

# *Sampling and estimation*

**Jacques van Helden**

**[Jacques.van-Helden@univ-amu.fr](mailto:Jacques.van-Helden@univ-amu.fr)**

Aix-Marseille Université, France  
Technological Advances for Genomics and Clinics  
(TAGC, INSERM Unit U1090)  
<http://jacques.van-helden.perso.luminy.univmed.fr/>

FORMER ADDRESS (1999-2011)  
Université Libre de Bruxelles, Belgique  
Bioinformatique des Génomes et des Réseaux (BiGRe lab)  
<http://www.bigre.ulb.ac.be/>

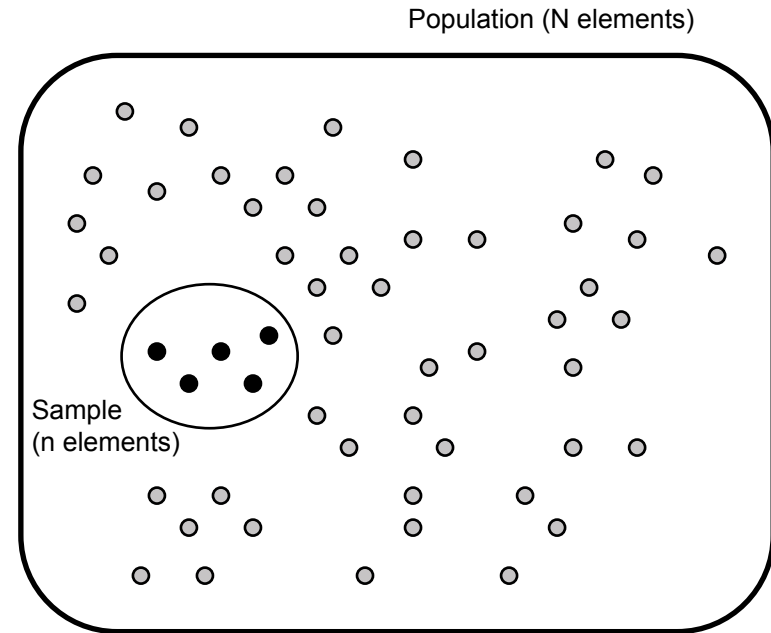
# Overview: sampling and estimation

- Definitions
  - Population and sample
  - Random sampling
  - Expectation
- Sampling distributions
  - Sample mean
  - Sample median
  - Sample variance
- Estimation
  - Population mean
  - Population median
    - Using sample median to estimate population mean
  - Population variance
- Confidence limits
  - Population mean
  - Population median
  - Population variance

# Population and sample

- Let us have a population of  $N$  elements,  $N$  being large.
- We select a sample of  $n$  elements in the population.
- We measure a certain characteristics on each element of the sample, and we obtain the sample values.

$$\{x_1, x_2, \dots, x_n\}$$



## *Examples of samples and populations*

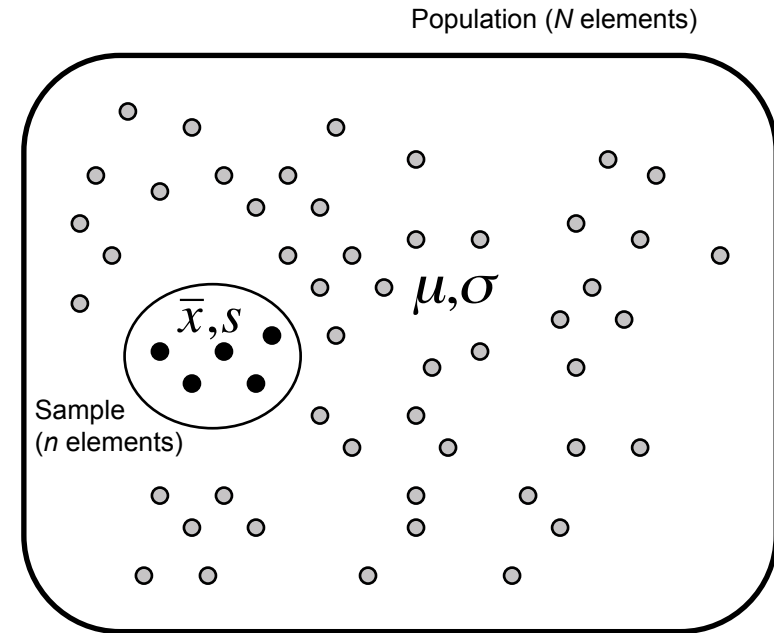
- **Saccharomyces cerevisiae genome**
  - The 3rd chromosome (sequenced in 1991) was a sample of the whole genome (sequenced in 1996)
- **Human genome**
  - The first human genome ever sequenced was a sample of 1 element taken from a population of 6 billion elements.
  - 1000 human genomes: project announced in 2008. A sample of 1000 genomes in a population of 6 billion human genomes.
- **Microarray expression profiling of cancer tissues**
  - Expression profiles taken from 30 patients suffering from a given cancer type.
  - Each tissue sample is a subset of cells of the cancerous tissue/organ.
  - Each patient is a sample of a population of all the persons suffering from the same cancer type in the world.

## *Example of the yeast genome*

- In 1991, first eukaryote chromosome completely sequenced: the chromosome III from the yeast *Saccharomyces cerevisiae*.
  - Sequence length: 316,616 bp
  - 173 ORFs
- We can consider this as a sample taken from a larger population (the set of all ORFs on the 16 yeast chromosomes), still unknown at that time.
- In 1996: the first eukaryote genome completely sequenced: *Saccharomyces cerevisiae*
  - 6,310 polypeptides (this can be considered as the population)
  - 16 chromosomes
  - Total size: 12,156,590 bp

# Estimating a parameter of the population from a sample

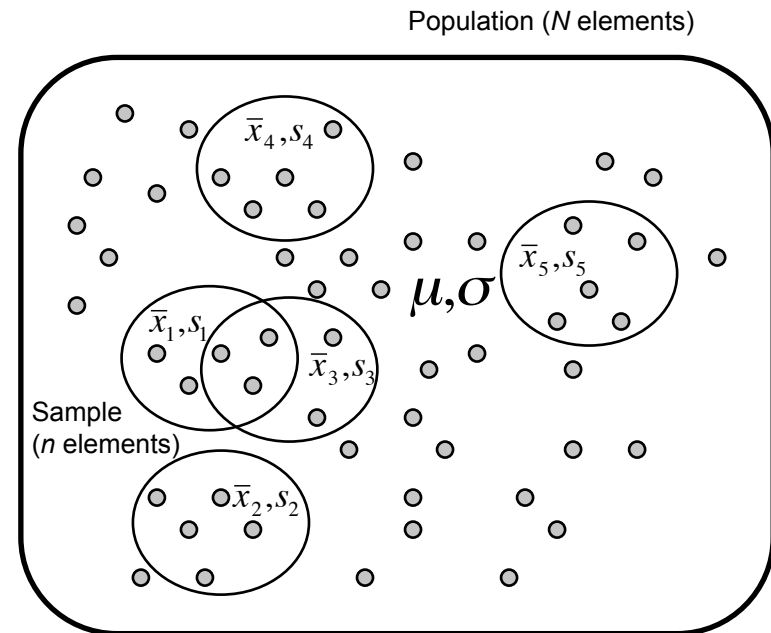
- On the basis of the sample, we want to estimate some parameters of the population.
- E.g.: mean, standard deviation.



Population mean (unknown)	$\mu$
Population sd (unknown)	$\sigma$
Sample	$\{x_1, x_2, \dots, x_n\}$
Sample mean	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
Sample sd	$s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$

# Estimating a parameter of the population from a sample

- Would we have chosen another sample, the sample mean and standard deviation would have been different.
- The population mean and standard deviation are however constant.
- Question: to which extent can we rely on the mean and standard deviation of the sample to estimate the mean and standard deviation of the population ?



Population mean (unknown)	$\mu$
Population sd (unknown)	$\sigma$
Sample	$\{x_1, x_2, \dots, x_n\}$
Sample mean	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
Sample sd	$s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$

# Expectation

*Discrete variables*

$$E(Y) = \sum_{x \in D} P(x)y(x)$$

*Continuous variables*

$$E(Y) = \int_{-\infty}^{+\infty} f(x)y(x)dx$$

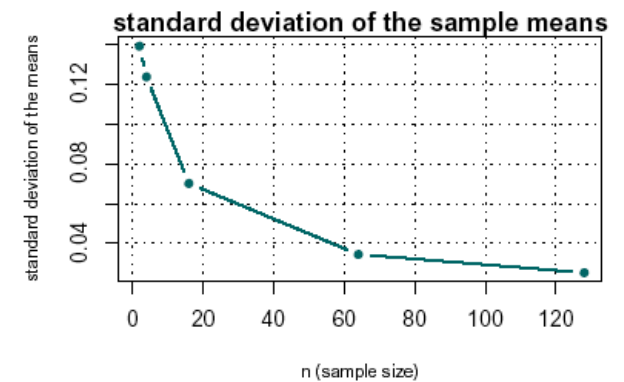
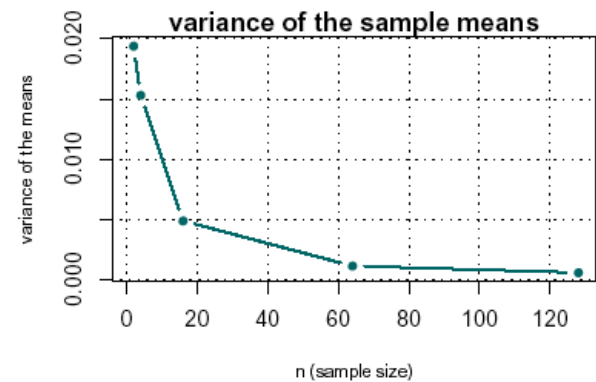
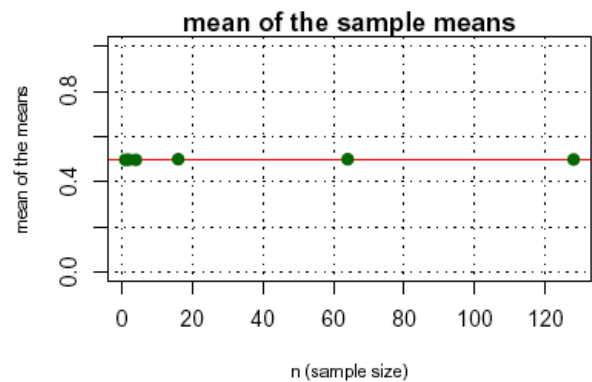
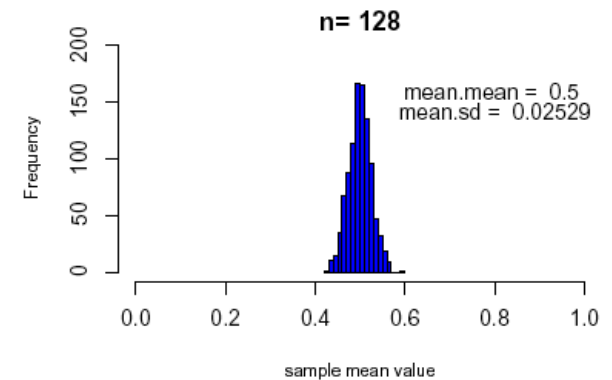
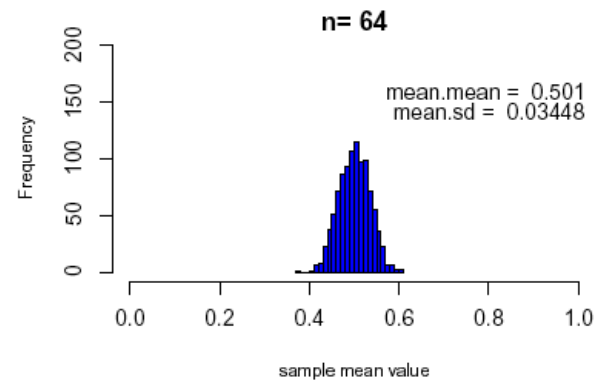
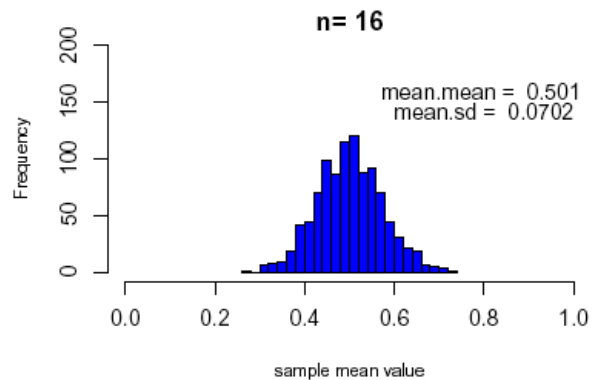
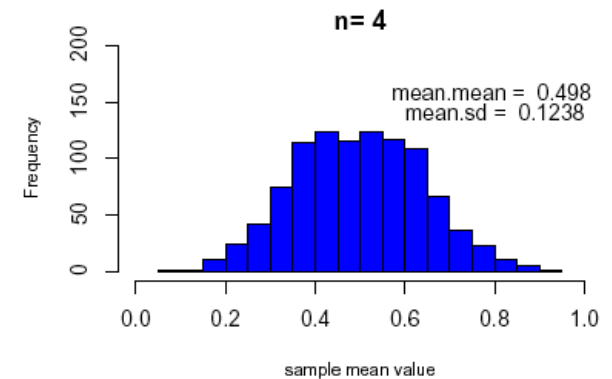
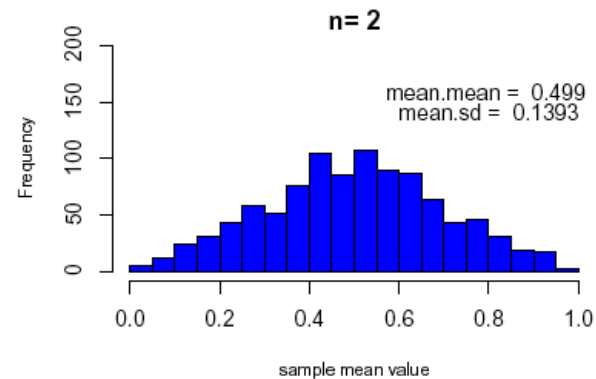
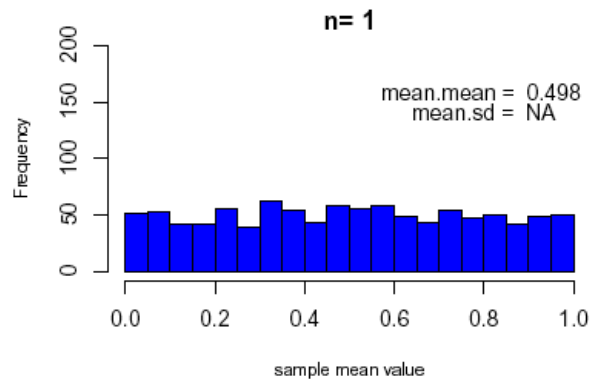
- Where
  - $P(x)$  is the probability to observe the value  $x$  (discrete random variables)
  - $f(x)$  is the density function (continuous random variables)
  - $f(x)dx$  is an element of probability
  - $y(x)$  can be any function of the random distribution  $x$ ,  
The mean, the variance, the median, ...
- $E(Y)$  is called the **expectation** for the random variable  $Y$  defined by the function  $y(x)$ .



## *Sampling distribution of the mean*

- Take  $r$  samples, each containing  $n$  elements.
- For each sample  $(x_1, x_2, \dots, x_n)$ , calculate the mean.
  - How big is the dispersion of the mean values ?
  - What is the mean value of the mean ?
  - What is the distribution of the mean ?

# Sampling distribution of the mean



## Sampling distribution of the mean

- In this simulation, the population is drawn randomly from a uniform distribution.
- When the sample size ( $n$ ) increases, the sample mean tends towards a normal distribution. This is an application of the **central limit theorem**.
- On the histograms of the previous slide, the distribution of the sample means is always centred around 0.5, irrespective of the sample size. The mean of the sample is an **unbiased estimate** of the population mean: its expected value equals the mean of the population.
- The variance and standard deviation of the sample mean decrease as the sample size ( $n$ ) increase.

## Sampling distribution - Sample variance

- The sample variance is a **biased** estimator of the population variance.

$$E(S^2) = \frac{(n-1)}{n} \sigma^2$$

$$E(S) = \sqrt{\frac{(n-1)}{n}} \sigma$$

- For this reason, one has to introduce a **corrective factor  $n/(n-1)$**  when one tries to estimate the population variance from the sample variance.

$$\hat{\sigma}^2 = \frac{n}{n-1} s^2$$

$$\hat{\sigma} = \sqrt{\frac{n}{n-1}} s$$

- Remarks
  - This correction only matters for small samples. For large samples,  $n/(n-1) \sim 1$ .
  - This correction is already included in some packages (e.g. R): when you compute the variance of a vector, the function `var()` returns the estimate for population variance rather than the actual variance of the input numbers (the sample).

# Sampling distribution - The standard error

- The expectation for the sample mean is the population mean. The sample mean is thus an **unbiased** estimator of the population mean.

$$E(\bar{X}) = m$$

$$\hat{m} = \bar{X} \quad (\text{the hat means "estimate"})$$

- The variance of the sample mean distribution **differs** from the population variance.

*for a finite population*

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \left( \frac{N-n}{N-1} \right)$$

*for an infinite population*

$$\sigma_{\bar{X}}^2 = \sigma^2 / n$$

- The standard deviation of the sample mean is called **standard error**. The standard error decreases when  $n$  increases. The larger is the sample, the more reliable is the estimation of the mean.

***For a finite population***

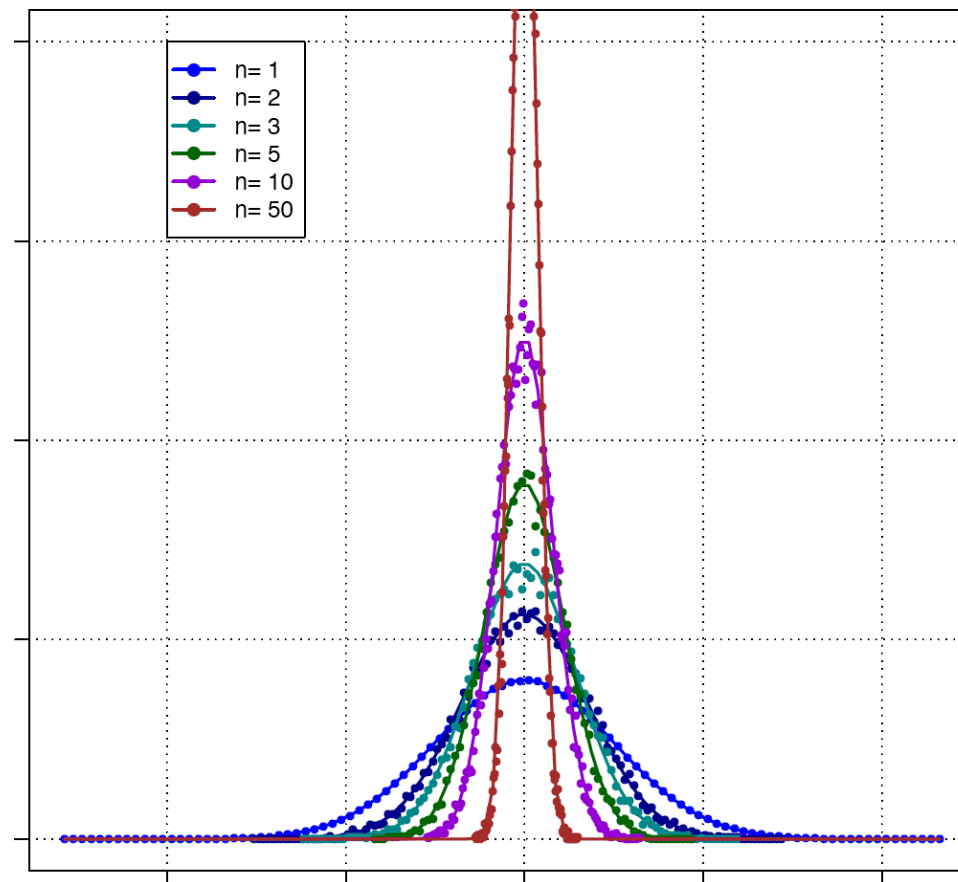
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\left( \frac{N-n}{N-1} \right)}$$

***For an infinite population***

$$\sigma_{\bar{X}} = \sigma / \sqrt{n}$$

## Standard error - simulation

- We generated 20,000 random samples with a normal distribution  $N(0,1)$ , and calculated the distribution of their means, for various sample sizes ( $n$ )
- The dots show the distribution of sample mean, for each sample size.
- The lines show the theoretical distribution, i.e.  $N(0, 1/\sqrt{n})$



## Sampling distribution – Sample median

- The expectation for the sample median is the population median. The sample median is thus an unbiased estimator of the population median
- In the case of symmetric populations,
  - The sample median is also an unbiased estimator of the population mean
  - The sample median is less efficient in the sense that its variance is higher than the variance of the sample mean. However it is more robust to the presence of outliers. When the sample is suspected to contain outliers, the sample median is thus preferable. This is typically the case with microarray data.

$$E(\tilde{X}) = \tilde{m}$$

***Confidence interval around the mean***



## Back to 1991

- Let us suppose that we are back in 1991.
  - The chromosome III from the yeast *Saccharomyces cerevisiae* has just been sequenced.
  - The rest of the genome is still unknown (the full genome sequence was published in 1996).
  - The researchers identified 173 genes in the chromosome III.
    - This can be considered as a sample from a wider population (all ORFs from the 16 chromosomes composing the genome).
- We would like to infer from this sample some characteristics of the population (the complete genome)
  - Mean gene length
  - Variance of gene length
- Questions:
  - *How can we estimate the mean length of all yeast ORFs ?*
  - *Can we define a confidence interval around our estimation ?*
  - *How can we estimate the variance of all yeast ORF lengths ?*
  - *Can we define a confidence interval around this estimation ?*
- After having performed these estimation, we will jump from 1991 to 1996 and evaluate our estimations.

## Confidence interval around the mean with a pre-defined variance

- The confidence interval is defined as a range around the mean estimate, having a probability  $1-\alpha$  to include the mean
- BEWARE
  - The mean of a population ( $m$ ) is NOT a stochastic event. Its value is *defined* by the population.
  - It is thus **INCORRECT** to say that “ $m$  has a X% probability to fall within the confidence interval”
  - The probability is a property of the interval : we can say that the interval  $[x_1, x_2]$  has a probability of 95% to include the mean.
- When the variance is known *a priori*, the confidence interval around the mean can be estimated using the normal distribution, with the standard error as parameters of dispersion.

$$P([x_1, x_2] \supset m) = 1 - \alpha$$

$$x_1 = \bar{x} - u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$x_2 = \bar{x} + u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- Exercise
  - search in the normal table the value of  $u$  corresponding to risk of error of 5%
- Warning
  - usually, the variance is a priori not known, see next slide

## *Confidence interval around the mean when the population variance has to be estimated from the sample itself*

- In most practical cases, the variance is priori not known.
- The population variance has thus to be estimated from the sample variance.
- This introduces an error, which modifies the theoretical distribution.
- Instead of a normal, use a Student distribution with  $k=n-1$  degrees of freedom.
- In practice, the Student distribution tends towards a normal when  $n$  tends towards infinity.

$$P([x_1, x_2] \supset m) = 1 - \alpha$$

$$x_1 = \bar{x} - t_{1-\alpha/2} \sqrt{\frac{\tilde{\sigma}^2}{n}} = \bar{x} - t_{1-\alpha/2} \sqrt{\frac{s^2}{(n-1)}}$$

$$x_2 = \bar{x} + t_{1-\alpha/2} \sqrt{\frac{\tilde{\sigma}^2}{n}} = \bar{x} + t_{1-\alpha/2} \sqrt{\frac{s^2}{(n-1)}}$$

<b>statistics</b>	<b>sample</b>
size	173
mean	1 302
standard error	
median	969
sample sd	1 118
sample var	1 249 251

statistics	sample	estimation
size	173	
mean	1 302	1 302
standard error		85
median	969	969
sample sd	1 118	1 121
sample var	1 249 251	1 256 514

## 1991: confidence interval around ORF length

statistics	sample	estimation	confidence interval	
			from	to
size	173			
mean	1 302	1 302	1 135	1 470
standard error		85		
median	969	969		
sample sd	1 118	1 121		
sample var	1 249 251	1 256 514		

$$x_1 = \bar{x} - t_{1-\alpha/2} \frac{s}{\sqrt{n-1}} = 1302 - 1.960 \frac{1121}{\sqrt{172}} = 1135$$

$$x_2 = \bar{x} + t_{1-\alpha/2} \frac{s}{\sqrt{n-1}} = 1302 + 1.960 \frac{1121}{\sqrt{172}} = 1470$$

1996: we can check the prediction

statistics	sample	estimation	confidence interval		population
			from	to	
size	173				6 310
mean	1 302	1 302	1 135	1 470	1 414
standard error		85			
median	969	969			1 140
sample sd	1 118	1 121			1 238
sample var	1 249 251	1 256 514			1 533 008