5- Introduction to Statistical Inference

Alex Sanchez, Miriam Mota, Mireia Ferrer and Santi Perez-Hoyos

Statistics and Bioinformatics Unit. Vall d'Hebron Institut de Recerca

Readme

- License: Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License http://creativecommons.org/licenses/by-nc-sa/4.0/
- You are free to:
  - **Share** : copy and redistribute the material
  - **Adapt** : rebuild and transform the material
- Under the following conditions:
  - **Attribution** : You must give appropriate credit, provide a link to the license, and indicate if changes were made.
  - **NonCommercial** : You may not use this work for commercial purposes.
  - **Share Alike** : If you remix, transform, or build upon this work, you must distribute your contributions under the same license to this one.

Outline

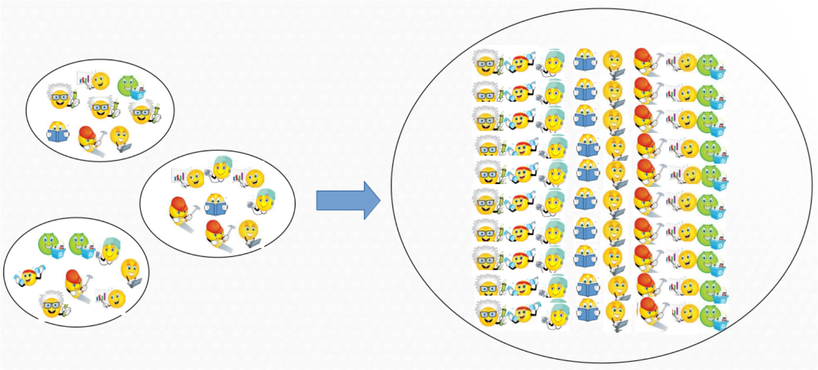
- The objectives of statistical inference
- Examples
- Point estimation. On incidence and prevalence
- Confidence intervals
- Sample size calculations

First of all

- Download `osteoporosis.csv` and `diabetes.xls` in a folder you choose.
- Ensure your working directory is the folder where the datasets are located.
- Download exercise notebook
- Open downloaded notebook

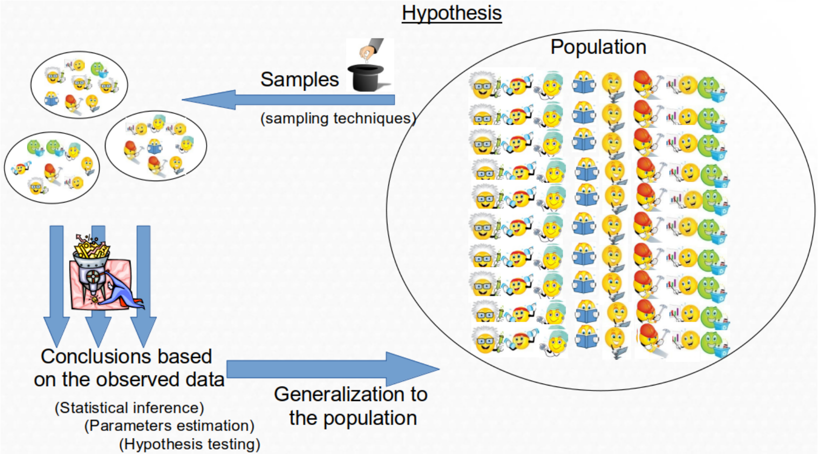
The objectives of Statistical Inference (I)

Taking the observed (measured) values of a group of samples. . .



we aim at determining the properties of the entire population.

The objectives of Statistical Inference (II)



Example

- Consider the data in the "osteoporosis.csv" dataset.
- It can be useful to provide information such as:
  - The percentage of menopausic women with osteoporosis
  - The mean bone density in menopausic or non-menopausic women
  - The existence of significant differences:
    - Observed % of osteoporosis vs "theoretical" population values
    - BUA in menopasuic vs non menopausic
- Answering these questions (and questions like these) is the main goal of Statistical Inference

Two types of statistical inference problems

- ESTIMATION
  - When we wish to *learn some characteristics of our population*, such as
    - The percentage of non osteopenic or menopausic women
    - The mean bone density in each of these groups
- HYPOTHESIS TESTING
  - When we wish to *check about some statement on some characteristic of the population* or we *wish to make some comparisons*
    - Is it true that the mean bone density is smaller than 75 in menopausic
    - Can we state that non menopausic women have a higher bone density than menopausic?

Estimators: Aproximating the value of population parameters

- Numerical values calculated on a sample that we believe to be a good approximation of a certain real value (parameter) in the population.
- Intuitively, we work with many estimators, such as the mean or a computed percentage of a given sample, that we assume that are somehow characterizing a population.
- It is **not always obvious to decide which is the best estimator for each parameter**
- In order to decide which estimator we use we can rely on the *properties* of the estimators such as **the bias** or the **precision (the variance)** of the estimator.

Estimation

The aim of estimation is to infer properties (parameters) of the distribution of population data from sample data

Some key concepts

- **Point estimate:** Give a numerical value to the parameter of interest
- **Estimator:** Mathematical function to obtain the estimate
- **Interval Estimation:** Give two values between which is the value of the population parameter with a preset confidence level (or probability)
- **Random error:** Difference between estimation and real value if the sample is random

Example. Computing estimations (1)

- Read the Osteoporosis dataset and turn factors into variables automatically with Rbase function `read.delim`
- Take a sample of size 100 from the original file. Call it 'osteo100' and work with this file from now on.
- Compute the mean value of the variable containing bone density values `BUA`
- Split the computation between all subgroups from variable `classific` and variable `menop`
- Compute the percentage of menopausic women from variable `menop`

Example. Computing estimations with R (1)

```
library(dplyr)
# Read data
osteoporosis <- read.delim2("datasets/osteoporosis.csv", stringsAsFactors=TRUE)
# Take subsample
osteo100 <- sample_n(osteoporosis, 100)
# mean bone density
buaMean <- mean(osteo100$bua)
print(buaMean)
```

```
## [1] 74.24
```

Example. Computing estimations with R (2)

```
# Mean bone density ny groups
osteo100 %>%
  group_by(menop) %>%
  summarize(m = mean(bua))
```

```
## # A tibble: 2 x 2
##   menop     m
##   <fct> <dbl>
## 1 NO     83.0
## 2 SI     70.5
```

```
# Proportion of menop women (Proportion  is a mean of 0-1 values
mean(ifelse(osteo100$menop=="SI",1,0))
```

```
## [1] 0.7
```

Exercise 1

- Read the diabetes dataset. Convert characters into factors before continuing.
- Provide an estimate of
    - The distribution of a numerical variable.
    - a proportion of at least one categorical variable and
    - the mean value of at least one numerical variable.
- Could you have used different estimators?
- How would you decide?

How precise is an estimator?

- We all are familiar with “forks” associated with voting results.
  - They usually start “wide” and tend to disappear as more votes are counted.
- Imagine you are given an estimate of 18% for the incidence of a certain disease.
- Is it a good estimate?
- Hard to know without more information
  - 18 ± 2 is probably useful
  - 18 ± 12 is probably too wide to be considered useful
- So given an estimator and a n estimation (a value) **how can we provide a measure of how precise this estimation is**?

The *Standard Error* of an estimator

- An obvious question when we choose an estimator is *how precise it is to approximate the value of the population parameter*.
- This can be answered using the **standard error of the estimator**
- The standard error is a great quantity :
  - It informs about the *precision* of our estimates
  - Helps build another type of estimators: *confidence intervals*
  - Helps find formulae to compute *sample size* for estimation

Normal approximation of sampling distributions

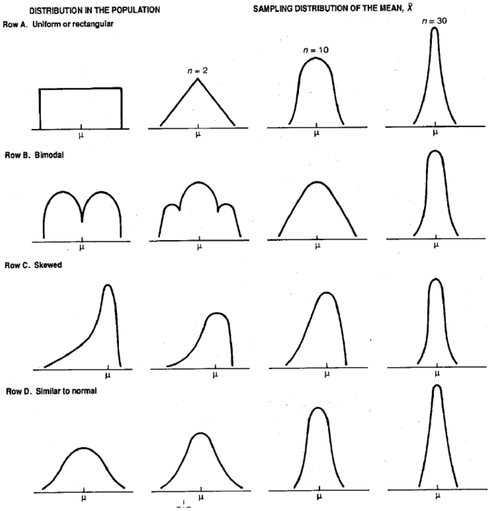


Figure 1: As sample size increases the distribution of sample means (=*Sampling Distribution of the mean*) tends to have a bell-shaped form

Some standard errors

- Standard error of the sample mean

$$SEM = \frac{\hat{s}}{\sqrt{n}}$$

- Standard error of the sample proportion

$$SEP = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Computing the standard error with R

- R does not include functions for standard errors, although it can be easily programmed.

- First create the functions

```
SEM <- function (x){sd(x)/sqrt(length(x))}

SEP <- function (x){
  ssize <- length(x)
  p <- sum(x)/ssize
  return(sqrt(p*(1-p)/ssize))
}
```

- Then apply them to your data

```
SEM (osteo100$"bua")
```

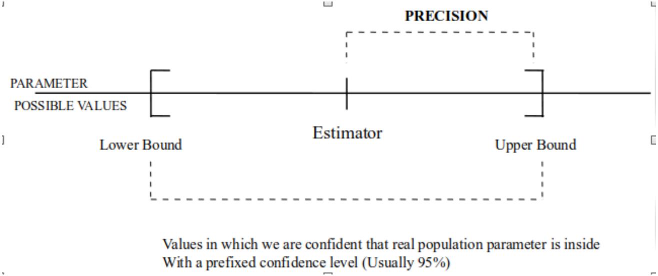
```
## [1] 1.573269
```

```
intMenop <- ifelse(osteo100$"menop"=="SI", 1, 0)
SEP (intMenop)
```

```
## [1] 0.04582576
```

Confidence intervals

- Confidence intervals are based on standard errors



Values in which we are confident that real population parameter is inside
With a prefixed confidence level (Usually 95%)

Formulae for confidence intervals

- Confidence interval for the mean

$$\overline{X} - \underbrace{t_{\epsilon/2}\frac{\hat{s}}{\sqrt{n}}}_{Precision} \leq \mu \leq \overline{X} + t_{\epsilon/2}\frac{\hat{s}}{\sqrt{n}} = \overline{\mathbf{X}} \pm \mathbf{t}_{\epsilon/2} \cdot \mathrm{SEM}$$

- Confidence interval for the proportion

$$\hat{p} - \underbrace{z_{\epsilon/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}_{Precision} \leq p \leq \hat{p} + z_{\epsilon/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \widehat{\mathbf{p}} \pm \mathbf{z}_{\epsilon/2} \cdot \mathrm{SEP}$$

Example 2. Computing Confidence Intervals with R

- In general R does not compute (has no functions) for the direct calculation of confidence intervals
- This can be done by calling the corresponding tests functions such as `t.test` or `prop.test`
- Some R packages incorporate direct calculations of confidence intervals.

Example 2. Computing Confidence Intervals with R (2)

```
t.test(osteo100[["bua"]])
```

```
##
##  One Sample t-test
##
## data:  osteo100[["bua"]]
## t = 47.188, df = 99, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  71.11829 77.36171
## sample estimates:
## mean of x
##     74.24
```

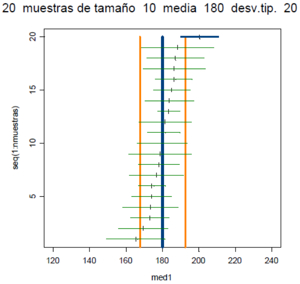
Example 2 . Computing Confidence Intervals with R (3)

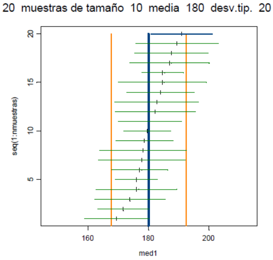
```
cntMenop <- table(osteo100[["menop"]])["SI"]
ssize <- length(osteo100[["menop"]])
prop.test (x=cntMenop, n=ssize)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  cntMenop out of ssize, null probability 0.5
## X-squared = 15.21, df = 1, p-value = 9.619e-05
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.5989396 0.7854574
## sample estimates:
##   p
## 0.7
```

Interpretation of Confidence Interval (1)

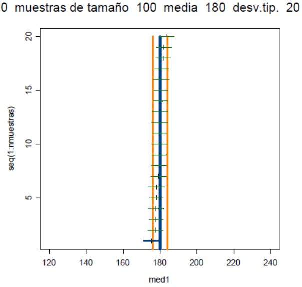
Sample size =10 , Mean=180, sd=20

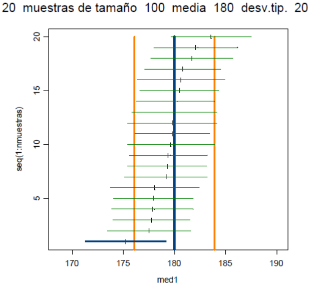




Interpretation of Confidence Interval (2)

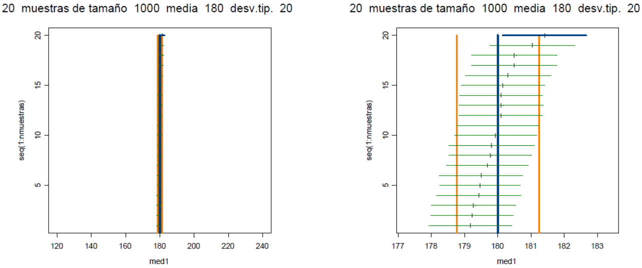
Sample size =100 , Mean=180, sd=20





Interpretation of Confidence Interval (3)

Sample size =100 , Mean=180, sd=20



Exercise 2.1 Computing Confidence intervals

- Read the file “osteoporosis.csv” into a dataset and call it “osteoporosis”
- Compute confidence intervals for the BUA mean and for the percentage of menopausic women with **all the individuals in the dataset**.
- Compare these confidence intervals with those that you obtained in example 2. How do they differ?

Exercise 2.2 Computing Confidence intervals

- Read the diabetes dataset. Convert characters into factors before continuing.
- Provide a confidence interval for:
  - a proportion of at least one categorical variable and
  - the mean value of at least one numerical variable.
- How would you find alternative approaches to compute these confidence intervals?
- Why would you want to do such a thing?

Sample Size for estimation (1)

- The standard error informs of how precise an estimation is **if one knows the variability and the sample size**

$$SE = \frac{\hat{\sigma}}{\sqrt{n}}$$

- We can proceed in the opposite sense: assuming we know:

  1. the variability (e.g. from a pilot study) and
  2. the highest precision we wish to attain ("arm length" of a confidence interval:

$$\Delta = z_{\epsilon_2} \cdot SE = z_{\epsilon_2} \cdot \frac{\hat{\sigma}}{\sqrt{n}}$$

Sample Size for estimation (2)

- The sample size needed to attain this precision can be isolated from the previous equation:

$$n = \frac{z_{\epsilon_2}^2 \hat{\sigma}^2}{\Delta^2}$$

Sample size formulae for estimating a mean or a proportion

The previous formula becomes, for specific questions:

$$n = \frac{t^2_{n-1,\epsilon_2}\, \hat{s}^2}{\Delta^2} \quad (1), \qquad n = \frac{z^2_{\epsilon_2}\, \hat{p}(1-\hat{p})}{\Delta^2} \quad (2), \qquad n = \frac{z^2_{\epsilon_2}}{4\,\Delta^2} \quad (3)$$

1. Mean of a normal population with a given precision Δ.
2. Proportion p, with a given precision Δ and with an estimate, $\hat{p}$ available, from a pilot study.
3. Proportion p, with a given precision Δ and assuming the *worst case* $p = q = 0.5$.

Sample size calculations with R

- There are many packages in R to compute sample size *for hypothesis testing*. This means thay have to account not only for “precision”, “variability” and “confidence”, but also with “power”.
- For the sake of examples it is straightforward to write simple functions to compute sample size.

```
ssize4Mean <- function (epsilon, sigma, precision){
  perc <- qnorm (1-epsilon/2)
  n <- ((perc*sigma)/prec)*2
}
```

Example 3. Sample size calculation

- Using the osteoporosis dataset, assume that the standard deviation is a good aproximation to σ.
- Find the sample size needed to achieve a margin of error equal to 5 with a 95% confidence interval.

Exercise 3. Sample size calculation

- Write a function to compute the sample size for proportions in the worst case (p=q=0.5) or assuming p is known.
- Using a 50% planned proportion estimate, find the sample size needed to achieve 5 margin of error for a survey at 95 confidence level.
- How would this result change if we are told that a pilot study suggests that $p = 10\%$?