HEADS

PhD PROGRAMME IN HEALTH DATA SCIENCE

Sampling and estimation

1. Sample vs population
2. Random sampling
3. Estimating quantities from a population
4. Confidence intervals

Basics of Health Intelligent Data Analysis

PhD Programme in Health Data Science

Cláudia Camila Dias

Pedro Pereira Rodrigues

U.PORTO
FMUP FACULDADE DE MEDICINA UNIVERSIDADE DO PORTO

MEDCIDS
DEPARTAMENTO DE MEDICINA DA COMUNIDADE, INFORMAÇÃO E DECISÃO EM SAÚDE

CINTESIS
Health.Research.

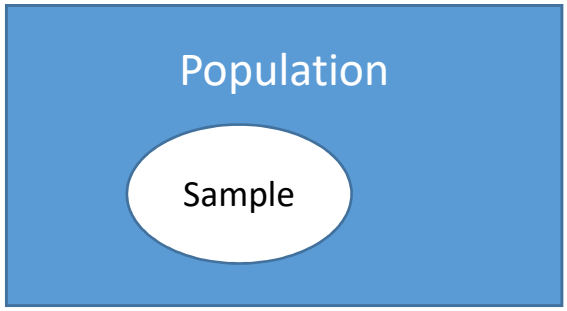
Sampling

Population:

Group of individuals that we care about

Sample:

We believe is representative of the population, we will use it to infer conclusions for it



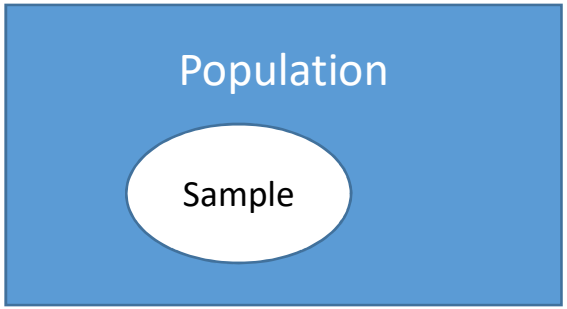
Sampling

Population:

Group of individual that we care about

Sample:

We believe is representative of the population, we will use it to infer conclusions for it



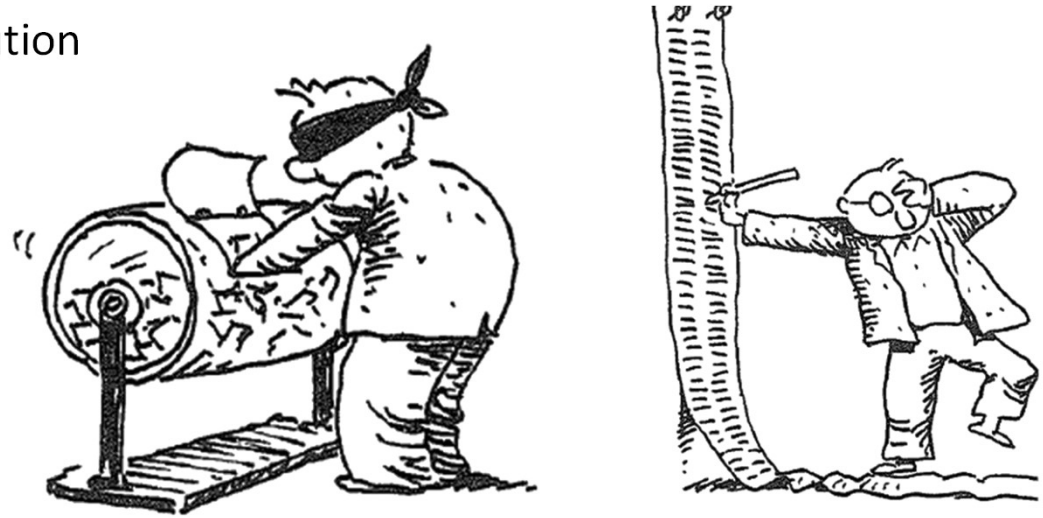
The quality of the sample (not skewed or representative) is as important as its size.

But how do you get a representative sample of population?

Random Sampling

Simple random samples:

All individuals have equal and independent probabilities of being selected, e.g. the probability of an individual being chosen follows a uniform distribution



Random Sampling

Stratified:

The population is divided into strata by a variable of interest, and within these strata, individuals are randomly chosen

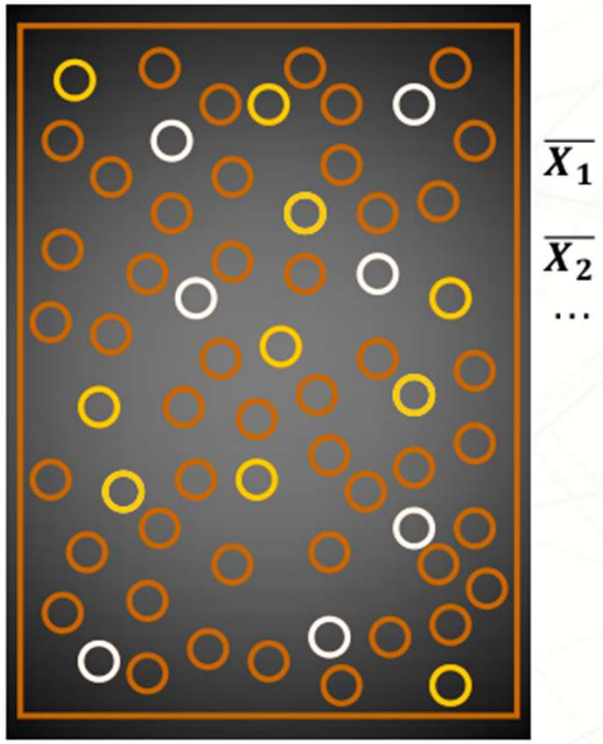
By groups:

There are 2 or more stages in the sampling process:

1) The groups are chosen randomly
2) Within these groups all individuals are chosen or only a few are randomly selected

Estimating the average of a population

- We cannot know the population average ***m***
- But we can **estimate** this average with a sample average ***x***
- We take repeated samples of the same size ***n*** and **calculate the sample mean** of each of theses samples



Distribution of sample means

If we look into the distribution of all possible sample averages it can be shown (using **Central Limit Theorem**) that:

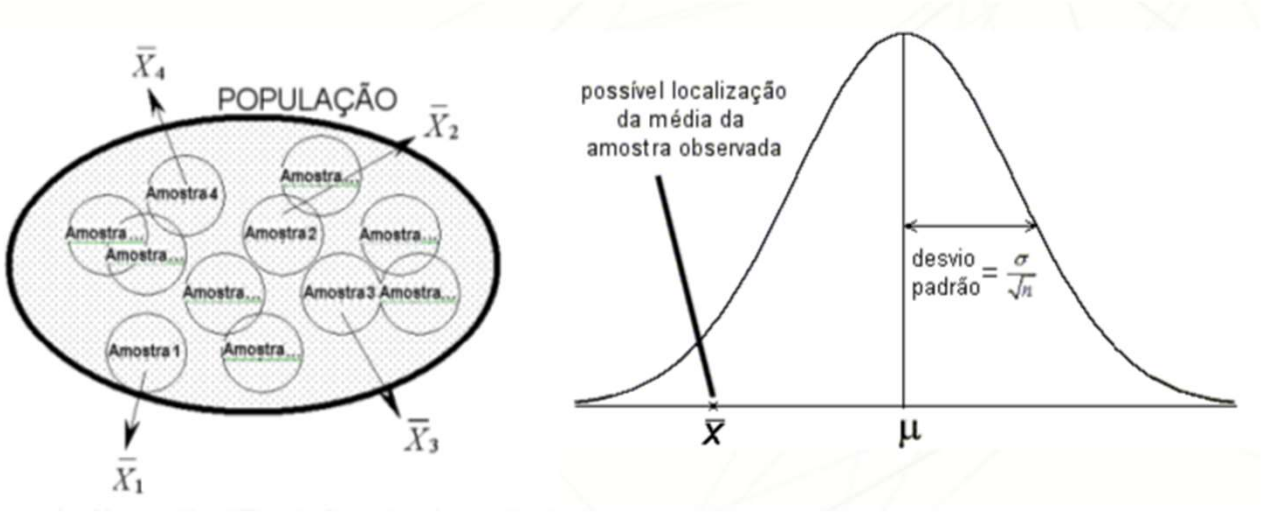
- this distribution of the sample means follows a Normal distribution regardless of the distribution of the original data;
- the average of the mean estimates is equal to the population mean (m);
- the standard deviation of the mean estimates is equal to the population standard deviation (s) of the original data divided by the square root of the sample size (n);
- the standard deviation of this distribution is usually called standard error of the mean.

Mean standard error

- Since we cannot calculate the population mean we estimate this average with an estimate (sample mean) that has a certain error: the **standard error**
  - A large standard error indicates an inaccurate estimate.
  - A small standard error indicates an accurate estimate.

  - The **standard error** decreases if:
    - sample size increases;
    - population data have a smaller standard deviation.

Confidence intervals for mean



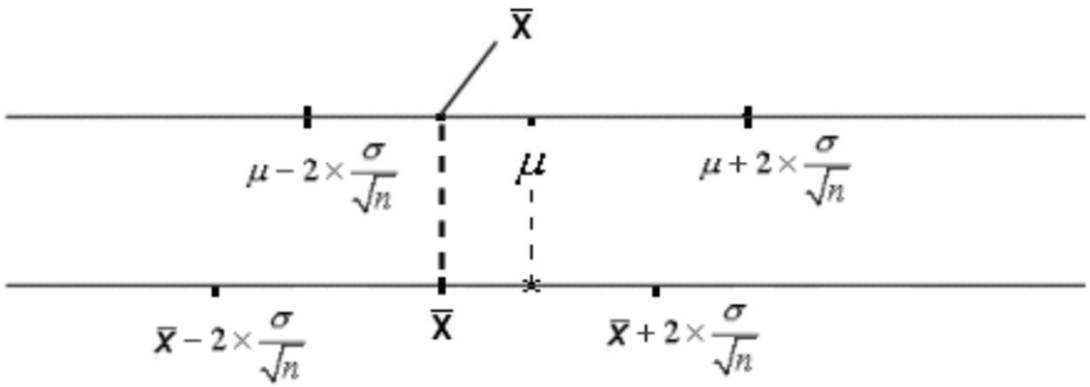
95% of sample means are between

$$\mu - 2 \times \frac{\sigma}{\sqrt{n}} \qquad \mu + 2 \times \frac{\sigma}{\sqrt{n}}$$

It can then be stated that with 95% confidence, the observed sample mean is in this range.

Confidence intervals for mean

$$\left(\mu - 2\times\frac{\sigma}{\sqrt{n}}\ ,\ \mu + 2\times\frac{\sigma}{\sqrt{n}}\right)$$



$$\left(\bar{X} - 2\times\frac{\sigma}{\sqrt{n}}\ ,\ \bar{X} + 2\times\frac{\sigma}{\sqrt{n}}\right)$$

Mean standard of a proportion

- We cannot know a proportion of population **p**, but we can estimate this proportion with a sample proportion **p** (estimate).

- If we make the distribution of all possible sample proportions it can be shown that this distribution of sample proportions follows a **Normal distribution** with mean **p**

- The **standard error** of an estimate of a proportion is the standard deviation of the sample proportions, e.g. $\sqrt{\frac{p(1-p)}{n}}$

Confidence intervals for proportion

Imagine a population of 5600 cases and 4400 controls (p = 56%), from which we took a sample of 100 individuals.

```
> pop <- factor(rep(0:1, c(4400,5600)), levels=0:1, labels=c("No","Yes"))
> sam <- sample(pop,100)
> summary(sam)
 No Yes
 39  61
```

The 95% confidence interval for proportion is

```
> prop.test(summary(sam)[2], 100)

        1-sample proportions test with continuity correction

data:  summary(sam)[2] out of 100, null probability 0.5
X-squared = 4.41, df = 1, p-value = 0.03573
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.5070114 0.7044326
sample estimates:
   p
0.61
```

Confidence intervals for proportion

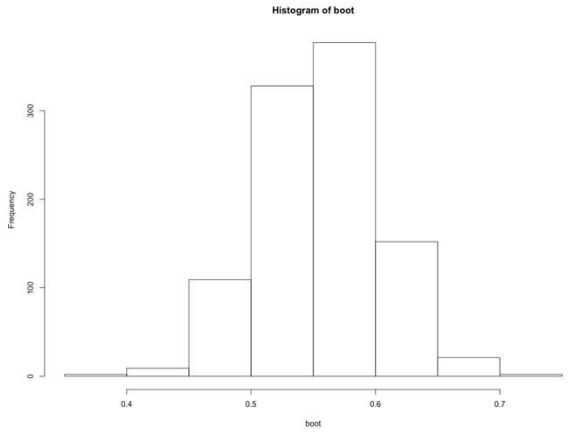
Imagine we could take 1000 samples like the previous one, of which we keep only the sample proportion:

```
> boot <- sapply(1:1000, function(i){prop.table(summary(sample(pop,100)))[2]})
> describe(boot)
boot
       n missing  unique    Info    Mean     .05     .10     .25     .50     .75     .90     .95
    1000       0      30       1  0.5608    0.48    0.50    0.53    0.56    0.59    0.62    0.64

lowest : 0.36 0.38 0.44 0.45 0.46, highest: 0.67 0.68 0.69 0.71 0.72
>
```

This set of sample estimates follows a distribution of ...

```
> quantile(boot, c(.025,.975))
 2.5% 97.5%
 0.47  0.65
```

i.e., the population average +- 9%.



Applied to the sample estimate we had before:

```
> .61 + (quantile(boot, c(.025,.975)) - .56)
 2.5% 97.5%
 0.52  0.70
```

Confidence intervals for mean

Imagine a population with 10,000 measurements following a uniform distribution from which we took a sample of 100 cases.

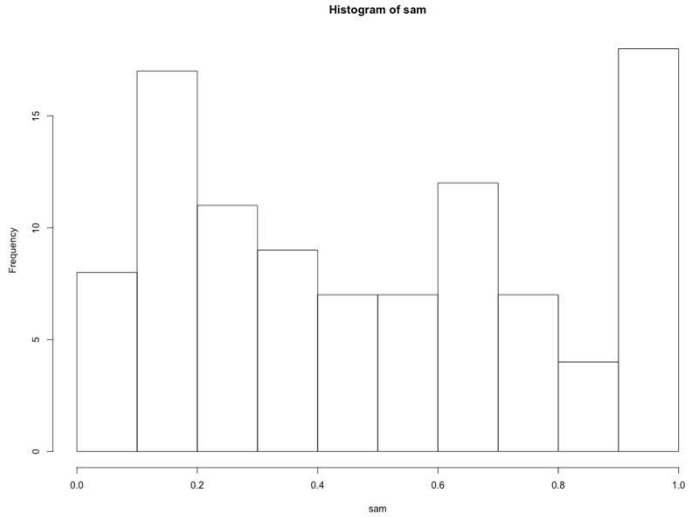
```
> pop <- runif(10000)
> sam <- sample(pop, 100)
> summary(sam)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0002712 0.2018000 0.4731000 0.4929000 0.7599000 0.9975000
```

```
> mean(pop)
[1] 0.5008168
```

The distribution of the sample is apparently irregular but not skewed

The 95% confidence interval for the mean is given by

```
> t.test(sam)

        One Sample t-test

data:  sam
t = 15.899, df = 99, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.4314234 0.5544608
sample estimates:
mean of x
0.4929421
```



Confidence intervals for mean

Imagine we could take 1000 samples like the previous one, of which we keep only the sample average:

```
> boot <- sapply(1:1000, function(i){mean(sample(pop,100))})
> describe(boot)
boot
      n missing  unique    Info    Mean     .05     .10     .25     .50     .75     .90     .95
   1000       0    1000       1  0.5002  0.4538  0.4654  0.4806  0.5003  0.5184  0.5369  0.5460

lowest : 0.4277 0.4297 0.4304 0.4309 0.4310, highest: 0.5751 0.5783 0.5805 0.5838 0.5862
```

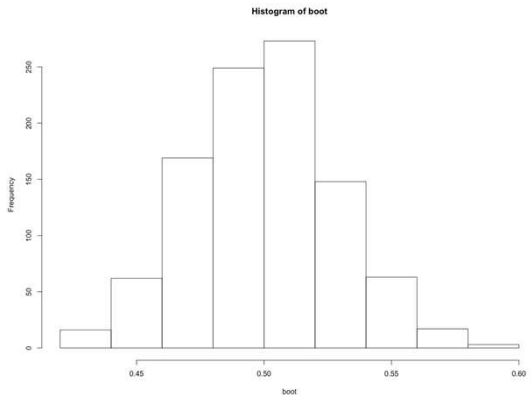
This set of sample estimates follows a distribution of ..

```
> quantile(boot, c(.025,.975))
     2.5%     97.5%
0.4477711 0.5556476
```

```
> sd(boot)
[1] 0.02806384
> sd(sam)/sqrt(100)
[1] 0.03100402
```

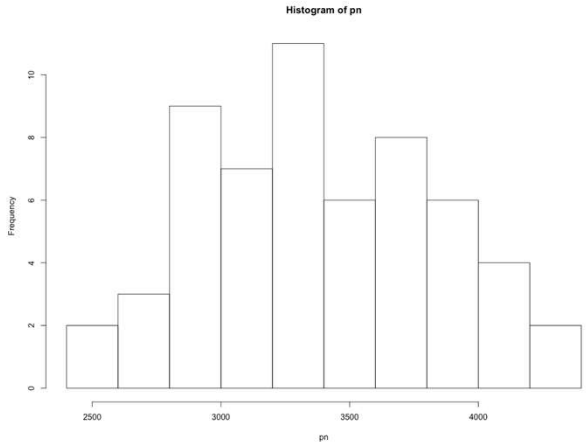
i.e. the population average +- 6%; applied to the sample estimate we had before:

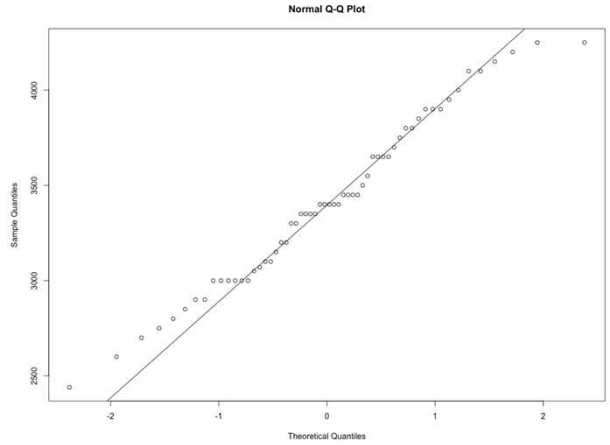
```
> mean(sam) + (quantile(boot, c(.025,.975)) - mean(boot))
     2.5%     97.5%
0.4405170 0.5483935
```



Confidence intervals for mean

```
nutricao <- as.data.frame(read.spss("nutricao.sav", use.missings=T))

attach(nutricao)

hist(pn)

qqnorm(pn)

qqline(pn)

t.test(pn)

detach(nutricao)
```






```
> t.test(pn)

        One Sample t-test

data:  pn
t = 58.504, df = 57, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 3294.616 3528.143
sample estimates:
mean of x
 3411.379
```

Confidence intervals for mean

So, we can **infer**

«With 95% confidence, the average weight of newborns in the population (m) is between 3294g and 3529g»

```
95 percent confidence interval:
 3294.616 3528.143
sample estimates:
mean of x
 3411.379
```

Notice how it is different from **describing**:

«95% of newborns in our sample weigh between 2642g and 4229g » **(sample distribution)**

```
> quantile(pn, c(.025,.975))
   2.5%   97.5%
2642.50 4228.75
```

Confidence intervals

If instead of 95% confidence we want 99% confidence?

just change the value that multiplies the standard error.

Is the range 99% wider or narrower than 95%?

```
> t.test(pn, conf.level=.99)

        One Sample t-test

data:  pn
t = 58.504, df = 57, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 3255.991 3566.767
sample estimates:
mean of x
 3411.379
```

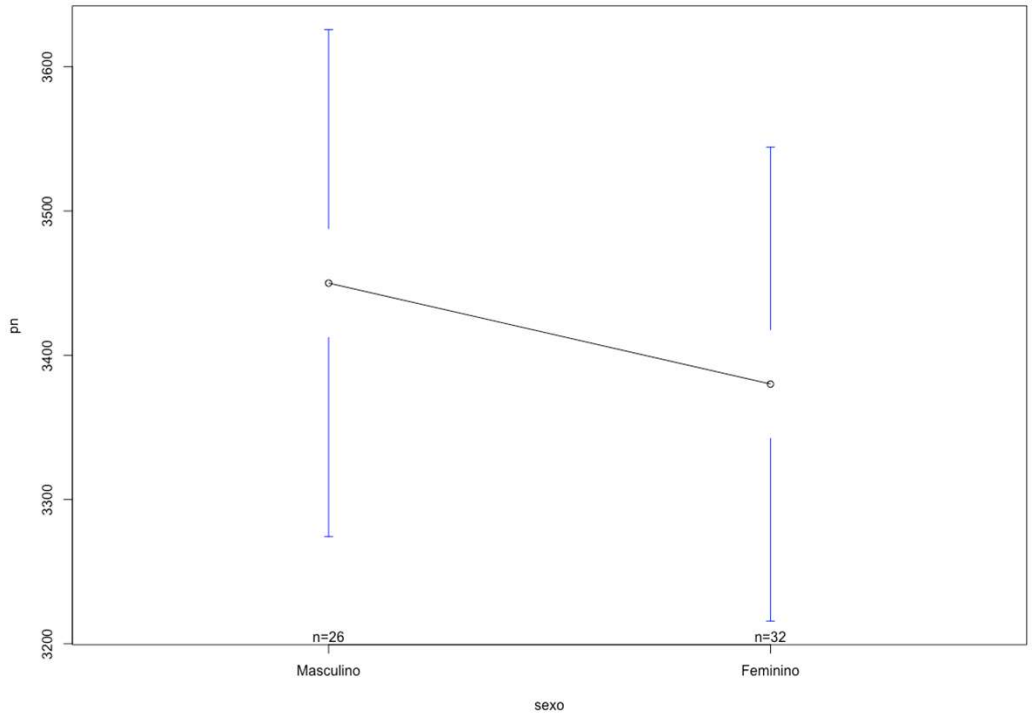
«With **95% confidence**, the average weight of newborns in the population (m) is between **3294g and 3529g** »

«With **99% confidence**, the average weight of newborns in the population (m) is between **3255g e 3567g**»

Confidence intervals (graphics)

library(gplots)

plotmeans(pn ~ sexo)



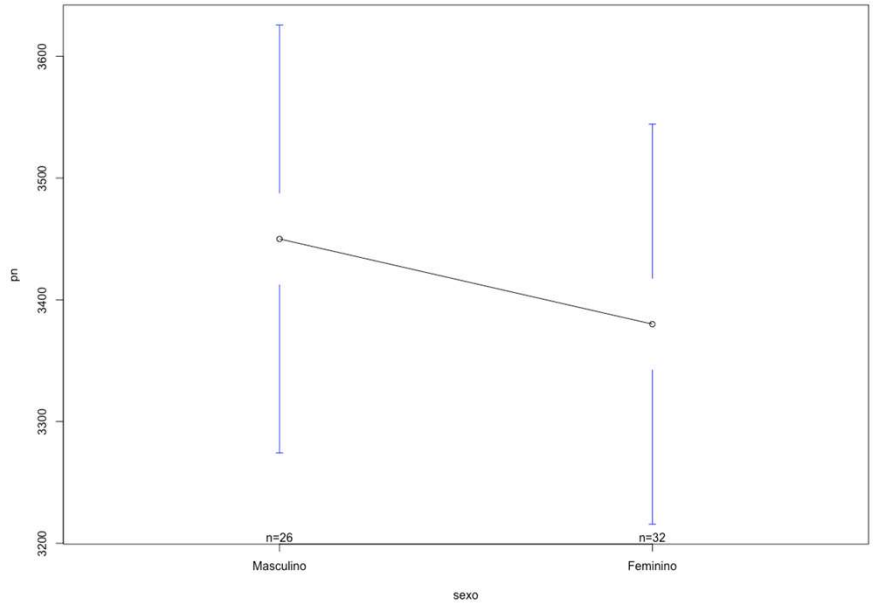
Confidence intervals (graphics)

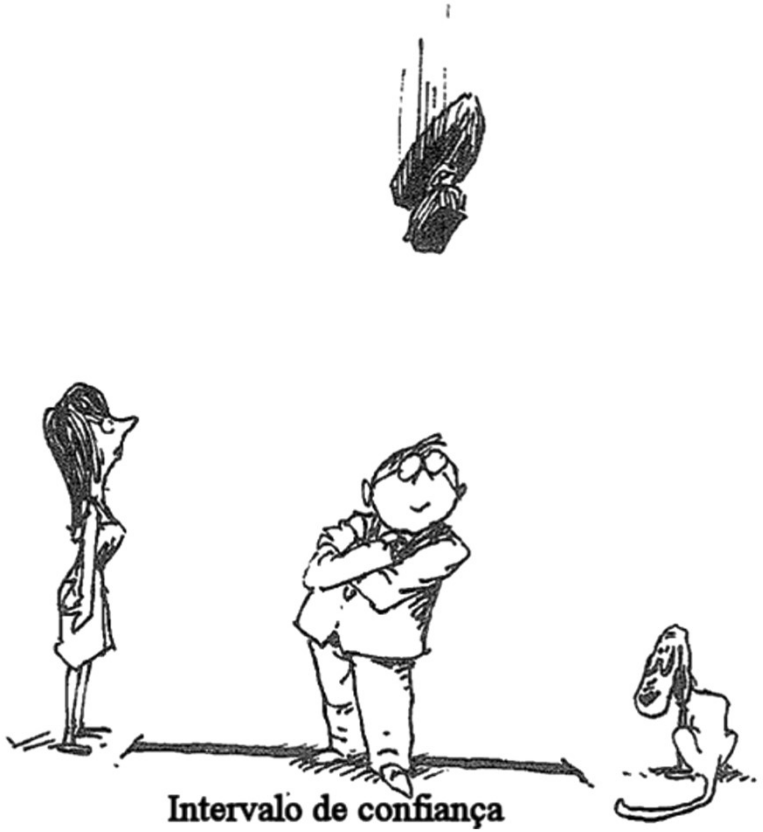
Error bar

The true average (population) weight of newborn boys may be equal to that of girls

Confidence intervals overlap

There is no evidence that in the population the weight of newborn boys and girls is on average different.





Intervalo de confiança