PRE1: APPLIED STATISTICS
Hypothesis Testing

Marie-Anne Poursat

AI Master
Université Paris-Saclay

October 6th 2022

Hypotheses

Statistical testing uses data to decide whether a statement ($=$ *null hypothesis*) is true.

- EX1 : *European regulations require that any presence of genetically modified organisms (GMOs) in food be labelled as soon as the level of GMOs exceeds the 0.9% threshold.*

  $\hookrightarrow$ Is the level of GMOs τ below the 0.9% threshold regulations ?
  Data : 20 packets of cereals, $\widehat{\tau} = 0.91$ is the observed percentage of OGMs

  $$H_0 : \ \tau \leq 0.9 \text{ versus } H_1 : \ \tau > 0.9$$

  Is the observed difference **a real difference** likely to appear in the larger population ? Or is it observed in the sample **by chance** ?

Hypotheses

- EX2 : Is the coin toss fair ?
  Data : $X = 18$ heads out of 30 tosses
  Model : $X \sim Bin(30, p)$,
  $H_0 : p = 0.5$ versus $H_1 : p \neq 0.5$

- EX3 : Is a given gene *differentially expressed* between two cell types (normal and tumor cells) ?
  Data : 2 Gaussian samples, $\overline{x_1} - \overline{x_2} = 1.25$,
  $H_0 : \mu_1 - \mu_2 = 0$ versus $H_1 : \mu_1 - \mu_2 \neq 0$

Hypotheses

The first step in the testing procedure : find H_0 and H_1

H_0 and H_1 are written according to the model parameters.

- "H_0 **is accepted**"= not rejected ! There is no evidence from the data that H_0 is wrong (but H_0 is not necessarily true)
- "H_0 **is rejected** in favor of H_1"= H_1 explains the data significantly better than H_0 so we decide H_1
- A hypothesis can be simple or composite : a *simple* hypothesis specifies the value of the unknown parameters

  $H_0 : p = 0.5$ and $H_1 : p \neq 0.5$, H_0 is *simple* and H_1 is *composite, two-sided*

  $H_0 : \tau \leq 0.9$ and $H_1 : \tau > 0.9$: H_0 and H_1 are *composite, one-sided*

2 types of errors

The decision to accept or reject H_0 is based on data observed from a random process
$\hookrightarrow$ the decision is random and may be incorrect

2 types of errors :

1. to reject H_0 when it is true : Type I error
2. to retain H_0 when it is false : Type II error

It is not possible to ensure that the probabilities of making a Type I error and a Type II error are both arbitrarily small

$\hookrightarrow$ Classical testing paradigm : *focus on the Type I error*

- the probability of a type I error is kept below α, the level of the test ($\alpha = 0.05,\ 0,01$)
- the probability of a type II error is **not** controlled

Test statistics

The second step in the testing procedure : determine a test statistic T
This is the quantity calculated from the data whose numerical value leads to acceptance or rejection of H_0.

- EX1 : suppose that $X_1, \ldots, X_{20}$ are i.i.d $\mathcal{N}(\tau, \sigma^2)$
  A reasonable choice is to reject H_0 if $(\overline{X} - 0.9) > c$.
  It is convenient to standardize the estimator and to define :
  $T = \dfrac{\overline{X} - 0.9}{S/\sqrt{20}}$.
  T is called the **test statistic** and c is a **critical value**.
- EX2 : $X \sim Bin(30, p)$.
  We reject H_0 if $|X - 15| > c$.

In some cases there are several possible choices for T (corresponding to different statistical tests) ; in more complicated cases, the choice of T is not straightforward.

Rejection regions

The third step in the testing procedure : determine the rejection region

The rejection region is the set of observed values of the test statistic that lead to reject H_0.

- EX1 : $\mathcal{R} = \{(X_1, \ldots, X_{20}) : \; T(X_1, \ldots, X_{20}) > c\}$
- EX2 : $\mathcal{R} = \{(X : \; T(X) > c\}$

Usually, the rejection region is of the form

$$\mathcal{R} = \{(X_1, \ldots, X_n) : \; T(X_1, \ldots, X_n) > c\}$$

$\hookrightarrow$ The value of c is determined by

$$\mathrm{P}(\text{reject } H_0 \text{ when } H_0 \text{ is true}) = \mathrm{P}_{H_0}(T > c) = \alpha$$

Example 1 (GMOs'rate)

EX1 : $X_1, \ldots, X_{20}$ i.i.d $\mathcal{N}(\tau, \sigma^2)$

1. $H_0 : \tau \leq 0.9$ and $H_1 : \tau > 0.9$
2. $T = \dfrac{\overline{X} - 0.9}{S/\sqrt{20}}$
3. reject H_0 if $T > c$

   $$\mathrm{P}_{H_0}(T > c) = \mathrm{P}\left(\frac{\overline{X} - 0.9}{S/\sqrt{20}} > c\right) = \alpha$$

   Thus, $c = t_{1-\alpha}(19)$ where $t_{1-\alpha}(19)$ is the quantile of the t-distribution with 19 degrees of freedom.

   The test using the t-quantile is called the *one-sample t-test*.
4. observed data : $\overline{X}^{obs} = 0.91$, $S^{obs} = .06$; $t_{.95}(19) = 1.73$

   Then $T^{obs} = 0.745 < 1.73$.
5. Decision : we do not reject H_0 ; according to these data, there is no evidence that the product does not respect the european regulations.

P-values

Rather than specifying α and computing c, we calculate the P-value of the test

Definition : The P-value for a sample is defined as the smallest value of α for which the null hypothesis is rejected.

$\hookrightarrow$ To perform the test, find the p-value of the sample and then H_0 is rejected if we decide to use α larger than the p-value :

$$\text{reject } H_0 \Longleftrightarrow \text{ p-value } < \alpha$$

Interpretation :

- a small p-value is evidence *against* H_0
- a large p-value shows that the *data are consistent* with H_0
- the p-value tells us whether the decision to reject or accept H_0 is close to α

Computation of p-values

Proposition

Suppose that the rejection region is of the form $T > c$. Then,

$$\text{p-value } = \mathrm{P}(T \geq T^{obs})$$

where T is the test statistic and T^{obs} is the observed numerical value of T on the data.

Statistical softwares calculate p-values.
Typically, the software output uses the evidence scale :

- p-value $< .001$ very strong evidence against H_0
- p-value $< .01$ strong evidence against H_0
- p-value $< .1$ weak evidence against H_0
- p-value $> .1$ little or no evidence against H_0

EX1 (GMOs'rate) : p-value

p-value$=\mathrm{P}(T \geq T^{obs})$

Answer : $\mathrm{P}(T \geq 0.745) = 0.23$ where T is a $t(19)$ variable.
At level $\alpha = 0.05$ (or 0.01, or even 0.02), there is no evidence against H_0 : H_0 is not rejected (but we don't know the type II error).

Testing for a mean

one-sample t-test

Normal model : $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$

- $H_0 : \mu = \mu_0 \quad H_1 : \mu > \mu_0$

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1) \quad \text{(exact Student distribution)}$$

$\hookrightarrow$ p-value $= \mathrm{P}(T \geq T^{obs}) =$ `1- stats.t.cdf(Tobs,df=n-1)`

- if $H_1 : \mu < \mu_0$,
  p-value $= \mathrm{P}(T \leq T^{obs}) =$ `stats.t.cdf(Tobs,df=n-1)`
- if $H_1 : \mu \neq \mu_0$,
  p-value $= \mathrm{P}(|T| \geq |T^{obs}|) =$ `2*(1- stats.t.cdf(Tobs,df=n-1))`

Testing for a mean

Non-normal model $X_1, \ldots, X_n$ i.i.d. $\mu = \mathrm{E}(X_i)$, $\sigma^2 = \mathrm{Var}(X_i)$

- $H_0: \ \mu = \mu_0 \quad H_1: \ \mu > \mu_0$

$$T = \frac{\overline{X} - \mu_0}{\widehat{\sigma}/\sqrt{n}} \sim \mathcal{N}(0,1) \quad \text{(approximate distribution via CLT)}$$

$\hookrightarrow$ p-value $= \mathrm{P}(T \geq T^{obs}) =$ `1- stats.norm.cdf(Tobs)`

- if $H_1: \mu \neq \mu_0$,
  p-value $= \mathrm{P}(|T| \geq |T^{obs}|) =$ `2*(1- stats.norm.cdf(Tobs))`
- if $H_1: \mu < \mu_0$,
  p-value $= \mathrm{P}(T \leq T^{obs}) =$ `stats.norm.cdf(Tobs)`

Testing for one parameter

$X_1, \ldots, X_n$ i.i.d. with distribution depending on θ, $\widehat{\theta}$ MLE, $\widehat{\text{s.e.}}$ estimated standard error of $\widehat{\theta}$

- $H_0 : \ \theta = \theta_0 \quad H_1 : \ \theta > \theta_0$

$$T = \frac{\widehat{\theta} - \theta_0}{\widehat{\text{s.e.}}} \sim \mathcal{N}(0,1) \quad \text{(approximate distribution for the MLE)}$$

  $\hookrightarrow$ p-value $= \mathrm{P}(T \geq T^{obs}) =$ `1- stats.norm.cdf(Tobs)`

- if $H_1 : \theta \neq \theta_0$,
  p-value $= \mathrm{P}(|T| \geq |T^{obs}|) =$ `2*(1- stats.norm.cdf(Tobs))`
- if $H_1 : \theta < \theta_0$,
  p-value $= \mathrm{P}(T \leq T^{obs}) =$ `stats.norm.cdf(Tobs)`

Duality CI-Tests

There is a relationship between the rejection regions of tests of level α and the $1-\alpha$ confidence intervals :

The test of $H_0 : \ \theta = \theta_0$ versus $H_1 : \ \theta \neq \theta_0$ with test statistic $T = \dfrac{\widehat{\theta} - \theta_0}{\widehat{\text{s.e.}}}$ and level α is the same test as the test that rejects H_0 if and only if $\theta_0 \notin [\widehat{\theta} - q_{1-\alpha/2}\widehat{\text{s.e.}}; \ \widehat{\theta} - q_{\alpha/2}\widehat{\text{s.e.}}]$

q_α is the α-quantile of the distribution of T.

Testing $\theta = \theta_0$ is equivalent to checking whether θ_0 is in the confidence interval.

The testing procedure

1. Specify the model and the hypotheses H_0 and H_1
2. Find an appropriate test statistic T :
   - T is calculable (does not depend on unknown parameters)
   - the distribution of T under H_0 is known
3. Find the form of the rejection region (*look at* H_1 !)
4. Calculate the p-value and make your decision
   - if you reject H_0, the risk of an incorrect decision is less than α (*type I error*)
   - if you don't reject H_0, the risk of an incorrect decision is usually unknown(*type II error*)

The **power of the test** is defined as the probability of rejecting H_0 when H_1 is true $=$ 1 - type II error.

The power depends on the actual value of the parameter under H_1 and is not calculable.

In practice, choosing the test (of given level) that maximizes the power is difficult.