(https://stanford.edu/~shervine/teaching/cme-106/cheatsheet-statistics#cme-106---introduction-to-probability-and-statistics-for-engineers)CME 106 - Introduction to Probability and Statistics for Engineers (teaching/cme-106)

English Français (l/fr/teaching/cme-106/pense-bete-statistiques)

Probability	**Statistics**

(https://stanford.edu/~shervine/teaching/cme-106/cheatsheet-statistics#cheatsheet)Statistics cheatsheet

Star 102

By Afshine Amidi (https://twitter.com/afshinea) and Shervine Amidi (https://twitter.com/shervinea)

(https://stanford.edu/~shervine/teaching/cme-106/cheatsheet-statistics#parameter-estimation) Parameter estimation

Definitions

Random sample — A random sample is a collection of n random variables $X_1, ..., X_n$ that are independent and identically distributed with X.

Estimator — An estimator is a function of the data that is used to infer the value of an unknown parameter in a statistical model.

Bias — The bias of an estimator $\hat{\theta}$ is defined as being the difference between the expected value of the distribution of $\hat{\theta}$ and the true value, i.e.:

$$\boxed{\mathrm{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta}$$

Remark: an estimator is said to be unbiased when we have $E[\hat{\theta}] = \theta$.

Estimating the mean

Sample mean — The sample mean of a random sample is used to estimate the true mean μ of a distribution, is often noted $\overline{X}$ and is defined as follows:

$$\boxed{\overline{X}=\frac{1}{n}\sum_{i=1}^n X_i}$$

Remark: the sample mean is unbiased, i.e $E[\overline{X}]=\mu$.

Characteristic function for sample mean — The characteristic function for a sample mean is noted $\psi_{\overline{X}}$ and is such that:

$$\boxed{\psi_{\overline{X}}(\omega)=\psi_X^n\left(\frac{\omega}{n}\right)}$$

Central Limit Theorem — Let us have a random sample $X_1,...,X_n$ following a given distribution with mean μ and variance σ^2, then we have:

$$\boxed{\overline{X}\underset{n\to+\infty}{\sim}\mathcal{N}\left(\mu,\frac{\sigma}{\sqrt{n}}\right)}$$

Estimating the variance

Sample variance — The sample variance of a random sample is used to estimate the true variance σ^2 of a distribution, is often noted s^2 or $\hat{\sigma}^2$ and is defined as follows:

$$\boxed{s^2=\hat{\sigma}^2=\frac{1}{n-1}\sum_{i=1}^n(X_i-\overline{X})^2}$$

Remark: the sample variance is unbiased, i.e $E[s^2]=\sigma^2$.

Chi-Squared relation with sample variance — Let s^2 be the sample variance of a random sample. We have:

$$\boxed{\frac{s^2(n-1)}{\sigma^2}\sim\chi_{n-1}^2}$$

Confidence intervals

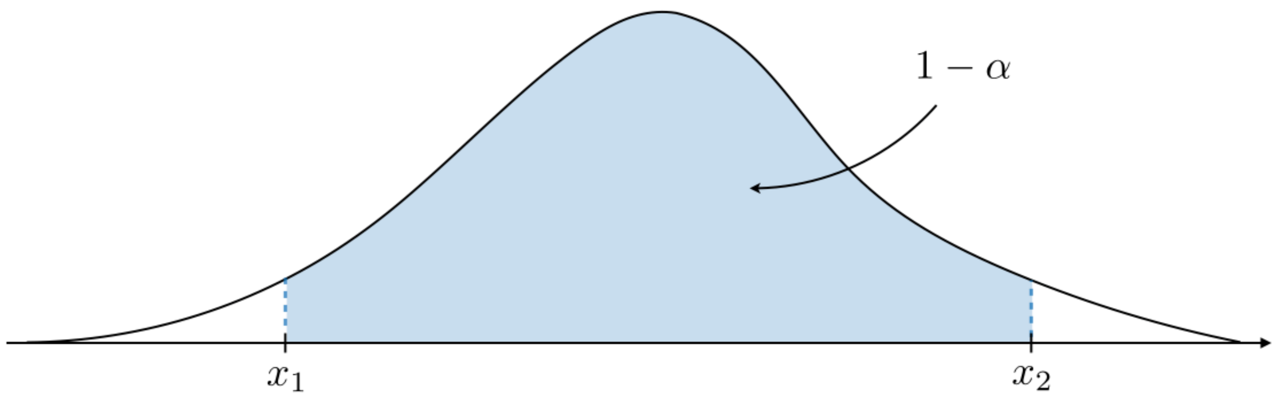
(https://stanford.edu/~shervine/teaching/cme-106/cheatsheet-statistics#confidence-intervals)

Definitions

Confidence level — A confidence interval with confidence level $1-\alpha$ is such that $1-\alpha$ of the time, the true value is contained in the confidence interval.

Confidence interval — A confidence interval $CI_{1-\alpha}$ with confidence level $1-\alpha$ of a true parameter θ is such that:

$$\boxed{P(\theta \in CI_{1-\alpha}) = 1-\alpha}$$



With the notation of the example above, a possible $1-\alpha$ confidence interval for θ is given by $CI_{1-\alpha} = [x_1, x_2]$.

Confidence interval for the mean

When determining a confidence interval for the mean μ, different test statistics have to be computed depending on which case we are in. The table below sums it up.

Distribution of X_i	**Sample size n**	**Variance σ^2**	**Statistic**	$1-\alpha$ **confidence interval**
$X_i \sim \mathcal{N}(\mu, \sigma)$	any	known	$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$	$\left[\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$
$X_i \sim$ any distribution	large	known	$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$	$\left[\overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$
$X_i \sim$ any distribution	large	unknown	$\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}} \sim \mathcal{N}(0,1)$	$\left[\overline{X} - z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right]$

$X_i \sim \mathcal{N}(\mu, \sigma)$	small	unknown	$\dfrac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$	$\left[\overline{X} - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \overline{X} + t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right]$
$X_i \sim$ any distribution	small	known or unknown	Go home!	Go home!

<u>Note</u>: a step by step guide to estimate the mean, in the case when the variance in known, is detailed here (teaching/cme-106/key-concepts#ci-mean).

Confidence interval for the variance

The single-line table below sums up the test statistic to compute when determining the confidence interval for the variance.

Distribution of X_i	**Sample size n**	**Mean μ**	**Statistic**	**$1-\alpha$ confidence interval**
$X_i \sim \mathcal{N}(\mu, \sigma)$	any	known or unknown	$\dfrac{s^2(n-1)}{\sigma^2} \sim \chi^2_{n-1}$	$\left[\frac{s^2(n-1)}{\chi_2^2}, \frac{s^2(n-1)}{\chi_1^2}\right]$

<u>Note</u>: a step by step guide to estimate the variance is detailed here (teaching/cme-106/key-concepts#ci-var).

(https://stanford.edu/~shervine/teaching/cme-106/cheatsheet-statistics#hypothesis-testing) Hypothesis testing

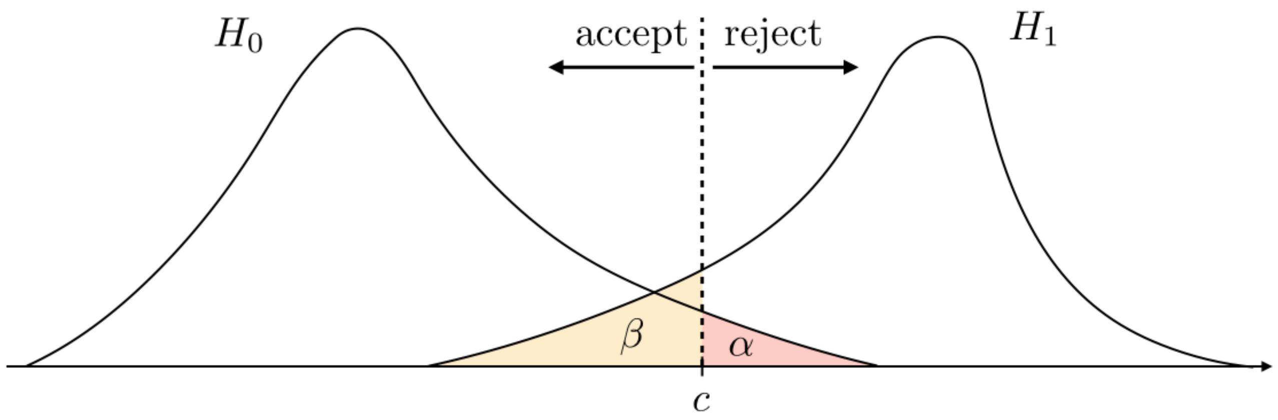
General definitions

Type I error — In a hypothesis test, the type I error, often noted α and also called "false alarm" or significance level, is the probability of rejecting the null hypothesis while the null hypothesis is true. If we note T the test statistic and R the rejection region, then we have:

$$\boxed{\alpha = P(T \in R | H_0 \text{ true})}$$

Type II error — In a hypothesis test, the type II error, often noted β and also called "missed alarm", is the probability of not rejecting the null hypothesis while the null hypothesis is not true. If we note T the test statistic and R the rejection region, then we have:

$$\boxed{\beta = P(T \notin R | H_0 \text{ not true})}$$

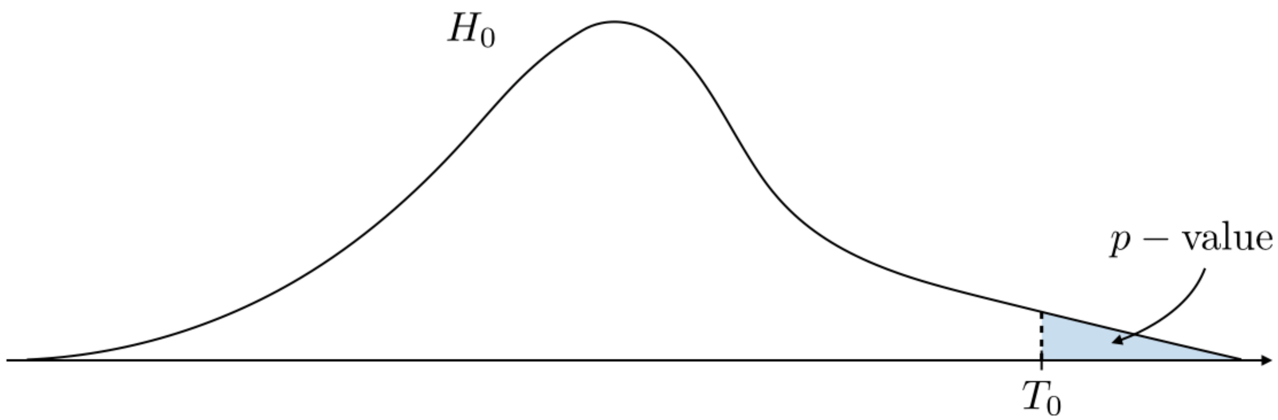


p-value — In a hypothesis test, the p-value is the probability under the null hypothesis of having a test statistic T at least as extreme as the one that we observed T_0. We have:

(left-sided) $\boxed{p\text{-value} = P(T \leqslant T_0 | H_0 \text{ true})}$ (right-sided) $p\text{-value} = P(T \geqslant T_0 | H_0 \text{ tru}$

(two-sided) $\boxed{p\text{-value} = P(|T| \geqslant |T_0| | H_0 \text{ true})}$

Remark: the example below illustrates the case of a right-sided p-value.



Non-parametric test — A non-parametric test is a test where we do not have any underlying assumption regarding the distribution of the sample.

Testing for the difference in two means

The table below sums up the test statistic to compute when performing a hypothesis test where the null hypothesis is:

$$H_0 \quad : \quad \mu_X - \mu_Y = \delta$$

Distribution of X_i, Y_i	**Sample size** n_X, n_Y	**Variance** σ_X^2, σ_Y^2	**Test statistic under** H_0
Normal	any	known	$\dfrac{(\overline{X}-\overline{Y})-\delta}{\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}} \underset{H_0}{\sim} \mathcal{N}(0,1)$
Normal	large	unknown	$\dfrac{(\overline{X}-\overline{Y})-\delta}{\sqrt{\frac{s_X^2}{n_X}+\frac{s_Y^2}{n_Y}}} \underset{H_0}{\sim} \mathcal{N}(0,1)$
Normal	small	unknown with $\sigma_X = \sigma_Y$	$\dfrac{(\overline{X}-\overline{Y})-\delta}{s\sqrt{\frac{1}{n_X}+\frac{1}{n_Y}}} \underset{H_0}{\sim} t_{n_X+n_Y-2}$

Testing for the mean of a paired sample

We suppose here that X_i and Y_i are pairwise dependent. By noting $D_i = X_i - Y_i$, the one-line table below sums up the test statistic to compute when performing a hypothesis test where the null hypothesis is:

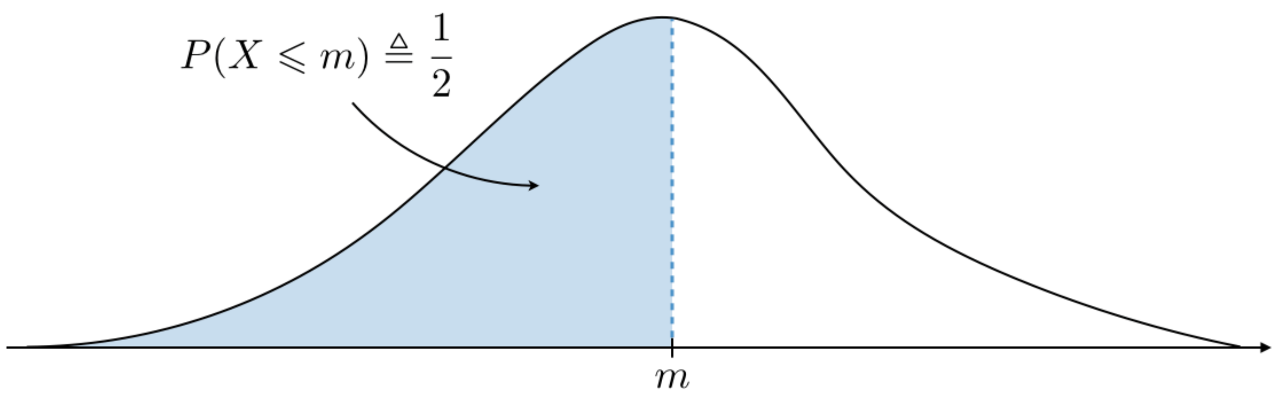
$$H_0 \quad : \quad \overline{D} = \delta$$

Distribution of X_i, Y_i	**Sample size** $n = n_X = n_Y$	**Variance** σ_X^2, σ_Y^2	**Test statistic under** H_0
Normal, paired	any	unknown	$\dfrac{\overline{D}-\delta}{\frac{s_D}{\sqrt{n}}} \underset{H_0}{\sim} t_{n-1}$

Testing for the median

Median of a distribution — We define the median m of a distribution as follows:

$$\boxed{P(X\leqslant m)=P(X\geqslant m)=\frac{1}{2}}$$



Sign test — The sign test is a non-parametric test used to determine whether the median of a sample is equal to the hypothesized median.

By noting $V\underset{H_0}{\sim}\mathcal{B}(n,p=\frac{1}{2})$ the number of samples falling to the right of the hypothesized median, we have:

— If $np\geqslant5$, we use the following test statistic:

$$\boxed{Z=\frac{V-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\underset{H_0}{\sim}\mathcal{N}(0,1)}$$

— If $np<5$, we use the following fact:

$$\boxed{V\underset{H_0}{\sim}\mathcal{B}\left(n,p=\frac{1}{2}\right)}$$

χ^2 test

Goodness of fit test — Let us have k bins where in each of them, we observe Y_i number of samples. Our null hypothesis is that Y_i follows a binomial distribution with probability of success being p_i for each bin.

We want to test whether modelling the problem as described above is reasonable given the data that we have. In order to do this, we perform a hypothesis test:

$$\boxed{H_0 : \textrm{good fit}} \quad \textrm{versus} \quad \boxed{H_1 : \textrm{not good fit}}$$

χ^2 **statistic for goodness of fit** — In order to perform the goodness of fit test, we need to compute a test statistic that we can compare to a reference distribution. By noting k the number of bins, n the total number of samples, if we have $np_i \geqslant 5$, the test statistic T defined below will enable us to perform the hypothesis test:

$$\boxed{T = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i} \underset{H_0}{\sim} \chi_{df}^2} \quad \textrm{with} \quad \boxed{df = (k-1) - \#(\textrm{estimated parameters})}$$

Trends test

Number of transpositions — In a given sequence, we define the number of transpositions, noted T, as the number of times that a larger number precedes a smaller one.

Example: the sequence $\{1, 5, 4, 3\}$ *has* $T = 3$ *transpositions because* $5 > 4, 5 > 3$ *and* $4 > 3$

Test for arbitrary trends — Given a sequence, the test for arbitrary trends is a non-parametric test, whose aim is to determine whether the data suggest the presence of an increasing trend:

$$\boxed{H_0 : \textrm{no trend}} \quad \textrm{versus} \quad \boxed{H_1 : \textrm{there is an increasing trend}}$$

If we note x the number of transpositions in the sequence, the p-value is computed as:

$$\boxed{p\textrm{-value} = P(T \leqslant x)}$$

Remark: the test for a decreasing trend of a given sequence is equivalent to a test for an increasing trend of the inversed sequence.

(https://stanford.edu/~shervine/teaching/cme-106/cheatsheet-statistics#regression-analysis) Regression analysis

In the following section, we will note $(x_1, Y_1), \ldots, (x_n, Y_n)$ a collection of n data points.

Simple linear model — Let X be a deterministic variable and Y a dependent random variable. In the context of a simple linear model, we assume that Y is linked to X via the regression coefficients α, β and a random variable $e \sim \mathcal{N}(0, \sigma)$, where e is referred as the error. We have:

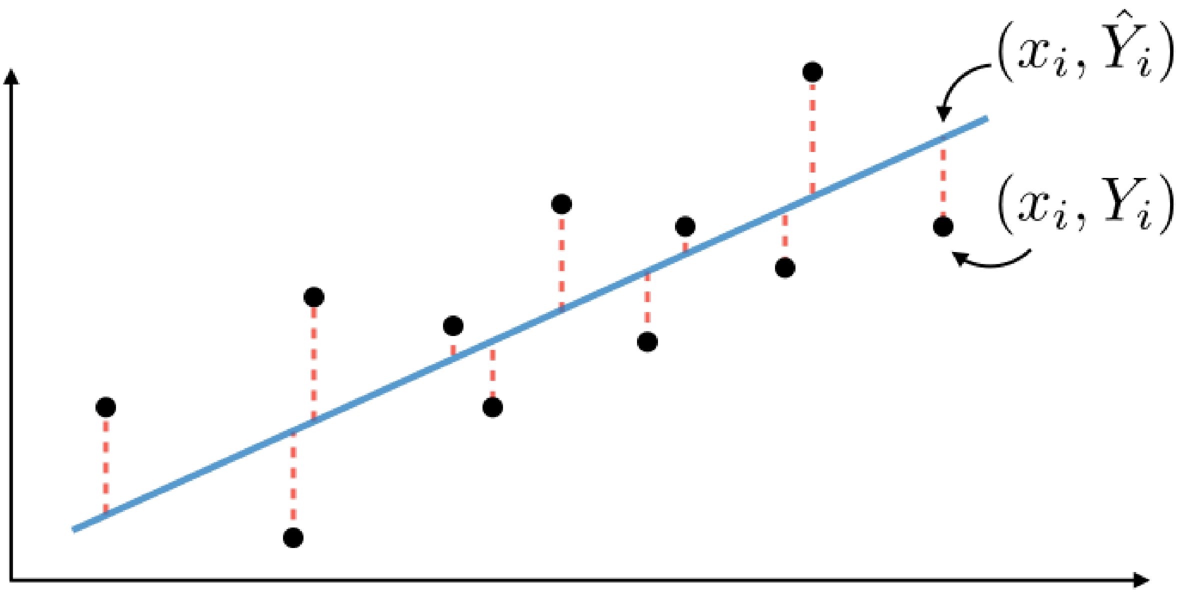
$$\boxed{Y = \alpha + \beta X + e}$$

Regression estimation — When estimating the regression coefficients α, β by A, B, we obtain predicted values $\hat{Y}_i$ as follows:

$$\boxed{\hat{Y}_i = A + Bx_i}$$

Sum of squared errors — By keeping the same notations, we define the sum of squared errors, also known as SSE, as follows:

$$\boxed{SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - (A + Bx_i))^2}$$



Method of least-squares — The least-squares method is used to find estimates A, B of the regression coefficients α, β by minimizing the SSE. In other words, we have:

$$\boxed{A, B = \underset{\alpha,\beta}{\arg\min}\sum_{i=1}^{n}(Y_i - (\alpha + \beta x_i))^2}$$

Notations — Given n data points (x_i, Y_i), we define S_{XY}, S_{XX} and S_{YY} as follows:

$$\boxed{S_{XY}=\sum_{i=1}^n(x_i-\overline{x})(Y_i-\overline{Y})\quad\textrm{and}\quad S_{XX}=\sum_{i=1}^n(x_i-\overline{x})^2\quad\textrm{and}\quad S_{YY}=\sum_{i=1}^n(Y_i-\overline{Y})^2}$$

Least-squares estimates — When estimating the coefficients α, β with the least-squares method, we obtain the estimates A, B defined as follows:

$$\boxed{A=\overline{Y}-\frac{S_{XY}}{S_{XX}}\overline{x}\quad\textrm{and}\quad B=\frac{S_{XY}}{S_{XX}}}$$

Sum of squared errors revisited — The sum of squared errors defined above can also be written in terms of S_{YY}, S_{XY} and B as follows:

$$\boxed{SSE=S_{YY}-BS_{XY}}$$

Key results

When σ is unknown, this parameter is estimated by the unbiased estimator s^2 defined as follows:

$$\boxed{s^2=\frac{S_{YY}-BS_{XY}}{n-2}}$$

The estimator s^2 has the following property:

$$\boxed{\frac{s^2(n-2)}{\sigma^2}\sim\chi_{n-2}^2}$$

The table below sums up the properties surronding the least-squares estimates A, B when σ is known or not:

Coefficient	Estimate	σ	Statistic	$1-\alpha$ confidence
α	A	known	$\frac{A-\alpha}{\sigma\sqrt{\frac{1}{n}+\frac{\overline{X}^2}{S_{XX}}}}\sim\mathcal{N}(0,1)$	$\left[A-z_{\frac{\alpha}{2}}\sigma\sqrt{\frac{1}{n}+\frac{\overline{X}^2}{S_{XX}}},A+\right.$
β	B	known	$\frac{B-\beta}{\frac{\sigma}{\sqrt{S_{XX}}}}\sim\mathcal{N}(0,1)$	$\left[B-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{S_{XX}}},B+\right.$

α	A	unknown	$\frac{A-\alpha}{s\sqrt{\frac{1}{n}+\frac{\overline{X}^2}{S_{XX}}}} \sim t_{n-2}$	$\left[A - t_{\frac{\alpha}{2}} s\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}}, A +\right.$
β	B	unknown	$\frac{B-\beta}{\frac{s}{\sqrt{S_{XX}}}} \sim t_{n-2}$	$\left[B - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{S_{XX}}}, B +\right.$

(https://stanford.edu/~shervine/teaching/cme-106/cheatsheet-statistics#correlation-analysis) Correlation analysis

Correlation coefficient — The correlation coefficient of two random variables X and Y is noted ρ and is defined as follows:

$$\boxed{\rho = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{E[(X-\mu_X)^2]E[(Y-\mu_Y)^2]}}}$$

Sample correlation coefficient — The correlation coefficient is in practice estimated by the sample correlation coefficient, often noted r or $\hat{\rho}$, which is defined as:

$$\boxed{r = \hat{\rho} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}}$$

Testing for correlation — In order to perform a hypothesis test with H_0 being that there is no correlation between X and Y, we use the following statistic:

$$\boxed{\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \underset{H_0}{\sim} t_{n-2}}$$

Fisher transformation — The Fisher transformation is often used to build confidence intervals for correlation. It is noted V and defined as follows:

$$\boxed{V = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)}$$

By noting $V_1 = V - \frac{z_{\frac{\alpha}{2}}}{\sqrt{n-3}}$ and $V_2 = V + \frac{z_{\frac{\alpha}{2}}}{\sqrt{n-3}}$, the table below sums up the key results surrounding the correlation coefficient estimate:

Sample size	**Standardized statistic**	$1-\alpha$ **confidence interval for** ρ
large	$\displaystyle\frac{V - \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)}{\frac{1}{\sqrt{n-3}}} \underset{n \gg 1}{\sim} \mathcal{N}(0,1)$	$\left[\dfrac{e^{2V_1}-1}{e^{2V_1}+1}, \dfrac{e^{2V_2}-1}{e^{2V_2}+1}\right]$

(https://twitter.com/shervinea) (https://linkedin.com/in/shervineamidi) (https://github.com/shervinea) (https://scholar.google.com/citations?user=nMnMTm8AAAAJ)