

Modèles stochastiques II

INFO 154

Gianluca Bontempi

Département d'Informatique
Boulevard de Triomphe - CP 212
<http://www.ulb.ac.be/di>

Testing hypothesis

- **Hypothesis testing** is the second major area of statistical inference.
- A **statistical hypothesis** is an assertion or conjecture about the distribution of one or more random variables.
- A **test** of a statistical hypothesis is a rule or procedure for deciding whether to reject the assertion on the basis of the observed data.
- The basic idea is formulate some statistical hypothesis and look to see if the data provides any evidence to reject the hypothesis.

An hypothesis testing problem

- Consider the model of the traffic in the boulevard.
 - Suppose that the measures of the inter-arrival times are $D_N = \{10, 11, 1, 21, 2, \dots\}$ seconds.
 - Can we say that the mean inter-arrival time θ is different from 10?
- Consider the grades of two different school sections.
 - Section A had $\{15, 10, 12, 19, 5, 7\}$.
 - Section B had $\{14, 11, 11, 12, 6, 7\}$.
 - Can we say that Section A had better grades than Section B?

A **statistical test** is a procedure that aims to answer such questions.

Types of hypothesis

We start by declaring the **working (basic, null) hypothesis** H to be tested, in the form $\theta = \theta_0$ or $\theta \in \omega \subset \Theta$.

The hypothesis can be

Simple. It fully specifies the distribution of \mathbf{z} .

Composite. It partially specifies the distribution of \mathbf{z} .

Example: if D_N constitutes a random sample of size N from $\mathcal{N}(\mu, \sigma^2)$ the hypothesis $H : \mu = \mu_0, \sigma = \sigma_0$, (with μ_0 and σ_0 known values) is simple while the hypothesis $H : \mu = \mu_0$ is composite since it leaves open the value of σ in $(0, \infty)$.

Types of statistical test

Suppose we have collected N samples $D_N = \{z_1, \dots, z_N\}$ from a distribution F_z and we have declared a null hypothesis H about F . Three are the most common types of statistical test:

Pure significance test: data D_N are used to assess the inferential evidence against H .

Significance test: the inferential evidence against H is used to judge whether H is inappropriate. In other words it is a rule for rejecting H .

Hypothesis test: data D_N are used to assess the hypothesis H against a specific **alternative hypothesis** \bar{H} . In other words this is a rule for rejecting H in favour of \bar{H} .

Pure significance test

- Suppose that the null hypothesis H is simple.
- Let $t(\mathbf{D}_N)$ be a statistic such that the larger its value the more it casts doubt on H .
- The quantity $t(\mathbf{D}_N)$ is called **test statistic** or **discrepancy measure**.
- Let $t_N = t(D_N)$ the value of t calculated on the basis of the sample data D_N .
- Let us consider the quantity

$$p = \text{Prob} \{t(\mathbf{D}_N) > t_N | H\}$$

- If p is small the sample data D_N are highly inconsistent with H and p (**significance probability** or **significance level**) is the measure of such inconsistency.

Some considerations

- p is the proportion of situations identical to the one who generated the sample data where we would observe a degree of inconsistency at least to the extent represented by t_N .
- t_N is the observed value of the statistic for a given D_N . Different D_N yield different values of $p \in (0, 1)$.
- it is essential that the distribution of $t(\mathbf{D}_N)$ under H is known.
- **We cannot say that p is the probability that H is true but better that p is the probability that the dataset D_N is observed given that H is true**
- Open issues
 1. What if H is composite?
 2. how to choose $t(\mathbf{D}_N)$.

Tests of significance

- Suppose that the value p is known. If p is small either a rare event has occurred or perhaps H is not true.
- Idea: if p is less than some stated value α , we reject H .
- We choose a **critical level** α , we observe D_N and we reject H at level α if

$$P\{t(\mathbf{D}_N) > t_N | H\} \leq \alpha$$

- This is equivalent to choose some **critical value** t_α and we reject H if $t_N > t_\alpha$.
- We obtain two regions in the space of sample data:
critical region S_0 where if $D_N \in S_0$ we reject H .
non-critical region S_1 where the sample data D_N gives us no-reason to reject H on the basis of the level- α test.

Some considerations

- The principle is that we will accept H unless we witness some event that has sufficiently small probability of arising when H is true.
- If H were true we could still obtain data in S_0 and consequently wrongly reject H with probability

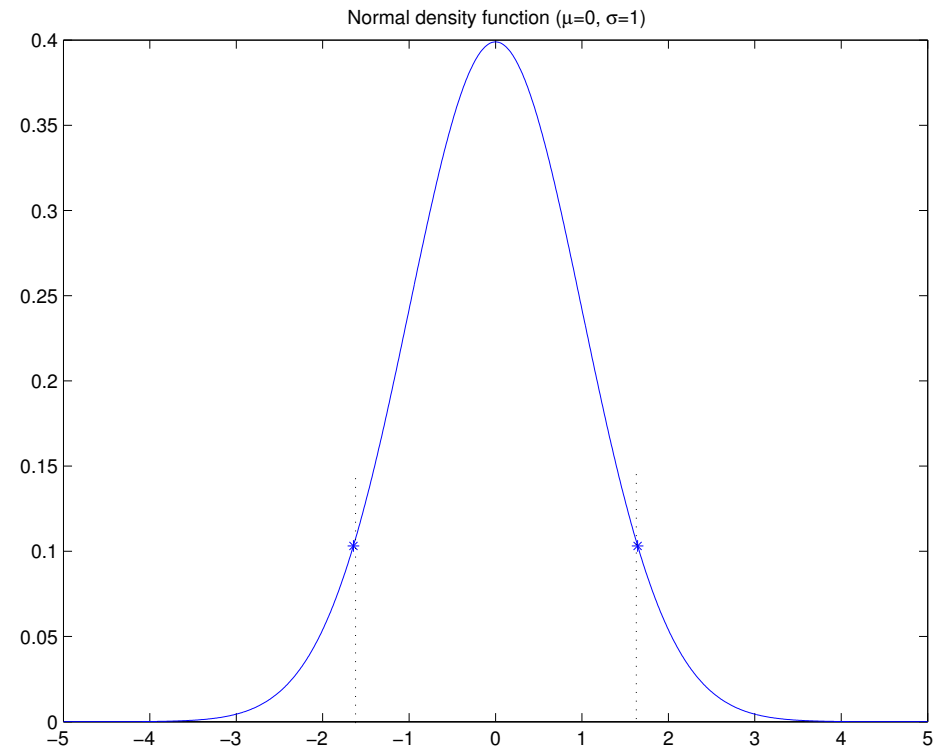
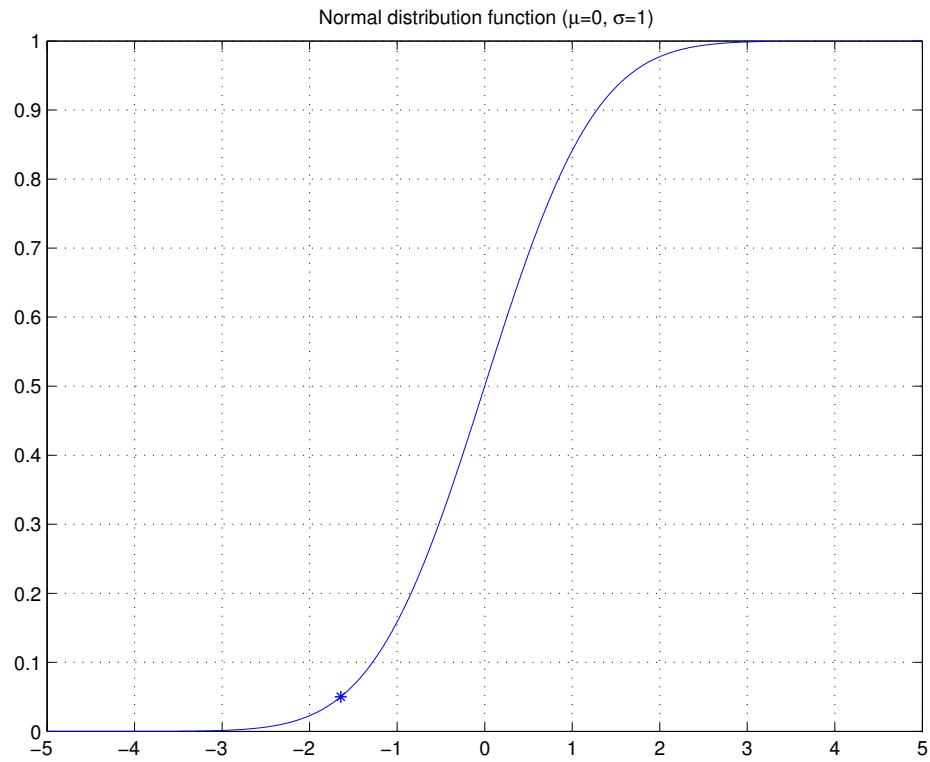
$$\text{Prob} \{ \mathbf{D}_N \in S_0 \} = \text{Prob} \{ t(\mathbf{D}_N) > t_\alpha | H \} \leq \alpha$$

- The significance level α provides an upper bound to the **maximum probability** of incorrectly rejecting H .
- Other interesting quantity is the **critical level** or **p-value** of the test

$$p_v = \text{Prob} \{ t(\mathbf{D}_N) \geq t_N | H \} = F_{t(\mathbf{D}_N)}(t_N | H)$$

This is the probability that the t-statistic is more extreme than its observed value.

Standard normal distribution



TP: example

- Let D_N consist of N independent observations of $\mathbf{x} \sim \mathcal{N}(\mu, \sigma^2)$, with known variance σ^2 .
- We want to test the hypothesis $H : \mu = \mu_0$ with μ_0 known.
- Consider as test statistic $t(\mathbf{D}_N)$, the quantity $|\hat{\mu} - \mu_0|$ where $\hat{\mu}$ is the sample average estimator . If H is true we know that $\hat{\mu} \sim \mathcal{N}(\mu_0, \sigma^2/N)$.
- Let us calculate the value $t(D_N) = |\hat{\mu} - \mu_0|$.
- Let us put a significance level $\alpha = 10\%$. This means that

$$\begin{aligned} \text{Prob} \{t(\mathbf{D}_N) > t_\alpha | H\} &= \text{Prob} \{|\hat{\mu} - \mu_0| > t_\alpha | H\} = \\ &= \text{Prob} \{(\hat{\mu} - \mu_0 > t_\alpha) \ \& \ (\hat{\mu} - \mu_0 < -t_\alpha) | H\} = 0.1 \end{aligned}$$

TP: example (II)

- For a standard normal variable $z = \frac{x-\mu}{\sigma}$

$$\text{Prob} \{z > 1.645\} = F_z(1.645) = 0.05 = z_{0.05} = z_{\alpha/2}$$

and consequently

$$\text{Prob} \{z > 1.645 \quad \& \quad z < -1.645\} = 0.05 + 0.05 = 0.1$$

- It follows that

$$t_\alpha = 1.645\sigma/\sqrt{N}$$

and that the critical region is

$$S_0 = \left\{ D_N : |\hat{\mu} - \mu_0| > 1.645\sigma/\sqrt{N} \right\}$$

TP: example (III)

- Suppose that $\sigma = 0.1$ and that we want to test if $\mu = \mu_0 = 10$ with a significance level 10%.
- After $N = 6$ observations we have $D_N = \{10, 11, 12, 13, 14, 15\}$.
- On the basis of the dataset we compute

$$\hat{\mu} = \frac{10 + 11 + 12 + 13 + 14 + 15}{6} = 12.5$$

and

$$t(D_N) = |\hat{\mu} - \mu_0| = 2.5$$

- Since $t_\alpha = 1.645 * 0.1 / \sqrt{6} = 0.0672$, and $t(D_N) > t_\alpha$, the observations D_N are in the critical region.
- **The hypothesis is rejected.**

Types of error

Type I error. It is the error we make when **we reject H if it is true.**

Significance level represents the probability of making the type I error.

Type II error. It is the error we make when **we accept H if it is false.**

In order to define this error, we are forced to declare an alternative hypothesis \bar{H} as a formal definition of what is meant by H being “false”.

The probability of type II error is the probability that the test leads to acceptance of H when in fact \bar{H} prevails.

An analogy

- Consider the analogy with a murder trial, where we have as suspect Mr. Bean.
- The null hypothesis H is “Mr. Bean is innocent”.
- The dataset is the amount of evidence collected by the police against Mr. Bean.
- The Type I error is the error that we make if, being Mr. Bean innocent, we send him to penalty death.
- The Type II error is the error that we make if, being Mr. Bean guilty, we acquit him.

Some considerations

- Suppose we have some data $\{z_1, \dots, z_N\} \sim F$ from a distribution F .
- H and \bar{H} represent two hypotheses about F .
- On the basis of the data, one is **accepted** and one is **rejected**.
- Note that the two hypotheses have different philosophical status (asymmetry).
- H is a conservative hypothesis, not to be rejected unless evidence is clear. This means that a type I error is **more serious** than a type II error (*benefit of the doubt*).
- It is often assumed that F belongs to a parametric family $F(z, \theta)$. The test on F becomes a test on θ .
- A particular example of hypothesis test is the **goodness of fit test** where we test $H : F = F_0$ against $\bar{H} : F \neq F_0$.

Parametric hypothesis tests

The **working hypothesis** H and an **alternative hypothesis** \bar{H} are declared.

The space Θ of parameters θ is partitioned in two complementary subspaces:

1. $H : \theta \in \Theta_H$
2. $\bar{H} : \theta \in \Theta_{\bar{H}} = \Theta - \Theta_H$

The sample space \mathcal{D} is partitioned by the test procedure into two complementary subspaces

1. **critical region** S_0 , i.e. reject H and accept \bar{H} if $D_N \in S_0$
2. **non critical region** S_1 , i.e. accept H if $D_N \in S_1 = \mathcal{D} - S_0$

Choice of test

The choice of test and consequently the choice of the partition $\{S_0, S_1\}$ is based on two steps

1. Define a significance level α , that is the probability of type I error

$$\text{Prob} \{ \mathbf{D}_N \in S_0 | H \} \leq \alpha$$

that is the probability of incorrectly rejecting H

2. Among the set of tests $\{S_0, S_1\}$ of level α , choose the test that minimizes the probability of type II error

$$\text{Prob} \{ \mathbf{D}_N \in S_1 | \bar{H} \}$$

that is the probability of incorrectly accepting H . This is equivalent to look for maximizing the **power of the test**

$$\text{Prob} \{ \mathbf{D}_N \in S_0 | \bar{H} \} = 1 - \text{Prob} \{ \mathbf{D}_N \in S_0 | \bar{H} \}$$

which is the probability of *correctly* rejecting H .

TP example

- Consider a r.v. $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$, where σ is known and a set of N iid observations are given.
- We want to test the null hypothesis $\mu = \mu_0 = 0$, with $\alpha = 0.1$
- Consider the 3 critical regions S_0
 1. $|\hat{\mu} - \mu_0| > 1.645\sigma/\sqrt{N}$
 2. $\hat{\mu} - \mu_0 > 1.282\sigma/\sqrt{N}$
 3. $|\hat{\mu} - \mu_0| < 0.126\sigma/\sqrt{N}$
- For all these tests $\text{Prob}\{\mathbf{D}_N \in S_0 | H\} \leq \alpha$, hence the significance level is the same.
- However if $\bar{H} : \mu = 10$ the type II error of the three tests is significantly different.
- What is the best one?

UMP level- α test

Given a significance level α we denote by **uniformly most powerful (UMP)** test, the test which

1. satisfies $P(\mathbf{D}_N \in S_0 | H) \leq \alpha$ and
2. for which $P(\mathbf{D}_N \in S_0 | \bar{H})$ is maximized simultaneously for all $\theta \in \Theta_{\bar{H}}$.

Open issue: how is it possible to find UMP tests?

An answer to the simplest case is given by the Neyman-Pearson lemma.

Likelihood ratio test

- Consider the simplest case $\Theta = \{\theta_0, \theta_1\}$, where $H : \theta = \theta_0$ and $\bar{H} : \theta = \theta_1$.
- Consider the two likelihoods $L_0(\theta)$ and $L_1(\theta)$.
- The idea of Neyman and Pearson was to base the acceptance/rejection of H on the relative values $L_0(\theta_0)$ and $L_1(\theta_1)$. In other terms we reject H if the **likelihood ratio**

$$\frac{L_0(\theta_0)}{L_1(\theta_1)}$$

is sufficiently small.

- We reject H only if the sample data D_N are sufficiently more probable when $\theta = \theta_1$ than when $\theta = \theta_0$.

Neyman-Pearson lemma

Lemma 1. Let $H : \theta = \theta_0$ and $\bar{H} : \theta = \theta_1$. If a partition $\{S_0, S_1\}$ of the sample space \mathcal{D} is defined by

$$S_0 = \{D_N : L_1(\theta_1) > kL_0(\theta_0)\} \quad S_1 = \{D_N : L_1(\theta_1) < kL_0(\theta_0)\}$$

with $\int_{S_0} p(z, \theta_0) dz = \alpha$, then $\{S_0, S_1\}$ is the most powerful level- α test of H against \bar{H} .

- This lemma demonstrates that among all tests of size $\leq \alpha$, the likelihood ratio test is the optimal procedure, i.e. it has the smallest probability of type II error.
- However, in a general case, the definition of an optimum tests is very difficult.
- An example of optimal likelihood ratio test is the z -test.

z-test (one-tailed)

Consider a random sample from $\mathbf{x} \sim \mathcal{N}(\mu, \sigma^2)$ with μ unknown et σ^2 known. Consider the two **simple hypotheses**

$$H : \mu = \mu_0; \quad \bar{H} : \mu = \mu_1 > \mu_0$$

If H is true then the distribution of $\hat{\mu}$ is $\mathcal{N}(\mu_0, \sigma^2/N)$. This means that the variable \mathbf{z} is

$$\mathbf{z} = \frac{(\hat{\mu} - \mu_0)\sqrt{N}}{\sigma} \sim \mathcal{N}(0, 1)$$

It is convenient to rephrase the test in terms of \mathbf{z} . This means that the hypothesis H is rejected if $z_N > z_\alpha$ where

$$z_N = \frac{(\hat{\mu} - \mu_0)\sqrt{N}}{\sigma}$$

and z_α is such that $\text{Prob} \{ \mathcal{N}(0, 1) > z_\alpha \} = \alpha$.

Ex: for $\alpha = 0.05$ we would take $z_\alpha = 1.645$ since 5% of the standard normal distribution lies to the right of 1.645.

TP: example z-test

- Consider a r.v. $\mathbf{x} \sim \mathcal{N}(\mu, 1)$.
- We want to test $H : \mu = 5$ against $\bar{H} : \mu = 6$ with significance level 0.05.
- Suppose that the data is $D_N = \{5.1, 5.5, 4.9, 5.3\}$.
- Then $\hat{\mu} = 5.2$ and $z_N = (5.2 - 5) * 2/1 = 0.4$.
- Since this is less than 1.645, we do not reject the null hypothesis.

Generalized likelihood ratio tests

This is the extension of the Neyman-Pearson idea to all possible test problems.

Suppose we test $H : \theta \in \Theta_H$ against $\bar{H} : \theta \in \Theta_{\bar{H}} = \Theta - \Theta_H$.

Let

$$L_N(\hat{\theta}_{\text{ml}}) = \max_{\theta \in \Theta} L_N(\theta) \text{ and } L_N(\hat{\theta}_H) = \max_{\theta \in \Theta_H} L_N(\theta)$$

then the critical region for rejection of the *level- α likelihood ratio test* is

$$S_0 = \{D_N : L_N(\hat{\theta}_H)/L_N(\hat{\theta}_{\text{ml}}) < c\}$$

where c is constrained by the condition

$$P\{D_N \in S_0 | H\} \leq \alpha$$

Note that $\hat{\theta}_{\text{ml}}$ is the m.l. estimate of θ and $\hat{\theta}_H$ is the constrained maximum likelihood estimator when $\theta \in \Theta_H$.

Parametric tests

- The generalized principle makes possible the definition of a number of parametric tests.
- In the single test we consider a r.v. $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$.
- In the two samples test we consider 2 r.v. $\mathbf{z}_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathbf{z}_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$.

Name	single/two	known	H	\bar{H}
z-test	single	σ^2	$\mu = \mu_0$	$\mu \neq \mu_0$
z-test	two	$\sigma_1^2 = \sigma_2^2$	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$
χ^2 -test	single	μ	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$
t-test	single		$\mu = \mu_0$	$\mu \neq \mu_0$
t-test	two		$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$
χ^2 -test	single		$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$
F-test	two		$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$

Useful lemma

Lemma 2. Consider the r.v. $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$. Suppose that we have observed D_N . Let $L_N(\mu, \sigma)$ be the likelihood of the observed dataset. Then the following relations hold

1.

$$\max_{\mu} L_N(\mu, \sigma) = \frac{1}{(2\pi\sigma^2)^{-n/2}} \exp \left[-\sum_{i=1}^N (z_i - \hat{\mu})^2 / 2\sigma^2 \right]$$

2.

$$\max_{\sigma^2} L_N(\mu, \sigma) = \left[2\pi \frac{\sum_{i=1}^N (z_i - \mu)^2}{N} \right]^{-N/2} \exp[-N/2]$$

3.

$$\max_{\mu, \sigma^2} L_N(\mu, \sigma) = \left[2\pi \frac{\sum_{i=1}^N (z_i - \hat{\mu})^2}{N} \right]^{-N/2} \exp[-N/2]$$

These relations will be useful to derive the formulas of the parametric tests.

The chi-squared distribution (replay)

For a N positive integer, a r.v. \mathbf{z} has a χ_N^2 distribution if

$$\mathbf{z} = \mathbf{x}_1^2 + \cdots + \mathbf{x}_N^2$$

where $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ are i.i.d. random variables $\mathcal{N}(0, 1)$.

- The probability distribution is a gamma distribution with parameters $(\frac{1}{2}N, \frac{1}{2})$.
- $E[\mathbf{z}] = N$ and $\text{Var}[\mathbf{z}] = 2N$.
- The distribution is called “a chi-squared distribution with N degrees of freedom”.

Student's t -distribution (replay)

If $\mathbf{x} \sim \mathcal{N}(0, 1)$ and $\mathbf{y} \sim \chi_N^2$ are independent then the **Student's t -distribution** with N degrees of freedom is the distribution of the r.v.

$$\mathbf{z} = \frac{\mathbf{x}}{\sqrt{\mathbf{y}/N}}$$

We denote this with $\mathbf{z} \sim t_N$.

Property 1. If $\mathbf{z}_1, \dots, \mathbf{z}_N$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ then

$$\frac{\sqrt{N}(\hat{\boldsymbol{\mu}} - \mu)}{\sqrt{\hat{\mathbf{S}}\mathbf{S}/(N-1)}} = \frac{\sqrt{N}(\hat{\boldsymbol{\mu}} - \mu)}{\hat{\boldsymbol{\sigma}}} \sim t_{N-1}$$

Single sample: the two-sided t-test

Consider a random sample from $\mathcal{N}(\mu, \sigma^2)$ with σ^2 unknown. Let

$$H : \mu = \mu_0; \quad \bar{H} : \mu \neq \mu_0$$

We have

$$\hat{\mu}_{\text{ml}} = \hat{\mu} = \frac{\sum_{i=1}^N z_i}{N}, \quad \hat{\sigma}_{\text{ml}}^2 = \frac{\sum_{i=1}^N (z_i - \hat{\mu})^2}{N}$$

and the constrained m.l. values are

$$\hat{\mu}_H = \mu_0, \quad \hat{\sigma}_H^2 = \frac{\sum_{i=1}^N (z_i - \mu_0)^2}{N}$$

Thus the likelihood ratio is

$$\lambda = \frac{L_N(\hat{\theta}_H)}{L_N(\hat{\theta}_{ml})} = \frac{\max_{\sigma^2} L_N(\mu_0, \sigma^2)}{\max_{\mu, \sigma^2} L_N(\mu, \sigma^2)} = \frac{\left[2\pi \frac{\sum_{i=1}^N (z_i - \mu_0)^2}{N}\right]^{-N/2}}{\left[2\pi \frac{\sum_{i=1}^N (z_i - \hat{\mu})^2}{N}\right]^{-N/2}} = \left(\frac{\hat{\sigma}_{ml}}{\hat{\sigma}_H}\right)^N$$

We reject H if $\lambda < c$ where c is such that

$$\text{Prob}\{\lambda < c | H\} = \alpha$$

We can put

$$\frac{\hat{\sigma}_{ml}}{\hat{\sigma}_H} = \left(1 + \frac{N(\hat{\mu} - \mu_0)^2}{\sum (z_i - \hat{\mu})^2}\right)^{-1/2} = \left(\frac{T^2}{N-1}\right)^{-1/2}$$

where

$$T^2 = \frac{N(N-1)(\hat{\mu} - \mu_0)^2}{\sum_{i=1}^N (z_i - \hat{\mu})^2}$$

If the hypothesis H holds, $\mathbf{T} \sim \mathcal{T}_{N-1}$ is a r.v. with a Student distribution with $N - 1$ degrees of freedom. Once we compute the statistic

$$t(D_N) = T = \frac{\sqrt{N}(\hat{\mu} - \mu_0)}{\sqrt{\frac{1}{N-1} \sum_{i=1}^N (z_i - \hat{\mu})^2}} = \frac{(\hat{\mu} - \mu_0)}{\sqrt{\frac{\hat{\sigma}^2}{N}}}$$

the test $\lambda < c$ is (approximately) equivalent to $|T| > k = t_{\alpha/2, N-1}$ where $t_{\alpha/2, N-1}$ is the upper α point of a \mathcal{T} -distribution on $N - 1$ degrees of freedom, i.e.

$$\text{Prob} \{ |t_{N-1}| > t_{\alpha/2, N-1} \} = \alpha/2.$$

where $t_{N-1} \sim \mathcal{T}_{N-1}$.

In other terms λ is small when T^2 and consequently $|T|$ is large.

TP example

Does jogging lead to a reduction in pulse rate? Eight non jogging volunteers engaged in a one-month jogging programme. Their pulses were taken before and after the programme

pulse rate before	74	86	98	102	78	84	79	70
pulse rate after	70	85	90	110	71	80	69	74
decrease	4	1	8	-8	7	4	10	-4

Suppose that the decreases are samples from $\mathcal{N}(\mu, \sigma^2)$ for some unknown σ^2 .

We want to test $H : \mu = \mu_0 = 0$ against $\bar{H} : \mu \neq 0$ with a significance $\alpha = 0.5$.

We have $N = 8$, $\hat{\mu} = 2.75$, $T = 1.263$, $t_{\alpha/2, N-1} = 2.365$

Since $|T| \leq t_{\alpha/2, N-1}$, the data is not sufficient to reject the hypothesis H . In other terms we have not enough evidence to show that there is a reduction in pulse rate.

Single sample: the χ^2 test

Consider a random sample from $\mathcal{N}(\mu, \sigma^2)$ with μ unknown. Let

$$H : \sigma^2 = \sigma_0^2; \quad \bar{H} : \sigma^2 \neq \sigma_0^2$$

The likelihood ratio is

$$\lambda = \frac{\max_{\mu} L_N(\mu, \sigma_0^2)}{\max_{\mu, \sigma^2} L_N(\mu, \sigma^2)} = \frac{[2\pi\sigma_0^2]^{-N/2} e^{-\frac{\sum_i (z_i - \hat{\mu})^2}{2\sigma_0^2}}}{\left[2\pi \frac{\sum_i (z_i - \hat{\mu})^2}{N}\right]^{-N/2} e^{-N/2}}$$

This is small when

$$\frac{\sum_i (z_i - \hat{\mu})^2}{N\sigma_0^2} = \frac{\hat{S}}{N\sigma_0^2}$$

differs substantially from 1. If H is true then $\hat{S}/\sigma_0^2 \sim \chi_{N-1}^2$

The size α test rejects H if $\hat{S}/\sigma_0^2 < a_1$ or $\hat{S}/\sigma_0^2 > a_2$ where

$$\text{Prob} \left\{ \hat{S}/\sigma_0^2 < a_1 \right\} + \text{Prob} \left\{ \hat{S}/\sigma_0^2 > a_2 \right\} = \alpha$$

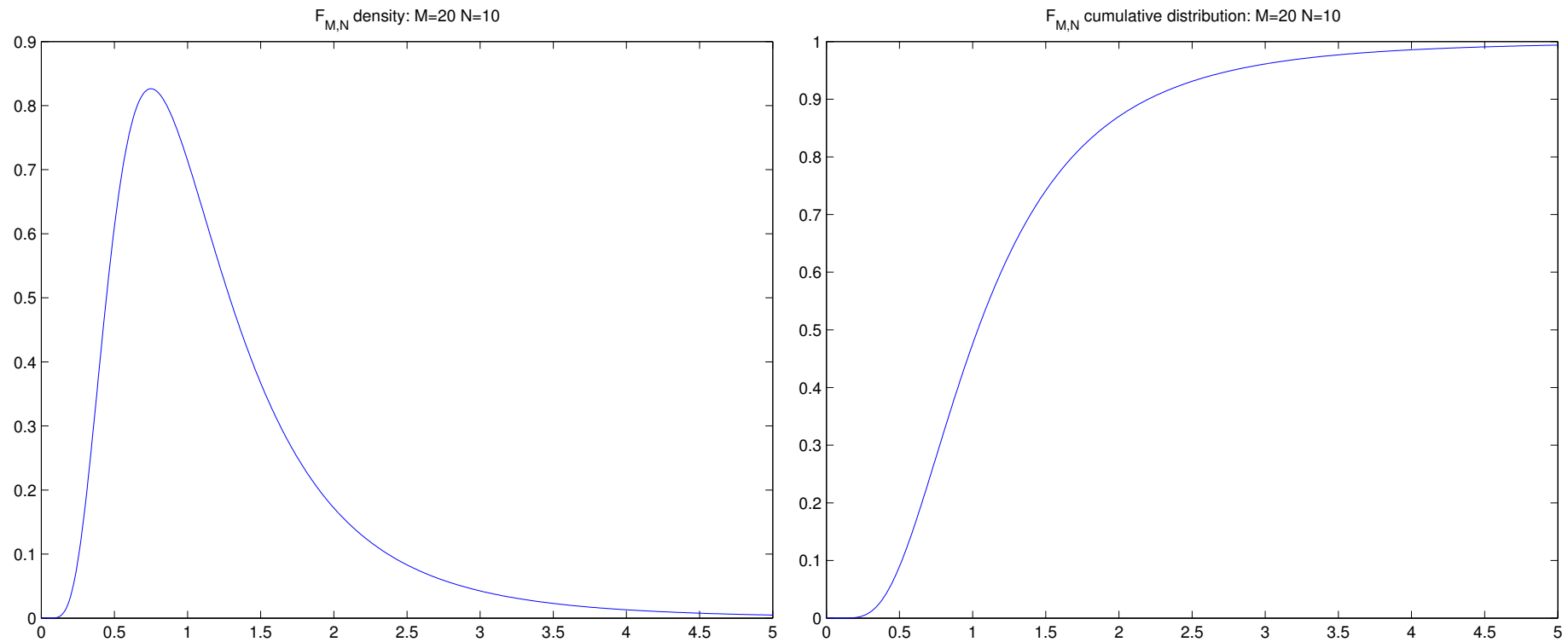
F-distribution

Let $x \sim \chi_M^2$ and $y \sim \chi_N^2$ be two independent r.v.. A r.v. z has a **F-distribution** $F_{m,n}$ with M and N degrees of freedom if

$$z = \frac{x/M}{y/N}$$

- If $z \sim F_{M,N}$ then $1/z \sim F_{N,M}$.
- If $z \sim \mathcal{T}_N$ then $z^2 \sim F_{1,N}$.

F-distribution



MATLAB script `s_fdis.m`.

Two samples: t-test

Consider two r.v.s $\mathbf{x} \sim \mathcal{N}(\mu_1, \sigma^2)$ and $\mathbf{y} \sim \mathcal{N}(\mu_2, \sigma^2)$ with the same variance. Let D_N^x and D_M^y two independent sets of samples .

We want to test $H : \mu_1 = \mu_2$ against $\bar{H} : \mu_1 \neq \mu_2$.

Let

$$\hat{\mu}_x = \frac{\sum_{i=1}^N x_i}{N}, \quad SS_x = \sum_{i=1}^N (x_i - \hat{\mu}_x)^2, \quad \hat{\mu}_y = \frac{\sum_{i=1}^M y_i}{M}, \quad SS_y = \sum_{i=1}^M (y_i - \hat{\mu}_y)^2$$

Once defined the statistic

$$T = \frac{\hat{\mu}_x - \hat{\mu}_y}{\sqrt{\left(\frac{1}{M} + \frac{1}{N}\right) \left(\frac{SS_x + SS_y}{M+N-2}\right)}} \sim \mathcal{T}_{M+N-2}$$

it can be shown that a test of size α rejects H if

$$|T| > t_{\alpha/2, M+N-2}$$

General formula t-test

From the previous results, it can be inferred that the general formula for a t-test is the following one:

$$T = \frac{\text{Statistic} - \text{Hypothesized value}}{\text{Estimated standard error of the statistic}} = \frac{\text{Statistic} - \text{Hypothesized value}}{\text{Estimated standard error of the numerator}}$$

If these calculations yield a value of T which is significantly different from the zero, the hypothesis is rejected.

Two samples: F test

Consider a random sample x_1, \dots, x_M from $\mathcal{N}(\mu_1, \sigma_1^2)$ and a random sample y_1, \dots, y_N from $\mathcal{N}(\mu_2, \sigma_2^2)$ with μ_1 and μ_2 unknown. Suppose we want to test

$$H : \sigma_1^2 = \sigma_2^2; \quad \bar{H} : \sigma_1^2 \neq \sigma_2^2$$

By the generalized likelihood ratio test we are led to consider the statistic

$$f = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{\hat{S}_1/(M-1)}{\hat{S}_2/(N-1)} \sim \frac{\sigma_1^2 \chi_{M-1}^2/(M-1)}{\sigma_2^2 \chi_{N-1}^2/(N-1)} = \frac{\sigma_1^2}{\sigma_2^2} F_{M-1, N-1}$$

Then if H is true, the ratio has a F-distribution $F_{M-1, N-1}$

We reject H if the ratio f is large, i.e. $f > F_{\alpha, M-1, N-1}$ where

$$\text{Prob} \{z > F_{\alpha, M-1, N-1}\} = \alpha$$

if $z \sim F_{M-1, N-1}$.

Region and interval estimate

- Unlike point estimation which is based on a one-to-one mapping from \mathcal{D} to Θ , region estimation maps D_N to a subset of Θ .
- Suppose we define an estimator $\underline{\theta}$ with the property that

$$\text{Prob} \{ \underline{\theta} \leq \theta | \theta \} \geq 1 - \alpha \text{ for all } \theta$$

- We call $\underline{\theta}$ a **lower confidence bound** for θ with confidence level $1 - \alpha$. This means that in the long run a proportion $1 - \alpha$ of values $\underline{\theta}(D_N)$ will be less than θ .
- In a similar manner we can obtain an **upper** $1 - \alpha$ confidence bound $\bar{\theta}$.
- Combining a lower bound estimator $\underline{\theta}$ at a confidence level $1 - \alpha_1$ and an upper bound estimator $\bar{\theta}$ at a confidence level $1 - \alpha_2$ we obtain a **two-sided** $1 - \alpha_1 - \alpha_2$ **confidence interval** satisfying

$$\text{Prob} \{ \underline{\theta} < \theta < \bar{\theta} | \theta \} \geq 1 - \alpha_1 - \alpha_2 \text{ for all } \theta$$

Some considerations

- Given a dataset D_N , we can obtain a $1 - \alpha$ confidence interval $[\underline{\theta}, \bar{\theta}]$ for θ .
- Notice that θ is a fixed unknown value and therefore the interval either does or does not contain the true θ .
- If we repeat the procedure of sampling D_N and constructing the confidence interval many times, then our confidence interval will contain the true θ at least $100(1 - \alpha)\%$ of the time.
- Notice that the endpoints of the interval $\underline{\theta}$ and $\bar{\theta}$ are r.v.s and not the parameter θ .

TP example: confidence interval of μ

- Consider a random sample D_N of a r.v. $\mathcal{N}(\mu, \sigma^2)$, with σ^2 known and the estimator $\hat{\mu} \sim \mathcal{N}(\mu, \sigma^2/N)$.

- It follows that

$$\frac{\hat{\mu} - \mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0, 1)$$

and consequently

$$\text{Prob} \left\{ -z_{\alpha/2} \leq \frac{\hat{\mu} - \mu}{\sigma/\sqrt{N}} \leq z_{\alpha/2} \right\} = 1 - \alpha$$

$$\text{Prob} \left\{ \hat{\mu} - z_{\alpha/2} \frac{\sigma}{\sqrt{N}} \leq \mu \leq \hat{\mu} + z_{\alpha/2} \frac{\sigma}{\sqrt{N}} \right\} = 1 - \alpha$$

- $\underline{\theta} = \hat{\mu} - z_{\alpha} \sigma / \sqrt{N}$ is a lower $1 - \alpha$ confidence bound for μ .
- $\bar{\theta} = \hat{\mu} + z_{\alpha} \sigma / \sqrt{N}$ is an upper $1 - \alpha$ confidence bound for μ .

Confidence region and hypothesis testing

- There is an interesting and useful duality between confidence intervals and hypothesis tests.
- In some circumstances, given an hypothesis test we can have a confidence interval. In other circumstances it may be the reverse.
- Let us see how to build a confidence interval from an hypothesis test.
- Consider a dataset D_N from a parameteric distribution with parameter θ . We want to construct a confidence region for θ .
- Suppose we construct, for a given value θ_0 , a level α -test of $H : \theta = \theta_0$. We call $S_1(\theta_0)$ its non-critical region, i.e. the set of values of D_N that do not significantly reject H .
- Let θ_0 be a variable. As we consider different values of θ_0 , we get a set of non-critical regions $S_1(\theta_0)$.

- Let us define the set of values of θ_0 for which D_N belongs to $S_1(\theta_0)$,

$$C(D_N) = \{\theta_0 : D_N \in S_1(\theta_0)\}$$

the confidence region for θ of level $1 - \alpha$.

- Indeed

$$\text{Prob} \{\theta_0 \in C(\mathbf{D}_N) | H\} = \text{Prob} \{\mathbf{D}_N \in S_1(\theta_0) | H\} \geq 1 - \alpha$$

- The confidence region for θ is the set of values θ_0 for which we would accept H , for a given D_N .

TP Example: Confidence region

- Let $\mathbf{x} \sim \mathcal{N}(\mu, 0.1)$ and $D_N = \{10, 11, 12, 13, 14, 15\}$.
- We want to estimate the confidence region of μ with level $\alpha = 0.1$.
- We have $\hat{\mu} = 12.5$, and

$$t_\alpha = z_{\alpha/2} * \sigma / \sqrt{N} = 0.0672$$

- For a given $\mu = \mu_0$ and a generic D_N , the noncritical region is

$$S_1 = \{D_N : |\hat{\mu} - \mu_0| \leq t_\alpha\}$$

- The confidence region for the given D_N is

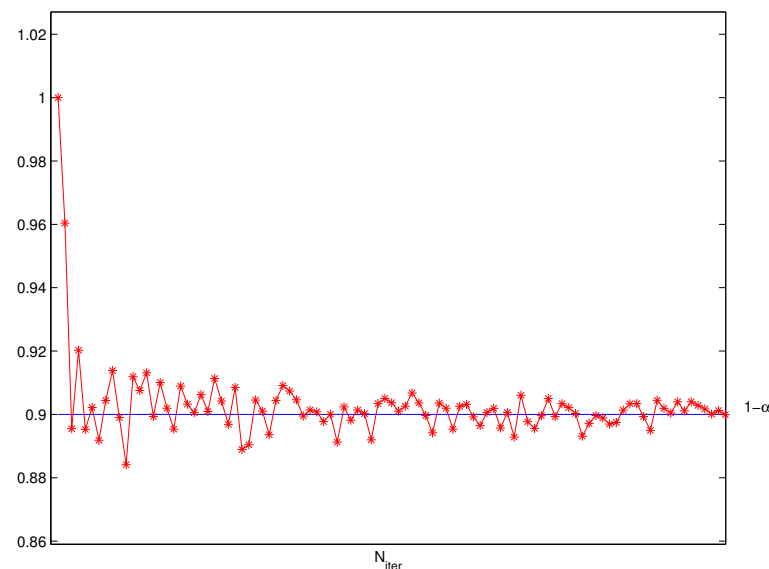
$$C(D_N) = \{\mu_0 : |\hat{\mu} - \mu_0| \leq t_\alpha\} = \{12.5 - 0.0672 \leq \mu_0 \leq 12.5 + 0.0672\}$$

TP Matlab script confidence.m

- The user sets μ , σ , N , α and a number of iterations N_{iter} .
- The script generates N_{iter} times $D_N \sim \mathcal{N}(\mu, \sigma^2)$ and computes $\hat{\mu}$.
- The script returns the percentage of times that

$$\hat{\mu} - \frac{Z_{\alpha/2}\sigma}{\sqrt{N}} < \mu < \hat{\mu} + \frac{Z_{\alpha/2}\sigma}{\sqrt{N}}$$

- We can easily check that this percentage converges to $(1 - \alpha)\%$ for $N_{\text{iter}} \rightarrow \infty$.



TP Example: Confidence region (II)

- Let $\mathbf{x} \sim \mathcal{N}(\mu, \sigma^2)$, with σ^2 unknown and $D_N = \{10, 11, 12, 13, 14, 15\}$.
- We want to estimate the confidence region of μ with level $\alpha = 0.1$.
- We have $\hat{\mu} = 12.5$, $\hat{\sigma}^2 = 3.5$. Since

$$\frac{\hat{\mu} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{N}}} \sim \mathcal{T}_{N-1}$$

we have

$$t_\alpha = t_{\{\alpha/2, N-1\}} \hat{\sigma} / \sqrt{N} = 2.015 * 1.87 / \sqrt{6} = 1.53$$

- The $(1 - \alpha)$ confidence interval of μ is

$$\hat{\mu} - t_\alpha < \mu < \hat{\mu} + t_\alpha$$

Opinion polls

- We want to estimate θ , the proportion of people who support the politics of Mr. Bush amongst a very large population. We want to define how many interviews are necessary to have a confidence interval of 3% with a significance of 5%.
- We interview N people and estimate θ as

$$\hat{\theta} = \frac{x_1 + \cdots + x_N}{N} = \frac{S}{N}$$

where $x_i = 1$ if the i th person supports Bush and $x_i = 0$ otherwise. Note that S is a binomial variable.

- We have

$$E[\hat{\theta}] = \theta, \quad \text{Var}[\hat{\theta}] = \text{Var}[S/N] = \frac{N(\theta)(1-\theta)}{N^2} = \frac{\theta(1-\theta)}{N} \leq \frac{1}{4N}$$

- We approximate the distribution of $\hat{\theta}$ by $\mathcal{N}(\theta, \frac{\theta(1-\theta)}{N})$.

Opinion polls (II)

- It follows that $\frac{\hat{\theta} - \theta}{\sqrt{\theta(1-\theta)/N}} \sim \mathcal{N}(0, 1)$.
- The following relation holds

$$\begin{aligned} \text{Prob} \left\{ \hat{\theta} - 0.03 \leq \theta \leq \hat{\theta} + 0.03 \right\} &= \\ \text{Prob} \left\{ -\frac{0.03}{\sqrt{\theta(1-\theta)/N}} \leq \frac{\hat{\theta} - \theta}{\sqrt{\theta(1-\theta)/N}} \leq \frac{0.03}{\sqrt{\theta(1-\theta)/N}} \right\} &= \\ \Phi \left(\frac{0.03}{\sqrt{\theta(1-\theta)/N}} \right) - \Phi \left(-\frac{0.03}{\sqrt{\theta(1-\theta)/N}} \right) &\geq \\ &\geq \Phi(0.03\sqrt{4N}) - \Phi(-0.03\sqrt{4N}) \end{aligned}$$

- In order to have this probability to be at least 0.95 we need $0.03\sqrt{4N} \geq 1.96$ or equivalently $N \geq 1068$.

Goodness of fit test

A **goodness of fit test** is a statistical hypothesis test that is used to assess formally whether the sequence of observations z_1, \dots, z_N are an independent sample from a particular distribution with distribution function F_z .

In other terms, a goodness-of-fit test is a test where the null hypothesis is

H : the z_i are iid r.v.'s with distribution function F_z

The most known goodness-of-fit tests are

1. the chi-square test,
2. the Kolmogorov Smirnov test

The Kolmogorov-Smirnov test

The **Kolmogorov-Smirnov test** is a goodness-of-fit test that uses the empirical estimator $\hat{\mathbf{F}}$ of the distribution function $F_{\mathbf{z}}(z)$.

Let us define the test statistic

$$\mathbf{q} = \max_z |\hat{\mathbf{F}}_{\mathbf{z}}(z) - F(z)|$$

For large N , \mathbf{q} is close to 0 if the hypothesis H is true. Then we must reject H if \mathbf{q} is larger than some constant c where

$$\text{Prob} \{ \mathbf{q} > c | H \} = \alpha \approx 1 - e^{-2Nc^2}$$

It results that the hypothesis H is accepted if

$$q < \sqrt{-\frac{1}{2N} \ln(1 - \alpha)}$$

Quantile-Quantile plot

- The **quantile-quantile (q-q) plot** is a graphical technique for determining if two data sets come from populations with a common distribution.
- A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
- By **quantile**, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.
- A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line.
- The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Matlab example: qq plot

We consider three data sets $D_x \sim \mathcal{N}(0, 1)$, $D_y \sim \mathcal{U}(0, 1)$ and $D_z \sim \mathcal{N}(0, 1)$, all of them containing 1000 samples.

We plot the qq-plot of x vs. y and x vs z .

