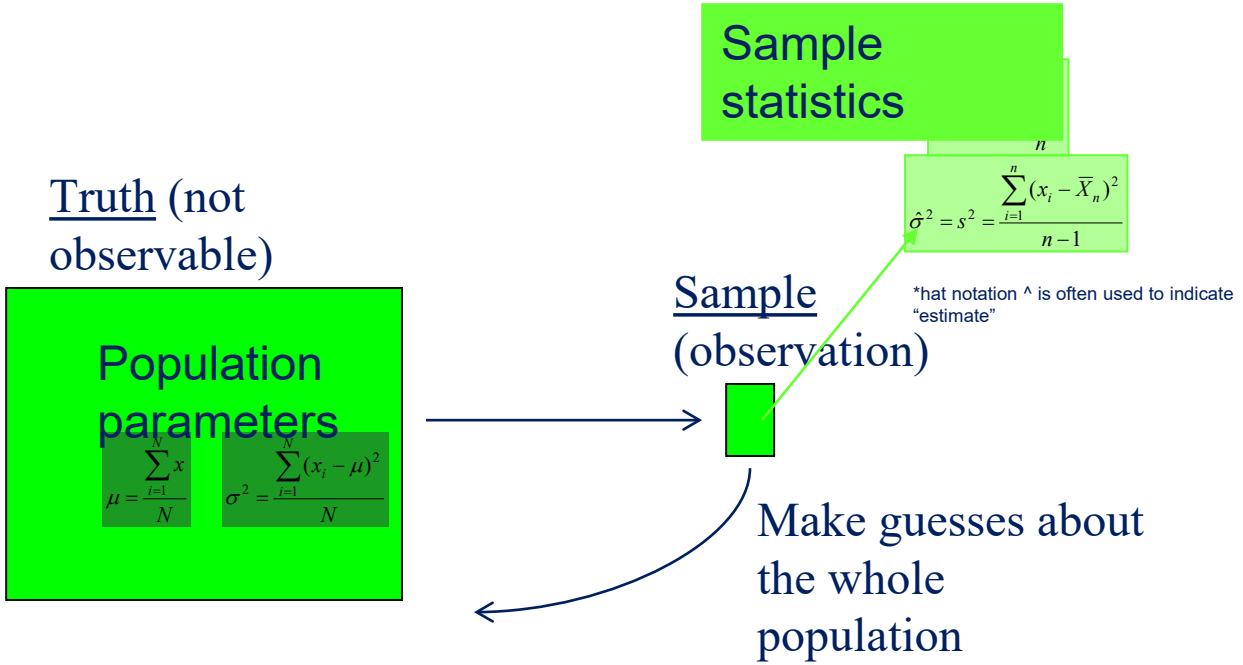
Co-financed by the Connecting Europe Facility of the European Union

Statistical inference: point and interval estimate

Data Driven Healthcare
Module B
Prof. Paola Cerchiello – University of Pavia

Statistical Inference

The process of making guesses about the truth from a sample.



Co-financed by the Connecting Europe Facility of the European Union

Inferential statistics (or Statistical inference)

- Assume that we are working with the sample and we calculate a *sample statistics* such: sample average, sample variance, sample standard deviation.

- Based on the sample we assume the properties of a population.

- This means , the values of a sample statistics are used to estimate the unknown values of population parameters
- Usually we estimate *parameters of population* such : population mean, population variance, standard deviation of population.

Co-financed by the Connecting Europe Facility of the European Union

Statistics vs. Parameters

- **Sample Statistic** – any summary measure calculated from data; e.g., could be a mean,
  - E.g., the mean vitamin D level in a sample of 100 men is 63 nmol/L

- **Population parameter** – the true value/true effect in the entire population of interest
  - E.g., the true mean vitamin D in all middle-aged and older European men is 62 nmol/L

Co-financed by the Connecting Europe Facility of the European Union

Graphicaly

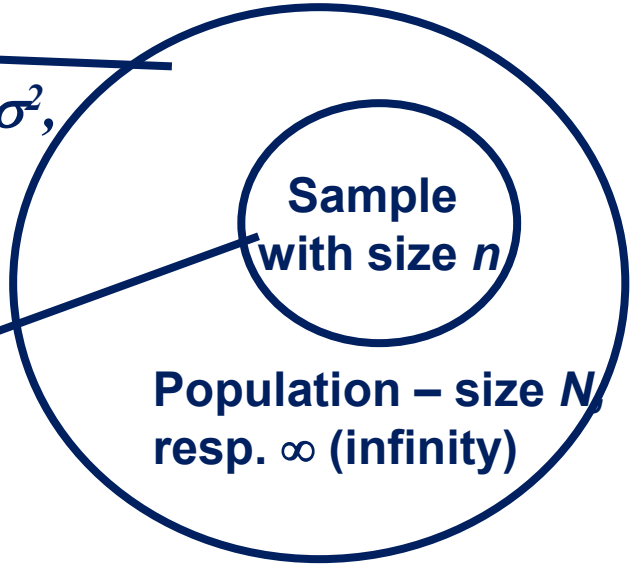
Symbols:

parameters of population: μ, σ^2, σ, generally Q

sample characteristics :

$\bar{x}$, s^2, s

Generally: u_n



What we have accomplished with sampling distributions

- Given a population parameter, we know that a sample statistic will produce a better estimate of the population parameter when the sample is larger. (Better means more accurate and normally distributed).

Estimation: definitions

- Point estimate: a single number, calculated from a set of data, that is the best guess for the parameter.
- Point estimator: the equation used to produce the point estimate.
- Interval estimate: a range of numbers around the point estimate within which the parameter is believed to fall. Also called a *confidence interval*.

Co-financed by the Connecting Europe Facility of the European Union

Mathematical Theory…
The Central Limit Theorem!

If all possible random samples, each of size *n*, are taken from any population with a mean μ and a standard deviation σ, the sampling distribution of the sample means (averages) will:

1. have mean: $\mu_{\bar{x}} = \mu$

2. have standard deviation: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

3. be approximately normally distributed regardless of the shape of the parent population (normality improves with larger *n*).

Co-financed by the Connecting Europe Facility of the European Union

Symbol Check

$\mu_{\bar{x}}$ The mean of the sample means.

$\sigma_{\bar{x}}$ The standard deviation of the sample means.
Also called "the standard error of the mean."

Co-financed by the Connecting Europe Facility of the European Union

The basics of point estimation

- The typical point estimator of a population mean is a sample mean:

$$est\ \hat{\mu} = \bar{x} = \frac{\Sigma x_i}{n}$$

- The typical point estimator of a population variance is a sample variance:

$$est\ \hat{\sigma}^2 = s_1^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

- The typical point estimator of a population standard deviation is a sample standard deviation:

$$est\ \hat{\sigma} = s_1 = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Co-financed by the Connecting Europe Facility of the European Union

Typical point estimators for standard errors

- Estimated *standard error* of samples drawn from a population:

$$\hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n}}$$

Co-financed by the Connecting Europe Facility of the European Union

Choosing a good estimator

You can technically use any equation you want as a point estimator, but the most popular ones have certain desirable properties.

- *Unbiasedness:* The sampling distribution for the estimator 'centers' around the parameter. (On average, the estimator gives the correct value for the parameter.)
- *Efficiency:* If at the same sample size one unbiased estimator has a smaller sampling error than another unbiased estimator, the first one is more efficient.
- *Consistency:* The value of the estimator gets closer to the parameter as sample size increases. Consistent estimators may be biased, but the bias must become smaller as the sample size increases if the consistency property holds true.

Examples for point estimates:

Given the following sample of seven observations:

5,2,5,2,4,5,5

- What is the estimator of the population mean?
- What is the estimate of the population mean?
- What is the estimator of the population standard error?
- What is the estimate of the population standard error for this sample?

Examples for point estimates:

Given the following sample of seven observations:

5,2,5,2,4,5,5

- What is the estimator of the population mean?

$$est\ \hat{\mu} = \bar{x} = \frac{\Sigma x_i}{n}$$

- What is the estimate of the population mean?

(5+2+5+2+4+5+5) / 7 = 28 / 7 = 4

- What is the estimator of the population standard error?

$$est\ \hat{\sigma}_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n}}$$

- What is the estimate of the population standard error for this sample?
  - =sqrt {[(5-4)^2+(2-4)^2+(5-4)^2+(2-4)^2+(4-4)^2+(5-4)^2+(5-4)^2]/(7-1)} / sqrt(7)
  - = sqrt { [1 + 4 + 1 + 4 + 0 + 1 + 1] / 6 } / sqrt(7)
  - = sqrt(2) / sqrt(7)
  - = 1.41 / 2.64
  - = 0.53

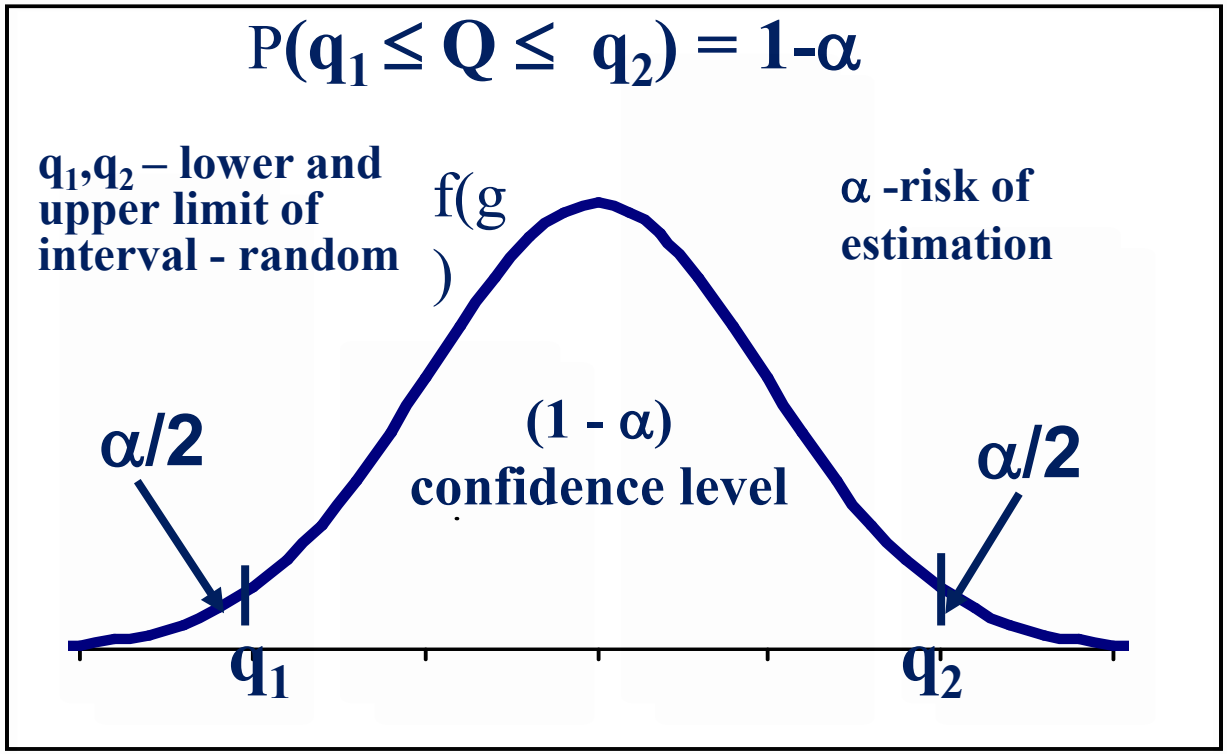
Co-financed by the Connecting Europe Facility of the European Union

Interval estimates

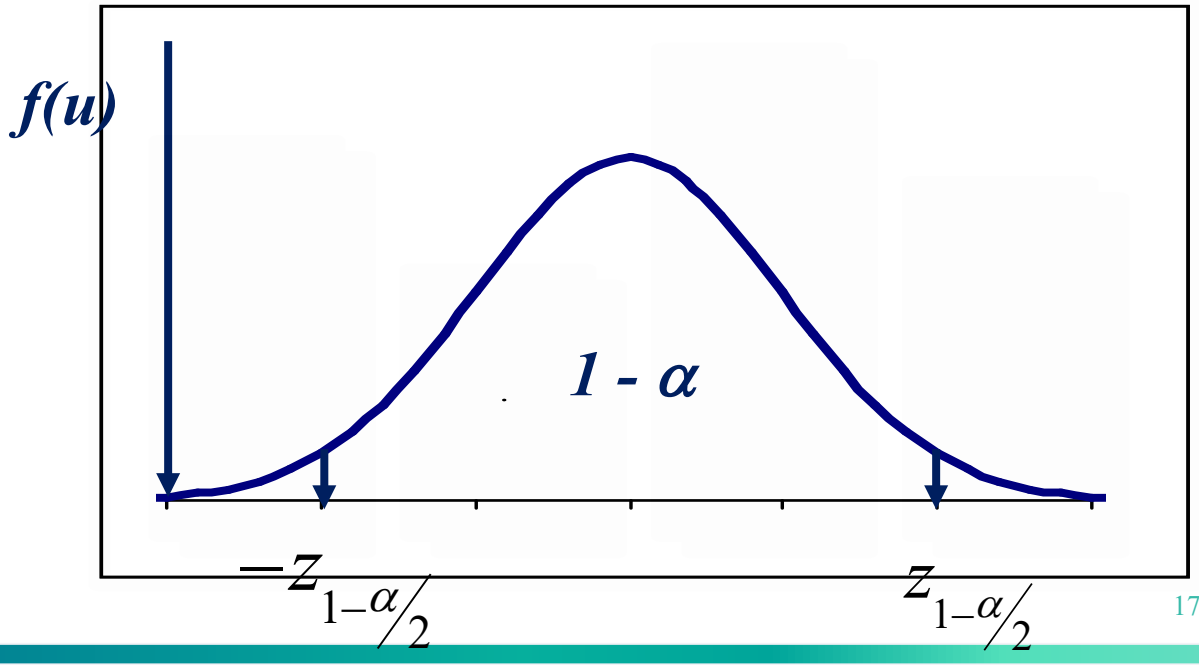
- Interval estimate (also called a confidence interval): a range of numbers that we think has a given probability of containing a parameter.

- Confidence coefficient: The probability that the interval estimate contains the parameter. Typical confidence coefficients are .95 and .99.

Interval estimate of parameter Q



$$P\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\frac{\alpha}{2}} \right) = 1-\alpha$$



Co-financed by the Connecting Europe Facility of the European Union

After transformation we get

$$P\left(\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) = 1\text{-}\alpha$$

Δ - sampling error

Δ - half of the interval,
determinates accurancy of the estimation,
Interval estimate is actually point estimate ± Δ, e.g. $\bar{x} \pm \Delta$

Co-financed by the Connecting Europe Facility of the European Union

Interval estimation of population mean μ

Confidence interval for μ depends on disponibility of information and sample size:

***a)* If the variance of population is known** (theoretical assumption) we can create standardized normal variables :

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

→ *z (u)* has N(0,1) independent on estimated value μ

Co-financed by the Connecting Europe Facility of the European Union

***b)* The population variance is unknown**

est $\sigma^2 = s_1^2$ **,** and the sample size is large, n > 30

$$\bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{s_1}{\sqrt{n}}$$

We can use N(0,1)

***c)* If the population variance is unknown**

est $\sigma^2 = s_1^2$ **, and the sample** size is small (less than 30), $n \leq 30$

$$\bar{x} \pm t_{\alpha(n-1)} \frac{s_1}{\sqrt{n}}$$

$t_{\alpha(n-1)}$ –critical value of Student's distribution at alfa level and at degrees of freedom

Example of confidence interval.

95% confidence interval for a sample mean:

$$95\%\ c.i. = \bar{x} \pm 1.96 * se(x)$$

$$95\%\ c.i. = \bar{x} \pm 1.96 * \hat{\sigma}_{\bar{x}}$$

example

```
. summarize age
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |    215754    27.34663    19.34841          0        116

. ci age
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
         age |     215754    27.34663    .0416549        27.26499    27.42827
```

Q: how is std. err. of age calculated?

Equations for interval estimates.

- Confidence interval of a mean

$$c.i. = \overline{x} \pm z\hat{\sigma}_{\overline{x}}$$

where…

$$\hat{\sigma}_{\overline{x}} = s / \sqrt{n}$$

and where you choose z, based on the p-value for the confidence interval you want

- Assumption: the sample size is large enough that the sampling distribution is approximately normal

Co-financed by the Connecting Europe Facility of the European Union

Notes on interval estimates

- Usually, we are not given *z*. Instead we start with a desired confidence interval (e.g., 95% confidence), and we select an appropriate *z* – score.

- We generally use a 2-tailed distribution in which ½ of the confidence interval is on each side of the sample mean.

Co-financed by the Connecting Europe Facility of the European Union

Equations for interval estimates

- Example: find c.i. when Y_{bar} =10.2, s=10.1, N=1055, interval=95%.
- z is derived from the 95% value: what value of z leaves 95% in the middle and 2.5 % on each end of a distribution?

  For p = .975, z = 1.96
- The standard error is s/SQRT(n) = 10.1/SQRT(1055) = .31095
- Top of the confidence interval is 10.2 + 1.96*.31095 = 10.8095
- The bottom of the interval is 10.2 – 1.96*.31095 = 9.5905
- Hence, the confidence interval is 9.59 to 10.81

Co-financed by the Connecting Europe Facility of the European Union

Confidence Intervals

The value of the statistic in my sample (eg., mean, odds ratio, etc.)

point estimate ± (measure of how confident we want to be) × (standard error)

From a Z table or a T table, depending on the sampling distribution of the statistic.

Standard error of the statistic.

Co-financed by the Connecting Europe Facility of the European Union

Common "Z" levels of confidence

- Commonly used confidence levels are 90%, 95%, and 99%

Confidence Level	*Z value*
80%	1.28
90%	**1.645**
95%	**1.96**
98%	2.33
99%	**2.58**
99.8%	3.08
99.9%	3.27

Co-financed by the Connecting Europe Facility of the European Union

Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

If s^2 is the sample variance from a random sample of n observations from a normal distribution with unknown variance σ^2, then **a $100(1-\alpha)\%$ confidence interval on σ^2 is**

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \qquad (8\text{-}21)$$

where $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$ are the upper and lower $100\alpha/2$ percentage points of the chi-square distribution with $n-1$ degrees of freedom, respectively. A **confidence interval for σ** has lower and upper limits that are the square roots of the corresponding limits in Equation 8-21.

Summary

Important notes regarding confidence intervals

- The width of the confidence interval is related to the significance level, standard error, and n (number of observations) such that the following are true:
- the higher the percentage of accuracy (significance) desired, the wider the confidence interval
- the larger the standard error, the wider the confidence interval
- the larger the n, the smaller the standard error, and so the narrower the confidence interval

Important note regarding the sample size and sampling error

- The lower the desired sampling error the larger the sample size must be.
- Note: Sampling error is influenced by:
- confidence level - we can influence the confidence level
- (sample) standard deviation - we can't influence the variability within the sample (chosen from the collected set of data)
- sample size - we can influence the sample size

Co-financed by the Connecting Europe Facility of the European Union

Point estimate of mean

$$est\ \mu = \overline{x}$$

- As for a point estimate, population mean (μ) can be estimated by means of a sample mean ($\overline{x}$).
- A sample mean can be calculated using a function AVERAGE in MS Excel.

Co-financed by the Connecting Europe Facility of the European Union

Interval estimate of mean

- $$P\left(\overline{x}-\Delta<\mu<\overline{x}+\Delta\right)=1-\alpha$$

- where Δ is the sampling error.

- If the sample size is greater than 30 (n>30) then the distribution of the random variable (sample statistic) will be approximated with a normal distribution – N(0,1). The sampling error will be calculated as follows:

$$\Delta = u_{(1-\alpha/2)} \cdot \frac{s_1}{\sqrt{n}}$$

- A critical value ($u_{(1-\alpha/2)}$) will be calculated using a function NORMSINV in MS Excel.

Co-financed by the Connecting Europe Facility of the European Union

Interval estimate of mean

- If the sample size is lower than 30 (n<30) then the distribution of the random variable (sample statistic) will be approximated with a Student t distribution). The sampling error will be calculated as follows:

$$\Delta = t_{(\alpha;n-1)} \cdot \frac{s_1}{\sqrt{n}}$$

- A critical value ($t_{(\alpha;n-1)}$)will be calculated using a function TINV in MS Excel.

Co-financed by the Connecting Europe Facility of the European Union

Point estimate of variance

- $$est\,\sigma^2 = s_1^2$$

- As for a point estimate, population variance (σ^2) can be estimated by means of a sample variance (s_1^2). A sample variance can be calculated using a function VAR in MS Excel.

Interval estimate of variance

$$P\left(\frac{(n-1)\cdot s_1^2}{\chi^2_{(\alpha/2;n-1)}} < \sigma^2 < \frac{(n-1)\cdot s_1^2}{\chi^2_{(1-\alpha/2;n-1)}}\right) = 1-\alpha$$

Critical values for both lower and upper limits($\chi^2_{(\alpha/2;n-1)}$; $\chi^2_{(1-\alpha/2;n-1)}$) can be calculated using a function CHIINV in MS Excel.

Co-financed by the Connecting Europe Facility of the European Union

Point estimate of standard deviation

- $$est\,\sigma = s_1$$

- As for a point estimate, population standard deviation(σ) can be estimated by means of a sample standard deviation (s_1). A sample standard deviation can be calculated using a function STDEV in MS Excel.

Co-financed by the Connecting Europe Facility of the European Union

Interval estimate of standard deviation

$$P\left(\sqrt{\frac{(n-1)\cdot s_1^2}{\chi^2_{(\alpha/2;n-1)}}} < \sigma^2 < \sqrt{\frac{(n-1)\cdot s_1^2}{\chi^2_{(1-\alpha/2;n-1)}}}\right) = 1-\alpha$$

- Critical values for both lower and upper limits($\chi^2_{(\alpha/2;n-1)}$; $\chi^2_{(1-\alpha/2;n-1)}$) can be calculated using a function CHIINV in MS Excel.
- It's much faster to calculate the interval estimate for variance and then calculate the square root of both lower and upper limits

Co-financed by the Connecting Europe Facility of the European Union

Calculating the sample size

- $$n = u^2_{(1-\alpha/2)} \cdot \frac{s_1^2}{\Delta^2}$$

where,

- Δ and s_1 are given in the text of the example
- A critical value ($u_{(1-\alpha/2)}$)will be calculated using a function NORMSINV in MS Excel.

Thank you!

Have a nice day!