Statistics and learning

Tests

Emmanuel Rachelson and Matthieu Vignes

ISAE SupAero

Wednesday 16^{th} October 2013

Motivations

When could tests be useful ?

- A statistical hypothesis is an assumption on the distribution of a random variable.
- Ex: test whether the average temperature in a holiday ressort is $28°\mathrm{C}$ in the summer.
- A test is a procedure which makes use of a sample to decide whether we can reject an hypothesis or whether there is nothing wrong with it (it's not really acceptance).
- Examples of applications: decide if a new drug can be put on market after adequate clinical trials, decide if items comply with predefined standards, which genes are significantly differentially expressed in pathological cells ...
- Typically, sources to build hypothesis stem from quality need, values from a previous experiment, a theory that need experimental confirmation or an assumption based on observations.

Outline and a motivating example

It's really about **decision making**; don't be fooled; tests shed light on a question, final results heavily depend on a human interpretation !

Today's goals:

- introduce basic concepts related to tests through 2 examples.
- A general presentation of tests.
- Some particular cases: one-sample, two-sample, paired tests; Z-tests, t-tests, χ^2-tests, F-tests...

Example 1: cheater detection

To introduce randomness, you are asked to throw a coin 200 times and write down the results. Why would I be suspicious about students that do not exhibit at least one HHHHHH or TTTTTT pattern ? Would I be (totally ?) fair if I was to blame (all of) them ?

Motivation 2

Example 2: rain makers

In a given area of agricultural interest, it usually rains 600 mm a year.

Suspicious scientists claim that they can locally increase rainfall, when spreading a revolutionary chemical (iodised silver) on clouds. Tests over the 1995-2002 period gave te following results:

Year	1995	1996	1997	1998	1999	2000	2001	2002
Rainfall (mm/year)	606	592	639	598	614	607	616	586

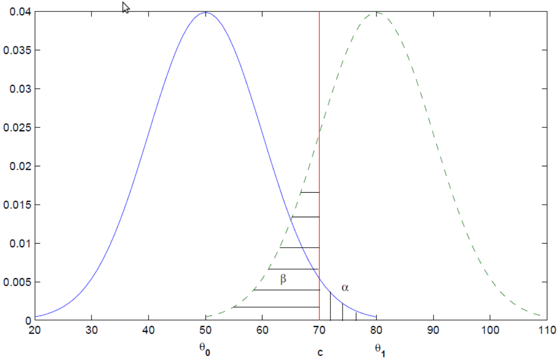
Does this sound correct to you ? Quantify the answer.

Bonus: what would have changed if you wanted to test if the increase was of say 30 mm ?

Motivation

Rain makers and possible errors

If you assume normality of rainfalls, had you applied the treatment or not



Hypothesis testing: (H0) $\theta = \theta_0$ and (H1) $\theta = \theta_1$.

Tests

Possible situations

		Realworld	
		(H0)	(H1)
Decision made	(H0)		
	(H1)		

Tests

Possible situations

		Realworld	
		(H0)	(H1)
Decision made	(H0)	 $(1-\alpha)$	 (β)
	(H1)	 (α)	 $(1-\beta)$

Tests

Possible situations

		Realworld	
		(H0)	(H1)
Decision made	(H0)	 $(1-\alpha)$	 (β)
	(H1)	 (α)	 $(1-\beta)$

Apply that to 'innoncent until proven guilty' and interpret the different situations. How do you want to control α and β ?

What about introducing a new drug on the market ??

Tests

General methodology

1. Modelling of the problem.
2. Determine alternative hypotheses to test (disjoint but not necessarily exhaustive).
3. Choose of a statistic which (a) can be computed from data and (b) which has a known distribution under (H0).
4. Determine the behaviour of statistics under (H1) and build critical region (where (H0) rejected)
5. Compute the region at a fixed error I threshold and compare to values obtained from data. Or compute p-value of the test from data.
6. Statistical conclusion: accept or reject (H0). Comment on p-value ?

opt. Can you say something about the power ?

7. Strategic conclusion: how do YOU decide thanks to the light shed by statistical result ?

Test methodology into details

- **Hypothesis**:= any subset of the family of all considered probability distributions $\mathcal{P}$. In practice, hypotheses are often on unknown parameters of distributions $\rightarrow$ parametric hypotheses, defined by equalities or inequalities: (H0) $\theta_0 \in \Theta_0$ and (H1) $\theta_1 \in \Theta_1$. In turn, they can be simple if only one value for the parameters is tested or multiple composite.

Test methodology into details

- **Hypothesis**:= any subset of the family of all considered probability distributions $\mathcal{P}$. In practice, hypotheses are often on unknown parameters of distributions $\rightarrow$ parametric hypotheses, defined by equalities or inequalities: (H0) $\theta_0 \in \Theta_0$ and (H1) $\theta_1 \in \Theta_1$. In turn, they can be simple if only one value for the parameters is tested or multiple composite.
- Choose a **test statistic** $T_n :=$ a random variable which only depends on $(\Theta_0; \Theta_1)$ and on observations of the (X_i)'s. Interesting if the distribution is known given (H0) is true. Note that it is an estimator...depending on (H0) and (H1).

Test methodology into details

- **Hypothesis**:= any subset of the family of all considered probability distributions $\mathcal{P}$. In practice, hypotheses are often on unknown parameters of distributions $\rightarrow$ parametric hypotheses, defined by equalities or inequalities: (H0) $\theta_0 \in \Theta_0$ and (H1) $\theta_1 \in \Theta_1$. In turn, they can be simple if only one value for the parameters is tested or multiple composite.
- Choose a **test statistic** $T_n :=$ a random variable which only depends on $(\Theta_0; \Theta_1)$ and on observations of the (X_i)'s. Interesting if the distribution is known given (H0) is true. Note that it is an estimator...depending on (H0) and (H1).
- How to choose a good test statistic ? Remember the typology of confidence intervals ? And explore R help ?!

Test methodology into details (cont'd)

- Determine the **rejection region** R. Usually of the form $(r; +\infty)$, $(-\infty; r)$ or $(-\infty; r) \cup (r'; +\infty)$. To decide, examine how the test statistic behaves under (H1).

Test methodology into details (cont'd)

- Determine the **rejection region** R. Usually of the form $(r;+\infty)$, $(-\infty;r)$ or $(-\infty;r)\cup(r';+\infty)$. To decide, examine how the test statistic behaves under (H1).
- **type I error**:=probability of rejecting (H0) whilst it is correct. Mathematically:

$$\alpha = \sup_{\theta\in\Theta_0} P(T_n \in R | X_1 \dots X_n \text{ iid} \sim P_\theta)$$

Test methodology into details (cont'd)

- Determine the **rejection region** R. Usually of the form $(r; +\infty)$, $(-\infty; r)$ or $(-\infty; r) \cup (r'; +\infty)$. To decide, examine how the test statistic behaves under (H1).
- **type I error**:=probability of rejecting (H0) whilst it is correct. Mathematically:

$$\alpha = \sup_{\theta \in \Theta_0} P(T_n \in R | X_1 \dots X_n \text{ iid} \sim P_\theta)$$

- Remark: useless (test) to try to get $\alpha = 0$!

Test methodology into details (cont'd)

- Determine the **rejection region** R. Usually of the form $(r; +\infty)$, $(-\infty; r)$ or $(-\infty; r) \cup (r'; +\infty)$. To decide, examine how the test statistic behaves under (H1).
- **type I error**:=probability of rejecting (H0) whilst it is correct. Mathematically:

$$\alpha = \sup_{\theta \in \Theta_0} P(T_n \in R | X_1 \dots X_n \text{ iid} \sim P_\theta)$$

- Remark: useless (test) to try to get $\alpha = 0$!
- **p-value**:= maximal value of α so that the test would accept the observed statistic to be drawn under (H0) $\approx$ credibility index on (H0). Alternative definition: probability to obtain a test statistic value at least as contradictory to (H0) as the observed value assuming (H0) is true (if we repeated the experiment a large number of times).

Test methodology into details (end)

- dissymetry between (H0) and (H1): (H0) tends to be kept unless good reasons to reject it.
  (H1) is only used to choose the form of the rejection region, not its bounds !
  It is then interesting to look at the

Test methodology into details (end)

- dissymetry between (H0) and (H1): (H0) tends to be kept unless good reasons to reject it.
  (H1) is only used to choose the form of the rejection region, not its bounds !
  It is then interesting to look at the
- **type II error**:= probability to wrongly keep (H0) (while (H1) is true).
  In mathematical terms:

$$\beta = \sup_{\theta \in \Theta_1} P(T_n \notin R | X_1 \dots X_n \text{ iid} \sim P_\theta)$$

Test methodology into details (end)

- dissymetry between (H0) and (H1): (H0) tends to be kept unless good reasons to reject it.
  (H1) is only used to choose the form of the rejection region, not its bounds !
  It is then interesting to look at the
- **type II error**:= probability to wrongly keep (H0) (while (H1) is true). In mathematical terms:

$$\beta = \sup_{\theta \in \Theta_1} P(T_n \notin R | X_1 \ldots X_n \text{ iid} \sim P_\theta)$$

- hence (H0) is chosen according to a firmly established theory (you don't want to make a fool of yourself), because caution is needed or...for subjective reasons (consumer choice is not that of manufacturers !)

Choosing hypotheses: launching a new drug

How does the pharmaceutical industry proceed to approve the commercialisation of a new drug ? Basically two possibilities

- test again a placebo; (H0) the new drug is better than the placebo. Do you like it ?

Choosing hypotheses: launching a new drug

How does the pharmaceutical industry proceed to approve the commercialisation of a new drug ? Basically two possibilities

- test again a placebo; (H0) the new drug is better than the placebo. Do you like it ?
- I don't: it's not difficult to find a chemical compound which makes better than empty pills (or sugar). At which price (= secondary effects.

Choosing hypotheses: launching a new drug

How does the pharmaceutical industry proceed to approve the commercialisation of a new drug ? Basically two possibilities

- test again a placebo; (H0) the new drug is better than the placebo. Do you like it ?
- I don't: it's not difficult to find a chemical compound which makes better than empty pills (or sugar). At which price (= secondary effects.
- you can also test again an existing drug. But then (H0) can be "the new drug is at least as efficient as the old one" (good for the company).

Choosing hypotheses: launching a new drug

How does the pharmaceutical industry proceed to approve the commercialisation of a new drug ? Basically two possibilities

- test again a placebo; (H0) the new drug is better than the placebo. Do you like it ?
- I don't: it's not difficult to find a chemical compound which makes better than empty pills (or sugar). At which price (= secondary effects.
- you can also test again an existing drug. But then (H0) can be "the new drug is at least as efficient as the old one" (good for the company).
- if the social healthcare hired me, I would test (H0) "the new drug does not improve over existing ones".

Choosing hypotheses: launching a new drug

How does the pharmaceutical industry proceed to approve the commercialisation of a new drug ? Basically two possibilities

- test again a placebo; (H0) the new drug is better than the placebo. Do you like it ?
- I don't: it's not difficult to find a chemical compound which makes better than empty pills (or sugar). At which price (= secondary effects.
- you can also test again an existing drug. But then (H0) can be "the new drug is at least as efficient as the old one" (good for the company).
- if the social healthcare hired me, I would test (H0) "the new drug does not improve over existing ones".
- Sadly enough, the first option if used most of the time ?! For fairness between new and existing molecules...

Choosing hypotheses: launching a new drug

How does the pharmaceutical industry proceed to approve the commercialisation of a new drug ? Basically two possibilities

- test again a placebo; (H0) the new drug is better than the placebo. Do you like it ?
- I don't: it's not difficult to find a chemical compound which makes better than empty pills (or sugar). At which price (= secondary effects.
- you can also test again an existing drug. But then (H0) can be "the new drug is at least as efficient as the old one" (good for the company).
- if the social healthcare hired me, I would test (H0) "the new drug does not improve over existing ones".
- Sadly enough, the first option if used most of the time ?! For fairness between new and existing molecules...

Historical note: statistics were of great help in modern medicine.

Tests you need to know

and we shall see during next session and use on practical examples

- Parametric tests (observations drawn from $\mathcal{N}$ or large samples so that C.L.Th. applies)
  - one sample: comparing the empirical mean to a theoretical value $\rightarrow$ Z-test or t-test
  - two independent samples $\rightarrow$ t-test, F-test
  - paired samples $\rightarrow$ paired t-test
  - several samples $\rightarrow$ ANOVA (not today).
- Adequation tests $\rightarrow$ χ^2-tests. Normality check $\rightarrow$ Kolmogorov or Shapiro-Wilks.
- Non-parametric tests (when small samples or non Gaussian distributions)
  - comparing 2 medians from independent samples $\rightarrow$ Mann-Whitney test.
  - two paired samples $\rightarrow$ Wilcoxon test on differences.
  - several samples $\rightarrow$ Kruskal-Wallis.

Exercises

Poisson arrival at a motorway toll booth

For two hours, at a motorway toll, we write down the number of cars arriving during each 2 minute intervals. We obtain:

#(cars)	0	1	2	3	4	5	6	7	8	9	10	11
#(intervals)	4	9	24	25	22	18	6	5	3	2	1	1

Test at a significance level of 0.1 the fit to a Poisson distribution with a parameter to be determined.

Evolution of purchasing power

In 2004, the total amount spent on products which are not essentials (*e.g.* travels, shows, *etc.* as opposed to food, hoosing, *etc.*) was 632 euros per month per household accoring to the INSEE during a partial survey over millions of households. In 2008, from a sample of $2,000$ interviewed by telephone, $1,837$ answers were obtained and the declared mean value was 598 euros (with sd 254 euros). If you assume a 2% inflation per year, would you say that the amount spent on non-essentials has significantly decreased ?

Finished

Next time: more tests and analysis of variance (ANOVA)