Scientific Method

Class 3

MASTER INFORMATIQUE - SIIA

G. Coppin

2021-2022

Course objectives

Objectives :

1. Master the main statistical tests

General test methodology (I)

- State the hypothesis to be tested. $H_0 : \theta = \theta_0$.
- Take a risk of error α
- Draw a sample of the population - be careful, methodology to follow (see lesson 1)
- Calculate an estimator of the parameter - for example $\overline{X}$ to estimate an average
- Study the difference between $\widehat{\theta}$ and θ_0. If this difference is large, the hypothesis is rejected.
- Draw a conclusion : We may therefore have rejected H_0 or not have been able to do so. So two types of possible error
  - reject even when true (risk α)
  - accept when it is false - difficult to evaluate ("degree of falsehood" of the hypothesis H_0)

General test methodology (II)

A cannery puts on the market cans of peas whose label mentions 400g. The production manager wants to check that the weight is respected.

- $H_0 : \mu = 400g$
- either H_0 is rejected for too heavy or too light weight. If boxes ok, unnecessary cost. Risk α
- is H_0 wrongly accepted but in reality possibility that the market is flooded with boxes too heavy (losses) or too light (scam). Probability of this type of uncontrolled error.

Hypothesis test on a proportion (II)

The null hypothesis is rejected if Z is too large or too small. or if Z is outside the interval $[-c_\alpha, +c_\alpha]$. In other words

- H_0 is rejected if $\mid Z \mid > c_\alpha$ or also $\mid \widehat{p} - p_0 \mid > \frac{\sqrt{p_0 q_0}}{\sqrt{n}}$
- H_0 is accepted if $\mid Z \mid \leq c_\alpha$ or also $\mid \widehat{p} - p_0 \mid \leq \frac{\sqrt{p_0 q_0}}{\sqrt{n}}$

Hypothesis test on a proportion (III)

We assume that 25% of people are left-handed. With $\alpha = 10\%$, we find 18 left-handed people out of 120 people.

- $p_0 = 0.25$ and $\widehat{p} = 0.15$
- $c_\alpha = 1.645$. So $c_\alpha \sqrt{p_0 q_0}/\sqrt{n} = 0.065$

Since $|\widehat{p} - p_0| = 0.25 - 0.15 = 0.10 > 0.065$ (we got a value "too extreme"), we reject the null hypothesis. There are certainly fewer left-handers.

With R language, it gives ...

```
> prop.test(18, 120, 0.25)
1-sample proportions test with continuity correction
data : 18 out of 120, null probability 0.25
X-squared = 5.8778, df = 1, p-value = 0.01533
alternative hypothesis : true p is not equal to 0.25
95 percent confidence interval :
0.09369541 0.22939185
sample estimates :
p
0.15
```

And also ...

```
> binom.test(18, 120, 0.25)
Exact binomial test
data : 18 and 120
number of successes = 18, number of trials = 120, p-value = 0.01102
alternative hypothesis : true probability of success is not equal to 0.25
95 percent confidence interval :
0.09138957 0.22666714
sample estimates :
probability of success
0.15
```

Test for equality on two proportions (I)

In an American study that looked at the death rate of 92 patients with severe cardiac problems, 53 of these patients had pets and of those 53 patients, 3 died within a year. Of the 39 who had no animals, 11 died within the year. Can we say that the two groups had the same chance to go there ? In other words, do $\widehat{p} = 0.057$ and $\widehat{p} = 0.282$ have a significant difference given the size of the sample ?

Test for equality on two proportions (II)

- we have two laws $B(n_x, p_x)$ and $B(n_y, p_y)$
- we want to test the hypothesis $H_0 : p_x = p_y$
- $\widehat{p}_x$ has law $N(p_x, \sigma^2_{\widehat{p}_x})$ with $\sigma^2_{\widehat{p}_x} = p_x q_x / n_x$
- ditto in y
- the two samples are a priori independent

Test for equality on two proportions (III)

Independent, so we can subtract and therefore

$$\frac{\widehat{p}_x - \widehat{p}_y - (p_x - p_y)}{\sqrt{\sigma^2_{\widehat{p}_x} + \sigma^2_{\widehat{p}_y}}}$$

has a distribution of $N(0.1)$.

If H_0 is true, we have $p_x = p_y$ and so $Z = \frac{\widehat{p}_x - \widehat{p}_y}{\sqrt{\sigma^2_{\widehat{p}_x} + \sigma^2_{\widehat{p}_y}}}$ is approx. law $N(0.1)$. We can apply a risk test α as before.

Test of equality in two proportions (IV) - the return of Doctor House

$$Z = \frac{\widehat{p}_x - \widehat{p}_y}{\sqrt{\widehat{\sigma^2_{\widehat{p}_x}} + \widehat{\sigma^2_{\widehat{p}_y}}}}$$

$$Z = \frac{0.057 - 0.282}{\sqrt{0.00101 + 0.00519}} = -2.86$$

Test of equality in two proportions (IV) - the return of Doctor House

$$Z = \frac{\widehat{p}_x - \widehat{p}_y}{\sqrt{\widehat{\sigma^2_{\widehat{p}_x}} + \widehat{\sigma^2_{\widehat{p}_y}}}}$$

$$Z = \frac{0.057 - 0.282}{\sqrt{0.00101 + 0.00519}} = -2.86$$

Even with $\alpha = 1\%$ (and $c_\alpha = 2.576$), we still have too extreme a value. So we reject H_0. Now it's Doctor House's turn to interpret the result ...

And what about R ?

```
> prop.test(c(3,11), c(53,39))
2-sample test for equality of proportions with continuity correction
data : c(3, 11) out of c(53, 39)
X-squared = 7.1899, df = 1, p-value = 0.007331
alternative hypothesis : two.sided
95 percent confidence interval :
-0.4020273 -0.0488677
sample estimates :
prop 1 prop 2
0.05660377 0.28205128
```

Hypothesis test on a mean (I)

We want to test an average, that is $H_0 : \mu = \mu_0$

- We know that the natural estimator of μ is $\overline{X}$ and that $\frac{\overline{X}-\mu}{\frac{\widehat{\sigma}}{\sqrt{n}}}$ practically has a distribution of $N(0,1)$.
- When the laws are normal, $T = \frac{\overline{X}-\mu}{\frac{\widehat{\sigma}}{\sqrt{n}}}$ is exactly Student's law at $n-1$ degrees of freedom. We will adopt this configuration.

Hypothesis test on a mean (II)

In that case

- H_0 is rejected if $\mid T \mid > t_\alpha$, or $\mid \overline{X} - \mu_0 \mid > \frac{t_\alpha \widehat{\sigma}}{\sqrt{n}}$
- H_0 is accepted if $\mid T \mid \leq t_\alpha$ either $\mid \overline{X} - \mu_0 \mid \leq \frac{t_\alpha \widehat{\sigma}}{\sqrt{n}}$

Hypothesis test on a mean (II) - example

We assume that the average sleep time is 7.7 hours. A lab sells a miracle sleeping pill and obtains the following results on a sample of size 10.

7.8	8.3	7.2	9.1	8.4	6.8	7.3	7.7	8.9	9.2

With $\alpha = 5\%$, we test $H_0 : \mu = \mu_0 = 7.7$

Hypothesis test on a mean (II) - example

- $n = 10$, $\sum X_i = 80.7$ and $\sum X_i^2 = 657.61$ which gives $\overline{X} = 8.07$ and $\widehat{\sigma} = 0.8407$
- $T = \frac{\overline{X} - \mu}{\frac{\widehat{\sigma}}{\sqrt{n}}} = \frac{\sqrt{10}(8.07 - 7.7)}{0.8407} = 1.392$

With $\nu = 9$ and $\alpha = 5\%$, $t_\alpha = 2,262$. So $\mid T \mid < t_\alpha$ and we do not reject the null hypothesis. This laboratory is a crook ! ...

With your favorite statistical framework ...

```
> t.test(c(7.8, 8.3, 7.2, 9.1, 8.4, 6.8, 7.3, 7.7, 8.9, 9.2), mu = 7.7,
conf.level = 0.95)
One Sample t-test
data : c(7.8, 8.3, 7.2, 9.1, 8.4, 6.8, 7.3, 7.7, 8.9, 9.2)
t = 1.3917, df = 9, p-value = 0.1974
alternative hypothesis : true mean is not equal to 7.7
95 percent confidence interval :
7.468599 8.671401
sample estimates :
mean of x
8.07
```

Hypothesis test on an average (II) - the Marcel Schblurb case

You are about to be hired in the Macheprot company. Marcel-Benoit Schblurb, from the previous promotion of your Master, tells you that the average salary of those hired is 35 keuros. After a little investigation, you get the following data :

Emb1	Emb2	Emb3	Emb4	Emb5
34.5	36	35.2	33	34.3

It's your turn ...

Test for equality of two means (I)

We compare two samples $X_1, .., X_n$ and $Y_1, .., Y_p$ coming from two populations and we want to test the hypothesis

$$H_0 : \mu_X = \mu_Y$$

If n and p are large enough, we know that $\overline{X}$ and $\overline{Y}$ respectively follow laws $N(\mu_X, \frac{\sigma_X^2}{n})$ and $N(\mu_Y, \frac{\sigma_Y^2}{p})$. If we estimate that the populations and therefore the variables are independent, we know that $\overline{X} - \overline{Y}$ has the law $N(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{p})$.

Test for equality of two means (II)

So $\frac{\overline{X}-\overline{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{p}}}$ has an approximate distribution of $N(0,1)$. If H_0 is true,

$$\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{p}}}$$

has an approximate distribution of $N(0,1)$ and, by approximating the unknown variances by the measurements taken from the observations,

$$Z=\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{\widehat{\sigma}_X^2}{n}+\frac{\widehat{\sigma}_Y^2}{p}}}$$

is too. H_0 will be rejected for $\mid Z \mid > c_\alpha$ too extreme.

Test for equality of two means (III) - equal variances

When the variances are assumed to be equal (very common assumption), the formula is simplified and the law $\frac{\overline{X}-\overline{Y}}{\sigma\sqrt{\frac{1}{n}+\frac{1}{p}}}$ is $N(0,1)$.

We estimate $\widehat{\sigma}$ from $\widehat{\sigma}_X$ and $\widehat{\sigma}_Y$ using the unbiased estimator :

$$\widehat{\sigma} = \frac{(n-1)\widehat{\sigma}_X^2 + (p-1)\widehat{\sigma}_Y^2}{n+p-2}$$

and

$$T = \frac{\overline{X}-\overline{Y}}{\widehat{\sigma}\sqrt{\frac{1}{n}+\frac{1}{p}}}$$

is exactly a Student's law with $n+p-2$ degrees of freedom. The hypothesis H_0 will be accepted if $\mid T \mid \leq t_\alpha$.

RRRRR - A French movie ...

```
> t.test(c(2,3,5,6,3), c(2,4,5,1,1), var.equal = FALSE)
Welch Two Sample t-test
data : c(2, 3, 5, 6, 3) and c(2, 4, 5, 1, 1)
t = 1.0954, df = 7.921, p-value = 0.3055
alternative hypothesis : true difference in means is not equal to 0
95 percent confidence interval :
-1.330505 3.730505
sample estimates :
mean of x mean of y
3.8 2.6

> t.test(c(2,3,5,6,3), c(2,4,5,1,1), var.equal = TRUE)
Two Sample t-test
data : c(2, 3, 5, 6, 3) and c(2, 4, 5, 1, 1)
t = 1.0954, df = 8, p-value = 0.3052
alternative hypothesis : true difference in means is not equal to 0
95 percent confidence interval :
-1.326101 3.726101
sample estimates :
mean of x mean of y
3.8 2.6
```

Test for equality of two means (IV) - paired data

If we want to measure the effect of a treatment on a population, we have the same individuals during the two tests ($n = p$). We can no longer assume independence and stay within the previous framework. It suffices to take the variable $W_i = X_i - Y_i$ and the null hypothesis becomes $H_0 : \mu_W = 0$. We can come back to the mean value test, with a Student variable at $n - 1$ degrees of freedom.
If we do not take these precautions, we overestimate the variance $(\sigma_X^2 + \sigma_Y^2)$ and we can end up with biased acceptances of the null hypothesis.

Test for equality of two means (V) - example

A couple of SIIA professors have finally decided to lose weight and are resolved to stop eating bread with their noodles. Their weights before and after the diet are as follows :

Subject	1	2	3	4	5	6
Front	64	54	73	59	64	68
After	61	54	71	58	61	66

Did they really do well to deprive themselves ($\alpha = 5\%$) ?

Test for equality of two means (V) - example

A couple of SIIA professors have finally decided to lose weight and are resolved to stop eating bread with their noodles. Their weights before and after the diet are as follows :

Subject	1	2	3	4	5	6
Front	64	54	73	59	64	68
After	61	54	71	58	61	66

Did they really do well to deprive themselves ($\alpha = 5\%$) ?
$H_0 : \mu_{av} = \mu_{ap}$, $\overline{X} - \overline{Y} = 1.833$, $\widehat{\sigma}^2 = 1.367$, $T = 3,84$, $\nu = 5$, $t_\alpha = 2.571$ so we can estimate that the regime is working.

with the letter following Q

```
> t.test(c(64, 54, 73, 59, 64, 68), c(61, 54, 71, 58, 61, 66), var.equal = TRUE, paired = FALSE)
Two Sample t-test
data : c(64, 54, 73, 59, 64, 68) and c(61, 54, 71, 58, 61, 66)
t = 0.502, df = 10, p-value = 0.6266
alternative hypothesis : true difference in means is not equal to 0
95 percent confidence interval :
-6.304374 9.971041
sample estimates :
mean of x mean of y
63.66667 61.83333

> t.test(c(64, 54, 73, 59, 64, 68), c(61, 54, 71, 58, 61, 66), var.equal = TRUE, paired = TRUE)
Paired t-test
data : c(64, 54, 73, 59, 64, 68) and c(61, 54, 71, 58, 61, 66)
t = 3.8414, df = 5, p-value = 0.01211
alternative hypothesis : true difference in means is not equal to 0
95 percent confidence interval :
0.6064956 3.0601710 sample estimates :
mean of the differences
1.833333
```