SD-TSIA204 - Statistics : linear models Confidence interval estimation and Hypothesis tests

Ekhiñe Irurozki
Télécom Paris

Confidence Intervals

- Confidence intervals provide a range of plausible values for a population parameter.
- They quantify the uncertainty associated with point estimates.
- A typical confidence interval is of the form : $\widehat{\theta} \pm$ margin of error.
- Margin of error depends on the desired confidence level (e.g., 95% confidence) and the sample data.
- The confidence level represents the probability that the interval contains the true parameter.
- Common confidence levels include 90%, 95%, and 99%.
- The formula for a confidence interval depends on the statistical distribution used.

Confidence interval

Context : regard an estimator $\widehat{g}(y_1, \ldots, y_n)$ for the value g. We would like to have an interval $\widehat{I}$ around $\widehat{g}$ which contains g with high probability.

We construct $\widehat{I} = [\underline{C}, \overline{C}]$ based on the observations $(y_1, \ldots, y_n)$: confidence interval is a random variable

$$\mathbb{P}(\widehat{I} \text{ contains } g) = \mathbb{P}(\underline{C} \leq g \text{ and } \overline{C} \geq g) = 95\%$$

Confidence interval of level $1-\alpha$

A confidence interval of **level** $1-\alpha$ for a value g is a function of the sample

$$\widehat{I} : (y_1, \dots, y_n) \mapsto \widehat{I} = \left[\underline{C}(y_1, \dots, y_n), \overline{C}(y_1, \dots, y_n)\right]$$

such that

$$\mathbb{P}\left[g \in \widehat{I}(y_1, \dots, y_n)\right] \geq 1-\alpha$$

or

$$\mathbb{P}\left[g \notin \widehat{I}(y_1, \dots, y_n)\right] \leq \alpha$$

Rem:usual choices $\alpha = 5\%, 1\%, 0.1\%$, etc. Defined often by the consideration data complexity / number of observations.

Rem:In the following we will denote confidence interval by CI.

Example : survey

Election survey with two candidates : A and B. Choice of the ith respondent follows Bernoulli distribution having parameter p, with $y_i = 1$ if he votes for A and 0 otherwise.

Aim : estimate p and give a CI

Sample of size n : a reasonable estimator is then $n = 1000$

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} y_i = \overline{y}_n \, .$$

The goal is to establish an oracle : is there a clear winner in this survey ?

What is the confidence interval for p ?

- is this estimator likely or not ?

Method 1 for CI : concentration inequalities

- Search for an interval $\widehat{I} = [\widehat{p} - \delta, \widehat{p} + \delta]$ such that $\mathbb{P}(p \in \widehat{I}) \geq 0.95 \Leftrightarrow$ search for δ such that $\mathbb{P}\left[\, |\widehat{p} - p| \ > \delta\right] \leq 0.05$
- Constituent : **Tchebyschev** inequality

$$\forall \delta > 0, \quad \mathbb{P}(|X - \mathbb{E}(X)| > \delta) \leq \frac{\mathbb{V}\mathrm{ar}(X)}{\delta^2}$$

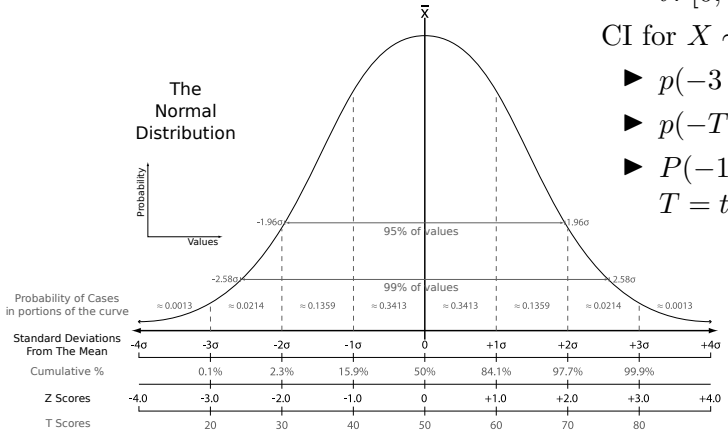
For $X = \widehat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ we know that $\mathbb{E}(\widehat{p}) = p$ and $\mathbb{V}\mathrm{ar}(\widehat{p}) = \frac{p(1-p)}{n}$:

$$\forall p \in (0,1), \forall \delta > 0, \quad \mathbb{P}(|\widehat{p} - p| > \delta) \leq \frac{p(1-p)}{n\delta^2} \leq \frac{1}{4n\delta^2}$$

Application : for a CI of 95%, find δ such that $\frac{1}{4n\delta^2} = 0.05$, *i.e.* $\delta = (0.2n)^{-1/2}$. For $n = 1000$, $\widehat{p} = 55\%$:

$$\delta = 0.07 \; ; \quad \widehat{I} = [0.48, 0.62]$$

Intervals - Gaussian case



- cumulative distribution (cdf) : $P(X \leq x)$
- quantile (ppf_α, t_α, z_α) : $t\colon [0,1] \to \mathbb{R}$, x s.t. $p(X \leq x) = \alpha$

CI for $X \sim N(0,1)$ are easy

- $p(-3 \leq X \leq 3) = 1 - 2 * cdf(-3)$
- $p(-T \leq X \leq T) = 1 - \alpha$, $T = t_{1-\alpha/2}$
- $P(-1.96 \leq X \leq 1.96) = 0.95$ or $T = t_{(1-.05)/2}$

The quantile function

The quantile function is a fundamental concept in probability and statistics. It is also referred to as the *quantile function* or *inverse cumulative distribution function* or *Percent Point Function*.

Definition : The quantile function of a RV X is a function that maps a probability p to the value x such that $P(X \leq x) = p$.

Notation :

- The quantile function of a distribution is denoted as t_p.
- Mathematically, $t_p = x$ iff $P(X \leq x) = p$.
- The quantile function is useful for finding critical values, confidence intervals, and performing hypothesis tests.

Usage : `norm.ppf(q, loc=0, scale=1)` in `scipy.stats`

Convergence in law

Is the weakest mode of convergence. It defines a relationship not between the RVs themselves but between their cumulative distribution functions.

Convergence in Law : A sequence of RVs $(X_n)_{n\in\mathbb{N}^*}$ converges in law to X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \quad \text{for all } x \text{ where } F_X \text{ is continuous.}$$

This convergence is denoted as : $X_n \xrightarrow{L} X$.

Central Limit Theorem (CLT)

If $(X_n)_{n\in\mathbb{N}^*}$ is a sequence of independent and identically distributed (i.i.d.) RVs with the same mean μ and the same standard deviation $\sigma > 0$, then, by defining $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, we have :

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{L} N(0,1)$$

In other words, the standardized sum of i.i.d. RVs with finite variance converges in law to the standard normal distribution $N(0,1)$.

In practice, this theorem is very useful because it allows us to approximate that, for a sufficiently large n, the sum of i.i.d. RVs approximately follows a normal distribution.

Case When n is Sufficiently Large

By "sufficiently large" we generally mean $n \geq 30$.

- The probability distribution of $\bar{X}$ depends on the distribution of X itself.
- The CLT asserts that the sequence of RV $U_n = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, converges in law to $N(0,1)$.
- In practice, this means that for a sufficiently large n, the RV $\bar{X}$ approximately follows the normal distribution $N(\mu, \sigma^2/n)$ even if the parent distribution is not normal.

Confidence interval of level 95% for μ when we know $\bar{X}$, n and σ

Setting : Let $X_i \sim P_{\mu,\sigma}$ where σ is known. Goal : give a CI for μ. Rem: for $U \sim N(0,1)$, we have $\alpha = .05$, $t_{1-\alpha/2} \approx 1.96$ and thus $P(-1.96 \leq U \leq 1.96) = 0.95$. Applying this result to the variable $U_n = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, we obtain an approximate CI for μ at the 95% confidence level :

$$P\left(-1.96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95.$$

Reordering

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

In summary, when σ is known and n is sufficiently large, we can construct a CI for μ at the 95% confidence level :

$$\mathrm{CI}_{0.95}(\mu) = [\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}; \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}]$$

What if the population variance is unknown ?

Method 2 for CI : Asymptotic confidence intervals

The survey example : $y_i \in \{0, 1\}$, $n = 1000$,

$$\widehat{p} = n^{-1} \sum_{i=1}^{n} y_i = 0.55$$

We assume that n is sufficiently large, such that

$$\sqrt{n} \left(\frac{\widehat{p} - p}{\widehat{\sigma}} \right) \sim \mathcal{N}(0, 1)$$

$$\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (y_i - \widehat{p})^2 = \widehat{p} - \widehat{p}^2$$

We know the quantiles of the normal distribution (numerically)
$t_{1-0.05/2} = norm.ppf(1 - 0.05/2) \simeq 1.96$

Following the CLT and approximation of the Gaussian quantiles

$$\mathbb{P} \left[-1.96 < \sqrt{n} \, \frac{0.55 - p}{\widehat{\sigma}} < 1.96 \right] \approx 0.95$$

new CI : $\widehat{I} = [0.52, 0.58]$: better ! (**more optimistic**)

The Student's t-distribution

- the quantile function is `t.ppf(q, df, loc=0, scale=1)` in `scipy.stats`
- It is similar in shape to the standard normal distribution but has heavier tails.
- The t-distribution is used when the population standard deviation is unknown and sample sizes are small.
- It depends on a single parameter called degrees of freedom (ν), which determines the shape of the distribution. For $\nu = 1$ t_ν becomes the standard Cauchy distribution, whereas for $\nu \to \infty$ it becomes the standard normal distribution.

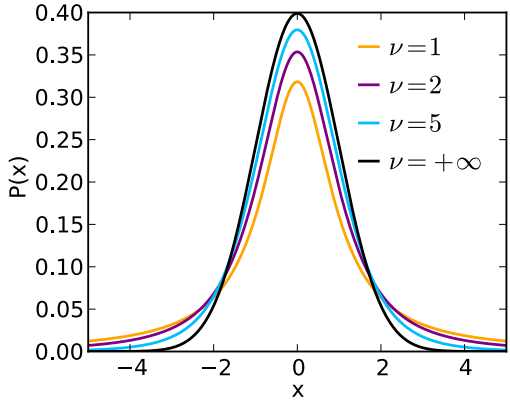


Figure – Probability density function of the Student's t-distribution with different degrees of freedom (Wikipedia).

The Chi-Squared Distribution (χ^2)

Definition : The chi-squared (χ^2) distribution is a continuous probability distribution that arises in various statistical applications.

Parameters : The χ^2 distribution depends on a single parameter, which is the degrees of freedom (ν). It is denoted as χ^2_ν.

Probability Density Function (PDF) : The probability density function of the χ^2 distribution is given by :

$$f(x;\nu) = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}, \quad x > 0$$

where $\Gamma(\cdot)$ represents the gamma function.

Chi-Squared Distribution with n Degrees of Freedom

Let $Y_1, Y_2, \ldots, Y_n$ be independent RVs, each following the standard normal distribution $N(0, 1)$. Then, the RV

$$Z = Y_1^2 + Y_2^2 + \ldots + Y_n^2 \sim \chi_n^2 \tag{1}$$

follows the chi-squared distribution with n degrees of freedom.

Relating distributions

Definition Let U be a RV following the standard normal distribution $N(0,1)$, and let Z be a RV, independent of U, following a chi-squared (χ^2_ν) distribution with ν degrees of freedom (where $\nu \in \mathbb{N}^*$). The t-Student RV, denoted as T, is defined as :

$$T = \frac{U}{\sqrt{Z/\nu}} \tag{2}$$

- T follows the t-Student distribution with ν degrees of freedom.
- It arises in statistical inference, particularly when dealing with small sample sizes and unknown population standard deviation.

Application

Let $X \sim N(\mu, \sigma^2)$ and $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ be the empirical variance of the sample, then the RV

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n-1}}$$

follows the Student's t-distribution, $T \sim T_{n-1}$.

Proof sketch : Consider the RV $U = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$. We know that $U \sim N(0,1)$.

Furthermore, we have that $\frac{nS^2}{\sigma^2}$ follows the χ^2_{n-1} distribution with $\nu = n-1$ degrees of freedom.

$$T = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{nS^2}{\sigma^2(n-1)}}} = \frac{\bar{X} - \mu}{S/\sqrt{n-1}}.$$

CI for the Mean - unknwn variance

For example, if $1-\alpha = 0.95$ and $n = 10$, then $t_{1-\alpha/2} = 2.262$.

- We can easily isolate μ by rearranging the equation :

$$-t_{1-\alpha/2} \leq T \leq t_{1-\alpha/2} \Leftrightarrow -t_{1-\alpha/2} \leq \frac{\bar{X}-\mu}{S/\sqrt{n-1}} \leq t_{1-\alpha/2}$$

$$\Leftrightarrow \bar{X} - t_{1-\alpha/2}\frac{S}{\sqrt{n-1}} \leq \mu \leq \bar{X} + t_{1-\alpha/2}\frac{S}{\sqrt{n-1}}.$$

- Therefore, we obtain a random CI for μ at the confidence level $1-\alpha$:

$$\mathrm{CI}_{1-\alpha}(\mu) = \left[\bar{X} - t_{1-\alpha/2}\frac{S}{\sqrt{n-1}}, \bar{X} + t_{1-\alpha/2}\frac{S}{\sqrt{n-1}}\right]$$

CI (method 2) for the regression coefficients (I)

Proposition

$$\text{If } \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \operatorname{Id}_n), \text{ then } \quad T_j = \frac{\widehat{\theta}_j - \theta_j^*}{\widehat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}}} \sim \mathcal{T}_{n-\operatorname{rang}(X)}$$

where $\mathcal{T}_{n-\operatorname{rang}(X)}$ is a Student-t distribution with $n - \operatorname{rang}(X)$ degrees of freedom.

Its density, quantiles, etc..., can be computed numerically and are accessible in any software.

proof Recall that $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \sim N(0, \sigma^2(X^T X)^{-1})$ and that $(n-p-1)\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p-1}$ for X full rank and $\widehat{\sigma}^2 := (n-p-1)^{-1}\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2$ an unbiased estimator for σ^2. It follows that

$$\frac{\frac{\widehat{\theta}_i - \theta_i}{\sqrt{\sigma^2 (X^T X)^{-1}_{jj}}}}{\sqrt{\frac{(n-p-1)\widehat{\sigma}^2}{(n-p-1)\sigma^2}}} = \frac{\widehat{\theta}_i - \theta_i}{\sqrt{\widehat{\sigma}^2 (X^T X)^{-1}_{jj}}} \sim T_{n-p-1}$$

CI for the regression coefficients (II)

Under the Gaussian assumption, since

$$T_j = \frac{\widehat{\theta}_j - \theta_j^*}{\widehat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}}} \sim \mathcal{T}_{n-p-1}$$

and noting $t_{1-\alpha/2}$ a quantile of order $1-\alpha/2$ of the distribution $\mathcal{T}_{n-p-1}$,

$$\left[\widehat{\theta}_j - t_{1-\alpha/2}\widehat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}}, \widehat{\theta}_j + t_{1-\alpha/2}\widehat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}}\right]$$

for the quantity θ_j^*.

Rem: $\mathbb{P}(|T_j| < t_{1-\alpha/2}) = 1-\alpha$ since the Student-t distribution is symmetric.

CI for the predicted values

Now we would like to construct a CI for the predicted value at a single (new) given point $x = (1, x_1, ..., x_p)^\top \in \mathbb{R}^{p+1}$.

The predicted value at x (under the true model) is defined as

$$y^* = x^\top \boldsymbol{\theta}^* .$$

Under the Gaussian assumption, with the same notation, the following confidence interval is of level $1 - \alpha$

$$\left[x^\top \widehat{\boldsymbol{\theta}} - t_{1-\alpha/2} \widehat{\sigma} \sqrt{x^\top (X^\top X)^{-1} x}, x^\top \widehat{\boldsymbol{\theta}} + t_{1-\alpha/2} \widehat{\sigma} \sqrt{x^\top (X^\top X)^{-1} x} \right]$$

for the quantity y^*.

CI for the new values (aka, Prediction interval)

The CI from above is for the regression hyperplane, i.e. it is reflecting uncertainty of the fitted values.

How to build a CI for a new value at a single (new) given point $x = (1, x_1, ..., x_p)^\top \in \mathbb{R}^{p+1}$?

A new predicted value at x (under the true model) is defined as

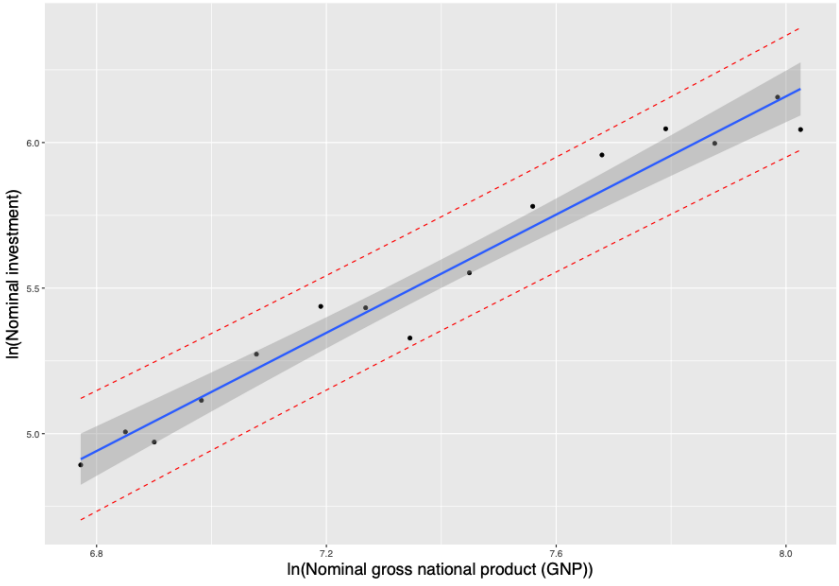
$$y = y^* + \epsilon\,.$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

One can show that, in this case, and with the same notation, the following confidence interval is of level α

$$\left[x^\top \widehat{\boldsymbol{\theta}} - t_{1-\alpha/2}\widehat{\sigma}\sqrt{1 + x^\top (X^\top X)^{-1} x}, x^\top \widehat{\boldsymbol{\theta}} + t_{1-\alpha/2}\widehat{\sigma}\sqrt{1 + x^\top (X^\top X)^{-1} x}\right]$$

Example : Investment data (II).



Hypothesys testing : General principle

Context

- We observe $X_1, \ldots, X_n$ from a common distribution $\mathcal{P}$
- We are interested in $\theta \in \Theta$, a parameter of $\mathcal{P}$

The goal is to decide whether an assumption on θ is likely (or not)

$$\mathcal{H}_0 = \{\theta \in \Theta_0\}$$

against some alternative

$$\mathcal{H}_1 = \{\theta \in \Theta_1\}$$

Call $\mathcal{H}_0$ the null hypothesis, $\mathcal{H}_1$ the alternative

Determine a test statistic $T(X_1, \ldots, X_n)$ and a region R such that if

$$T(X_1, \ldots, X_n) \in R \;\Rightarrow\; \text{we reject } \mathcal{H}_0$$

In other words the observed data discriminates between $\mathcal{H}_0$ and $\mathcal{H}_1$

General principle : Hypothesis Testing for "Heads or Tails"

Scenario : You are given a fair coin, and you want to test whether it's indeed fair or biased towards heads.

Hypotheses :

- Null Hypothesis ($\mathcal{H}_0$) : The coin is fair, and the probability of getting heads ($P(\text{Heads})$) is 0.5.
- Alternative Hypothesis ($\mathcal{H}_1$) : The coin is biased towards heads, and $P(\text{Heads}) > 0.5$.

Test Statistic : You decide to flip the coin 100 times and record the number of heads (X).

Statistical Test : Using a significance level of $\alpha = 0.05$, perform a one-sided hypothesis test to determine if there is enough evidence to reject the null hypothesis in favor of the alternative hypothesis based on the observed number of heads.

Conclusion : You may "reject" the null hypothesis and conclude that the coin is biased towards heads or not reject it.

Do we reject or do we accept ?

In most practical situations, $\mathcal{H}_0$ is simple, i.e.,

$$\Theta_0 = \{\theta_0\}$$

and $\Theta_1 = \Theta \backslash \Theta_0$ is large

($\mathcal{H}_0$ is often an hypothesis on which we care particularly, e.g., something acknowledged to be true, easy to formulate)

We only reject $\mathcal{H}_0$: If $\mathcal{H}_0$ is not rejected we cannot conclude $\mathcal{H}_0$ is true because $\mathcal{H}_1$ is too general

e.g. $\{p \in [0, 0.5[\cup]0.5, 1]\}$ can not be rejected !

2 types of error

	$\mathcal{H}_0$	$\mathcal{H}_1$
$\mathcal{H}_0$ is not rejected	Correct (True positive)	Wrong (False negative)
$\mathcal{H}_0$ is rejected	Wrong (False positive)	Correct (True negative)

- Type I : probability of a wrong reject
$$P(T(X_1, \ldots, X_n) \in R \mid \mathcal{H}_0)$$
- Type II : probability of wrong non-reject
$$P(T(X_1, \ldots, X_n) \notin R \mid \mathcal{H}_1)$$

Significance level and power

Significance level α if $\limsup_{n\to+\infty} P(T(X_1,\dots,X_n)\in R \mid \mathcal{H}_0) \leq \alpha$

We speak of 95%-test when α is 0.05%

Consistency : A test statistics (given by $T(X_1,\dots,X_n)$ and a region R) is said to be α-consistent if the significant level is α and if the power goes to one, i.e.,

$$\limsup_{n\to+\infty} P(T(X_1,\dots,X_n)\in R \mid \mathcal{H}_0) \leq \alpha$$

$$\lim_{n\to\infty} P(T(X_1,\dots,X_n)\in R \mid \mathcal{H}_1) = 1$$

Test statistic and reject region

Goal : to build a α-consistent test

(1) Define the test statistic $T(X_1, \ldots, X_n)$ and the level α you wish

(2) Do some maths to determine a reject region R that achieves a significance level α

(3) Prove the consistency

(4) Rule decision : reject whenever $\boxed{T_n(X_1, \ldots, X_n) \in R}$

Famous tests

- Test of the equality of the mean for 1 sample
- Test of the equality of the means between 2 samples
- Chi-square test for the variance
- Chi-square test of independence
- Regression coefficient non-effects test

Conformity Test for the Mean of a Normal RV with Known Variance

Let $X_i \sim N(\mu, \sigma^2)$ and $\bar{X} = \sum_{i=1}^{n} X_i$. Is $\mu = \mu_0$? We will proceed in 4 steps :

Step 1 : Formulate Hypotheses Let's start with the null hypothesis $\mathcal{H}_0$: $\mu = \mu_0$, where the alternative hypothesis is $H_1 : \mu \neq \mu_0$. We assume $\mathcal{H}_0$ is true.

Step 2 : Distribution of X Under $\mathcal{H}_0$ As a result of $\mathcal{H}_0$, X follows a normal distribution $N(\mu_0, \sigma^2)$, and consequently :

$$U = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Step 3 : Define the reject region Let's choose a significance level α, which we consider negligible. This leads to an interval $[-t_{1-\alpha/2}, t_{1-\alpha/2}]$ within which the variable U (our decision variable) has a probability of $(1 - \alpha)$ of falling if the null hypothesis is true. Consequently, if $\mathcal{H}_0$ is true, we have $P(|U| > t_{1-\alpha/2}) = \alpha$. Neglecting the probability α means considering it very unlikely to find U outside the interval $[-t_{1-\alpha/2}, t_{1-\alpha/2}]$ if the null hypothesis is true. The reject region is thus $R =]-\infty, -t_{1-\alpha/2}[\cup]t_{1-\alpha/2}, \infty[$

Interpreting the Test Results

Step 4 : Interpretation of Results Check whether $u \in R$:

- If $u \in R$ we prefer to reject the hypothesis $\mathcal{H}_0$. However, it's important to acknowledge that by doing so, we are accepting the risk α of making a Type I error, meaning we might reject $\mathcal{H}_0$ incorrectly.
- If $u \notin R$ it does not imply that $\mathcal{H}_0$ is true. Rather, it indicates that the collected data is not in contradiction with the hypothesis. In other words, we are unable to conclude in favor or against the hypothesis. In practical applications, this is often less problematic than it may seem because the focus is on avoiding the incorrect rejection of $\mathcal{H}_0$, while maintaining the status quo corresponds to retaining the hypothesis.

Step 5(*) : Calculating the p-value Calculate the probability of observing a test statistic as extreme as $|u|$ in both tails of the distribution. The resulting p-value represents the likelihood of observing such an extreme result under the null hypothesis.

Hypothesis Testing Example : Gaussian Mean (Two-Sided Test)

Scenario : Suppose a manufacturer produces light bulbs, and the claimed mean lifespan of these bulbs is 1000 hours with a known standard deviation of 50 hours. To test the manufacturer's claim, a random sample of 36 light bulbs is selected and tested. The sample has a mean lifespan of 990 hours. We want to determine if there is enough evidence to reject the manufacturer's claim at a 5% significance level.

Hypotheses :

- Null Hypothesis ($\mathcal{H}_0$) : The mean lifespan of the bulbs produced by the manufacturer is equal to 1000 hours, i.e., $\mu = 1000$ hours.
- Alternative Hypothesis ($\mathcal{H}_1$) : The mean lifespan of the bulbs is not equal to 1000 hours, i.e., $\mu \neq 1000$ hours (Two-Sided Test).

Step 1 : Formulate Hypotheses

- $\mathcal{H}_0 : \mu = 1000$ hours
- $H_1 : \mu \neq 1000$ hours (Two-Sided Test)

Hypothesis Testing Example : Gaussian Mean (Two-Sided Test, Continued)

Step 2 : Distribution of X Under $\mathcal{H}_0$

Calculate the test statistic (z) using the sample data, population mean (μ), and standard deviation (σ).

$$U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{990 - 1000}{50/\sqrt{36}} = -1.2$$

Step 3 : Define reject region Determine the two-tailed critical values at the 5% level of significance :

$$t_{1-\alpha/2} = -1.96 \text{ and } 1.96 \text{ (from the standard normal distribution)}$$

$$R =]-\infty, -t_{1-\alpha/2}[\cup]t_{1-\alpha/2}, \infty[=]-\infty, -1.96[\cup]1.96, \infty[$$

Step 4 : Interpretation of Results Since $u \notin R$ we don't reject $\mathcal{H}_0$.

Step 5 : p-value, type-I and type-II errors

One-Sided vs. Two-Sided Tests

One-Sided Test :

- Used to detect an effect in one specific direction (greater than or less than).
- Has a single critical region in one tail of the distribution.
- Hypotheses :
  - $\mathcal{H}_0$: No effect or no difference ($\mu = \mu_0$).
  - $\mathcal{H}_1$: Effect or difference in a specific direction ($\mu > \mu_0$ or $\mu < \mu_0$).

Two-Sided Test :

- Used to detect an effect in either direction (greater than or less than).
- Has two critical regions in both tails of the distribution.
- Hypotheses :
  - $\mathcal{H}_0$: No effect or no difference ($\mu = \mu_0$).
  - $\mathcal{H}_1$: Effect or difference in either direction ($\mu \neq \mu_0$).

Example :

- One-Sided Test : Testing if a new drug increases blood pressure ($\mathcal{H}_0 : \mu \leq \mu_0$ vs. $\mathcal{H}_1 : \mu > \mu_0$).
- Two-Sided Test : Testing if a scale is accurate ($\mathcal{H}_0 : \mu = \mu_0$ vs. $\mathcal{H}_1 : \mu \neq \mu_0$).

The p-value

Quantifies the strength of evidence against the null hypothesis ($\mathcal{H}_0$). The p-value represents the probability of obtaining a result as extreme as, or more extreme than, the one observed, assuming that $\mathcal{H}_0$ is true.

- Calculate the test statistic u based on the sample data and $\mathcal{H}_0$ assumptions.
- Determine the direction of the test (two-tailed, left-tailed, or right-tailed) based on the alternative hypothesis ($\mathcal{H}_1$).
- For a two-tailed test, calculate the probability of observing a test statistic as extreme as $|u|$ in both tails of the distribution.
- For a one-tailed test (left-tailed or right-tailed), calculate the probability of observing a test statistic as extreme as u in the specified tail.

Interpreting the p-value :

- If the p-value is less than the chosen significance level α, it suggests strong evidence against the null hypothesis, and we may reject $\mathcal{H}_0$.
- If the p-value is greater than or equal to α, it implies that the observed data is consistent with $\mathcal{H}_0$, and we do not have sufficient evidence to reject it.

Test of no-effect : Gaussian case

Gaussian Model

$$y_i = \theta_0^\star + \sum_{k=1}^{p} \theta_k^\star x_{i,k} + \varepsilon_i$$

$$x_i^\top = (1, x_{i,1}, \ldots, x_{i,p}) \in \mathbb{R}^{p+1} \text{ (deterministic)}$$

$$\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2), \text{ for } i = 1, \ldots, n$$

Rem: Let $X = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n \times (p+1)}$ of full rank, and $\widehat{\sigma}^2 = \|\mathbf{y} - X\widehat{\boldsymbol{\theta}}\|_2^2/(n-(p+1))$, then

$$\widehat{T}_j = \frac{\widehat{\theta}_j - \theta_j^*}{\widehat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}}} \sim \mathcal{T}_{n-(p+1)}$$

Rem: $\theta_j^* = 0$ then column X_j has no effect on Y

Test of no-effect : Gaussian case

Goal : Develop a test of significance level α to check whether $\theta_j^* = 0$

Null hypothesis, $\mathcal{H}_0 : \theta_j^* = 0$, equivalently, $\Theta_0 = \{\theta \in \mathbb{R}^p : \theta_j = 0\}$

Under $\mathcal{H}_0$, we know the value of $\widehat{T}_j$:

$$T_j := \frac{\widehat{\theta}_j}{\widehat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}}} \sim \mathcal{T}_{n-(p+1)}$$

Choosing $R = [-t_{1-\alpha/2}, t_{1-\alpha/2}]^c$ with $t_{1-\alpha/2}$ the $1-\alpha/2$-quantile of $\mathcal{T}_{n-(p+1)}$, we decide to reject $\mathcal{H}_0$ whenever

$$|\widehat{T}_j| > t_{1-\alpha/2}$$

Link between IC and test

Reminder (Gaussian model) :

$$IC_\alpha := \left[\widehat{\theta}_j - t_{1-\alpha/2}\widehat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}}, \widehat{\theta}_j + t_{1-\alpha/2}\widehat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}}\right]$$

is a CI at level α for θ^*_j. Stating "$0 \in IC_\alpha$" means

$$|\widehat{\theta}_j| \leq t_{1-\alpha/2}\widehat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}} \quad \Leftrightarrow \quad \frac{|\widehat{\theta}_j|}{\widehat{\sigma}\sqrt{(X^\top X)^{-1}_{j,j}}} \leq t_{1-\alpha/2}$$

It is equivalent to accepting the hypothesis $\theta^*_j = 0$ at level α. The smallest α such that $0 \in IC_\alpha$ is called the p-**value**.

Rem: Taking α close to zero IC_α covers the full space, hence one can find (by continuity) an α achieving equality in the aforementioned equations.

“Diabetes” data set

	age	sex	bmi	bp	Serum measurements						Resp
patient	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	y
1	59	2	32.1	101	157	93	38	4	4.9	87	151
2	48	1	21.6	87	183	103	70	3	3.9	69	75
...	...										...
...	...										...
441	36	1	30.0	95	201	125	42	5	5.1	85	220
442	36	1	19.6	71	250	133	97	3	4.6	92	57

$n = 442$ patients having diabetes, $p = 10$ variables “baseline” body mass index (bmi), average blood pressure (bp), etc have been measured.

Goal : predict disease progression one year in advance after the “baseline” measurement.

- Each variable of the data set from *sklearn* has been previously standardized.
- We apply an “expensive” version of the **forward variable selection** method

“Diabetes” data set

- We define a vector of covariates with intercept $\tilde{X} = (\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_{10})$.

Step 0

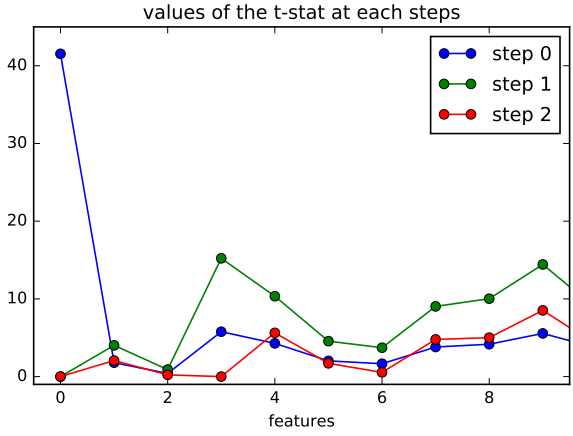
- for each variable $\tilde{X}_k$, $k = 1, \ldots, 11$, we consider the model
$$\mathbf{y} \simeq \theta_k \mathbf{x}_k$$
- we test whether its regression coefficient equals zero, *i.e.*
$$H_0 : \theta_k = 0$$
using the statistic $\frac{\widehat{\theta}_k}{\widehat{s}_k}$ with $\widehat{s}_k$ being the estimated standard deviation.
- we compare all of the p-values, and keep the one possessing the smallest p-value. We save the residuals in the vector V_0.

“Diabetes” data set

Step ℓ We have selected ℓ variable(s) : $\tilde{X}^{(\ell)} \in \mathbb{R}^{\ell}$. Those not selected are noted $\tilde{X}^{(-\ell)} \in \mathbb{R}^{p-\ell}$. We possess the vector of residuals $V_{\ell-1}$ calculated on the previous step.

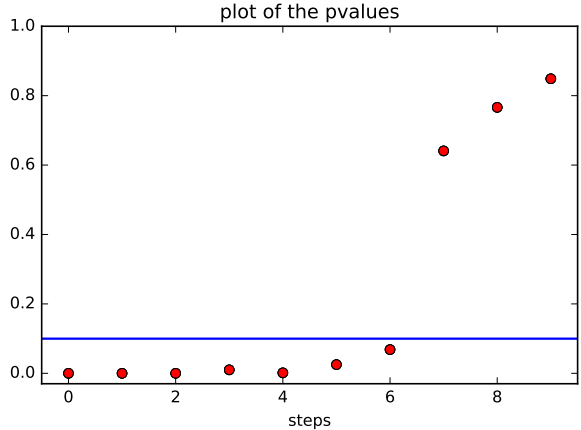
- for each variable $\mathbf{x}_k$ in $\tilde{X}^{(-\ell)}$, we consider the model
$$V_{\ell-1} \simeq \theta_k \mathbf{x}_k$$
- we test if its regression coefficient equal zero, *i.e.*
$$H_0 : \theta_k = 0$$
using the test statistic $\frac{\widehat{\theta}_k}{\widehat{s}_k}$ with $\widehat{s}_k$ being the estimated standard deviation.
- we compare all of the p-values, and keep the one possessing the smallest p-value. We save the residuals in the vector V_{ℓ}.

Values of the test statistics at each step



- The test statistic of the selected variable is 0 on the following steps.
- The intercept is the first selected variable, then x_3, etc

Values of the test statistics at each step



- Sequence of the selected variables wit the test size 0.1 :

[0, 3, ,9 ,5 ,4 ,2 ,7]

ROC curve, Medical context

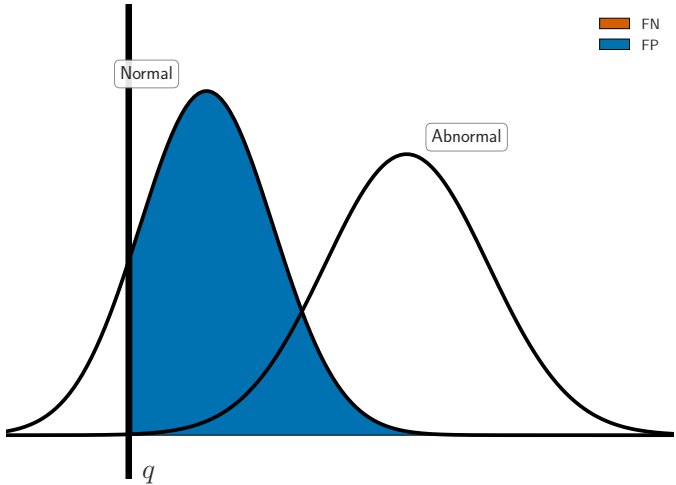
- A group of patients $i = 1, \ldots, n$ is followed for disease screening.
- For each individual, the test relies on a random variable $X_i \in \mathbb{R}$ and a threshold $q \in \mathbb{R}$

$$\begin{array}{rcl} \text{as soon as} & X_i > q & \text{the test is } \textbf{positive} \\ \text{o.w.} & & \text{the test is } \textbf{negative} \end{array}$$

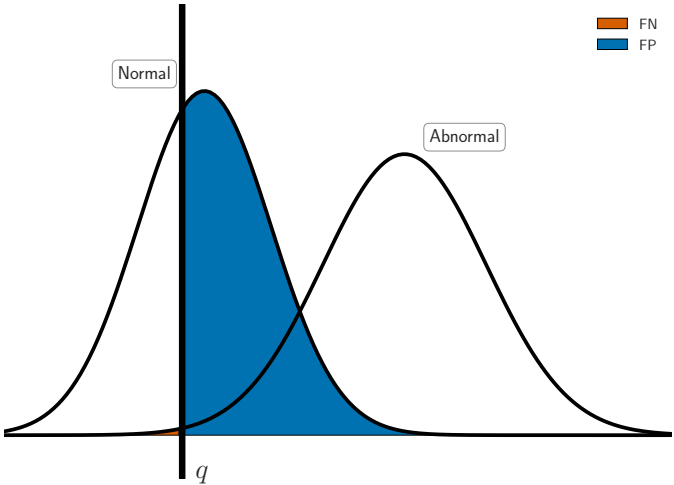
Set of possible configurations

	Normal H_0	Sick H_1
negative	true negative	false negative
positive	false positive	true positive

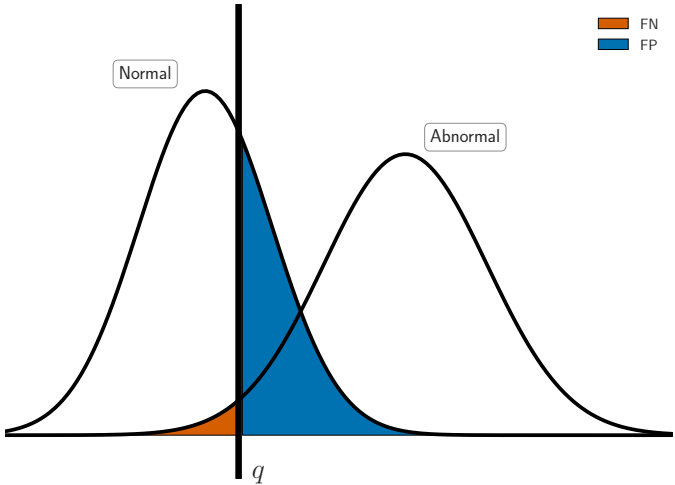
False positive vs. false negative



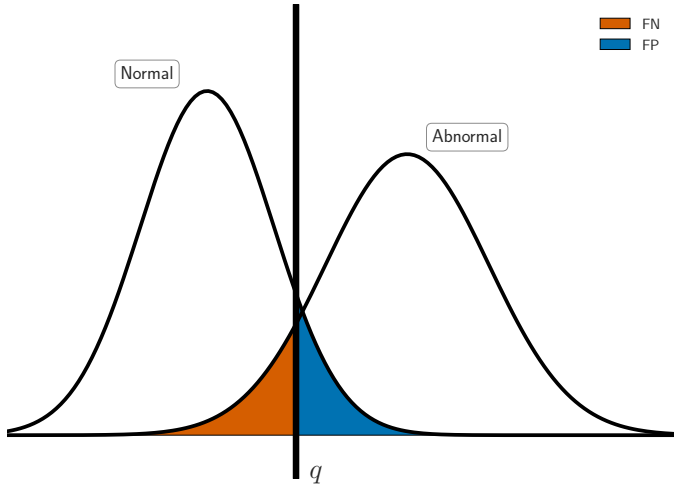
False positive vs. false negative



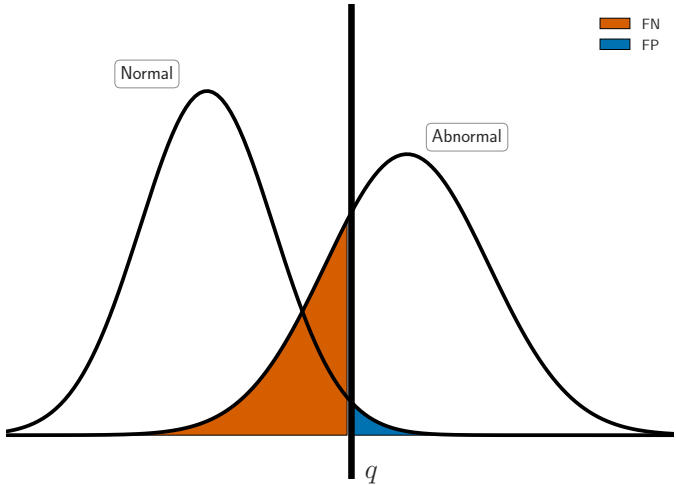
False positive vs. false negative



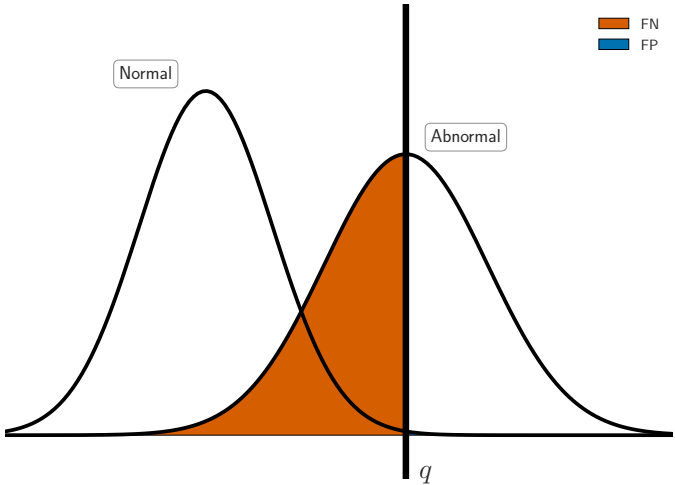
False positive vs. false negative



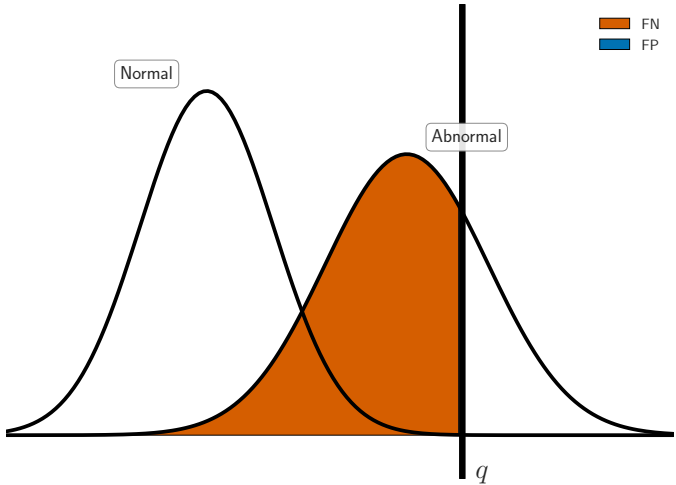
False positive vs. false negative



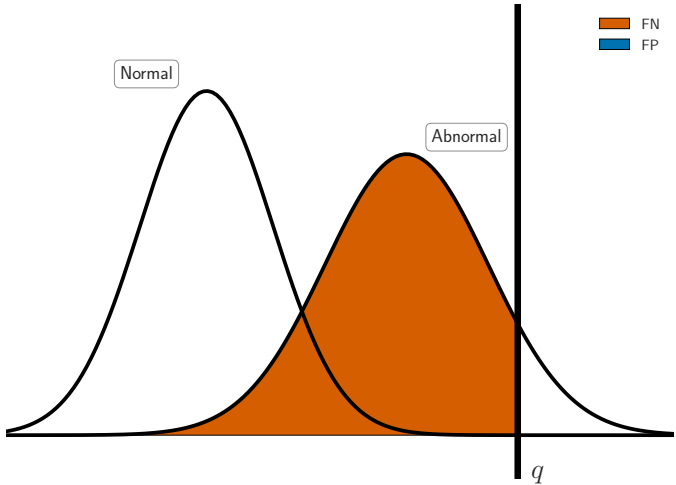
False positive vs. false negative



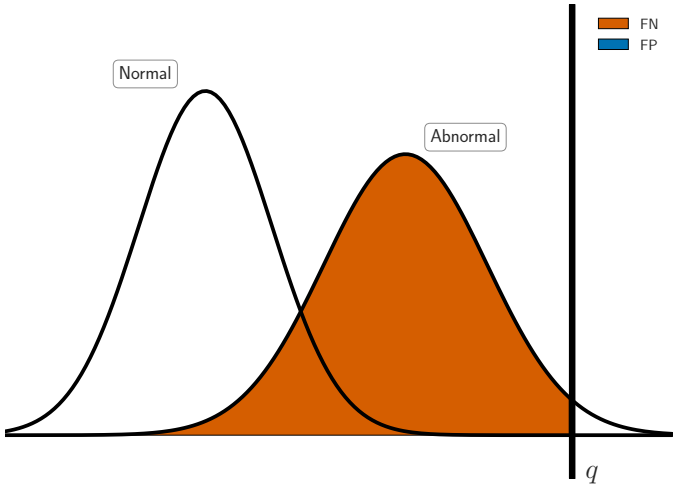
False positive vs. false negative



False positive vs. false negative



False positive vs. false negative



Sensitivity - Specificity

- Assumption : Normal individuals have the same c.d.f. F
- Assumption : Sick individual have the same c.d.f G

Definition

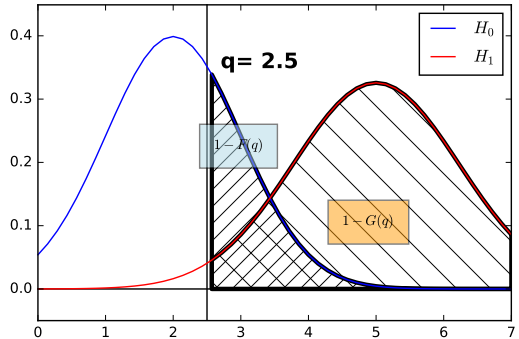
- Sensitivity : $sen(q) = 1 - G(q)$ (1− type 2nd error)
- Specificity : $spe(q) = F(q)$ (1− type 1st error)

ROC curve

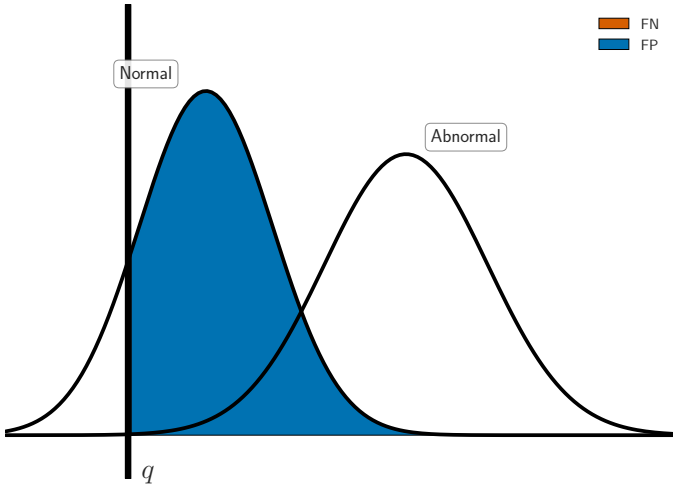
Definition The ROC curve is the curve described by $(1 - spe(q), sen(q))$, when $q \in \mathbb{R}$. Hence, it is the function $[0, 1] \to [0, 1]$

$$\text{ROC}(t) = 1 - G(F^{-}(1 - t))$$

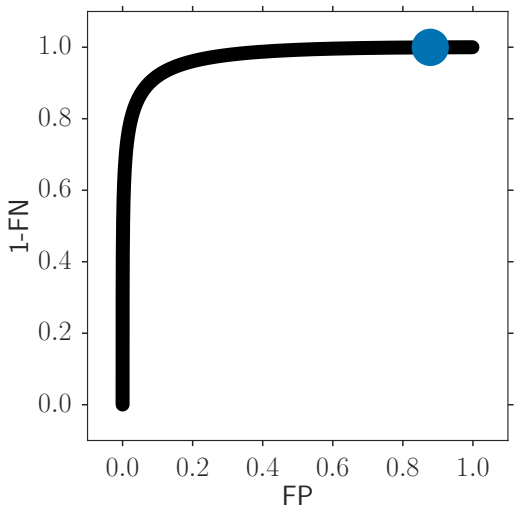
where $F^{-}(1 - t) = \inf\{x \in \mathbb{R} : F(x) \geq 1 - t\}$.



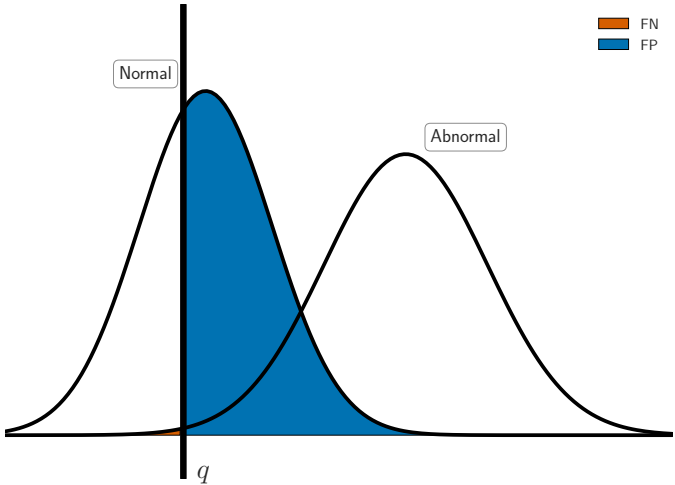
ROC Curve



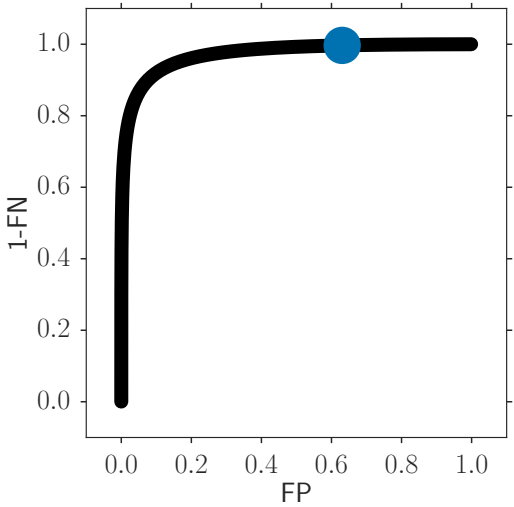
ROC Curve



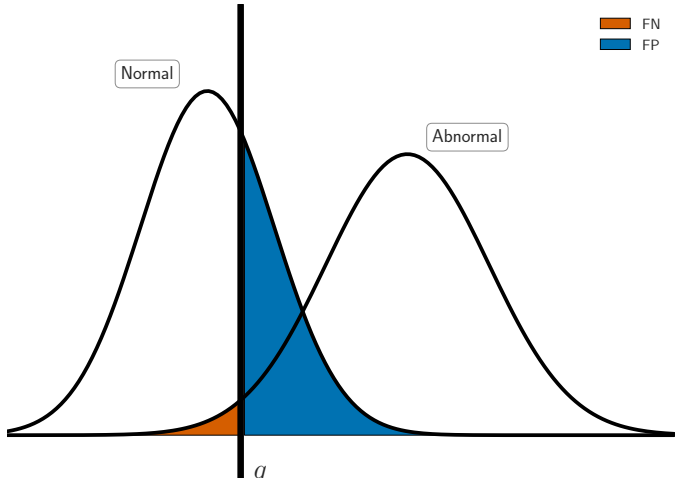
ROC Curve



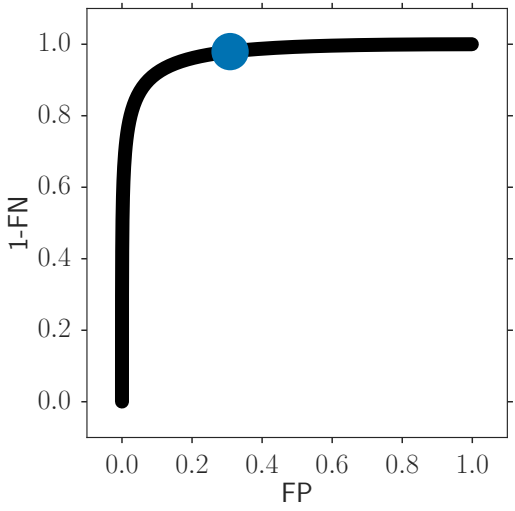
ROC Curve



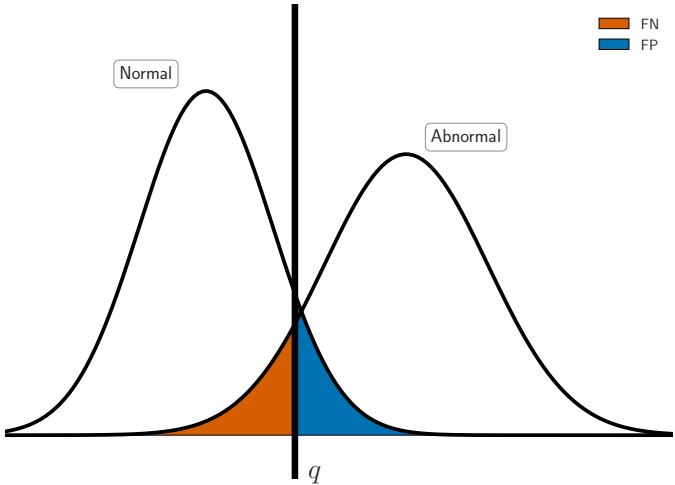
ROC Curve



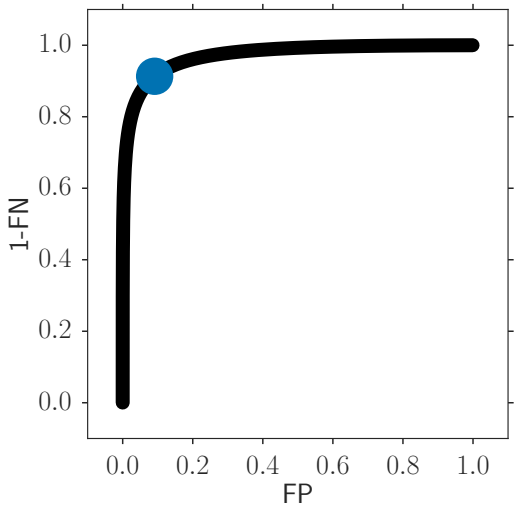
ROC Curve



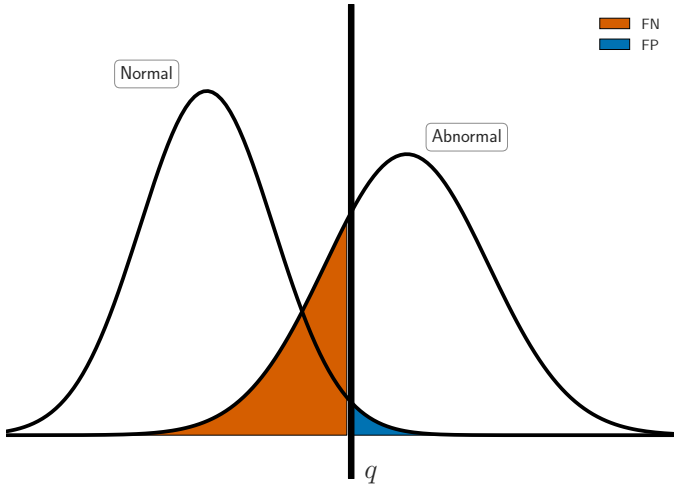
ROC Curve



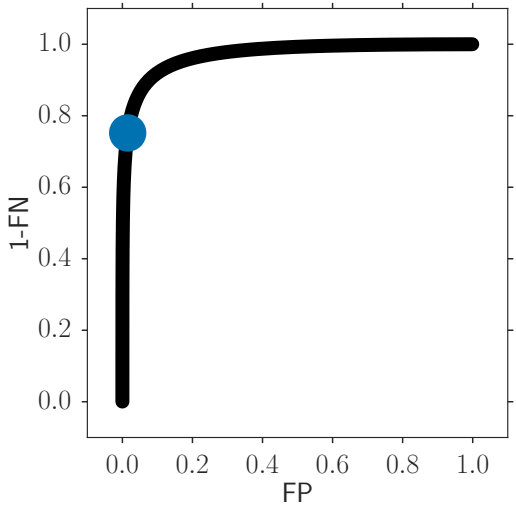
ROC Curve



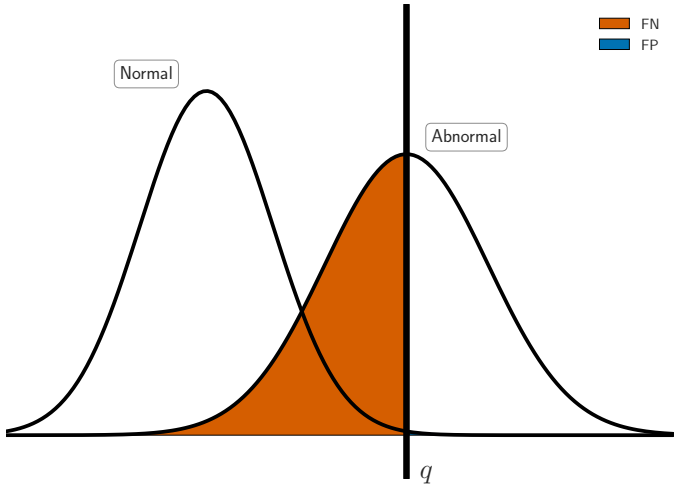
ROC Curve



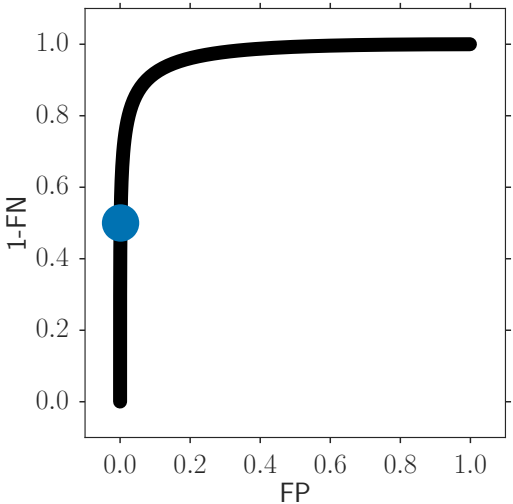
ROC Curve



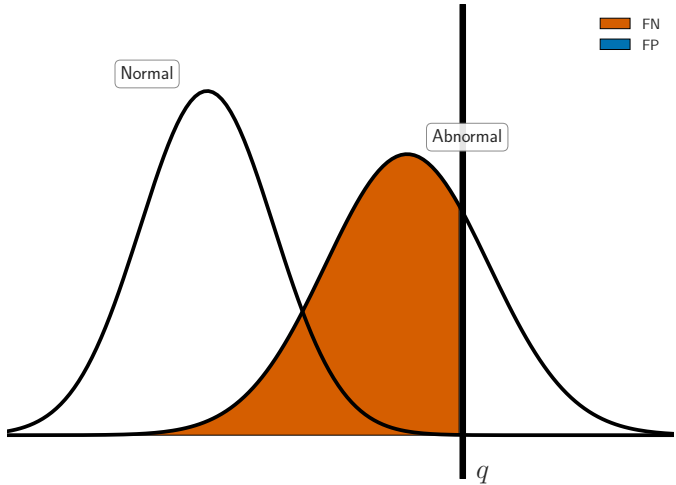
ROC Curve



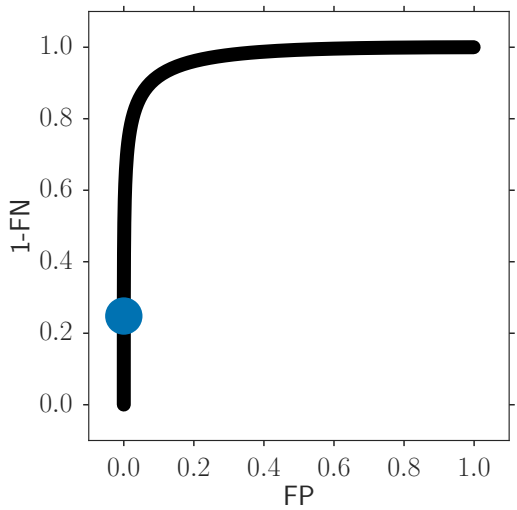
ROC Curve



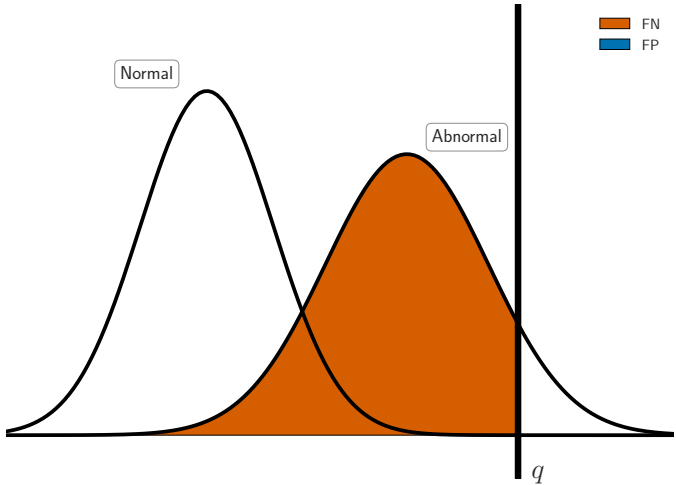
ROC Curve



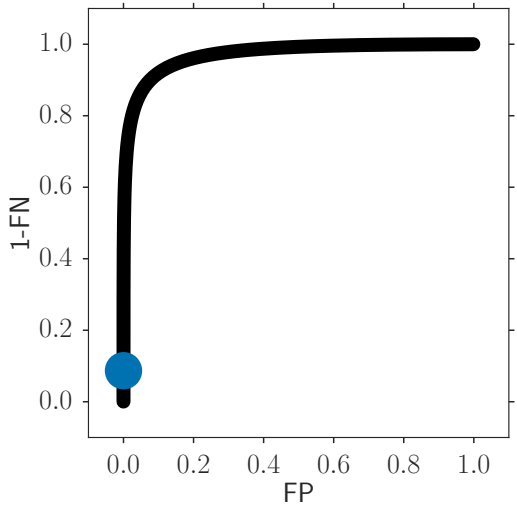
ROC Curve



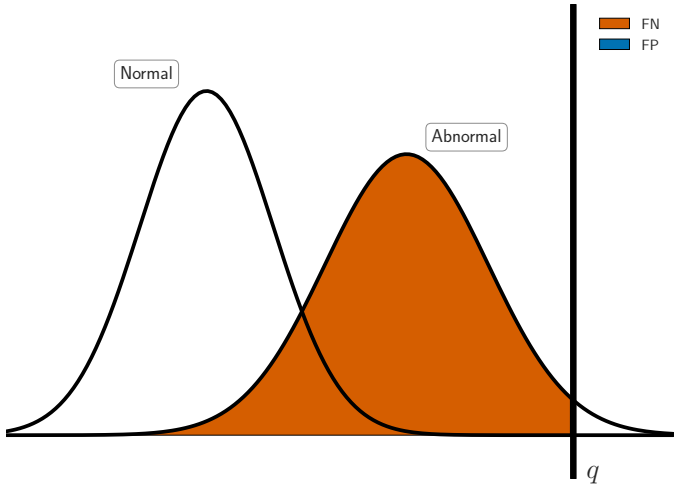
ROC Curve



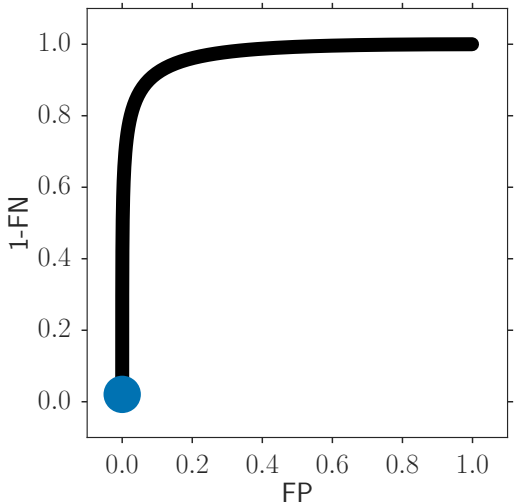
ROC Curve



ROC Curve

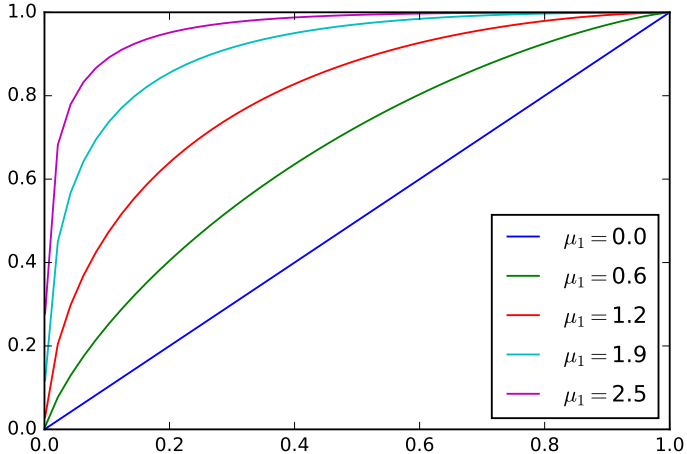


ROC Curve



ROC curves for bi-normal case

- F and G are Gaussian with parameter μ_0, σ_0 and μ_1, σ_1, respectively.
- Here $\mu_0 = 0$, $\sigma_0 = \sigma_1 = 1$, and μ_1 varies



Estimation–application

ROC curve estimation

- Maximum likelihood
- Non-parametric
- Bayesian with latent variables
- Estimation of the area under the ROC curve (AUC)

Application

- To compare different statistic tests
- To compare different (supervised) learning algorithm
- To compare variable selection methods (*e.g.* Lasso, OMP, etc.)

nb : ROC = Receiver Operating Characteristic