GM – Probabilités et Statistique

http://moodle.epfl.ch/course/view.php?id=18431

Lecture 8

- Hypothesis tests : Introduction
- Hypothesis test for μ
- Hypothesis test for a proportion
- (Hypothesis test for 2 *independent* means / proportions – video)

Hypothesis tests – introduction

- Let's consider the experiment of tossing a coin
- In particular, consider the question of whether the coin is *fair* (equally likely to show 'Heads' or 'Tails')
- Formally : we want to make some *inference* about P(Heads)
- **Activity : Take a coin**, toss it several times (let's say 10 times) and count how many times it shows Heads
- Assuming that it is fair (P(Heads)=0.5), we can try to assess whether this assumption is *compatible* with our observations

10 tosses

Number of heads

Number of tosses	0	1	2	3	4	5	6	7	8	9	10
1	0.5 1	0.5 0.5									
2	0.25 1	0.5 0.75	0.25 0.25								
3	0.125 1	0.375 0.875	0.375 0.5	0.125 0.125							
4	0.0625 1	0.25 0.9375	0.375 0.6875	0.25 0.3125	0.0625 0.0625						
5	**0.03125** **1**	0.15625 0.96875	0.3125 0.8125	0.3125 0.5	0.15625 0.1875	**0.03125** **0.03125**					
6	**0.01563** **1**	0.09375 0.98438	0.23438 0.89063	0.3125 0.65625	0.23438 0.34375	0.09375 0.10938	**0.01563** **0.01563**				
7	**0.00781** **1**	0.05469 0.99219	0.16406 0.9375	0.27344 0.7734	0.27344 0.5	0.16406 0.22656	0.05469 0.0625	0.00781 0.00781			
8	**0.00391** **1**	**0.03125** **0.99609**	0.10939 0.96484	0.21875 0.85547	0.27438 0.63672	0.21875 0.36328	0.10939 0.14453	**0.03125** **0.03516**	**0.00391** **0.00391**		
9	**0.00195** **1**	**0.01758** **0.99804**	0.07031 0.98047	0.16406 0.91016	0.24609 0.74609	0.24609 0.5	0.16406 0.25391	0.07031 0.08984	**0.01757** **0.01953**	**0.00195** **0.00195**	
10	**0.00098** **1**	**0.00977** **0.99902**	0.04394 0.98926	0.11719 0.94531	0.20508 0.82812	0.24609 0.62304	0.20508 0.37695	0.11719 0.17188	0.04395 0.05469	0.00977 0.01074	0.00098 0.00098

Significant evidence ($p<0.05$) that the coin is biased towards head or tail.

8

10 | **0.04395**
0.05469

Probability of obtaining 8 heads in 10 tosses

Probability of obtaining at least 8 heads in 10 tosses

Statistical hypothesis testing

Definition : A (statistical) **hypothesis** is a *statement about a population* **parameter**

- 2 competing *hypotheses*
  - H (or H_0) : the *NULL hypothesis*, usually more conservative
  - A (or H_A) : the *ALTERNATIVE hypothesis*, in general the one we are interested in
- Examples of NULL hypothesis :
  - The coin is fair
  - This new machine is no better (or worse) than a placebo
- Examples of alternative hypothesis :
  - The coin is biased (either towards Tails or Heads)
  - The coin is biased towards Heads
  - The new machine is better than the old one

Test statistic

- In order to decide between the hypotheses, we need to measure how far the observed value is from what we expect to see if the NULL hypothesis H is true – that is, we need a **test statistic** (TS) T
- The statistic T is chosen so that 'unusual' values (too big and/or too small) suggest that the NULL hypothesis H is false
- We compute T based on our sample, and denote the observed value as t_{obs}

Example

Example 8.1 On a SRS of 25 farms in a particular county, the effect of spraying against a bug was evaluated by measuring crop yields (bushels per acre) on sprayed and unsprayed strips in a field on each farm.

The data :

sample mean difference $\overline{x}$= 4.7 bushels per acre
sample SD of differences : s = 6.5 bushels per acre

Assume that a gain of 2 bushels per acre would pay for the cost of spraying. Does the sample furnish strong evidence that spraying is profitable *? ?*

Now, we will give the *concrete reasoning* to address this question.

Steps in hypothesis testing (I)

1. **Identify** the population parameter being tested
   - Here, the parameter being tested is the population mean difference in yield μ
2. **Formulate** the NULL and ALT hypotheses
   - $H : \mu = 2$ (or $\mu \leq 2$)
     $A : \mu > 2$
3. Compute the **test statistic (TS)**
   - $t_{obs} = (4.7 - 2)/(6.5/\sqrt{25}) = 2.08$

Hypothesis truth vs. decision

Truth \ Decision	not rejected	rejected
true H	☺ specificity	X Type I error (False +) α
false H	X Type II error (False -) β	☺ Power $1 - \beta$; sensitivity

Some terminology

- The chance of rejecting a NULL which is *true* is α; this type of mistake is called a *Type I error* or *false positive*
- The chance of *NOT* rejecting a NULL which is *false* is β; this type of mistake is called a *Type II error* or a *false negative*
- The probability of correctly rejecting a NULL hypothesis (that is, when it is really false) is called the *power*

p-value

- We decide on whether or not to *reject* the NULL hypothesis H based on the chance of obtaining a value of T *as or more extreme* (as far away from what we expected or even farther, in the direction of the ALT) than the one we got, *ASSUMING THE NULL IS TRUE*
- This chance is called the *observed significance level*, or *p-value* p_{obs}
- The smaller the value of p_{obs}, the more doubt thatd the NULL hypothesis H is true
- A result (TS) with a *p*-value less than some pre-specified false positive *level* (or *size*) α is said to be 'statistically significant' at that level
- **Note :**

statistical significance ≠ practical significance ≠ scientific significance

p-value – interpretation

- Just as in the interpretation of a CI, the interpretation of a p-value is a little tricky
- In particular, the *p*-value does **NOT** tell us the probability that the NULL hypothesis is true
- The *p*-value represents the chance that we would see a difference as big as we saw (or bigger) **IF** there were really nothing happening other than chance variability
- (Frequentist interpretation of probability)

Steps in hypothesis testing (II)

4. Compute the p-value
   Assuming the CLT, under H : $p_{obs} = P(Z > 2.08) \approx 0.02$

5. *Decision rule and practical interpretation* : REJECT the NULL hypothesis H if $p_{obs} \leq \alpha$
   (This is a type of argument by contradiction)

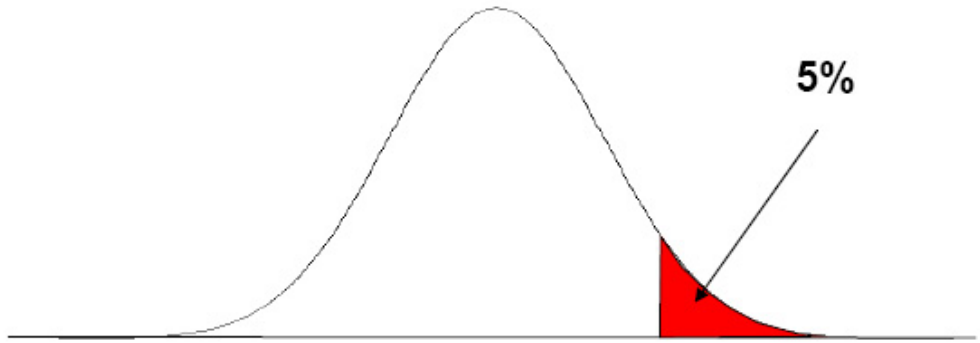
- A typical value of α is 0.05, due mainly to historical reasons
- In practice, you should choose a value of α appropriate to the situation
- Here, if we use $\alpha = 0.05$, the decision will be to **REJECT** H
- If we instead use $\alpha = 0.01$, the decision is **DO NOT REJECT** H
- **Interpretation :** _____

Unilateral vs. bilateral test

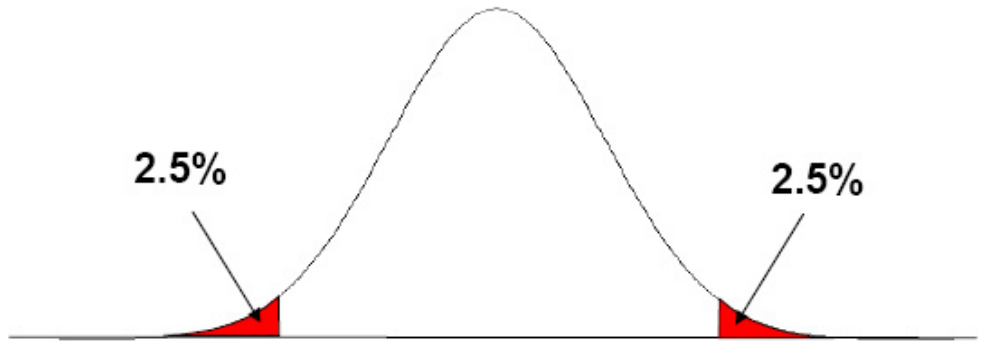
- The choice of alternative hypothesis influences the conclusion
- If the ALT is "the coin is biased", we do not specify the direction of the bias
- This is a *bilateral test*
- If α is, for example, 0.05, that means that we must allow $\alpha/2$ (0.025) for bias towards 'Heads' and $\alpha/2$ (0.025) for bias towards 'Tails'
- If the ALT is "the coin is biased towards Heads", we are specifying the direction of bias and the test is *unilateral*

Unilateral vs. bilateral test

One-sided
e.g. H_A: $\mu>0$



Two-sided
e.g. H_A: $\mu\neq0$



Example

Example 8.2 From long experience with a process for manufacturing an alcoholic beverage, it is known that the yield distribution has a mean of 500 units and a standard deviation (SD) of 100 units.

A modification of the process is suggested for which it is claimed that the mean yield will increase (the SD will not change). With a random sample of 49 observations, a mean yield of 535 units is observed.

Test the claim using a hypothesis test. What do you conclude (use $\alpha = 0.05$) ?

Test

1

2

3

4

5

Another example

Example 8.3 The mean yield of corn in the US is about 120 bushels per acre. A random sample of 50 farmers from Illinois (a state in the US) yielded a sample mean of $\bar{x} = 123.6$ bushels per acre. Assume that the SD of the yield for this population is $\sigma = 10$ bushels per acre.

Test whether the mean yield μ for Illinois is the same as the national average. (Specify your assumptions.) Use $\alpha = 0.05$.

Test

1

2

3

4

5

PAUSE

Hypothesis test for a proportion p

- We have considered the *z-test* for the population mean μ
- This test is based on the *CLT*
- Test statistic for the mean μ : $\boxed{T = \dfrac{\overline{X} - \mu_H}{S/\sqrt{n}}}$
- We can also test *the population proportion p* of individuals possessing a particular characteristic
- The CLT also applies for an observed proportion $\hat{p} = X/n$
- We saw how to find the SE of a proportion : $\sqrt{p_H \cdot (1 - p_H)/n}$
- Thus, the test statistic for the test for a proportion p is :

$$\boxed{T = \frac{\hat{p} - p_H}{\sqrt{p_H \cdot (1 - p_H)/n}}}$$

$$\left(\text{or : } \quad \frac{\hat{p} - p_H}{\sqrt{\hat{p} \cdot (1 - \hat{p})/n}}, \text{ which is asymptotically equivalent to T} \right)$$

Example : Hypothesis test for a proportion

Example 8.4 A new surgical intervention has been developed to correct cardiac malformation for infants younger than 1 month. Previously, the procedure had been used on infants older than 1 month, with a success rate of 91%. A study is carried out to determine whether the success rate of the new procedure is higher than 91% : $n = 200$, with 187 successes ($= 93.5\%$).

Is this new procedure more effective for infants younger than 1 month (than for infants older than 1 month) **??**

Test

1

2

3

4

5

Another example

Example 8.5 An Editor-in-Chief thinks that 50% of young brides are younger than their husbands.

She takes a random sample of 100 brides, of which 53 are younger than their husbands.

Carry out a hypothesis test in order to determine if this percentage is compatible with being 50% or not. Use $\alpha = 1\%$.

Test

1

2

3

4

5

Standard error for a difference

- Variance of the difference between two (independant) means :

- Variance of the difference between two (independant) proportions :

Test of comparison on 2 independent samples

- Until now, we have been interested by *a single population*. Often, however, we are interested in the **comparison of two populations**. In this case, we carry out a test on *two independent samples*.
- When we compare two *means* (or *proportions*) the basic notion is the same as above : for T, we use the *standardized difference* between the sample means (or proportions).
- TS for the *difference in means* from two independent populations : $$\frac{\overline{X_1} - \overline{X_2}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$
  (use s instead of σ if σ is unknown)
- TS for the *difference in proportions* from two independent populations : $$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}}$$

Example

Example 8.6 The amount of a certain blood trace element varies with a SD of 16 ppm (parts per million) for men and 10 ppm pour women. Random samples of 64 men and 20 women have mean concentrations of 28 and 33 ppm, respectively.

Test whether the mean concentrations are the same for men and women (use $\alpha = 0.05$).

Test

1

2

3

4

5

Example

Example 8.7 During a study at a certain large university, a researcher is interested in whether women study engineering less frequently than men. A random sample of 100 men found 28 (28%) studying engineering, while a random sample of 64 women found 12 (18,75%) studying engineering.

Test his theory that women study engineering less frequently (use $\alpha = 0.05$).

Test

1

2

3

4

5

Test of comparison for 2 proportions, again

- The standard error (SE) for the difference of two proportions (from two independent samples) is :

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- However, the null hypothesis specifies *equality of proportions*, such that we would be able to use an estimate based on the *global proportion* (by combining the samples) :

$$p_{\text{agrégé}} = \frac{x_1 + x_2}{n_1 + n_2},$$

where x_i = number of 'success' from population i ($i = 1, 2$) and n_i are the respective sample sizes

Test of comparison for 2 proportions, again (cont)

- Thus, **under the null hypothesis H :**

$$SE_{ag} = \sqrt{\frac{\hat{p}_{ag}(1-\hat{p}_{ag})}{n_1} + \frac{\hat{p}_{ag}(1-\hat{p}_{ag})}{n_2}}$$

$$= \sqrt{\hat{p}_{ag}(1-\hat{p}_{ag})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

- (For me, either of these methods is ok : under the NULL hypothesis, the distribution of T will be correct either way.)
- TS for the difference of *proportions* from two independent populations :

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_{ag}(1-\hat{p}_{ag})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Relation between test and CI

- In a study, the confidence interval shows the precision with which the treatment effect is known
- When the CI contains the value assumed under the null hypothesis, it is not possible to exclude the possibility that *the true value equals this null value*
- Thus : the null value *cannot be considered as statistically significant*
- Inversely, if a test is significant at level α, it means that the $100(1-\alpha)\%$ CI *does not contain the null value*
- Summary :
  - *significant* test (level α) $\Leftrightarrow$ $100(1-\alpha)\%$ CI *does not contain* the null value
  - *non-significant* test (level α) $\Leftrightarrow$ $100(1-\alpha)\%$ CI *contains* the null value

Regarding small samples...

- The z-test that we have studied assumes that the sampling distribution of the test statistic T is *Normal*
  - exactly, or
  - approximately, by the CLT
- However, IF :
  1. the data are Normally distributed, AND
  2. 4the population SD σ is *unknown*, AND
  3. the sample size is *small* (for example, under 30)

  THEN : the true sampling distribution of T has *heavier tails* than the Normal distribution
- In this case, you should use the *t-test* (next week)

Pitfalls in hypothesis testing

Pay attention :

- Difficulty interpreting tests based on *arbitrary asmples* (not random) and observational (not experimental dataa)
  - in practice, most samples are not exactly random
  - the p-value of this kind of sample should be considered as an *approximate indicator* of importance
- Statistical significance vs. practical significance
  - a samll p-value can come from a very small deviation from the null hypothesis if the sample size is very large
- Risk of *searching for* significance (multiple testing)
- Ignoring the *absence* of significance