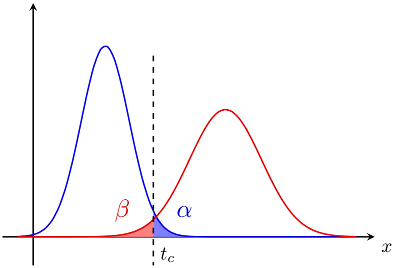
Hypothesis testing

EE209 - Eléments de Statistiques pour les Data Sciences

Telling apart two distributions based on an observation



Let's assume that X can follow two distributions:

- Under the *null hypothesis* $H_0 : X \sim \mathcal{N}(\mu_0, \sigma_0^2)$
- Under the *alternate hypothesis* $H_1 : X \sim \mathcal{N}(\mu_1, \sigma_1^2)$,

with $\mu_0 < \mu_1$ and σ_0, σ_1 not too large.

Can we try to decide based on an observation x of X which hypothesis is the correct one?

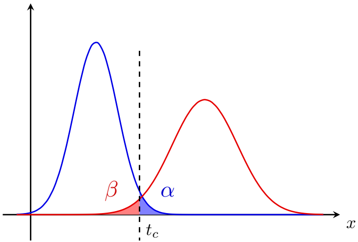
We can choose a *critical value* t_c on the value x, and

- if $x \leq t_c$, decide that H_0 is correct
- if $x > t_c$, decide that H_1 is correct

We can define

- $\alpha = \mathbb{P}_0(X > t_c)$ where $\mathbb{P}_0$ is "the probability **if** H_0 is true"
- $\beta = \mathbb{P}_1(X \leq t_c)$ where $\mathbb{P}_1$ is "the probability **if** H_1 is true"

Types of errors



With the same setting as on previous slide, let's denote by Δ our decision with

- $\Delta = 0$ if we decide that H_0 is correct
- $\Delta = 1$ if we decide that H_1 is correct

We have:

- $\{\Delta = 0\} = \{X \leq t_c\}$ and
- $\{\Delta = 1\} = \{X > t_c\}$

	$\Delta = 0$	$\Delta = 1$
H_0	☺	Type I-error ☹
H_1	Type II-error ☹	☺

$$\mathbb{P}_0(\Delta = 1) = \mathbb{P}_0(X > t_c) = \alpha$$

$$\mathbb{P}_1(\Delta = 0) = \mathbb{P}_0(X \leq t_c) = \beta$$

The probabilities of the configurations are

	$\Delta = 0$	$\Delta = 1$
H_0	$1 - \alpha$	α
H_1	β	$1 - \beta$

Telling apart two distributions based on a sample

Let's assume that $X_1, \ldots, X_n$ are i.i.d. but can follow two distributions:

- Hypothesis $H_0 : X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_0, \sigma_0^2)$
- Hypothesis $H_1 : X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$,

Can we try to decide based on a sample $x_1, \ldots, x_n$ which hypothesis is the correct one?

We can for example compute $\bar{x}$ and use the fact that

- Hypothesis $H_0 : \bar{X} \sim \mathcal{N}(\mu_0, \frac{\sigma_0^2}{n})$
- Hypothesis $H_1 : \bar{X} \sim \mathcal{N}(\mu_1, \frac{\sigma_1^2}{n})$,

Since the variance decrease with n, with a well chosen t_c, the probability of error should decrease with n.

Testing an alternative with one hypothesis to privilege by default

When deciding between hypotheses, the situation is very often asymmetric: there is one hypothesis which should be privileged by default.

Ham vs spam.
If a spam filter has to decide between two hypotheses

- This email is valid correspondence ("ham")
- This email is spam

it is much worse to classify ham as spam than the opposite.

By default we would rather consider that a mail is ham. This will be the null hypothesis, H_0.

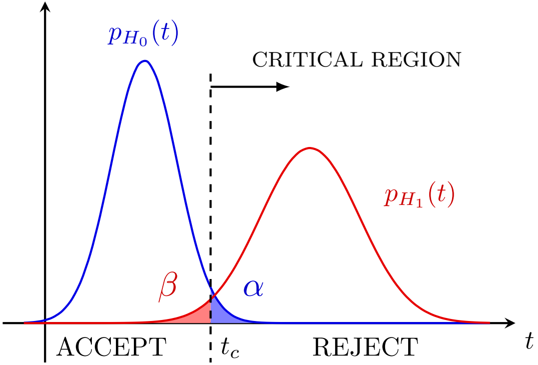
Tumor vs not.
If the result of an analysis based on a radio or a CT-SCAN has to detect the presence of a tumor, it is much worse to fail to detect an existing tumor than to detect something which will turn out later not to be. So here the null hypothesis H_0 will be "there is a tumor."

The Neyman-Pearson hypothesis testing framework

We assume that

- the data follows a distribution $p(\cdot;\theta)$ from a statistical model parameterized by $\theta \in \Theta$.
- Under the null hypothesis $\theta \in H_0 \subset \Theta$, and under the alternate hypothesis, $\theta \in H_1 \subset \Theta$.
- $H_0 \cap H_1 = \varnothing$.
- We assume that there is a *statistic* of the data $T = T(X_1, \ldots, X_n)$ which tends to be small under H_0 and larger under H_1
- The null hypothesis H_0 is privileged by default
- Our priority is to make sure that the Type-I error $\alpha = \mathbb{P}_0(\Delta = 1)$ is low.
- We will thus choose the *critical value* t_c on T to guarantee that α is low.

The Neyman-Pearson testing framework: vocabulary



- α is the significance level (Type-I error)
- $1-\alpha$ is the confidence level
- β is the Type-II error level
- $1-\beta$ is the power

- We will decide that H_1 is correct (i.e. set $\Delta = 1$) typically if $T \in [t_c, +\infty)$ which is called the *critical region* of the test. This set can take other forms.
- if $\Delta = 1$ we say that "we reject the null hypothesis" and that the result is "statistically significant.

One-sided Gaussian test

We assume that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ where σ is known.

- We consider the simple alternative $H_0 : \mu = \mu_0$ *vs* $H_1 : \mu = \mu_1$ with $\mu_1 > \mu_0$.
- We consider the *test statistic*

$$T(X_1, \ldots, X_n) = T := \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

We have $T \overset{H_0}{\sim} \mathcal{N}(0,1)$, so $\mathbb{P}_0(T > z_{1-\alpha}) = \alpha$ and we can choose $t_c = z_{1-\alpha}$ to control the type-I error.

We will reject the null hypothesis if T falls in the *critical region* $[z_{1-\alpha}, +\infty)$. In that case, we also say that $\bar{X}$ is *significantly larger* than μ_0.

We have $T - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \overset{H_1}{\sim} \mathcal{N}(0,1)$, so

$$\beta = \mathbb{P}_1(T \leq t_c) = \mathbb{P}_1\Big(T - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \leq t_c - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\Big) = \Phi\big(-\sqrt{n}\big(\tfrac{\mu_1 - \mu_0}{\sigma}\big) + t_c\big)$$

Simple hypothesis vs composite hypothesis

A simple hypothesis is a hypothesis $H_k = \{\theta_k\}$ which specifies a single value for θ. A non-simple hypothesis is called a *composite* hypothesis.

Simple alternative

$$H_0 : \theta = \theta_0 \qquad \text{vs} \qquad H_1 : \theta = \theta_1.$$

Composite alternative

Assuming that $\theta \in \mathbb{R}$,

$$H_0 : \theta = \theta_0 \qquad \text{vs} \qquad H_1 : \theta > \theta_0.$$

$$H_0 : \theta = \theta_0 \qquad \text{vs} \qquad H_1 : \theta < \theta_0.$$

$$H_0 : \theta = \theta_0 \qquad \text{vs} \qquad H_1 : \theta \neq \theta_0.$$

Other alternatives leading to the one-sided Gaussian test

$H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$

Given that t_c is only determined by the distribution under H_0, the *critical region* is again $[z_{1-\alpha}, +\infty)$

$H_0 : \mu = \mu_0$ vs $H_1 : \mu = \mu_1$ with $\mu_1 < \mu_0$.

In this case, we can reject if T is lower than an *critical value* such that $\mathbb{P}_0(T < t_c) = \alpha$, which entails $t_c = z_\alpha$. Of course, the *critical region* is now $(-\infty, z_\alpha]$.

$H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$

Given that t_c is only determined by the distribution under H_0, this case is the same as the case just before for the determination of the *critical region*.

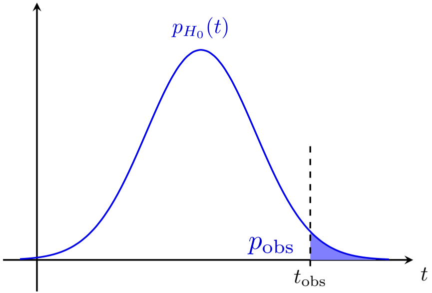
p-value

One limitation of the test methodology that we have to choose a *significance level* α. I could be useful to report a value such that one can easily assess whether the test would be rejected at other levels and which would directly measure the significance of the value t_{obs}.

p-value definition

If t_{obs} is the observed value of the *test statistic* T then the associated p-value is

$$p_{\text{obs}} = \mathbb{P}_0(T \geq t_{\text{obs}}).$$



Interpretations of the p-value

The p-value is

- the probability to observe a more extreme value of T than t_{obs} under H_0.
- the smallest significance level such that the null would be rejected for $T = t_{\text{obs}}$.
- the significance level of the test with $t_c = t_{\text{obs}}$.
- a measure of significance of the test statistic value t_{obs}.

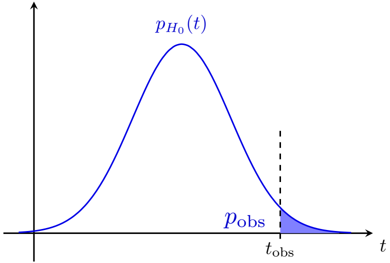
Test decision in terms of the p-value

By definition H_0 is rejected iff $(t_{\text{obs}} > t_c) \Leftrightarrow (p_{\text{obs}} < \alpha)$.

Example: p-value for a one-sided Gaussian test.

We have $T \overset{H_0}{\sim} \mathcal{N}(0,1)$, so

$$p_{\text{obs}} = \mathbb{P}_0(T \geq t_{\text{obs}}) = 1 - \Phi(t_{\text{obs}}).$$



Two-sided Gaussian test

We assume that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ where σ is known.

- We consider the composite alternative $H_0 : \mu = \mu_0$ *vs* $H_1 : \mu \neq \mu_0$.
- We consider the *test statistic*

$$|T(X_1, \ldots, X_n)| = |T| := \left| \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \right|.$$

We have $T \overset{H_0}{\sim} \mathcal{N}(0,1)$, so $\mathbb{P}_0\big(|T| > z_{1-\alpha/2}\big) = 1 - \mathbb{P}_0\big(z_{\alpha/2} \leq T \leq z_{1-\alpha/2}\big) = \alpha$ and we can choose $t_c = z_{1-\alpha/2}$ to control the type-I error.

In case of rejection of the null hypothesis, we say that $\bar{X}$ is *significantly different* from μ_0.

The p-value is $p_{\text{obs}} = \mathbb{P}_0(|T| \geq |t_{\text{obs}}|) = 2\big(1 - \Phi(|t_{\text{obs}}|)\big)$.

Relationship between Gaussian confidence intervals and Gaussian tests

Two-sided test:

The null hypothesis is **not** rejected iff as $|T| \leq t_c$ but

$$-t_c \leq \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq t_c \quad \Leftrightarrow \quad \bar{X} - t_c \frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + t_c \frac{\sigma}{\sqrt{n}}.$$

But $t_c = z_{1-\alpha/2}$, so the null hypothesis is rejected at the level of significance α iff

$$\mu_0 \notin \left[\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \ \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right].$$

In other words:

The hypothesis that $\mu = \mu_0$ is rejected at a level of significance α
if and only if
μ_0 is not inside the (symmetric) Gaussian confidence interval of level $1 - \alpha$.

Relationship between Gaussian confidence intervals and Gaussian tests

One-sided test:

The null hypothesis is **not** rejected iff as $T \leq t_c$ but

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq t_c \quad \Leftrightarrow \quad \bar{X} \leq \mu_0 + t_c \frac{\sigma}{\sqrt{n}} \quad \Leftrightarrow \quad \bar{X} - t_c \frac{\sigma}{\sqrt{n}} \leq \mu_0.$$

But $t_c = z_{1-\alpha}$, so the null hypothesis is rejected at the level of significance α iff

$$\mu_0 \notin \left[\bar{X} - z_{1-\alpha} \frac{\sigma}{\sqrt{n}}, +\infty\right).$$

In other words:

The hypothesis that $\mu = \mu_0$ is rejected at a level of significance α
if and only if
μ_0 is not inside the semi-infinite upper Gaussian confidence interval of level $1 - \alpha$.

One-sided Student test

We assume that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ where σ is **unknown**.

- We consider the simple alternative $H_0 : \mu = \mu_0$ *vs* $H_1 : \mu = \mu_1$ with $\mu_1 > \mu_0$.
- We consider the *test statistic*

$$T(X_1, \ldots, X_n) = T := \frac{\bar{X} - \mu_0}{S/\sqrt{n}}.$$

We have $T \overset{H_0}{\sim} \mathsf{St}_{n-1}$, so $\mathbb{P}_0(T > t_{1-\alpha}^{(n-1)}) = \alpha$ and we can choose $t_c = t_{1-\alpha}^{(n-1)}$ to control the type-I error.

We have $T - \frac{\mu_1 - \mu_0}{S/\sqrt{n}} \overset{H_1}{\sim} \mathsf{St}_{n-1}$, so

$$\beta = \mathbb{P}_1(T \le t_c) = \mathbb{P}_1\Big(T - \frac{\mu_1 - \mu_0}{S/\sqrt{n}} \le t_c - \frac{\mu_1 - \mu_0}{S/\sqrt{n}}\Big) = F_{\mathsf{St}_{n-1}}\big(- \sqrt{n}\big(\tfrac{\mu_1 - \mu_0}{S}\big) + t_c\big).$$

The p-value is $p_{\text{obs}} = \mathbb{P}_0(T \ge t_{\text{obs}}) = \big(1 - F_{\mathsf{St}_{n-1}}(t_{\text{obs}})\big).$

Two-sided Student test

We assume that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ where σ is **unknown**.

- We consider the composite alternative $H_0 : \mu = \mu_0$ *vs* $H_1 : \mu \neq \mu_0$.
- We consider the *test statistic*

$$|T(X_1, \ldots, X_n)| = |T| := \left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right|.$$

We have $T \overset{H_0}{\sim} \mathsf{St}_{n-2}$, so $\mathbb{P}_0\big(|T| > t^{(n-1)}_{1-\alpha/2}\big) = 1 - \mathbb{P}_0\big(t^{(n-1)}_{\alpha/2} \leq T \leq t^{(n-1)}_{1-\alpha/2}\big) = \alpha$ and we can choose $t_c = t^{(n-1)}_{1-\alpha/2}$ to control the type-I error.

The p-value is $p_{\text{obs}} = \mathbb{P}_0(|T| \geq |t_{\text{obs}}|) = 2\big(1 - F_{\mathsf{St}_{n-1}}(|t_{\text{obs}}|)\big)$

One-sided asymptotic Gaussian test

We assume that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$ where P is unknown, but we assume that $\mathbb{E}[X_1^2] < \infty$.

- We consider the simple alternative $H_0 : \mu = \mu_0$ *vs* $H_1 : \mu = \mu_1$ with $\mu_1 > \mu_0$.
- We consider the *test statistic*

$$T(X_1, \ldots, X_n) = T := \frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{n}},$$

where $\hat{\sigma}$ is a consistent estimator of σ, like S for example.

By the CLT, under H_0, $T \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$, so $\mathbb{P}_0(T > z_{1-\alpha}) \xrightarrow[n\to\infty]{} \alpha$ and we can choose $t_c = z_{1-\alpha}$ to *asymptotically* control the type-I error.

Symmetrically, under H_1, $T - \frac{\mu_1 - \mu_0}{\hat{\sigma}/\sqrt{n}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$, so

$$\beta = \mathbb{P}_1(T \le t_c) = \mathbb{P}_1\Big(T - \frac{\mu_1 - \mu_0}{\hat{\sigma}/\sqrt{n}} \le t_c - \frac{\mu_1 - \mu_0}{\hat{\sigma}/\sqrt{n}}\Big) \approx \Phi\big(-\sqrt{n}\big(\tfrac{\mu_1 - \mu_0}{\hat{\sigma}}\big) + t_c\big).$$

We can define similarly two-sided asymptotic Gaussian tests.

One-sided χ^2 test for the variance σ^2

We assume that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ where μ is **unknown**.

- We consider the simple alternative $H_0 : \sigma = \sigma_0$ *vs* $H_1 : \sigma = \sigma_1$ with $\sigma_1 > \sigma_0$.
- We consider the *test statistic*

$$T(X_1, \ldots, X_n) = T := (n-1)\frac{S^2}{\sigma_0^2} = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

We have $T \overset{H_0}{\sim} \chi^2_{n-1}$, so if $\chi^2_{n-1,1-\alpha}$ is the quantile of level $1-\alpha$ of a χ^2_{n-1} distribution,$\mathbb{P}_0(T > \chi^2_{n-1,1-\alpha}) = \alpha$ and we can choose $t_c = \chi^2_{n-1,1-\alpha}$ to control the type-I error.

We have $\frac{\sigma_0^2}{\sigma_1^2} T \overset{H_1}{\sim} \chi^2_{n-1}$, so $\beta = \mathbb{P}_1(T \le t_c) = \mathbb{P}_1\Big(\frac{\sigma_0^2}{\sigma_1^2} T \le \frac{\sigma_0^2}{\sigma_1^2} t_c\Big) = F_{\chi^2_{n-1}}\Big(\frac{\sigma_0^2}{\sigma_1^2} t_c\Big)$.

The p-value is $p_{\text{obs}} = \mathbb{P}_0(T \ge t_{\text{obs}}) = \big(1 - F_{\chi^2_{n-1}}(t_{\text{obs}})\big)$, with $F_{\chi^2_{n-1}}$ the cdf of a χ^2_{n-1} r.v.
We could define similarly a two-sided χ^2 test.

Two-sided Wald test

We assume that $\hat{\theta}$ is the MLE for the parameter θ based on an i.i.d. sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p(\cdot; \eta)$ with $\theta = \psi(\eta)$. We consider the log-likelihood $\ell(\theta)$, the Fisher information matrix $I(\theta)$.

- We consider the composite alternative $H_0 : \theta = \theta_0$ *vs* $H_1 : \theta \neq \theta_0$.
- We consider the *test statistic* $|T|$ with

$$T(X_1, \ldots, X_n) = T := \sqrt{I(\hat{\theta})}(\hat{\theta} - \theta_0).$$

By the CLT with Slutsky, under H_0, $T \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$, so $\mathbb{P}_0(|T| > z_{1-\alpha/2}) \xrightarrow[n\to\infty]{} 1 - \alpha$ and we can choose $t_c = z_{1-\alpha/2}$ to *asymptotically* control the type-I error.

We can define an asymptotic p-value $\mathbb{P}_0(|T| > |t_{\text{obs}}|) = 2(1 - \Phi(|t_{\text{obs}}|))$.

Summary

- In the Neyman-Pearson framework a null hypothesis H_0 is the default hypothesis.
- We can *reject* the null hypothesis in favor of an alternative if the value of a *test statistics* is larger than a *critical value*.
- We focus on controling the Type-I error level α, aka the *significance level*.
- Instead of setting the *critical value* based on a *significance level*.
- The p-value p_{obs} is the probability $\mathbb{P}_0(T \geq t_{\mathsf{obs}})$.
- It is possible to construct one and two-sided Gaussian and Student tests.
- It is possible to construct asymptotic Gaussian tests.
- One form of asymptotic test for $\hat{\theta}_{\mathsf{MLE}}$ is the Wald test.
- The null is rejected at the confidence level α in a two-sided test iff the parameter μ_0 or θ_0 is not in the corresponding (symmetric) confidence interval.
- The same holds for one-sided test, but with one-sided confidence intervals.