & Combination

Permutations – the number of different ways that r objects picked out of n objects can be ordered

$$P(n,r) = \frac{n!}{(n-r)!}$$

Example: A code have 4 digits in a specific order, the digits are between 0-9. How many different permutations are there if one digit may only be used once?

$$P(n,r) = \frac{10!}{(10-4)!} = \frac{10\cdot 9\cdot 8\cdot 7\cdot 6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1}{6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1} = 5040$$

Answer: There are 5040 different lock combinations for a 4 digit passcode.

Combinations – the number of different groups of r objects that can be picked out of n objects. Order does not matter.

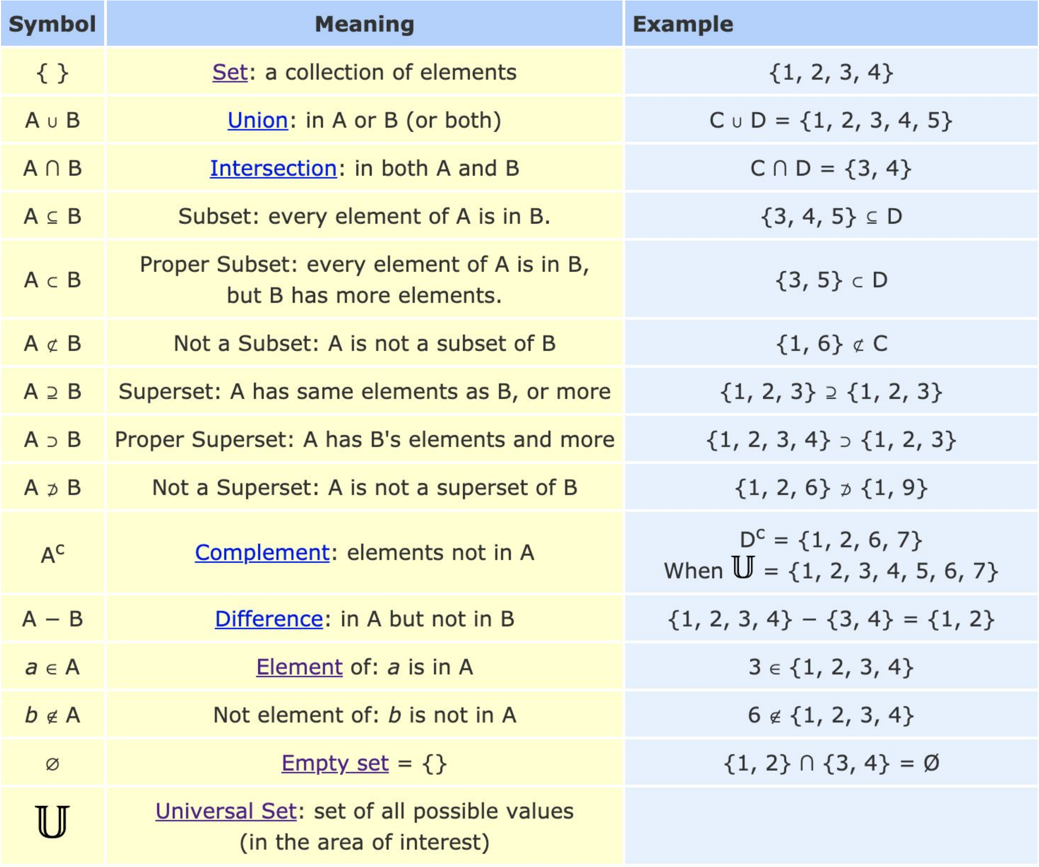
$$C(n,r) = \frac{n!}{r!(n-r)!}$$

Example: 2 students out of a group of 5 will be picked to do a presentation. How many possible unique pairs of presenters are there?

$$C(n,r) = \frac{5!}{(5-2)!2!} = \frac{5\cdot 4\cdot 3\cdot 2\cdot 1}{(3\cdot 2\cdot 1)(2\cdot 1)} = 10$$

Answer: There are 10 different possible pairs of presenters.

Set Theory Notations



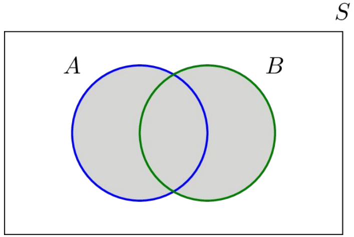
Symbol	Meaning	Example
$\{\ \}$	Set: a collection of elements	$\{1, 2, 3, 4\}$
$A \cup B$	Union: in A or B (or both)	$C \cup D = \{1, 2, 3, 4, 5\}$
$A \cap B$	Intersection: in both A and B	$C \cap D = \{3, 4\}$
$A \subseteq B$	Subset: every element of A is in B.	$\{3, 4, 5\} \subseteq D$
$A \subset B$	Proper Subset: every element of A is in B, but B has more elements.	$\{3, 5\} \subset D$
$A \not\subset B$	Not a Subset: A is not a subset of B	$\{1, 6\} \not\subset C$
$A \supseteq B$	Superset: A has same elements as B, or more	$\{1, 2, 3\} \supseteq \{1, 2, 3\}$
$A \supset B$	Proper Superset: A has B's elements and more	$\{1, 2, 3, 4\} \supset \{1, 2, 3\}$
$A \not\supset B$	Not a Superset: A is not a superset of B	$\{1, 2, 6\} \not\supset \{1, 9\}$
A^c	Complement: elements not in A	$D^c = \{1, 2, 6, 7\}$ When $\mathbb{U} = \{1, 2, 3, 4, 5, 6, 7\}$
$A - B$	Difference: in A but not in B	$\{1, 2, 3, 4\} - \{3, 4\} = \{1, 2\}$
$a \in A$	Element of: *a* is in A	$3 \in \{1, 2, 3, 4\}$
$b \notin A$	Not element of: *b* is not in A	$6 \notin \{1, 2, 3, 4\}$
$\varnothing$	Empty set = {}	$\{1, 2\} \cap \{3, 4\} = \varnothing$
$\mathbb{U}$	Universal Set: set of all possible values (in the area of interest)	

Source: https://www.mathsisfun.com/sets/symbols.html

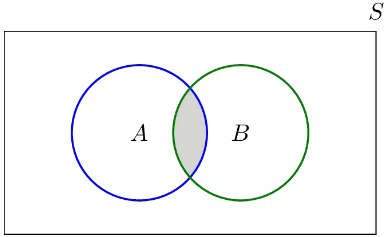
Set operations

Universal set (U or S): the set of all elements under consideration

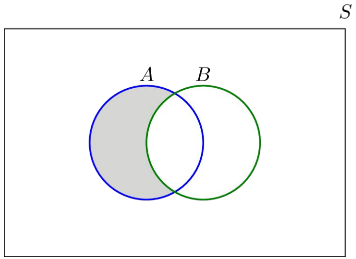
Disjointed / Mutually exclusive: if two sets do not have any shared elements



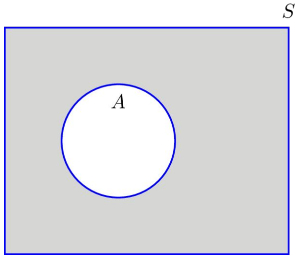
Union of two sets ($A \cup B$):
a set of all elements that are in A or in B (possibly both).



Intersection of two sets ($A \cap B$):
A set of elements that are in both A and B



Difference of two sets ($A - B$):
a set of all elements that are in A but not in B



Complement of a set (A^c or $\bar{A}$):
the set of all elements that are in the universal set but are not in A

Independent / Dependent Events

Independent events – events whose probability of occurring is not affected by another event's chances of happening

Examples:

- Coin toss
- Getting a parking ticket and running out of milk

Dependent events – When two events are dependent events, one event influences the probability of another event

Examples:

- Being late to work and being reprimanded by your boss
- Parking illegally and getting a parking ticket

If two events A and B are **independent**, then the probability of both A and B happening is

$$P(A \cap B) = P(A)P(B)$$

If two events A and B are **dependent**, then the probability of both A and B happening is

$$P(A \cap B) = P(A)P(B|A)$$

(more info about conditional probability is available at the end of the subsection)

Mutually Exclusive

Mutually exclusive – two or more events that cannot happen simultaneously.

Examples:

- Turning left and right in 1 intersection
- Getting heads and tails in 1 coin toss
- Running forwards and backwards at the same time

If two events A and B are **mutually exclusive,** then the probability of both A and B happening is

$$P(A \cap B) = 0$$

Sampling

Sampling with replacement – when an element of the set can be chosen more than once (element is returned to the set after being chosen)

Example: We have a box of 5 red, 2 yellow and 3 green balls. A ball is chosen at random, and then it is returned before a second ball is chosen. What is the probability that a red ball is chosen both times?

A = red is chosen in first pick; B = red is chosen in second pick

$$P(A) = P(B) = \frac{5}{10} = \frac{1}{2}$$

$$P(A \cap B) = P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

Answer: the probability that a red ball is chosen both times with replacement is 1/4

Sampling

Sampling without replacement – when an element of the set cannot be chosen more than once (element is not returned to the set after being chosen)

Example: We have a box of 5 red, 2 yellow and 3 green balls. A ball is chosen at random, put away and a second ball is chosen from the box. What is the probability that a red ball is chosen both times?

Total # of balls before 1st pick = 10; Total # of balls before 2nd pick = 9

Total # of red balls before 1st pick = 5; Total # of red balls before 2nd pick = 4

A = red is chosen in first pick; B = red is chosen in second pick

$$P(A) = \frac{5}{10} \qquad P(B) = \frac{4}{9}$$

$$P(A \cap B) = P(A) \cdot P(B) = \frac{5}{10} \cdot \frac{4}{9} = \frac{2}{9}$$

Answer: the probability that a red ball is chosen both times without replacement is 2/9

Addition Rule

The probability that **either event A or event B** happening is:

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

If A and B are **independent**, then

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

Since $P(A \cap B) = 0$ for independent events

Example:

The probability that it rains on Monday is 30% and on Tuesday is 40%. The probability that it rains on both days is 25%. What is the probability that it rained on either Monday or Tuesday?

A = it rains on Monday; B = it rains on Tuesday

P(A) = 0.3; P(B) = 0.4; P(A and B) = 0.25

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.3 + 0.4 - 0.25 = 0.45$$

Answer: there is a 45% chance that it rains on either Monday or Tuesday.

Conditional Probability

Conditional probability, noted as $P(A|B)$, is the probability of event A happening given that event B has happened.

Bayes Rule:

$P(A \mid B)$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B)}$$

If A and B are independent, then $P(A|B) = P(A)$

Example:

The probability that it rains on Monday is 30% and on Tuesday is 40%. If it rains on Monday, then the probability that it rains on Tuesday increased to 50%. What is the probability that it rained on Monday, given that it rains on Tuesday?

A = it rains on Monday; B = it rains on Tuesday

P(A) = 0.3; P(B) = 0.4; P(B|A) = 0.5

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{(0.5)(0.3)}{0.4} = 0.375 = 37.5\%$$

Answer: there is a 37.5% chance that it rained on Monday.