Data Science – Images

Descriptive Statistics

The formula for the population mean is as follows:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{N}$$

where:

μ = the population mean (pronounced *mu*, as in "I hope you find this a*mu*sing")

$\sum_{i=1}^{n} x_i$ = the sum of all the data values in the population

N = the number of data values in the population

$$\sigma = \sqrt{\sigma^2}$$

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N}$$

where:

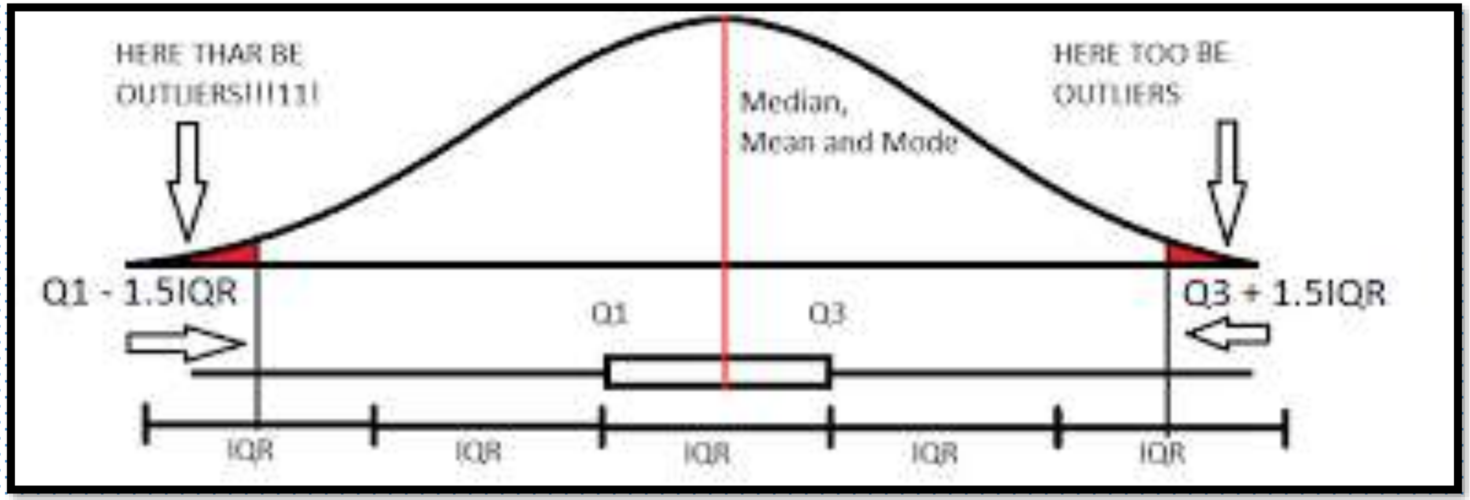
σ^2 = the variance of the population (pronounced "sigma squared")

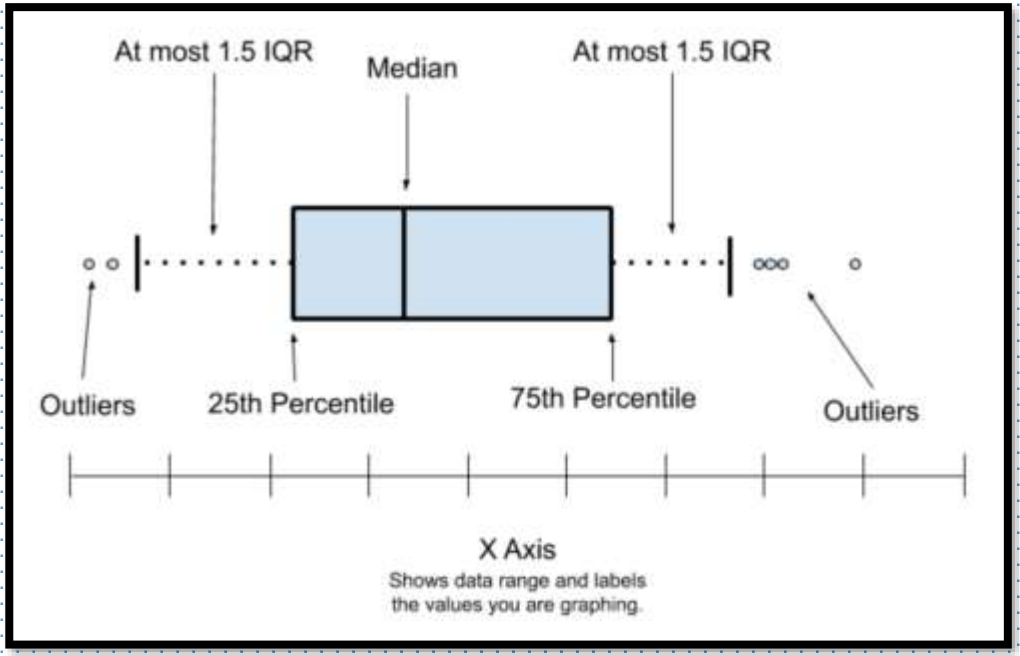
x_i = the measurement of each item in the population

μ = the population mean

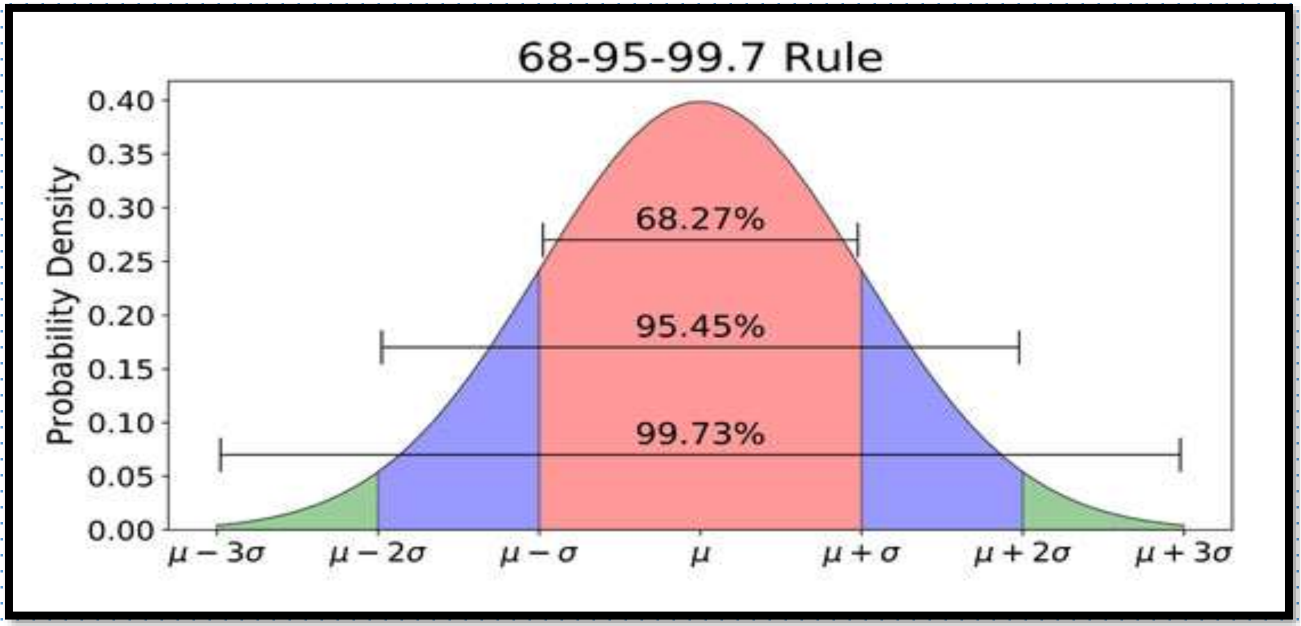
N = the size of the population

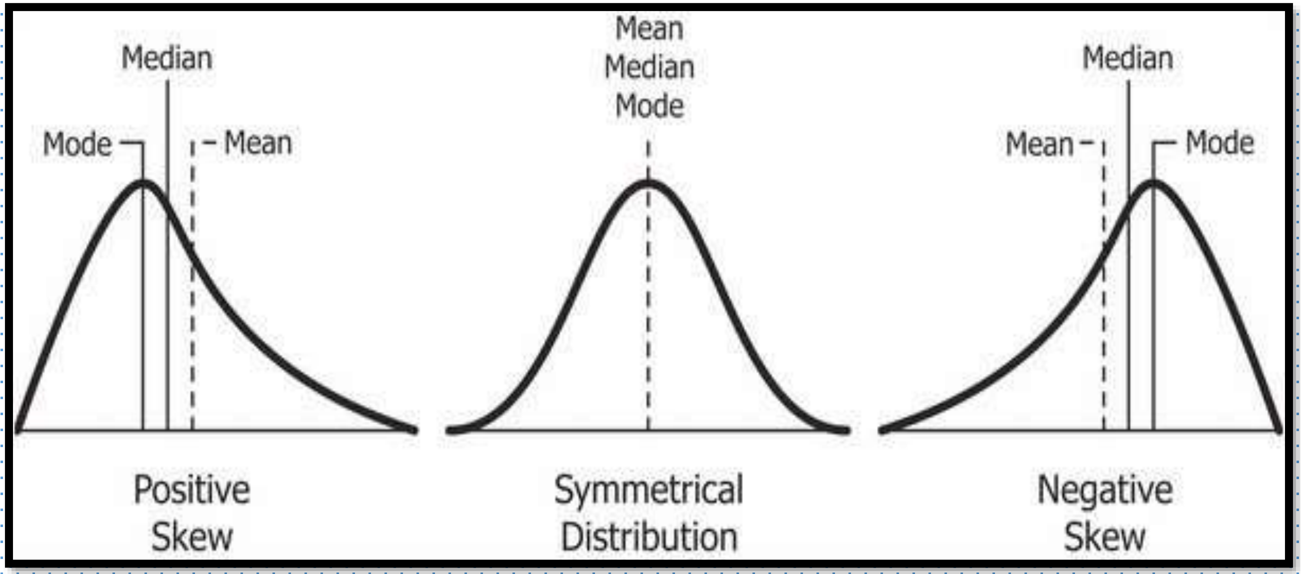
Descriptive Statistics



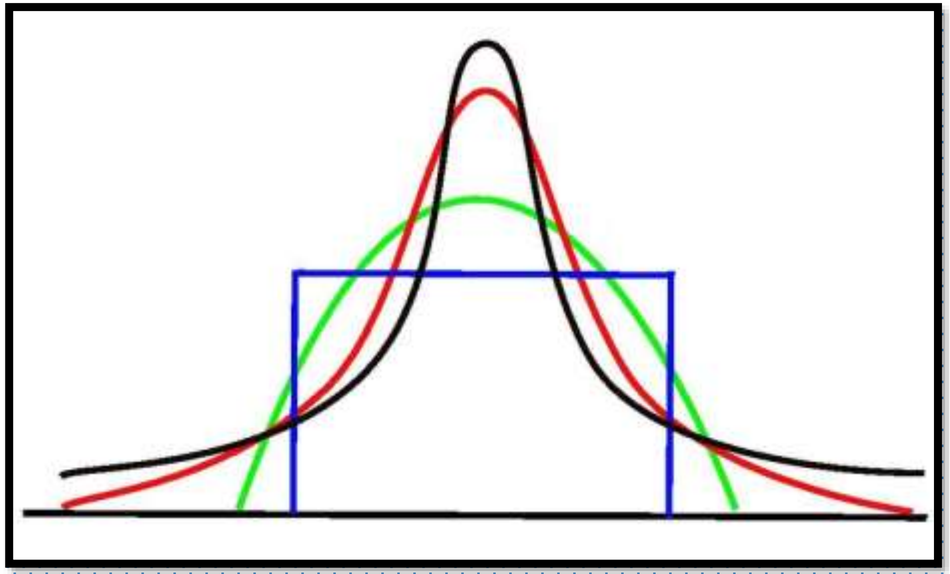


Descriptive Statistics





- **High kurtosis** - Black
- **Low kurtosis** - Green



Probability

def•i•ni•tion

Permutations are the number of different ways in which objects can be arranged in order. The number of permutations of n objects taken r at a time can be found by ${}_nP_r = \frac{n!}{(n-r)!}$.

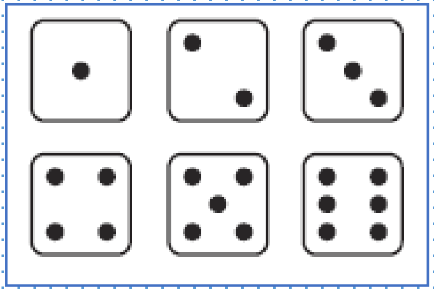
def•i•ni•tion

Combinations are the number of different ways in which objects can be arranged without regard to order. The number of combinations of n objects taken r at a time can be found by ${}_nC_r = \frac{n!}{(n-r)!r!}$.

$$P[A] = \frac{\text{Number of possible outcomes in which Event A occurs}}{\text{Total number of possible outcomes in the sample space}}$$

where:

P[A] = the probability that Event A will occur



Probability

Contingency Table for the Tennis Example

Warm-Up Time	Debbie Wins (A)	Bob Wins (A')	Total
Less than 10 min (B)	4	9	13
10 min or more (B')	5	2	7
Total	9	11	20

The events of interest are ...

- Event A = Debbie wins the tennis match.
- Event B = the warm-up time is less than 10 minutes.
- Event A' = Bob wins the tennis match.
- Event B' = the warm-up time is 10 minutes or more.

A z-score measures exactly how many standard deviations above or below the mean a data point is.

Here's the formula for calculating a z-score:

$$z = \frac{\text{data point} - \text{mean}}{\text{standard deviation}}$$

Here's the same formula written with symbols:

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

Confidence Intervals – for sample size >30

- Let's say from a sample of 32 customers the average order is $78.25 and the population standard deviation is $37.50. (This represents the variation among orders within the population.). The SD of the sample can be used if population SD is not known

- Put simply, the **standard error** of the sample mean is an estimate of how far the sample mean is likely to be from the population mean
- We can do this by increasing the sample size.

- **Problem statement** : Where does the population mean lie (lower and upper limits) if 90% is the confidence level?

In general, we can construct a *confidence interval* around our sample mean using the following equations:

$\bar{x} + z_c\sigma_{\bar{x}}$ (upper limit of confidence interval)

$\bar{x} - z_c\sigma_{\bar{x}}$ (lower limit of confidence interval)

where:

$\bar{x}$ = the sample mean

z_c = the critical z-score, which is the number of standard deviations based on the confidence level

$\sigma_{\bar{x}}$ = the standard error of the mean (remember our friend from Chapter 13?)

The term $z_c\sigma_{\bar{x}}$ is referred to as the *margin of error*, or *E*, a phrase often referred to in polls and surveys.

Confidence Intervals with Various Confidence Levels

Confidence Level	z_c	$\sigma_{\bar{x}}$	Sample Mean	Lower Limit	Upper Limit
90	1.64	\$6.63	\$78.25	\$67.38	\$89.12
95	1.96	\$6.63	\$78.25	\$65.26	\$91.24
99	2.57	\$6.63	\$78.25	\$61.21	\$95.29

$\bar{x} = \$78.25$

$n = 32$

$\sigma = \$37.50$

$z_c = 1.64$

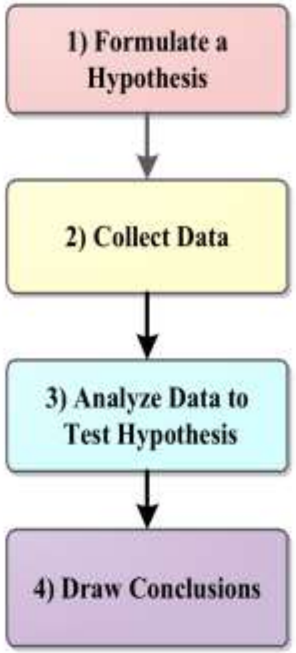
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\$37.50}{\sqrt{32}} = \$6.63$$

$$\text{Upper limit} = \bar{x} + 1.64\sigma_{\bar{x}} = \$78.25 + 1.64(\$6.63) = \$89.12$$

$$\text{Lower limit} = \bar{x} - 1.64\sigma_{\bar{x}} = \$78.25 - 1.64(\$6.63) = \$67.38$$

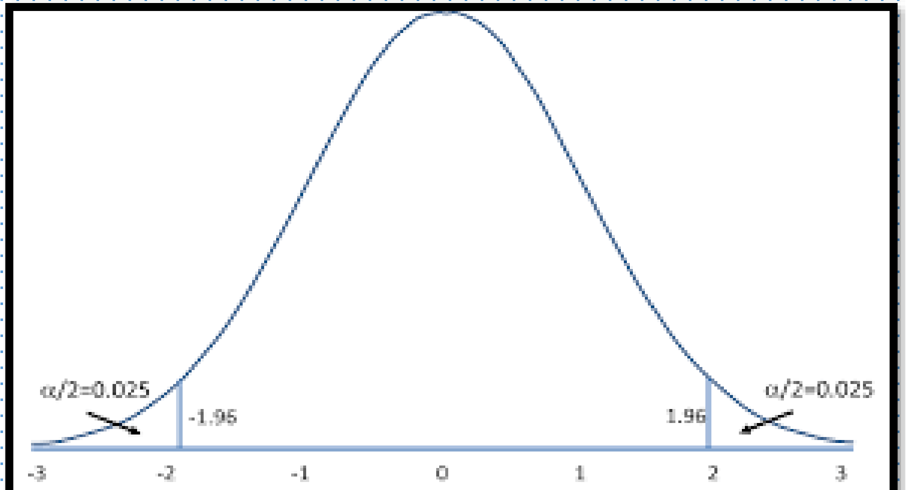
Hypothesis testing

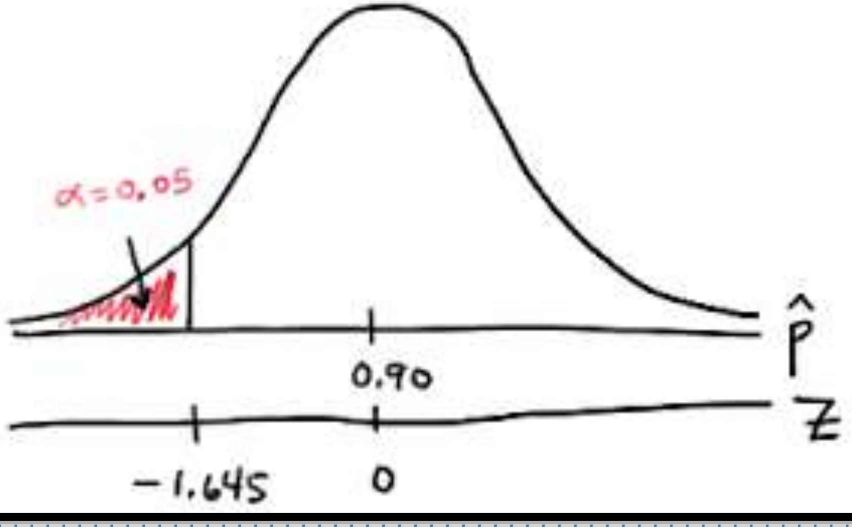
Null Hypothesis	Alternative Hypothesis
$H_0 : \mu = 6.0$	$H_1 : \mu \neq 6.0$
$H_0 : \mu \geq 6.0$	$H_1 : \mu < 6.0$
$H_0 : \mu \leq 6.0$	$H_1 : \mu > 6.0$



The following table summarizes the two types of hypothesis errors.

	H_0 Is True	H_0 Is False
Reject H_0	Type I Error P[Type I Error] = α	Correct Outcome
Do Not Reject H_0	Correct Outcome	Type II Error





Let us define the test statistic z in terms of the sample mean, the sample size and the population standard deviation σ :

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Chi-square distribution

The chi-square statistic is found using the following equation:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where:

O = the number of observed frequencies for each category

E = the number of expected frequencies for each category

Observed Frequencies for Tennis Example

	0–10 Min	11–20 Min	More than 20 Min	Total
Debbie wins	4	10	9	23
Bob wins	14	9	4	27
Total	18	19	13	50

First we state the hypotheses as:

H_0: Warm-up time is independent of performance

H_1: Warm-up time affects performance

H_0: The actual rating distribution can be described by the expected distribution.

H_1: The actual rating distribution differs from the expected distribution.

Expected Movie-Rating Distribution

Number of Stars	Percentage
5	40%
4	30%
3	20%
2	5%
1	5%
Total	100%

After its debut, a sample of 400 moviegoers were asked to rate the movie, with the results shown in the following table.

Observed Movie-Rating Distribution

Number of Stars	Number of Observations
5	145
4	128
3	73
2	32
1	22
Total	400

ANOVA, f-test

To test the hypothesis for ANOVA, we need to compare the calculated test statistic to a critical test statistic using the F-distribution. The calculated F-statistic can be found using the equation:

$$F = \frac{MSB}{MSW}$$

where MSB is the *mean square between*, found by:

$$MSB = \frac{SSB}{k-1}$$

and MSW is the *mean square within*, found by:

$$MSW = \frac{SSW}{N-k}$$

hypothesis statement would look like the following:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_1 : \text{not all } \mu\text{'s are equal}$$

Data for Lawn Clippings

	Fertilizer 1	Fertilizer 2	Fertilizer 3
	10.2	11.6	8.1
	8.5	12.0	9.0
	8.4	9.2	10.7
	10.5	10.3	9.1
	9.0	9.9	10.5
	8.1	12.5	9.5
Mean	9.12	10.92	9.48
Variance	1.01	1.70	0.96

Thank You