Statistics Formulae Collection

Wolfgang W. Stoettner mail@stoettner.net

August 26, 2021

The following collection represents an excerpt of the statistics formulae taken from the perennial text book *Lind, Douglas A. et. al. (2015): Statistical Techniques in Business and Economics, Sixteenth edition, New York, NY 2015.*

Frequency distributions

$$\tag{1}$$

Constructing frequency distributions where k is the smallest number of classes and n the number of observations:

1. decide on the number of classes k where $2^k > n$
2. determine the class interval i by $i \geq \frac{\text{max value - min value}}{k}$
3. set individual class limits.
4. tally the values into the classes.
5. count the number of items in each class.

Population mean (raw data)

$$\mu = \frac{\sum x}{N} \tag{2}$$

where:

- μ denotes the population mean.
- x denotes any value.
- N denotes the number of the values in the population.

Sample mean

$$\bar{x} = \frac{\sum x}{n} \tag{3}$$

where:

- $\bar{x}$ denotes the sample mean.
- n denotes the number of values in the sample.
- x denotes any value.

median

$$\tag{4}$$

The midpoint of the values after they have been ordered from the minimum to the maximum values. The data must be at least an ordinal level of measurement.

mode

$$\tag{5}$$

The value of the observation that appears most frequently. It is especially useful in summarizing ordinal level data.

Weighted mean

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + \cdots + w_nx_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum\limits_{i=1}^{n} w_ix_i}{\sum\limits_{i=1}^{n} w_i} \tag{6}$$

where:

- $\bar{x}_w$ denotes the weighted mean.
- w denotes the corresponding weight.

Geometric mean

$$\text{Geometric mean} = \sqrt[n]{(x_1)(x_2)\cdots(x_n)} \tag{7}$$

Note:
Useful in finding the *average change* – in contrast to equation (8) – of percentages, ratios, indices or growth rates over time. The geometric mean will always be less than or equal (never more than) the arithmetic mean. Also, all the data values must be *positive*. It is applied in Fisher's Ideal Index as in formula (111).

Geometric mean for a rate increase over time

$$\text{GM for a rate increase} = \sqrt[n]{\frac{\text{value at end of period}}{\text{value at start of period}}} - 1 \tag{8}$$

Used to find an average percentage *increase* (in contrast to (7)) over time.

Range

$$\text{Range} = \text{maximum value} - \text{minimum value} \tag{9}$$

Population variance

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N} \tag{10}$$

Note: The population variance σ^2 in essence mitigates the dilemma of a single sample. In the first sample the deviation between observed value x and the population mean μ might differ to a great extent, in a second sample the deviation might well be very different again. Here, σ^2 provides a measure for the average variance accounting for all samples for one unit of the population.

where:

- σ^2 is the population variance.
- x is the value of a particular observation in the population.
- μ is the arithmetic mean of the population.
- N is the number of observations in the population.

Population standard deviation

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}} \tag{11}$$

Note: The population standard deviation σ in essence mitigates the dilemma of a single sample (ref. (10) . In the first sample the deviation between observed value x and the population mean μ might differ to a great extent, in a second sample the deviation might well be very different again. Here, σ provides a measure for the average deviation accounting for all samples for one unit of the population in the same unit of measure as in the sample.

where:

- σ is the population standard deviation.
- x is the value of a particular observation in the population.
- μ is the arithmetic mean of the population.
- N is the number of observations in the population. By taking the square root of the variance the deviation is now of the same unit of measurement as the original data.

Sample variance

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1} \tag{12}$$

where:

- s^2 is the sample variance.
- x is the value of a each observation in the sample.
- $\bar{x}$ is the mean of the sample.
- n is the number of observations in the sample.
- The denominator $(n-1)$ corrects its tendency for underestimation.

Sample standard deviation

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} \tag{13}$$

where:

- s is the sample standard deviation.
- x is the value of a each observation in the sample.
- $\bar{x}$ is the mean of the sample.
- n is the number of observations in the sample.
- By taking the square root of the variance the deviation is now of the same unit of measurement as the original data.

Chebyshev's Theorem

(14)

Chebyshev's Theorem
For any set of observations (sample or population), the proportion of the values that lie within k standard deviations of the mean is at least $1-\frac{1}{k^2}$, where k is any value greater than 1.

Arithmetic mean of grouped data

$$\bar{x} = \frac{\sum fM}{n} \tag{15}$$

where:

- $\bar{x}$ is the sample mean.
- M is the midpoint of each class.
- f is the frequency in each class.
- n is the total number of frequencies.

Standard deviation of grouped data

$$s = \sqrt{\frac{\sum f(M-\bar{x})^2}{n-1}} \tag{16}$$

where:

- s is the sample standard deviation.
- M is the midpoint of the class.
- f is the the class frequency.
- $\bar{x}$ is the mean of the sample.
- n is the number of observations in the sample.

Location of percentile

$$L_p = (n+1)\frac{P}{100} \tag{17}$$

where:

- P is the percentile.
- n is the number of observations.
- L_p is the location of the percentile.

Coefficient of Variation[1]

$$\text{CV} = \frac{s}{\bar{x}}(100) \tag{18}$$

Note: multiplying by 100 converts the decimal to a percent where:

- s is the sample standard deviation.
- $\bar{x}$ is the sample mean.

Pearson's coefficient of skewness

$$\text{sk} = \frac{3(\bar{x}-\text{median})}{s} \tag{19}$$

where:

- s is the sample standard deviation.
- $\bar{x}$ is the sample mean.

[1] (Lind et al., 2002, ISBN 0-07-112318-0, p. 115)

Software coefficient of skewness

$$\text{sk} = \frac{n}{(n-1)(n-2)}\left[\Sigma(\frac{x-\bar{x}}{s})^3\right] \tag{20}$$

where:

- n is the number of observation in the sample.
- s is the sample standard deviation.
- $\bar{x}$ is the sample mean.

Classical probability

$$\text{Probability of an event} = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}} \tag{21}$$

Empirical probability

$$\text{Probability of an event} = \frac{\text{Number of times event occurred in the past}}{\text{Total number of observations}} \tag{22}$$

Special Rule of Addition

$$\text{P(A or B)} = P(A) + P(B) \tag{23}$$

Events must be *mutually exclusive.*

Complement Rule

$$P(A) = 1 - P(\sim A) \tag{24}$$

Events A and ∼A are mutually exclusive and collectively exhaustive.

General Rule of Addition

$$\text{P(A or B) = P(A) + P(B) - P(A and B)} \tag{25}$$

Events that are *not* mutually exclusive.

Special Rule of Multiplication

$$\text{P(A and B) = P(A)P(B)} \tag{26}$$

Requires that two events are *independent*, that is, the occurrence of one event has no effect on the probability of the occurrence of the other event.

General Rule of Multiplication

$$\text{P(A and B) = P(A)}P(B|A) \tag{27}$$

where:

- P(B|A) stands for the probability of B will occur given that A has already occurred (conditional probability).
- two events are *not independent.*

Bayes' Theorem

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2)} \tag{28}$$

Events A_1 and A_2 are mutually exclusive and collectively exhaustive, and A_i refers to either event $A_1 or A_2$.

Multiplication Formula

$$\text{Total number of arrangements} = (m)(n) \tag{29}$$

Permutation

$${}_nP_r = \frac{n!}{(n-r)!} \tag{30}$$

Any arrangement of r objects selected from a single group of n possible objects.

Combination Formula

$${}_nC_r = \frac{n!}{r!(n-r)!} \tag{31}$$

The order of the selected objects is *not important.*

Mean of a Probability Distribution

$$\mu = \Sigma\left[xP(x)\right] \tag{32}$$

where:

- μ is the population mean.
- $P(x)$ is the probability.
- x is a particular value.

Variance of a Probability Distribution

$$\sigma^2 = \Sigma\left[(x-\mu)^2 P(x)\right] \tag{33}$$

where:

- μ is the mean.
- $P(x)$ is the probability.
- x is a particular value.

Binomial Probability Distribution (with replacement)

$$P(x) = {}_nC_x \pi^x (1-\pi)^{n-x} \tag{34}$$

where:

- C denotes a combination.
- n is the number of trials.
- x is the number of successes.
- π is the probability of a success on each trial.

Mean of a Binomial Distribution (with replacement)

$$\mu = n\pi \tag{35}$$

where:

- μ is the probability mean.
- n is the number of trials.
- π is the probability of a success on each trial.

The formula is identical to (39).

Variance of a Binomial Distribution (with replacement)

$$\sigma^2 = n\pi(1-\pi) \tag{36}$$

where:

- n is the total number of trials.

- π is the probability of a success on each trial.

Hypergeometric Distribution (without replacement)

$$P(x) = \frac{({}_SC_x)(C_{n-x}^{N-S})}{{}_NC_n} \tag{37}$$

where:

- N is the size of the population.
- S is the number of successes in the population.
- x is the number of successes in the sample. It may be 0, 1, 2, 3, ...
- n is the size of the sample or the number of trials.
- C stands for a combination.

Poisson Distribution

$$P(x) = \frac{\mu^x e^{-\mu}}{x!} \tag{38}$$

where:

- μ is the mean number of occurrences (successes) in a particular interval.
- e is the constant 2.71828 (base of the Naperian logarithmic system).
- x is the number of occurrences (successes).
- $P(x)$ is the probability for a specified value of x.

Mean of a Poisson Distribution

$$\mu = n\pi \tag{39}$$

where:

- π is the probability of success.
- n is the number of trials.
- n is the total number of trials.

The formula is identical to (35).

Mean of the Uniform Distribution

$$\mu = \frac{a+b}{2} \tag{40}$$

where:

- μ is mean.

Standard Deviation of the Uniform Distribution

$$\sigma = \sqrt{\frac{(b-a)^2}{12}} \tag{41}$$

where:

- μ is the mean.
- a is minimum value of the interval.
- b is maximum value of the interval.

Uniform Distribution Probability

$$P(x) = (height)(base) = \frac{1}{b-a}(b-a) \tag{42}$$

where:

- if $a \leq x \geq b$ and 0 elsewhere.
- μ is the mean.
- a is minimum value of the interval.
- b is maximum value of the interval.

Normal Probability Distribution

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{(x-\mu)^2}{2\sigma^2}\right]} \tag{43}$$

where:

- σ refers to the standard deviation.
- μ refers to the mean.
- e is a constant, respectively, the base of the natural log system and approximately equals to 2.718.
- π a constant with an approximate value of $\frac{22}{7}$ or 3.1416.
- x refers to the value of the random variable.

Standard Normal Value (One Observation - σ Known)

$$z = \frac{x-\mu}{\sigma} \tag{44}$$

Notice an important difference to equation (49). Whereas equation (49) uses the sample mean $\bar{x}$, the population mean μ, and the sample standard deviation $\frac{\sigma}{\sqrt{n}}$, this formula is used in cases where the z value for **only one observation** is calculated.

where:

- z denotes the signed distance between a selected value x and the population mean μ divided by the population standard deviation σ.
- x is the value of any particular observation or measurement.
- μ is the mean of the distribution.
- σ is the standard deviation of the distribution.

Continuity Correction Factor

$$\begin{aligned}
&\text{If for } P(x) \geq x \text{ then use } (x-0.5)\\
&\text{If for } P(x) > x \text{ then use } (x+0.5)\\
&\text{If for } P(x) \leq x \text{ then use } (x+0.5)\\
&\text{If for } P(x) < x \text{ then use } (x-0.5)
\end{aligned} \tag{45}$$

The value 0.5 is subtracted or added to a selected value when a discrete probability distribution is approximated by a continuous probability distribution.

Exponential Distribution

$$P(x) = \lambda e^{-\lambda x} \tag{46}$$

where:

- λ refers to the rate parameter and $\lambda = \frac{1}{\mu}$ or $\mu = \frac{1}{\lambda}$
- e is a constant, respectively, the base of the natural log system and approximately equals to 2.718.
- x refers to the value of the random variable.
- Both the mean (μ) and the standard deviation (σ) are equal to $\frac{1}{\lambda}$.

Probability of Exponential Distribution

$$P(\text{arrival time} < x) = 1 - e^{-\lambda x} \tag{47}$$

where:

- λ refers to the rate parameter.
- e is a constant, respectively, the base of the natural log system and approximately equals to 2.718.
- x refers to the value of the random variable.
- Both the mean (μ) and the standard deviation (σ) are equal to $\frac{1}{\lambda}$.

Standard Error of the Mean

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{48}$$

where:

- $\sigma_{\bar{x}}$ refers standard deviation of the sample means indicated by $\bar{x}$.
- n refers to the sample size.
- σ refers to the population standard deviation.

Standard Normal Value (More than One Observation - σ known)

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{49}$$

Notice an important difference to equation (44). Whereas equation (44) uses the random variable x, the population mean μ, and the population standard deviation σ, this formula is used in cases where the research refers to a **sample rather than to just one observation**.

where:

- z refers to the distance between a selected value, designated $\bar{x}$, and the population mean μ divided by the standard error of the mean $\frac{\sigma}{\sqrt{n}}$ as in formula (48).
- $\bar{x}$ refers to the sample mean.
- μ refers to the population mean.
- σ refers to the population standard deviation.
- n refers to the sample size.

Confidence Interval for a Population Mean (σ Known)

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}} \tag{50}$$

where:

- z the standardized distance from the mean μ.
- $\bar{x}$ refers for the population standard deviation.
- σ refers to the population standard deviation.
- n refers to the sample size.

Confidence Interval for a Population Mean (σ Unknown)

$$\bar{x} \pm t\frac{s}{\sqrt{n}} \tag{51}$$

where:

- t refers to the t distribution.
- $\bar{x}$ refers for the population standard deviation.
- s refers to the sample standard deviation.
- n refers to the sample size.

Sample Proportion

$$p = \frac{x}{n} \tag{52}$$

where:

- p refers to the sample proportion.
- x refers for the number of successes.
- n refers to the sample size.

Confidence Interval for a Population Proportion

$$p \pm z\sqrt{\frac{p(1-p)}{n}} \tag{53}$$

where:

- p refers to the sample proportion which is an estimate for the population proportion π.
- z refers for the standard distance from the mean μ.
- n refers to the sample size.

Sample Size for Estimating the Population Mean

$$E = z\frac{\sigma}{\sqrt{n}} \quad \text{solved for n yields} \quad n = (\frac{z\sigma}{E})^2 \tag{54}$$

where:

- n refers to the sample size.
- z refers for the standard distance from the mean μ.
- σ refers to the population standard deviation.
- E is the maximum allowable error.

Sample Size for the Population Proportion

$$E = z\sqrt{\frac{\pi(1-\pi)}{n}} \quad \text{solved for n yields} \quad n = \pi(1-\pi)(\frac{z}{E})^2 \tag{55}$$

where:

- n is the size of the sample.
- z is the standard normal value corresponding to the desired level of confidence.
- π is the population proportion.
- E is the maximum allowable error.

Finite-Population Correction Factor

$$FPC = \sqrt{\frac{N-n}{N-1}} \tag{56}$$

To be used if the sample is a significant part of a finite population. where:

- n is the size of the sample.
- N is the size of the population.

Testing a Mean (σ Known)

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{57}$$

Refer to equation (49) that computes a z value on the basis of a sample.
The z value is based on the sampling distribution of the sample mean $\bar{x}$, which follows the normal distribution with the sampling mean $\mu_{\bar{x}}$ equal to the population mean μ and a standard error of the mean $\sigma_{\bar{x}}$, with is equal to $\frac{\sigma}{\sqrt{n}}$.
where:

- z refers to the distance between a selected value, designated $\bar{x}$, and the mean μ divided by the standard error of the mean $\frac{\sigma}{\sqrt{n}}$ as in formula (48).
- $\bar{x}$ refers for the sample mean.
- μ refers to the population mean.
- σ refers to the population standard deviation.
- n refers to the sample size.

Testing a Mean (σ Unknown)

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{58}$$

with $n - 1$ degrees of freedom where:

- $\bar{x}$ is the sample mean.
- μ is the hypothesized population mean.
- s is the sample standard deviation.
- n is the number of observations in the sample.

Type II Error

$$z = \frac{\bar{x}_C - \mu_1}{\frac{\sigma}{\sqrt{n}}} \tag{59}$$

Refer to equation (49) that computes a z value on the basis of a sample.
where:

- $\bar{x}_C$ is the sample mean of region C.
- μ_1 is the hypothesized population mean of region C.
- σ is the population standard deviation.
- n is the number of observations in the sample.

Two-Sample Test – Variance of the Distribution of Differences (σ Known)

$$\sigma^2_{\bar{x_1} - \bar{x_2}} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \tag{60}$$

where:

- $\sigma^2_{\bar{x_1} - \bar{x_2}}$ is variance of the differences in means.
- $\bar{x}_1$ and $\bar{x}_2$ are the sample means of the first and second sample, respectively.
- n_1 and n_2 are the sample sizes.

Two-Sample Test – Test of Means (σ Known)

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{61}$$

Note: above formula refers to formula (62). Using it stipulates the following[2]:

1. the sampled populations are approximately normally distributed.
2. the sampled populations are independent.
3. the standard deviations of the two populations are **known**.

where:

- z refers to the standard value.
- $\bar{x}$ refers for the sample mean.
- σ^2 refers to the population variance.
- n refers to the sample size.

Two-Sample Test – Pooled Variance (σ Unknown)

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{62}$$

Note: using the pooled variance formula assumes following important assumptions for the test.

1. the sampled populations are approximately normally distributed.
2. the sampled populations are independent.
3. the standard deviations of the two populations are **equal**.

In essence, the formula computes a weighted mean of the two sample standard deviations using the degrees of freedom that each sample provides. The resulting value serves then as an estimate for the unknown population standard deviation. The reason for pooling arises from the assumption that the two populations have *equal standard deviations* and best estimate we can make of that value is to combine or pool all the sample information we have about the value of the population standard deviation. In contrast, formula (67) assumes related or *paired* samples. If such an assumption is reasonable the resulting hypothesis test is much more sensitive to detecting a significant difference than a hypothesis test based on independent samples compared to independent samples since we are able to *reduce the variation* in the sampling distribution[3].
where:

- s_p^2 is pooled variance. Further used in equation (63) due to the assumption in that the two populations sampled have the same standard deviations.
- s_1^2 is the variance of the first sample.
- s_2^2 is the variance of the second sample.
- n_1 is the sample size of the first sample.
- n_2 is the sample size of the second sample.
- $n_1 + n_2 - 2$ is the degree of freedom usually denoted as *df*.

Two-Sample Test – Test of Means (σ Unknown)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} \tag{63}$$

Note: the pooled variance s_p^2 is to be computed using equation (62) beforehand. Therefore the same prerequisite applies here:

1. the sampled populations are approximately normally distributed.
2. the sampled populations are independent.

[2] Lind et al., p. 351

[3] Lind et al., p. 368

3. the standard deviations of the two populations are **equal**.

Note: the **Wilcoxon rank-sum test** (ref. formula (102)) is an alternative to the Two-Sample t test. While the Two-Sample t test requires two populations follow the normal distribution and have equal population variances, these prerequisites **do not** apply for the Wilcoxon rank-sum test.

where:

- $\bar{x}_1$ is the mean of the first sample.
- $\bar{x}_2$ is the mean of the second sample.
- n_1 is the number of observations in the first sample.
- n_2 is the number of observations in the second sample.
- s_p^2 is the pooled estimate of the population variance.

Two-Sample Test – Test Statistic for No Difference in Means, Unequal Variances (σ Unknown)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{64}$$

Note: this equation differs only in one aspect from equation (63) in that the pooled variance s_p^2 could be used due to the assumption that both sample variances are equal. Since this formula assumes that both sample variances are **unequal** the denominator now splits into two components using s_1^2 and s_2^2 as variances.

where:

- $\bar{x}_1$ is the mean of the first sample.
- $\bar{x}_2$ is the mean of the second sample.
- s_1^2 is sample variance of the first sample.
- s_2^2 is sample variance of the second sample.
- n_1 is the number of observations in the first sample.
- n_2 is the number of observations in the second sample.

Two-Sample Test – Degrees of Freedom for Unequal Variance Test (σ Unknown)

$$df = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2 - 1}} \tag{65}$$

where:

- df is the degree of freedom.
- s_1^2 is sample variance of the first sample.
- s_2^2 is sample variance of the second sample.
- n_1 is the number of observations in the first sample.
- n_2 is the number of observations in the second sample.

Two-Sample Test – Standard Deviation of Differences

$$s_d = \sqrt{\frac{\sum(d - \bar{d})^2}{n - 1}} \tag{66}$$

where:

- $\bar{d}$ is the mean of the difference between the paired or related observations.
- s_d is the standard deviation of the differences between the paired or related observations.
- n is the number of paired observations.

Two-Sample Test – Paired t Test

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} \tag{67}$$

Note: whereas formula (62) assumes independent samples and thus incurs a much greater variation through the process of pooling, tests from paired samples allow to greatly reduce the variation from the sampling distribution. Also note the possible disadvantage that the degrees of freedom for paired samples are usually lower than their independent counterparts[4].

Note: An alternative to test with *dependent samples* is the **Wilcoxon signed-rank test**. For this test the normality assumption is not required.

There are $n - 1$ degrees of freedom and

- $\bar{d}$ is the mean of the difference between the paired or related observations.
- s_d is the standard deviation of the differences between the paired of related observations.
- n is the number of paired observations.

ANOVA – Test Statistic for Comparing Two Variances

$$F = \frac{s_1^2}{s_2^2} \tag{68}$$

where:

- s_1^2 is the sample variance of the first sample.
- s_1^2 is the sample variance of the second sample.
- F represents the F distribution.

If the null hypothesis is true, the test statistic follows the F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

ANOVA – Critical Value for F Statistic

$$F = \frac{k - 1}{n - k} \tag{69}$$

where:

- F represents the F distribution.
- k is the number of treatments.
- n is the total number of observations.

In ANOVA the critical value of the F statistic decides whether the null hypothesis H_0 can be rejected.

ANOVA – Components

The sum of the squared differences between (a) and (b) defines (c):

(70)

and		
(a)	(b)	(c) ANOVA term
each observation	overall mean	**total variation**
each treatment mean	overall mean	**treatment variation**
each observation	treatment mean	**random variation**

Note: to use the ANOVA test the following assumptions need to be met:

[4] Lind et al., p. 369

- The samples are from independent populations.
- The population variances must be equal.
- The samples are from normal populations.

If these assumptions are not reasonable the **Kruskal-Wallis** nonparametric test as in formula (103) is advisable.

ANOVA: One-Way – Table

$$\text{ANOVA Table (One-Way)} \tag{71}$$

where:

ANOVA				
source of variation	SS	df	mean square	F
treatments	SST	$k-1$	$\frac{SST}{(k-1)} = MST$	$\frac{MST}{MSE}$
error	SSE	$n-k$	$\frac{SSE}{(n-k)} = MSE$	
total	SS total	$n-1$		

- n denotes the total number of observations.
- k denotes the number of treatments.
- SS denotes the sum of squares.
- SST denotes the sum of squares due to treatments. It is the sum of the squared differences each treatment mean $\bar{x}_C$ and the overall mean $\bar{x}_G$. It can also be calculated as the difference of SS total $- SSE$.
- SSE denotes the sum of squares due to errors (random error). It is calculated by $\text{SSE} = \sum (x - \bar{x}_C)^2$ with $\bar{x}_C$ being the sample mean for treatment C.
- SS total denotes the total variation. It is $\text{SS total} = \sum (x - \bar{x}_G)^2$ with x being each sample observation and $\bar{x}_G$ being the overall mean.
- MST denotes mean square for treatments. Mean square is another term for an estimate of variance.
- MSE denotes mean square for errors.

ANOVA: Confidence Interval for the Difference in Treatment Means

$$(\bar{x}_1 - \bar{x}_2) \pm t\sqrt{\text{MSE}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{72}$$

where:

- $\bar{x}_1$ is the mean of the first sample.
- $\bar{x}_2$ is the mean of the second sample.
- t refers to the t distribution with degrees of freedom equal to $n - k$.
- MSE is the mean square error term obtained from the ANOVA table $\left[\frac{SSE}{n-k}\right]$.
- n_1 is the number of observations in the first sample.
- n_2 is the number of observations in the second sample.

ANOVA: Two-Way – Sum of Squares Blocks

$$\text{SSB} = k \sum (\bar{x}_b - \bar{x}_G)^2 \tag{73}$$

where:

- k is the number of treatments.
- b is the number of blocks.
- $\bar{x}_b$ refers to the sample mean of block b.
- $\bar{x}_G$ is the overall mean.

ANOVA: Two-Way – Table

ANOVA Table (Two-Way) (74)

ANOVA				
source of variation	SS	df	mean square	F
treatments	SST	$k-1$	$\frac{SST}{(k-1)} = MST$	$\frac{MST}{MSE}$
blocks	SSB	$b-1$	$\frac{SSB}{(b-1)} = MSB$	$\frac{MSB}{MSE}$
error	SSE	$(k-1)(b-1)$	$\frac{SSE}{(k-1)(b-1)} = MSE$	
total	SS total	$n-1$		

where:

- n denotes the total number of observations.
- k denotes the number of treatments.
- SS denotes the sum of squares.
- SST denotes the sum of squares due to treatments. It is the sum of the squared differences of each treatment mean $\bar{x}_C$ and the overall mean $\bar{x}_G$. It can also be calculated as the difference of SS total $- SSE - SSB$.
- SSB denotes the sum of squares blocks. It is calculated by $SSB = k\sum(\bar{x}_b - \bar{x}_G)^2$ with $\bar{x}_b$ being the number of blocks and $\bar{x}_G$ being the overall mean.
- SSE denotes the sum of squares due to errors (random error). It is calculated by SSE $= \sum(x - \bar{x}_C)^2$ with $\bar{x}_C$ being the sample mean for treatment C.
- SS total denotes the total variation. It is SS total $= \sum(x - \bar{x}_G)^2$ with x being each sample observation and $\bar{x}_G$ being the overall mean.
- MST denotes mean square for treatments. Mean square is another term for an estimate of variance.
- MSB denotes mean square blocks.
- MSE denotes mean square for errors.

The following is an illustration how the different variation relate.

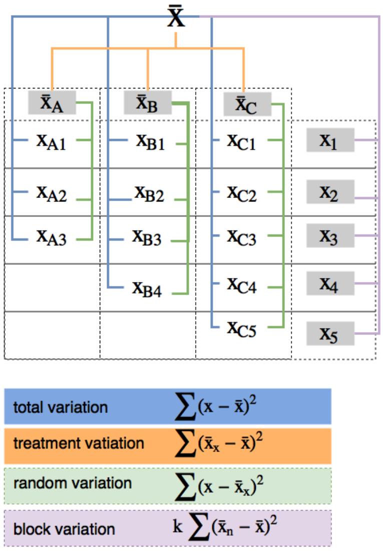
x̄
x̄A
x̄B
x̄C
xA1
xA2
xA3
xB1
xB2
xB3
xB4
xC1
xC2
xC3
xC4
xC5
x1
x2
x3
x4
x5
total variation ∑(x – x̄)²
treatment vatiation ∑(x̄x – x̄)²
random variation ∑(x – x̄x)²
block variation k ∑(x̄n – x̄)²

Figure 1: *Illustration of a Two-Way ANOVA*

ANOVA: Two-Way with Interaction (Factors) – Table

ANOVA Table with Interaction (Factors) (75)

ANOVA				
source of variation	SS	df	mean square	F
Factor A	SSA	$k-1$	$\frac{SSA}{(k-1)} = MSA$	$\frac{MSA}{MSE}$
Factor B	SSB	$b-1$	$\frac{SSB}{(b-1)} = MSB$	$\frac{MSB}{MSE}$
Interaction	SSI	$(k-1)(b-1)$	$\frac{SSI}{[(k-1)(b-1)]} = MSI$	$\frac{MSI}{MSE}$
Error	SSE	$n-kb$	$\frac{SSE}{(n-kb)} = MSE$	
total	SS total	$n-1$		

Linear Regression: Correlation Coefficient

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{(n-1)s_x s_y} \tag{76}$$

Subject: A measure of strength of the linear relationship between two variables. It ranges from -1 up to and including $+1$.

Note: The correlation coefficient becomes independent of the scale used if the term $\sum(x-\bar{x})(y-\bar{y})$ is divided by the sample standard deviations s_x and s_y. Similarly, the term becomes independent of sample size once divided by $(n-1)$.

Note: Correlation between two independent variables is considered uncritical if $-0.70 < r < 0.70$. A more precise test provides the variance inflation factor (VIF) as in formula (93).

where:

- r denotes the correlation coefficient.
- x denotes the variable value of the x population.
- y denotes the variable value of the y population.
- $\bar{x}$ denotes the mean of variable values in the x population.
- $\bar{y}$ denotes the mean of variable values in the y population.
- n denotes the number of observations in the sample.
- $(n-1)$ denotes the degree of freedom.
- s_x denotes the standard deviation of the x population.
- s_y denotes the standard deviation of the y population.

Linear Regression: t Test for the Correlation Coefficient r

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{77}$$

Subject: Resolves the question about whether there could be zero correlation in the population from which the sample was selected.

with $n-2$ degrees of freedom where:

- r denotes the correlation coefficient.
- n denotes the number of observations in the sample.

Linear Regression Equation: General Form

$$\hat{y} = a + bx \tag{78}$$

Subject: An equation that expresses the linear relationship between two variables.

where:

- $\hat{y}$ is the estimated value of the y variable for a selected x value.
- a is the y-intercept. It is the estimated value of Y when $x = 0$.
- b is the slope of the line, or the average change in $\hat{y}$ for each change of one unit (either increase or decrease) in the independent variable x.
- x is any value of the independent variable that is selected.

Linear Regression: Slope of Regression Line

$$b = r(\frac{s_y}{s_x}) \tag{79}$$

where:

- r denotes the correlation coefficient.
- s_y denotes the standard deviation of y (the dependent variable).
- s_x denotes the standard deviation of x (the independent variable).

Linear Regression: Y-Intercept

$$a = \bar{y} - b\bar{x} \tag{80}$$

where:

- $\bar{y}$ is the mean of y (the dependent variable).
- $\bar{x}$ is the mean of x (the independent variable).

Linear Regression: t Test for the Slope b

$$t = \frac{b-0}{s_b} \tag{81}$$

Subject: Conducts a test on whether the slope of the regression line is different from zero. In such a circumstance we can reasonably conclude that the regression line adds to the predictive ability of the regression equation.

with $n-2$ degrees of freedom where:

- b is the estimate of the regression line's slope calculated from the sample information.
- s_b is the standard error of the slope estimate, also determined from sample information.

Linear Regression: Standard Error of Estimate

$$s_{y\cdot x} = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}} = \sqrt{\frac{\text{SSE}}{n-2}} \tag{82}$$

Subject: It is a relative measure of a regression equation's ability to predict.

where:

- $s_{y\cdot x}$ denotes the standard error of estimate with $y\cdot x$ to be interpreted as the standard error of y for a given value of x. It is the same concept as the standard deviation in formula (13) which measures the dispersion around a mean.
- y denotes the observed value.
- $\hat{y}$ denotes the predicted value.
- $\sum(y-\hat{y})^2$ denotes the sum of squares error or residuals referred to in the ANOVA equation (71) as SSE.

Linear Regression: Coefficient of Determination

$$r^2 = \frac{\text{SSR}}{\text{SS Total}} = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2} = 1 - \frac{\text{SSE}}{\text{SS Total}} = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} \tag{83}$$

Subject: The proportion of the total variation in the dependent variable Y that is explained, or accounted for, by the variation in the independent variable X.

where:

- *SS Total* denotes total variation, that is, the sum of squares total.
- *SSR* denotes the sum of squares regression.
- *SSE* denotes the sum of squares errors or residuals, respectively.

Linear Regression: Confidence Interval for the Mean of Y given X

$$\hat{y} \pm t s_{y\cdot x}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x-\bar{x})^2}} \tag{84}$$

Subject: Applied when the regression equation is used to predict the *mean value of y* for a given value of x

where:

- x denotes the given value.
- $\bar{x}$ denotes the sample mean.
- $\hat{y}$ denotes the predicted value.
- $s_{y\cdot x}$ denotes the standard error of estimate with $y\cdot x$ to be interpreted as the standard error of y for a given value of x. It is the same concept as the standard deviation in formula (13) which measures the dispersion around a mean.

Linear Regression: Prediction Interval for Y given X

$$\hat{y} \pm t s_{y\cdot x}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x-\bar{x})^2}} \tag{85}$$

Subject: Applied when the regression equation is used to predict an *individual value of y* ($n = 1$) for a given value of x. Refer to equation (84).

where:

- x denotes the given value.
- $\bar{x}$ denotes the sample mean.
- $\hat{y}$ denotes the predicted value.
- $s_{y\cdot x}$ denotes the standard error of estimate with $y\cdot x$ to be interpreted as the standard error of y for a given value of x. It is the same concept as the standard deviation in formula (13) which measures the dispersion around a mean.

Multiple Regression: General Equation

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 + ...b_kx_k \tag{86}$$

Subject: Enhanced equation of formula (78) for more than one dependent variable.

where:

- a is the intercept, the value of $\hat{y}$ when all the x's are zero.
- b_j is the amount by which $\hat{y}$ changes when that particular x_j increases by one unit, with the values of all other independent variables held constant.
- k represents the number of independent variables.

Multiple Regression: Standard Error of Estimate

$$s_{y\cdot 123\cdot k} = \sqrt{\frac{\sum(y-\hat{y})^2}{n-(k+1)}} = \sqrt{\frac{\text{SSE}}{n-(k+1)}} = \sqrt{\text{M}SE} \tag{87}$$

Subject: It is a relative measure of a regression equation's ability to predict for more than one independent variable.

where:

- y is the actual observation.
- $\hat{y}$ is the estimated value computed from the regression equation.
- n is the number of observations in the sample.
- k is the number of independent variables.
- SSE is the residual sum of squares from an ANOVA table. It is equal to the term $\sum(y-\hat{y})^2$ as used also in the ANOVA formula (71).

Multiple Regression ANOVA – Table

$$\text{Multiple Regression ANOVA Table} \tag{88}$$

ANOVA				
source	SS	df	MS	F
Regression	SSR	k	$MSR = \frac{SSR}{k}$	$\frac{MSR}{MSE}$
Residual or error	SSE	$n-(k+1)$	$MSE = \frac{SSE}{n-(k+1)}$	
total	SS total	$n-1$		

Multiple Regression: Coefficient of Multiple Determination

$$R^2 = \frac{\text{SSR}}{\text{SS Total}} \tag{89}$$

Subject: The percent of variation in the dependent variable, y, explained by the set of independent variables, x_1, x_2, x_3, ... x_k.

where:

- *SS Total* denotes total variation, that is, the sum of squares total.
- *SSR* denotes the sum of squares regression.

Multiple Regression: Adjusted Coefficient of Multiple Determination

$$R^2_{adj} = \frac{\frac{SSE}{n-(k+1)}}{\frac{SStotal}{n-1}} \tag{90}$$

Subject: The percent of variation in the dependent variable, y, explained by the set of independent variables, x_1, x_2, x_3, ... x_k. As more independent variables are added to the multiple regression model R^2 of formula (89) tends to increase. In fact, if the number of variables, k, and the sample size, n, are equal, the coefficient of determination is 1. To avoid this trend R^2 is adjusted.

where:

- *SS Total* denotes total variation, that is, the sum of squares total.
- *SSE* denotes the sum of squares error or residual.

Multiple Regression: Global Test

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{n-(k+1)}} \tag{91}$$

Subject: The Global Test investigates whether it is possible all the independent variables have zero regression coefficients. As in formula (69) it expresses this as sum of squares regression per unit of sum of squares residuals. The higher the explained variances compared to the residual variances, the more positive the value of the F distribution.

where:

- *SSR* denotes the sum of squares regression.

- SSE denotes the sum of squares error or residual.
- n is the number of observations in the sample.
- k is the number of independent variables.

Multiple Regression: t Test Individual Coefficients b

$$t = \frac{b_j - 0}{s_{b_j}} \tag{92}$$

SUBJECT: Conducts a test on the independent variables individually to determine whether the regression coefficients differ from zero. If a regression coefficient is likely to be zero it does not contribute to the regression equation's ability to predict.

where:

- b_j refers to any one of the regression coefficients.
- s_b is the standard error of the slope estimate, also determined from sample information.

Multiple Regression: Variance Inflation Factor

$$VIF = \frac{1}{1 - R_j^2} \tag{93}$$

SUBJECT: A VIF greater than 10 is considered unsatisfactory, indicating that the independent variable should be removed from the analysis.

where:

- R_j^2 refers to the coefficient of determination

Test of Hypothesis: One Proportion

$$z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \tag{94}$$

where:

- π is the population proportion.
- p is the sample proportion.
- n is the sample size.

Test of Hypothesis: Pooled Proportion

$$p_c = \frac{x_1 + x_2}{n_1 + n_2} \tag{95}$$

where:

- p_c is the pooled proportion possessing the trait in the combined samples. It is called the pooled estimate of the population proportion.
- x_1 is the number possessing the trait in the first sample.
- x_2 is the number possessing the trait in the second sample.
- n_1 is the number of observations in the first sample.
- n_2 is the number of observations in the second sample.

Test of Hypothesis: Two-Sample Test of Proportions

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_c(1-p_c)}{n_1} + \frac{p_c(1-p_c)}{n_2}}} \tag{96}$$

where:

- n_1 is the number of observations in the first sample.
- n_2 is the number of observations in the second sample.
- p_1 is the proportion in the first sample possessing the trait.
- p_2 is the proportion in the second sample possessing the trait.
- p_c is the pooled proportion possessing the trait in the combined samples. It is called the pooled estimate of the population proportion.

Chi-Square Test Statistic

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right] \tag{97}$$

with $k - 1$ degrees of freedom, where:

- k is the number of categories.
- f_o is an observed frequency in a particular category.
- f_e is an expected frequency in a particular category.

Expected Frequency

$$f_e = \frac{\text{(row total) (column total)}}{\text{(grand total)}} \tag{98}$$

where:

- f_e is an expected frequency in a particular category.

Sign Test: n > 10

$$z = \frac{(x \pm .50) - \mu}{\sigma} \tag{99}$$

where:

- z refers to the standard value.
- ± 0.50 is the *continuity correction factor* as in formula (45).
- x denotes the number of plus $(+)$ or minus $(-)$ signs.
- μ denotes the population mean.
- σ denotes the population standard deviation.

Sign Test: n > 10, + Signs *more* than $n/2$

$$z = \frac{(x - .50) - \mu}{\sigma} = \frac{(x - .50) - .50n}{.50\sqrt{n}} \tag{100}$$

where:

- z refers to the standard value.
- ± 0.50 is the *continuity correction factor* as in formula (45).
- x denotes the number of plus $(+)$ or minus $(-)$ signs.
- n denotes the sample size.
- μ denotes the population mean.
- σ denotes the population standard deviation.

Sign Test: n > 10, + Signs less than $n/2$

$$z = \frac{(x + .50) - \mu}{\sigma} = \frac{(x + .50) - .50n}{.50\sqrt{n}} \tag{101}$$

where:

- z refers to the standard value.
- ± 0.50 is the *continuity correction factor* as in formula (45).
- x denotes the number of plus (+) or minus (−) signs.
- n denotes the sample size.
- μ denotes the population mean.
- σ denotes the population standard deviation.

Wilcoxon Rank-Sum Test

$$z = \frac{W - \frac{n_1(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}} \qquad (102)$$

SUBJECT: this test is specifically designed to determine whether two *independent samples* came from equivalent populations.

Note: this test is an alternative to the Two-Sample t test as in formula (63), however it does **not** require that the two populations follow the normal distribution and have equal population variances.

where:

- n_1 is the number of observations of the first population.
- n_2 is the number of observations of the second population.
- W is the sum of the ranks from the first population.

Kruskal-Wallis Test

$$H = \frac{12}{n(n+1)}\left[\frac{(\sum R_1)^2}{n_1} + \frac{(\sum R_2)^2}{n_2} + ... + \frac{(\sum R_k)^2}{n_k}\right] - 3(n+1) \qquad (103)$$

Note: for the Kruskal-Wallis test to be applied, the samples selected from the populations must be independent. If in addition the following prerequisites are met an ANOVA analysis as in formula (70) can be applied instead:

- The samples are from independent populations.
- The population variances must be equal.
- The samples are from normal populations.

with $k-1$ degrees of freedom (k is the number of populations), where:

- $\sum R_1, \sum R_2, ..., \sum R_k$ are the sums of the ranks of samples 1, 2, ... k, respectively.
- $n_1, n_2, ..., n_k$ are the sizes of samples 1, 2, ..., k, respectively.
- n is the combined number of observations for all samples.

Spearman's Coefficient of Rank Correlation

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} \qquad (104)$$

where:

- r_s is Spearman's coefficient of rank correlation.
- d is the difference between the ranks for each pair.
- n is the number of paired observations.

Hypothesis Test: Rank Correlation

$$t = r_s\sqrt{\frac{n-2}{1-r_s^2}} \qquad (105)$$

where:

- r_s is Spearman's coefficient of rank correlation.

Index Numbers: Simple Index

$$P = \frac{p_t}{p_0} * 100 \qquad (106)$$

where:

- p_0 is the base-period price.
- p_t is the given period price.

Index Numbers: Simple Average of the Price Relatives

$$P = \frac{\sum P_i}{n} \qquad (107)$$

Note:

- Advantage: simple price indices are not dependent on the unit of measure of the item quantified.
- Disadvantage: simple price indices do not the account for the relative importance of the items included.

The aforementioned shortcomings of not accounting for the relative importance, i.e. the marginal quantities consumed in each item category, are mitigated by the Laspeyres price index as in formula (109) and the Paasche price index as in formula (110).

where:

- P_i refers to the simple index for each to the items.
- n refers to the number of items.

Index Numbers: Simple Aggregate Index

$$P = \frac{\sum p_t}{\sum p_0} * 100 \qquad (108)$$

Note: since the aggregate index is influenced by the unit of measure, it is not used frequently.

where:

- p_0 is the base-period price.
- p_t is the given period price.

Index Numbers: Laspeyres Price Index

$$P = \frac{\sum p_t q_0}{\sum p_0 q_0} * 100 \qquad (109)$$

Note: the Laspeyres price index thus assumes that the base period quantities have still important bearing on the current price index and thus are realistic. Hence it contrasts the assumption of the Paasche price index as in formula (110).

where:

- P is the price index.
- p_t is the current price.
- p_0 is the price in the base period.
- q_0 is the quantity used in the base period.

Index Numbers: Paasche Price Index

$$P = \frac{\sum p_t q_t}{\sum p_0 q_t} * 100 \qquad (110)$$

Note: the Paasche price index assumes current period quantity levels as base to account for changed preferences in consumed quantities. Hence, it contrasts the assumption of the Laspeyres price index as in formula (109).

where:

- P is the price index.
- p_t is the current price.
- p_0 is the price in the base period.
- q_0 is the quantity used in the base period.

Index Numbers: Fisher's Ideal Index

$$\text{Fisher's Ideal Index} = \sqrt{(\text{Laspeyres index})(\text{Paasche index})} \tag{111}$$

Note: Fisher's Ideal Index is actually a geometric mean (7) of Laspeyres (109) and Paasche (110) price indices.

Index Numbers: Value Index

$$V = \frac{\sum p_t q_t}{\sum p_0 q_0} * 100 \tag{112}$$

where:

- P is the price index.
- p_t is the current price.
- p_0 is the price in the base period.
- q_0 is the quantity used in the base period.

Index Numbers: Real Income

$$\text{Real income} = \frac{\text{Money income}}{\text{CPI}} * 100 \tag{113}$$

where:

- CPI denotes consumer price index.

Index Numbers: Index as a Deflator

$$\text{Deflated sales} = \frac{\text{Actual sales}}{\text{An appropriate index}} * 100 \tag{114}$$

Index Numbers: Index for Purchasing Power

$$\text{Purchasing power of dollar} = \frac{\$1}{\text{CPI}} * 100 \tag{115}$$

Time Series & Forecasting: Linear Trend Equation

$$\hat{y} = a + bt \tag{116}$$

where:

- $\hat{y}$ is the projected value of the y variable for a selected value of t.
- a is the y-intercept. It is the estimated value of y when $t = 0$.
- b is the slope of the line, or the average change in $\hat{y}$ for each increase of one unit in t.
- t is any value of time that is selected.

Time Series & Forecasting: Log Trend Equation

$$\log \hat{y} = \log a + \log b(t) \tag{117}$$

Time Series & Forecasting: Correction Factor for Adjusting Quarterly Means

$$\text{Correction factor} = \frac{4.00}{\text{Total of four means}} \tag{118}$$

Time Series & Forecasting: Durbin-Watson Statistic

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n}(e_t)^2} \tag{119}$$