Applied Biostatistics

https://moodle.epfl.ch/course/view.php?id=15590

- Course organization
- Quiz
- Reproducible Research
- Hypothesis testing - review of basic notions

Organisation

- **Instructor :** *Darlene Goldstein* (me)
- **Assistants :** Francesco Spadaro (+ a few others, I hope)
- **Course meeting time :** Monday 8.15 – 10.00, CM 1 120 (here)
- **Lab/Exercice session :** Go to **one** meeting per week :
  - Monday 10.15 – 12.00, CM 0 11, **OR**
  - Tuesday 10.15 – 12.00 in CH B3 30
- **Course note :**
  - 2 short reports ~ 5 pages (1/6 each) : 1 data analysis, 1 article review
  - 1 longer report~ 15 pages (2/3) : data analysis report
- **Software :** R Statistical Software.
  `http://cran.r-project.org/`

Reproducible research principle

- Claerbout : 'An article about computational science in a scientific publication is **not** the scholarship itself, it is merely **advertising** of the scholarship. The **actual scholarship** is the complete software development environment and the complete set of instructions which generated the figures.'
- Wavelet community, Stanford University
  - Buckheit and Donoho : 'When we publish articles containing figures which were generated by computer, we also publish the *complete software environment* which generates the figures.'
- Anecdotes
  - 'Final' versions of figs for publication
  - Lost or stolen work
  - Communication
  - Applying old/existing methods on new data
  - Reconstructing work of others

Steps leading to a report

- Data **entry** and storage
- Data **cleaning** – check, resolve, correct data entry errors
- **Prepare** data for analysis – transform/recode variables, create new variables, *etc.*
- Carry out **statistical analyses**
- Save desired results/graphs
- Write the results report, which may include *documentation text, tables and/or graphs*

Report preparation

- A common approach is to write the report around the results
- Results commonly obtained via 'point and click' approach (*e.g.* MS Excel, SPSS,)
- Then copy/paste or – worse – type by hand the results into the word processor used to create the report
- **NOT A GOOD METHOD – DON'T DO THIS!!!!** :
  - no documentation on how the results were obtained, how missing data are handled, *etc.*
  - unreliable results

Problems with this approach : examples

- You need to run an additional analysis ; when you re-run the primary analysis, the *results don't match* what you have in your manuscript
- You go to the project folder to run additional analyses and find *multiple* data files, multiple analysis files, multiple results files and can't remember which ones are relevant
- You have spent a week running your analysis and creating a results report (including tables and graphs) to present to your collaborators ; you then receive an email from your PI asking you to **regenerate the report** based on a subset of the original data set and including an additional set of analyses – **AND** she would like it by tomorrow's meeting **! !**

Problems with this approach : specifics

- With point and click programs, *no way to record/save* the steps that generated the documented results
- Common to keep analysis code, results, reports as separate files and save various versions of each of these separately ; after several modifications, *unclear which version* corresponds to the desired analysis/results
- Every time analyses and/or results change, have to regenerate the results report by hand – *wastes time* ! !
- Easy to introduce *human error* into report – typing in results by hand, copying/pasting the wrong tables/graphs, *etc.*

Research practice

- *Discipline* in software building
- From the start, *expect* it to be made available to others as part of the publication of their work
- *Avoid copy/paste/editing* in a way that is not reproducible
- (Also think in terms of program re-use)

Literate Programming

- Donald Knuth
- Combining the use of a text formatting language (such as TeX) and a conventional programming language (like C or R) so as to maintain documentation and source code together, the art of writing computer programs for the human reader
- may use *inverse comment convention*
- A kind of literate programming where the program code is marked to distinguish it from the text, rather than the other way around as in normal programs
- Literate programming paradigm :
  1. **parse** the source document and separate code from narrative
  2. **execute** source code and return results
  3. **mix results** from the source code with the original narrative

WEB (not www)

- WEB (Donald Knuth), noweb (Norman Ramsey)
- a WEB system consists of two processors, called *WEAVE* and *TANGLE*
  - WEAVE "weaves" the document for a human reader, producing TeX output
  - TANGLE "tangles" the document for a computer, producing a plain programming language file to be compiled, linked and executed
- WEB (and variants) are not the only environments for Literate Programming
- We will focus on using **knitr with R**

Good/bad practices (1)

- Manage all source files under the *same directory* and use *relative path names* whenever possible – absolute paths can break code/reproducibility
- *Do not* change the working directory after computing started ; if necessary, set at *beginning* of R session, and if absolutely unavoidable then *restore* the directory later
- Compile documents in a *'clean' R session* : existing objects in a current session may contaminate the code
- (OK to do interactive data analysis while checking results for code chunks, but at end, compile report in batch mode with a new R session so that all results are freshly generated from code)

Good/bad practices (2)

- Avoid commands that need *human interaction*, since human input can be unpredictable (and therefore not reproducible) ; instead, explicitly code for the required input
- Avoid environment variables for data analysis ; if you need to set up options, do it *inside* the source document
- Attach `sessionInfo()` and instructions on how to compile the document

Barriers to reproducible research

- Huge data
- Data confidentiality issues
- Software version and configuration – changing versions/availability
- Competition

Tools in R

- CRAN Task Views : https://cran.r-project.org/web/views/
- Reproducible research in R : https://cran.r-project.org/web/views/ReproducibleResearch.html
- Compendium concept
  - dynamic document
  - data
  - auxiliary software

Editor

- Could use *ANY* text editor with the **knitr** package, since the documents are *plain text files*
- Special text editors are *more useful* :
  - input R code chunks more easily
  - more convenient to call R and knitr to compile source documents to pdf/html within an editor, as well as sending R code chunks to R from within the editor directly
- Several editors available, *e.g.* :
  - **RStudio** – has the most comrehensive support for knitr (and Sweave)
  - **LyX** – front end for LaTeX with a GUI to help with document writing
  - **Emacs/ESS** (Emacs Speaks Statistics) – supports statistical software packages, including R

PAUSE

Statistical hypothesis testing - review

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*, the one we are actually interested in
- Examples of NULL hypothesis :
  - The coin is fair
  - This new drug is no better (or worse) than a placebo
  - There is no difference in weight between two given strains of mice
- Examples of Alternative hypothesis :
  - The coin is biased (either towards tail or head)
  - The coin is biased towards tail
  - The coin has probability 0.6 of landing on tail
  - The drug is better than a placebo

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 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 H is false
- T is computed based on the sample ; we denote the observed value as t_{obs}

Example

On 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.

Data :
sample mean difference = 4.7 bushels per acre
sample SD of differences = 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 *??*

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 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*
- In other contexts, these quantities are sometimes referred to with other terminology :
  - The *specificity* of a test is the chance that the test result is negative given that the subject is negative ; this is just $1 - \alpha$
  - The *sensitivity* of a test is the chance that the test result is positive given that the subject is positive ; this is just $1 - \beta$, also called *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 that H is true
- A 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 $\neq$ practical significance $\neq$ scientific significance

p-value interpretation

- 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

Steps in hypothesis testing (II)

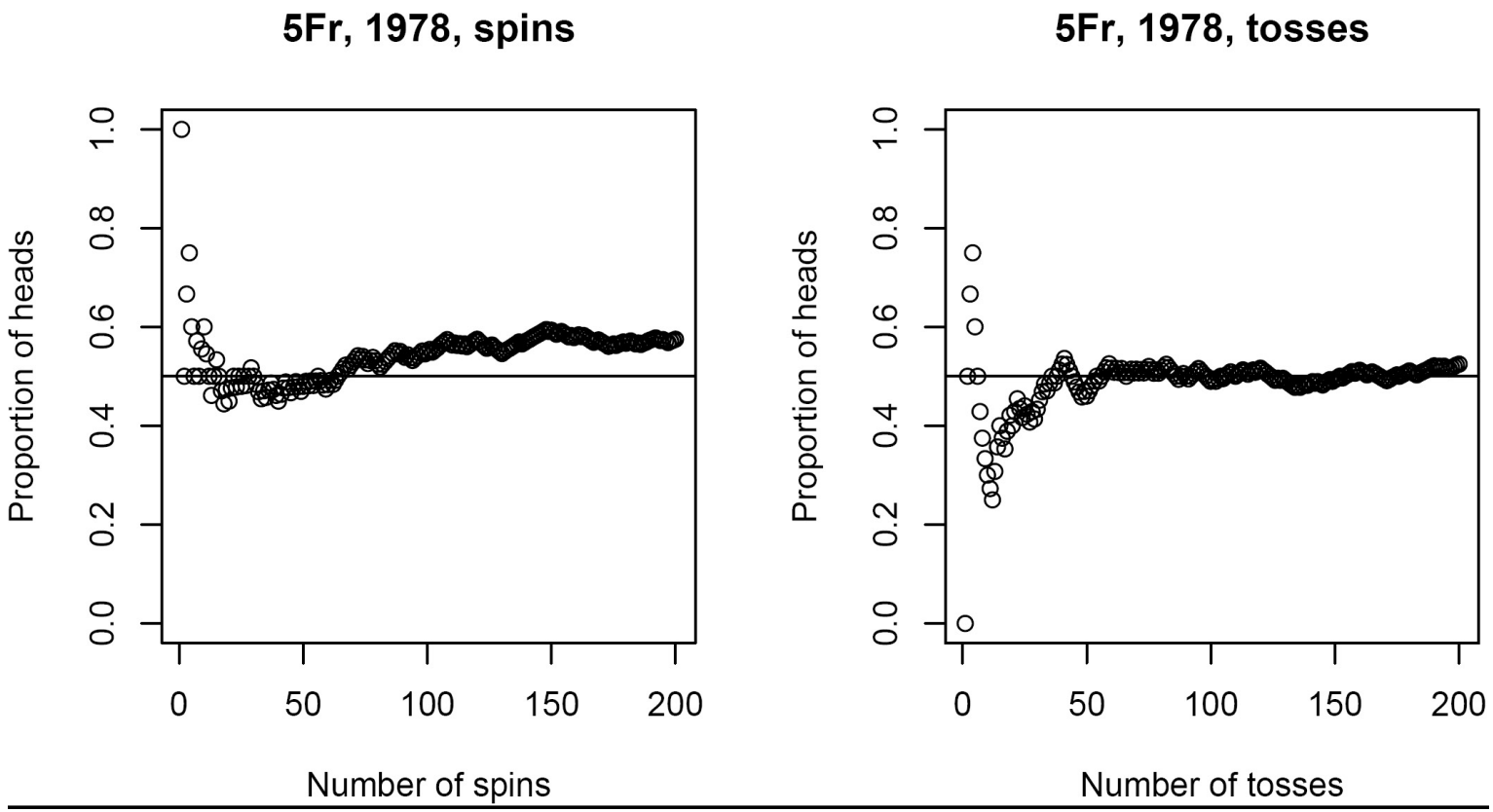
4. Compute the p-value
   Here, $p_{obs} = P(Z > 2.08) = 0.02$
5. (Optional) *Decision Rule* : REJECT H if $p_{obs} \leq \alpha$
   (This is a type of argument by contradiction)

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 here will be REJECT H; if we instead use $\alpha = 0.01$, the decision is DO NOT REJECT H

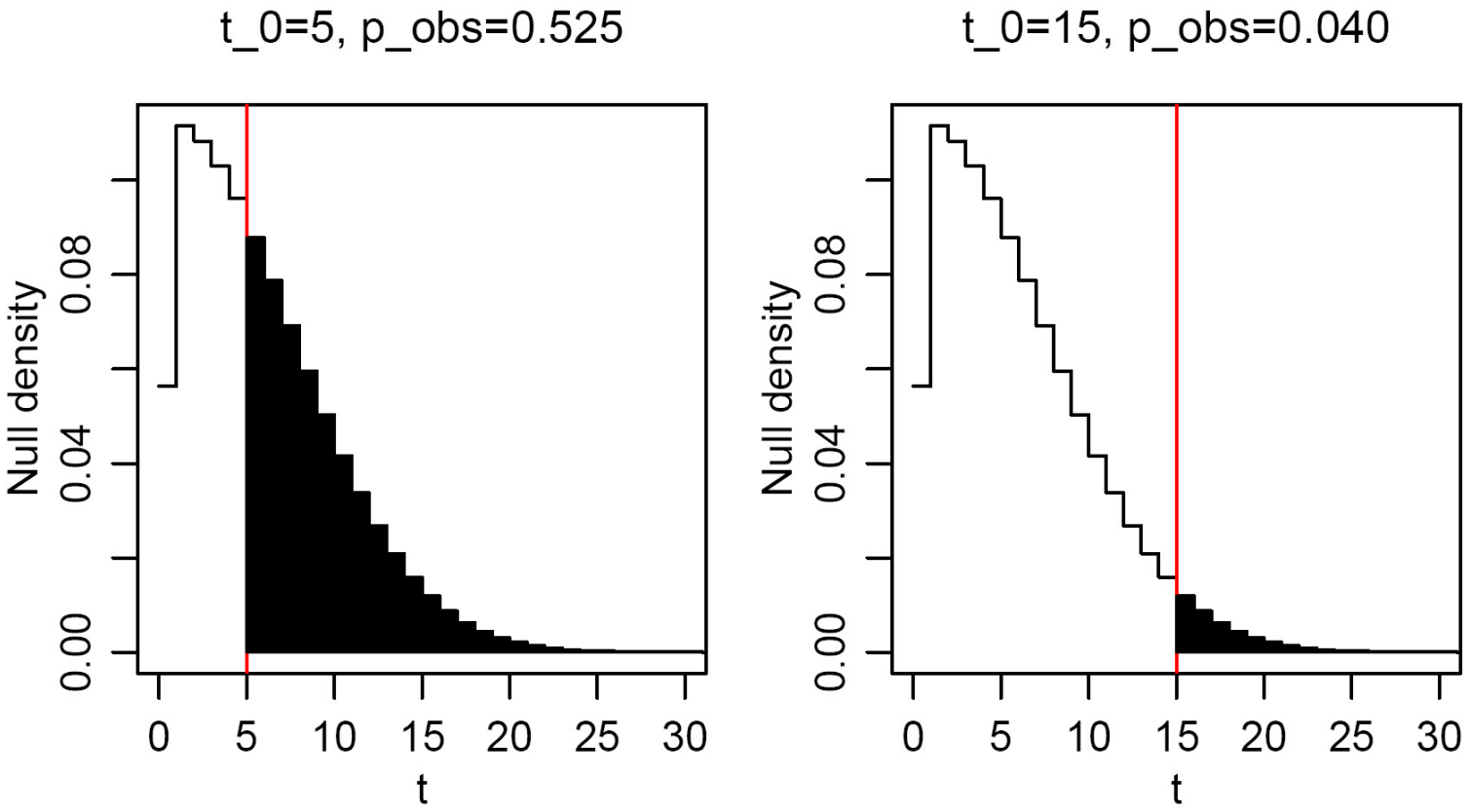
Example – Spinning a 5 Fr coin

Does $P(Heads) = 0.5$ when we *spin* the coin ?
200 trials : $x_{obs} = 115$ when spinning ; $x_{obs} = 105$ when tossing.



Is the coin/process fair ??

Null distribution for the coin



Interpretation of p_{obs}

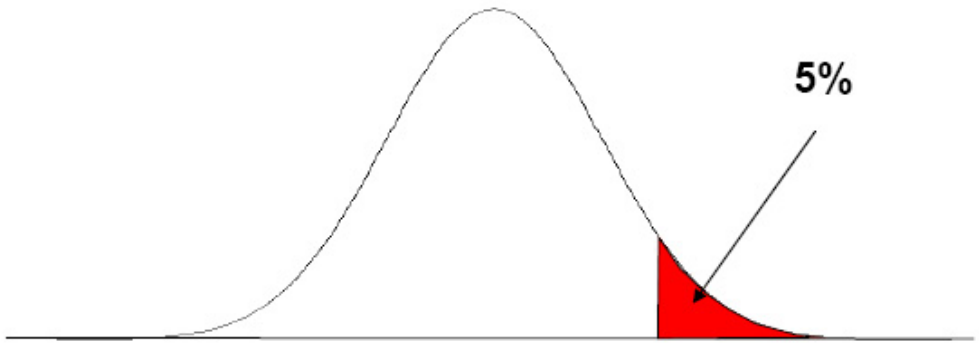
- The smaller the p-value (p_{obs}), the more we doubt the NULL hypothesis H
- There are 2 possibilities :
  - H is TRUE, and a rare event has occurred
  - H is FALSE
- The decision about whether or not to REJECT H depends on our judgement of the importance of the two types of possible errors :
  - **Type I error** : H is TRUE, but we REJECT it
  - **Type II error** : H is FALSE, but we DO NOT REJECT it
- The choice depends on the consequences of the two types of errors, and therefore on *the context of the problem*

Unilateral vs. bilateral tests

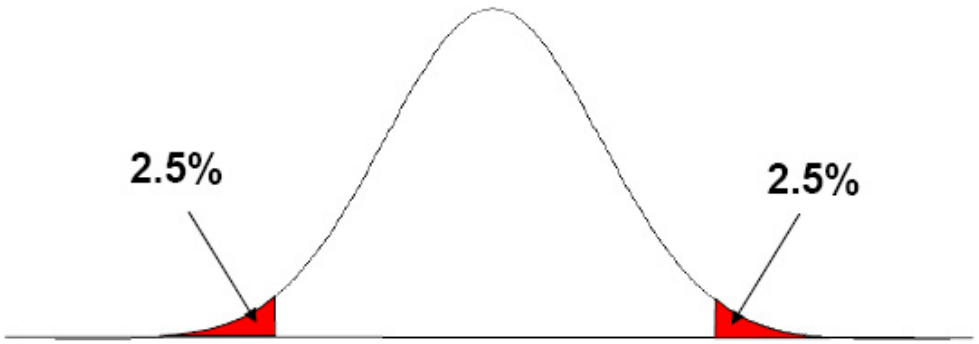
- The choice of hypotheses influences the conclusion
- If the ALTis "la coin is biased", we haven't specified the direction of the bias
- Here we would carry out a *bilateral test*
- If α is, *e.g.* 0.05, then we have $\alpha/2$ (0.025) for bias towards HEADS and $\alpha/2$ (0.025) for bias towards TAILS
- If the 'ALT is "the coin is biased towards HEADS", we have specified the direction and the test is *unilateral*

Test unilatéral vs. bilatéral

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



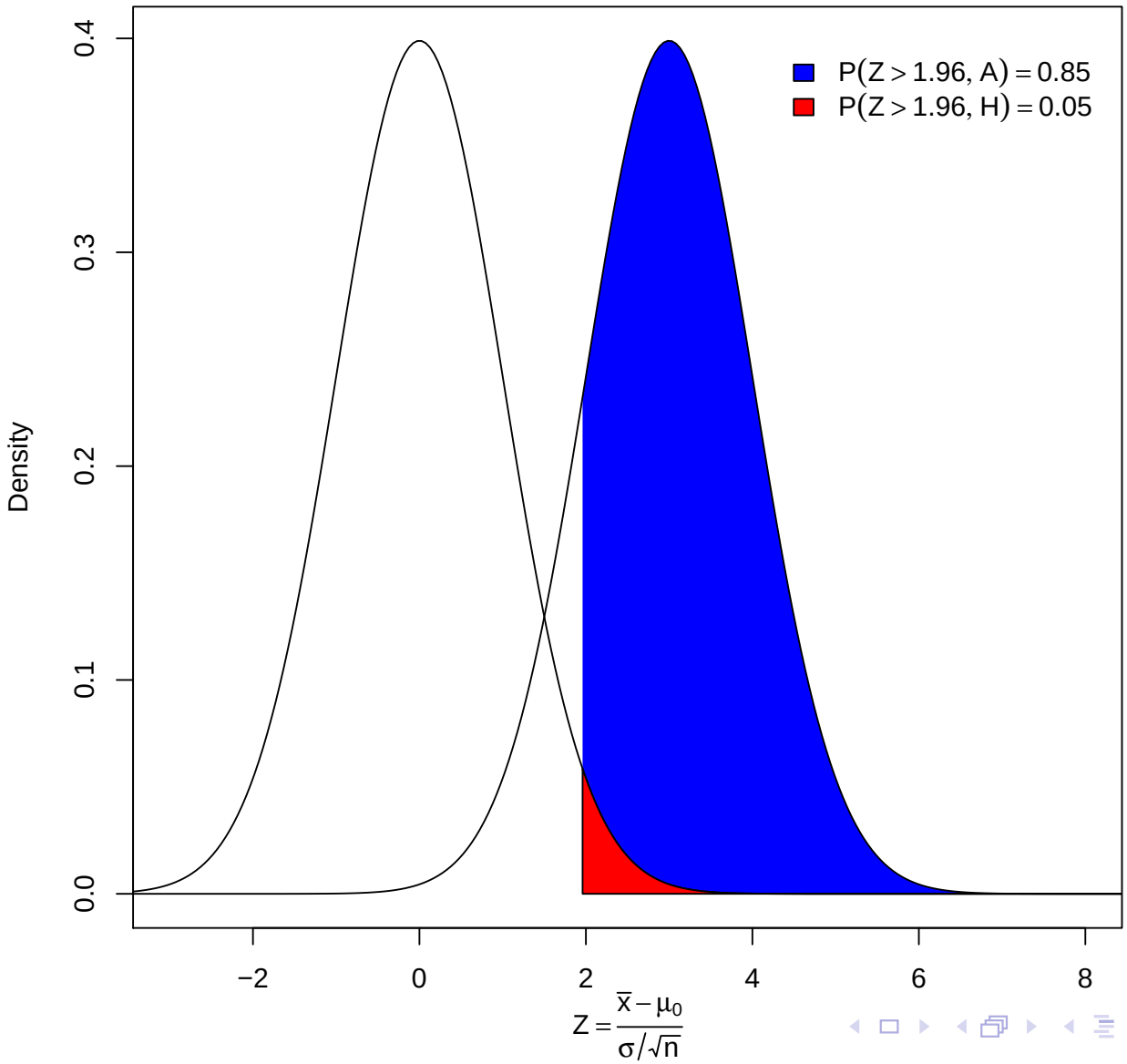
Two-sided
e.g. H_A: $\mu\neq0$



Power of a test

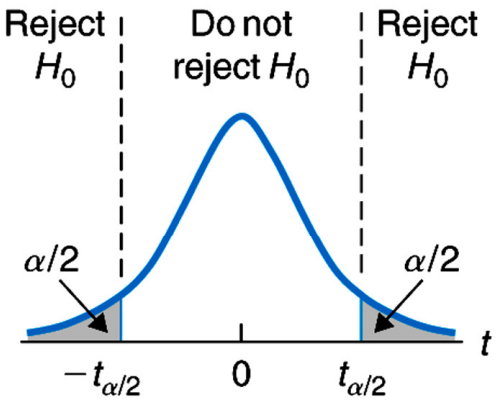
- Not only do you want to have a low FALSE positive rate, but you would also like to have a high TRUE positive rate – that is, high **power**, the chance to find an effect (or difference) if it is really there
- Statistical tests will not be able to detect a true difference if the sample size is too small compared to the effect size of interest
- To compute or estimate power of a study, you need to be able to specify the α level of the test, the sample size n, the effect size d, and the SD σ (or at least an estimate s)

Power


P(Z > 1.96, A) = 0.85
P(Z > 1.96, H) = 0.05
Density
0.0
0.1
0.2
0.3
0.4
−2
0
2
4
6
8
$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

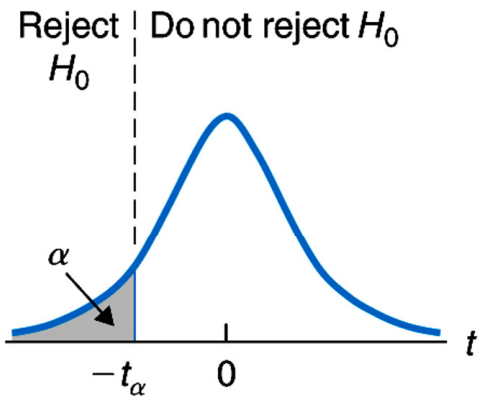
Rejection region in direction of ALT

H: $\mu = \mu_H$
A: $\mu \neq \mu_H$



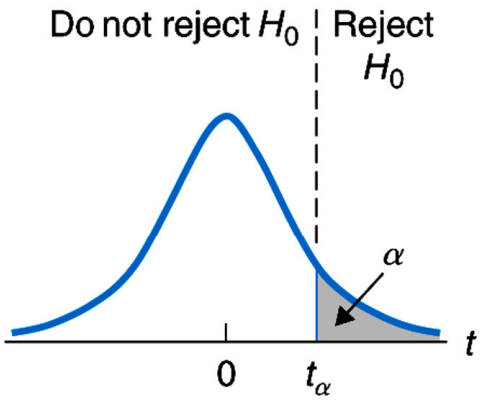
Two-tailed

H: $\mu = \mu_H$
A: $\mu < \mu_H$



Left-tailed

H: $\mu = \mu_H$
A: $\mu > \mu_H$



Right-tailed

Large-sample tests : CLT

- **Central Limit Theorem (CLT)** : Suppose $X_1, X_2, \ldots$ are independent and identically distributed (iid) such tat $E[X_i] = \mu < infty$ and $Var(X_i) = \sigma^2 < \infty$ exist. Then the distribution of
$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$
approaches a normal distribution as $n \to \infty$.
- This means that for n 'sufficiently large', *the distribution of the sum (or the mean)* is approximately normal
- A test based on the CLT is called a z-test
- Power calculations for the z-test are straightforward (distribution of T under the ALT hypothesis is normal)

Test for a single mean or proportion

- Testing a population mean μ :
  $H : \mu = \mu_H$
  $A_1 : \mu \neq \mu_H \quad \text{or} \quad A_2 : \mu > \mu_H \quad \text{or} \quad A_3 : \mu < \mu_H,$
  with $T = \frac{\hat{\mu} - \mu_H}{\sigma/\sqrt{n}}$.

- Testing a population proportion p :
  $H : p = p_H$
  $A_1 : p \neq p_H \quad \text{ou} \quad A_2 : p > p_H \quad \text{ou} \quad A_3 : p < p_H,$
  with $T = \frac{\hat{p} - p_H}{\sqrt{\frac{p_H(1-p_H)}{n}}}$.

Two-sample tests

- Above, we have been interested in a *single population*; Often, however, we are interested in *comparing two (independent) populations*
- In this case, we carry out a *two-sample test*
- When comparing two *means* (or *proportions*) the basic idea is the same as above : for T we use the *standardized difference* of the sample difference in means (or proportions)
- T for difference of independent means : $\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$

  (use s instead of σ if σ is unknown)
- T for difference of independent proportions : $\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p^*(1-p)^*\left(\frac{1}{n_1 + \frac{1}{n_2}}\right)}}$, where $p^* - \frac{X_1 + X_2}{n_1 + n_2}$

What about small samples ?

- The z-test that we have covered assumes the sampling distribution of the test statistic T is normal, either exactly or by the CLT
- However, if the population SD is *not known* and the sample size is small (less than about 30, say) then the true sampling distribution of T has *heavier tails* than the normal distribution – in this case, we use the t-test
- The test statistic for the t-test is also the standardized sample mean (using the estimated SD in the denominator), and in the one-sample case follows a t-distribution with $n-1$ *degrees of freedom*

Student (= William Sealy Gosset)

W. S. Gosset

Guinness

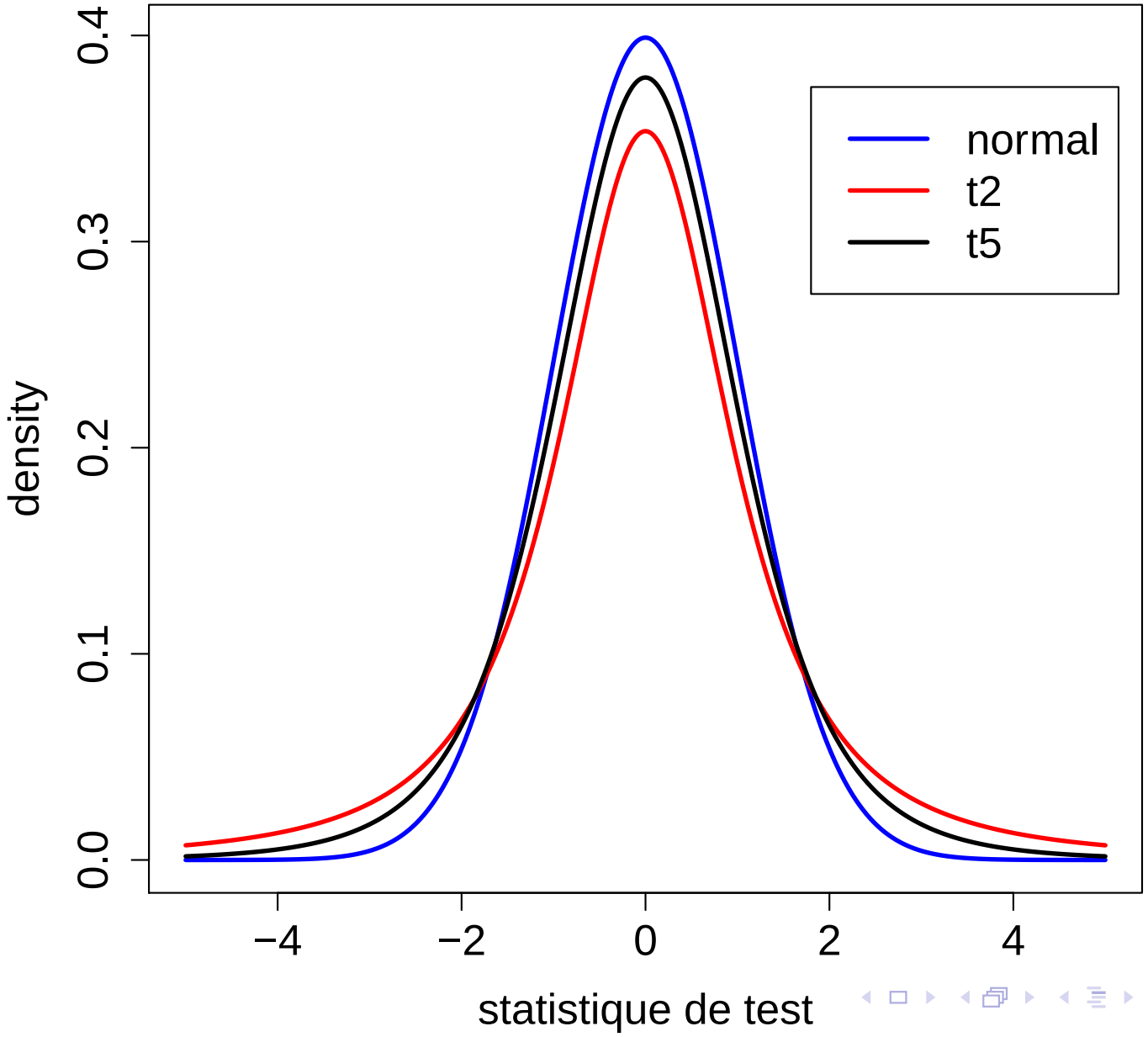
t-test for a single mean

- For small samples of normally distributed observations with σ unknown, the CLT is not applicable – there is *additional uncertainty* introduced into the null distribution due to variability of the estimator $S = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$
- In the case where :
  (1) the observations are normally distributed
  (2) σ is unknown, and
  (3) n is small,

  the standardized mean $T = \frac{\overline{X}}{s/\sqrt{n}} \sim t_{n-1}$
- Testing a population mean μ :
  $H : \mu = \mu_H$
  $A_1 : \mu \neq \mu_H$ or $A_2 : \mu > \mu_H$ or $A_3 : \mu < \mu_H$,

  with $T = \frac{\hat{\mu} - \mu_H}{s/\sqrt{n}}$.

Distribution t de Student



Two-sample t-test

- T for difference of independent means, when the observations are normally distributed, σ is the *same* for both populations (but unknown), and sample sizes are small :

$$T = \frac{\overline{X}_1 - \overline{X}_2}{s\sqrt{1/n_1 + 1/n_2}}, \text{ where } s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

  Under the null, $T \sim t_{n_1+n_2-2}$.
- When the population variances are *different* (Welch test), then

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}},$$

  in which case the null distribution is t_ν, where

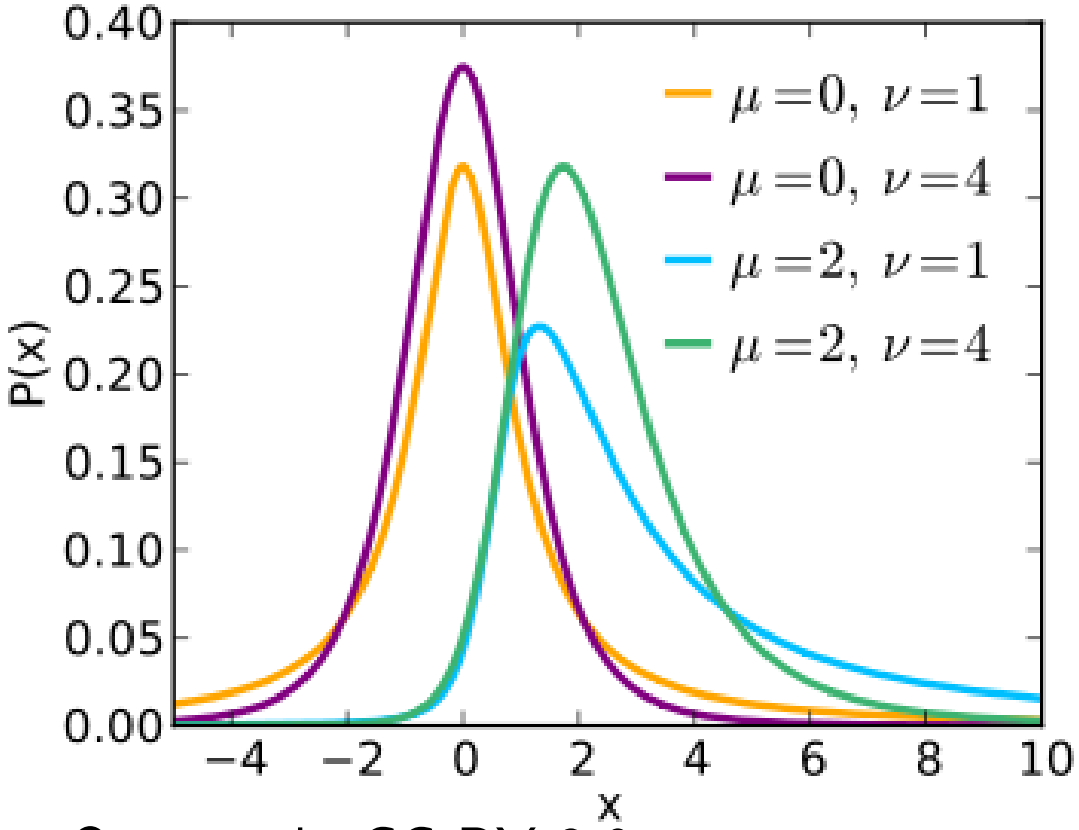
$$\nu = \left(\frac{c}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1}\right)^{-1}, \text{ with } c = \frac{s_1^2/n_1}{s_1^2/n_1 + s_2^2/n_2}.$$

Paired observations

- When there are *2 measures for each subject*, then the observations are *not* independent, but are instead *paired*
- Here, we consider the *differences* between observations for each individual
- The most typical NULL in this case is that the mean difference is 0 : $H : \mu = 0$.
- In this case, $T = \frac{\overline{d}}{s/\sqrt{n}} \sim t_{n-1}$, where $\overline{d}$ is the mean difference between the paired measurements and s is its standard deviation (the standard deviation of the differences of paired measures)

Power of the *t*-test

- Power of the *t*-test is based on the *non-central t distribution*
- Difficult to calculate 'by hand'
- Use software (R) to do power calculations



By Skbkekas - Own work, CC BY 3.0,
https ://commons.wikimedia.org/w/index.php ?curid=9745650

Review – Steps in hypothesis testing

1. Identify the population parameter being tested
2. Formulate the NULL and ALT hypotheses
3. Compute t_{obs}
4. Compute the p-value (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*)
5. (Optional) *Decision Rule* : REJECT H_0 if $p_{obs} \leq \alpha$
   (This is a type of argument by contradiction)

Pitfalls in hypothesis testing

There are a few things we need to watch out for in hypothesis testing

- Difficulties of interpreting tests on *nonrandom samples* and *observational data*
  - in practice, most samples nonrandom
  - p-values computed on such samples are generally not very meaningful ; should be viewed only as *rough* indicators of significance
- Statistical vs. practical significance
  - Was the difference *important* – a small p-value can come from a very small deviation from the null if the sample size is very large
- Perils of *searching* for significance
- Ignoring *lack* of significance

Hypothesis testing summary

- We use statistical tests to assess whether data $y_1, \ldots, y_n$ support a hypothesis
- There are 3 key components to a test :
  - a **NULL hypothesis** H, that constrains the model for how the data arise ; we usually also have an *ALTERNATIVE hypothesis* A
  - a **test statistic** T, with observed value t_{obs} ; 'unusual' values of T suggest that $y_1, \ldots, y_n$ are not compatible with H
  - an **observed significance level (p-value)** p_{obs}, such that small values suggest (but cannot *prove*) that H is false