Phd Program in Transportation Systems

Transport Demand Modeling

Filipe Moura

Session 2

Basic statistics and Sampling

(Acknowledgements to Prof. João Abreu e Silva who initially prepared these slides)

Statistical Inference

- **Confidence intervals, hypothesis tests** and **population comparisons** are statistical tools used in transportation planning (or at least they should be)
- They could be used to answer questions as the examples bellow
  - Does crash occurrence at a particular intersection support the notion that it is a hazardous location?
  - Do traffic calming measures reduce traffic speeds?
  - Does route guidance information implemented via a variable message sign system successfully divert motorists from congested areas?
  - Does altering the levels of operating subsidies to transit systems change their operating performance?

Random variable

- It corresponds to the mapping outcomes from random processes
  - Flipping coins; weather events; pedestrian flows; etc.
- Examples of random variables definition

$$X = \begin{cases} 1 \; if \; heads \\ 0 \; if \; tails \end{cases} \qquad Y = \textit{Total mass of students of random class}$$

- Why do we need to do this?
  - Allows for using mathematical notation and tools to quantify random processes
  - What is the probability of some outcome of a random process?

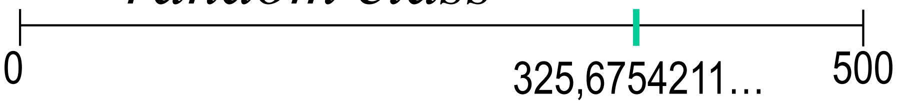
$$P(X = 1) = 1 - \alpha \qquad P(Y \leq 500) = 1 - \alpha$$

Discrete vs. Continuous Random Variables

- Discrete variable (X1)
  - Variable that can only take on a certain number of distinct or separate values
- Continuous variable (X2)
  - Variable can have an infinite number of values within an interval
- Examples

$$X1 = \begin{cases} 1 \; if \; heads \\ 0 \; if \; tails \end{cases}$$

$X2 =$ *Total mass of students of random class*



$X1 =$ *Year that a random student was born*

$X1 =$ *# of pedestrians crossing the street over 15'*

$X2 =$ *Exact winning time of 100m run*

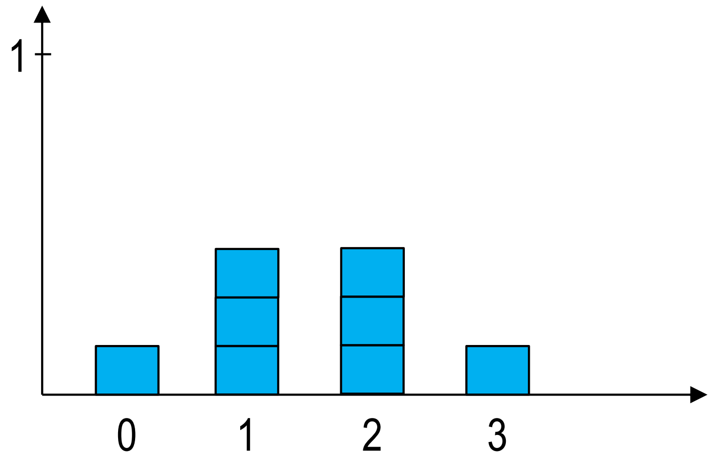
Probability distributions of discrete random variables

- X= # of “heads” after 3 flips of a fair coin (where Heads =0; Tails =1)
- 8 possible outcomes: HHH; HHT;HTH;HTT;THH;THT;TTH;TTT
- P(X=0)=1/8

  P(X=1)=3/8

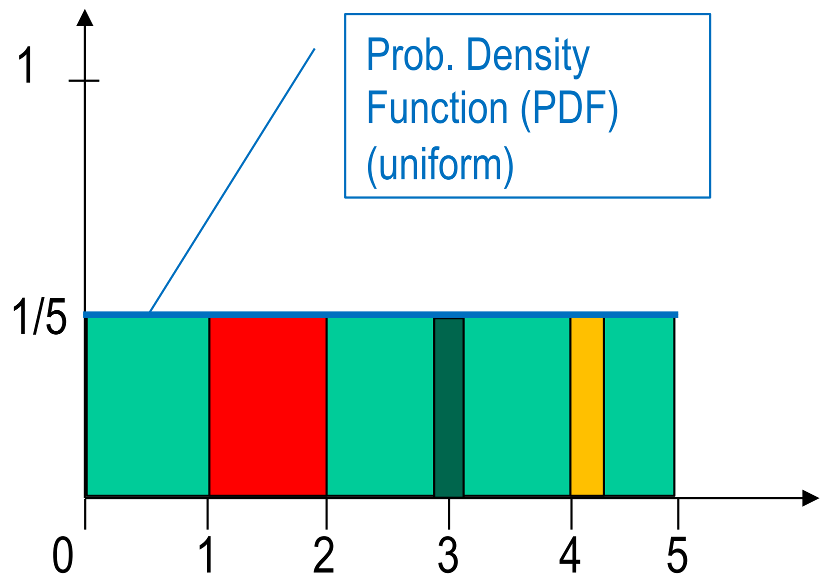
  P(X=2)=3/8

  P(x=3)=1/8



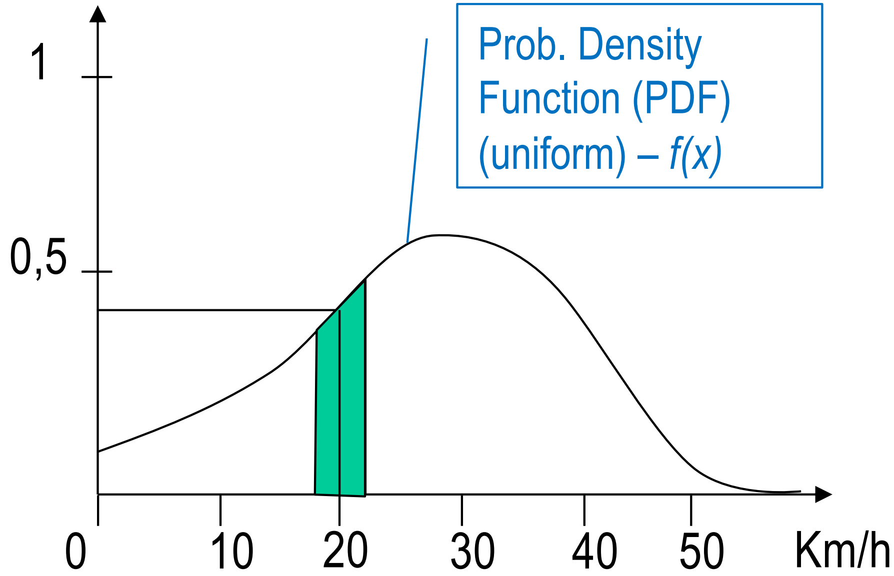
Probability distributions of continuous random variables

- X is a continuous random variable
- P(1=<X=<2)=?=1x1/5=1/5
- P(4=<X=<4.1/3)=?=1/3x1/5=1/15
- P(2,9=<X=<3,1)=?=0,2x1/5=1/5x1/5=1/25
- P(2,99=<X=<3,01)=1/50x1/5=1/250
- P(2,999=<X=<3,001)=1/500x1/5=1/2500
- **P(X=3)=?**



Probability distributions of random variables

- Let X = exact time mean speed of a traffic flow
- What is the prob. of the speed being exactly 20km/h?
  - P(X=20)=0,45???
  - NO!!!

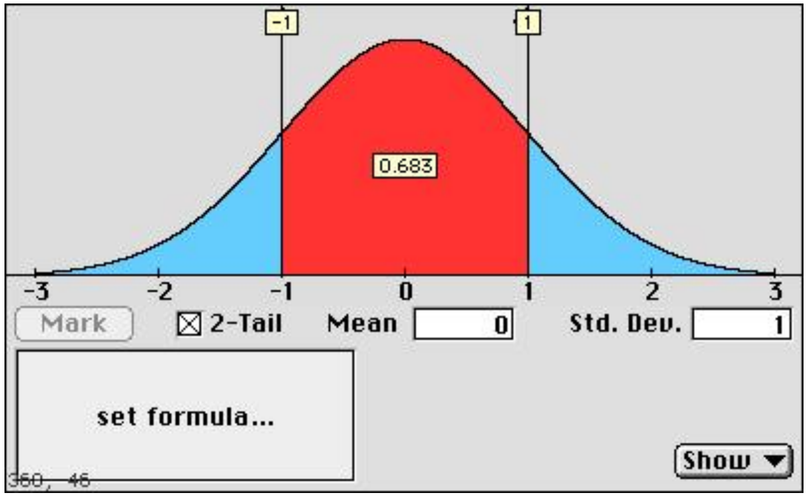


- What is the prob. of the speed being approx. 20km/h?
- P(|X-2|<2)=?
- Integral of the PDF from 18 to 22 (Green area)
- $P(|X-20|<2)=\int_{18}^{22} f(x).dx$

Confidence Intervals

- An **interval** calculated using **sample data** that **contains the true population parameter** with some level of confidence
  - There is a *X*% probability that it contains the true parameter
- This is called a **confidence interval** (CI) and can be constructed for an array of **levels of confidence**
  - Lower confidence limit (LCL)
  - Upper confidence limit (UCL).

- The wider a confidence interval, the more confidence exists that it contains the true population parameter (e.g., mean, variance, etc.)



Source: www.sciencesoftware.com

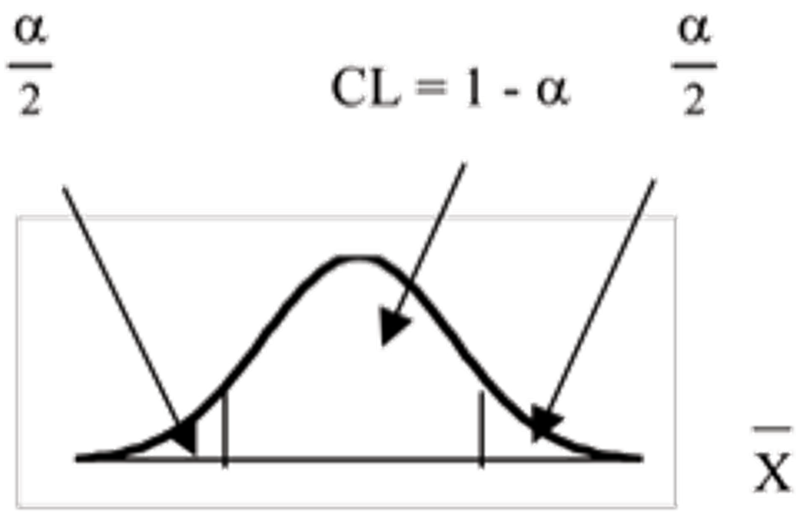
Confidence Interval for μ with known σ^2

❒ Central Limit Theorem

- Whenever a **sufficiently large random sample** is drawn from any population with mean μ and standard deviation σ, the sample mean is **approximately normally distributed** with **mean** $\bar{X}$ and **standard deviation** $\sigma/\sqrt{n}$.
- **Standardization** of the variable X is

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$, where $Z \sim N(0,1)$

Confidence Interval for μ with known σ^2



Source: www.cnx.org

- The **confidence interval is $(1-\alpha)$** , and $Z_{\alpha/2}$ is the value of Z such that the **area in each of the tails** under the standard normal curve is $(\alpha/2)$.
- The confidence interval estimator of μ can be written as:

$$\overline{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Example 1

- A 95% confidence interval is desired for the mean vehicular speed on a specific road. The assumption of normality is assumed. The sample size is $n = 1296$, and the sample mean is 58.86. Suppose a long history of prior studies has shown the population standard deviation as $\sigma = 5.5$. Calculate the Confidence Interval for μ.

Example 1 - Answer

- Useful formula: $$\overline{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$
- Let *X* be the continuous variable of the "*vehicular speed on a specific road*", with mean μ and standar deviation σ.
- It is said that:
  - n = 1296; $\overline{X}$ = 58,86; σ = 5,5.
- The confidence interval is the following, for $\alpha = 0{,}05$:

$$\overline{X} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} \Leftrightarrow 58{,}86 \pm 1{,}96 \times \frac{5{,}5}{\sqrt{1296}} \Leftrightarrow 58{,}86 \pm 0{,}30 \Leftrightarrow [58{,}56;59{,}16]$$

where, $Z_{\alpha/2}$=*1,96* for $\alpha = 0{,}05$, assuming that X follows a Normal Distribution.

Confidence Interval for the Mean with Unknown Variance

- In most cases the population variance is not known. On the contrary, it is estimated from the data (estimated from the sample data).
- When the population variance is unknown and the population is normally distributed, a $(1 - \alpha)$ confidence interval for μ is given by:

$$\overline{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

, where s is the standard deviation and $t_{\alpha/2}$ is the value of the t distribution with n-1 degrees of freedom.

Example 2 - Answer

Assuming the previous example what would be the confidence interval when one considers that the population variance is not known?

Answer:

- Useful formula: $\overline{X} \pm t_{\alpha/2} \times \frac{s}{\sqrt{n}}$
- Let *X* be the continuous variable of the "*vehicular speed on a specific road*", with mean μ and standar deviation σ.
- It is said that:
  - $n = 1296$; $\overline{X} = 58{,}86$; $s = 4{,}41$ (if you go back to your calculation of sample standard deviation of speeds database of the exercise from previour lecture).
- The confidence interval is the following, for $\alpha = 0{,}05$:

$$\overline{X} \pm t_{\alpha/2} \times \frac{s}{\sqrt{n}} \Leftrightarrow 58{,}86 \pm 1{,}96 \times \frac{4{,}41}{\sqrt{1296}} \Leftrightarrow 58{,}86 \pm 0{,}24 \Leftrightarrow [58{,}61;59{,}10]$$

where, $t_{\alpha/2}$=*1,96* for $\alpha = 0{,}05$ and $n\text{-}1=1295$ Degress of Freedom.

Confidence Interval for a Population Proportion

- We might be interested in the **relative frequency of some characteristic** in a population
  - e.g. % of people who uses public transport
- An estimate of the population proportion, p, whose estimator is $\hat{p}$ has an approximate normal distribution when *n* is sufficiently large. The mean of the sampling distribution $\hat{p}$ is the population proportion p and the standard deviation is $\sqrt{pq/n}$ (where $q=1-p$).
- The **$(1-\alpha)$ confidence interval for the population proportion**, *p* is given by

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

, where *p* is the number "sucesses" devided by the sample size.

Example 3

- A transit planning agency wants to estimate, at a 95% confidence level, the share of transit users in the daily commute "market" (% of commuters using transit). A random sample of 100 commuters is obtained and it is found that 28 people in the sample are transit users. Calculate the confidence interval of the average proportion *p* of transit users.

Example 3 - Answer

- Useful formula: $\hat{p} \pm Z_{\alpha/2} \times \sqrt{\frac{\hat{p}\hat{q}}{n}}$
- Let *p* be the continuous variable of the "*proportion of transit users*".
- It is said that:
  - $n = 100; p = 28/100 = 0,28$ and $q = 1 - p = 0,72$.
- The confidence interval is the following, for $\alpha = 0,05$:

$$\hat{p} \pm Z_{\alpha/2} \times \sqrt{\frac{\hat{p}\hat{q}}{n}} \Leftrightarrow 0,28 \pm 1,96 * \sqrt{\frac{0,28 \times 0,72}{100}} \Leftrightarrow 0,28 \pm 0,088 \Leftrightarrow [0,192; 0,368]$$

where, $Z_{\alpha/2}=1,96$ for $\alpha = 0,05$.

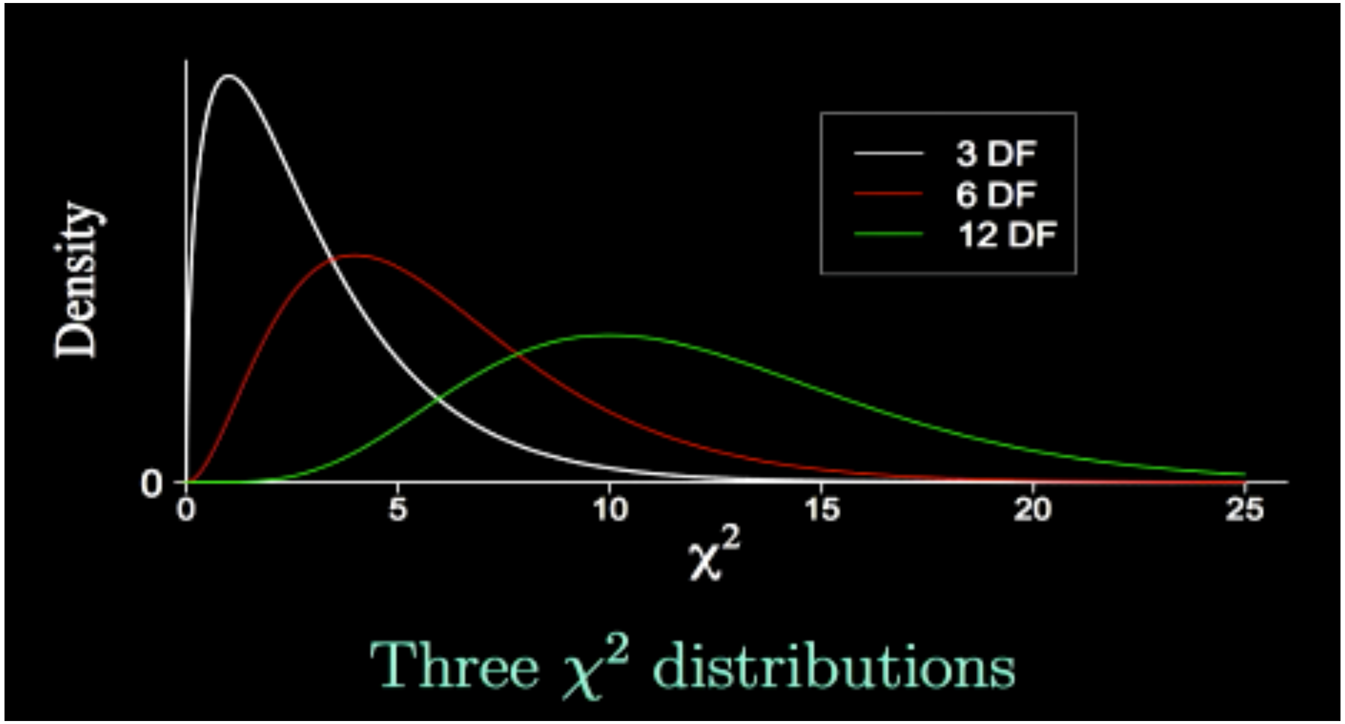
Confidence Interval for the Population Variance

- Sometimes (e.g. traffic safety), interest is on the **population variance**.
  - E.g.,variability in speeds is correlated with the frequency of crashes
- A **confidence interval for s^2**, assuming the population is normally distributed, is given by

$$X = \frac{(n-1)s^2}{\sigma^2} \quad \text{and } X \sim \chi^2 \text{ then} \quad \left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \right]$$

  - $\chi^2_{\alpha/2}$ is the value of the χ^2 distribution with *n-1* degrees of freedom
  - The area in the right-hand tail of the distribution is $\chi^2_{\alpha/2}$, while the area in the left-hand tail of the distribution is $\chi^2_{1-\alpha/2}$

χ^2 Distribution



Example 4

A 95% confidence interval for the variance of speeds on the road of example 1 is desired.

Answer:

- Useful formula: $\left[\dfrac{(n-1)s^2}{\chi^2_{\alpha/2}}, \dfrac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right]$
- Let s^2 be the continuous variable of "*sample variance of vehicular speed on a specific road*".
- It is said that:
  - $n = 100$; $s^2 = 19{,}51$ (if you go back to your calculation of sample standard deviation of speeds database of the exercise from previous lecture).
- The confidence interval is the following, for $\alpha = 0{,}05$:

$$\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}};\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right]=\left[\frac{(100-1)19{,}51}{129{,}56};\frac{(100-1)19{,}51}{74{,}22}\right]=\left[15{,}05;26{,}02\right]$$

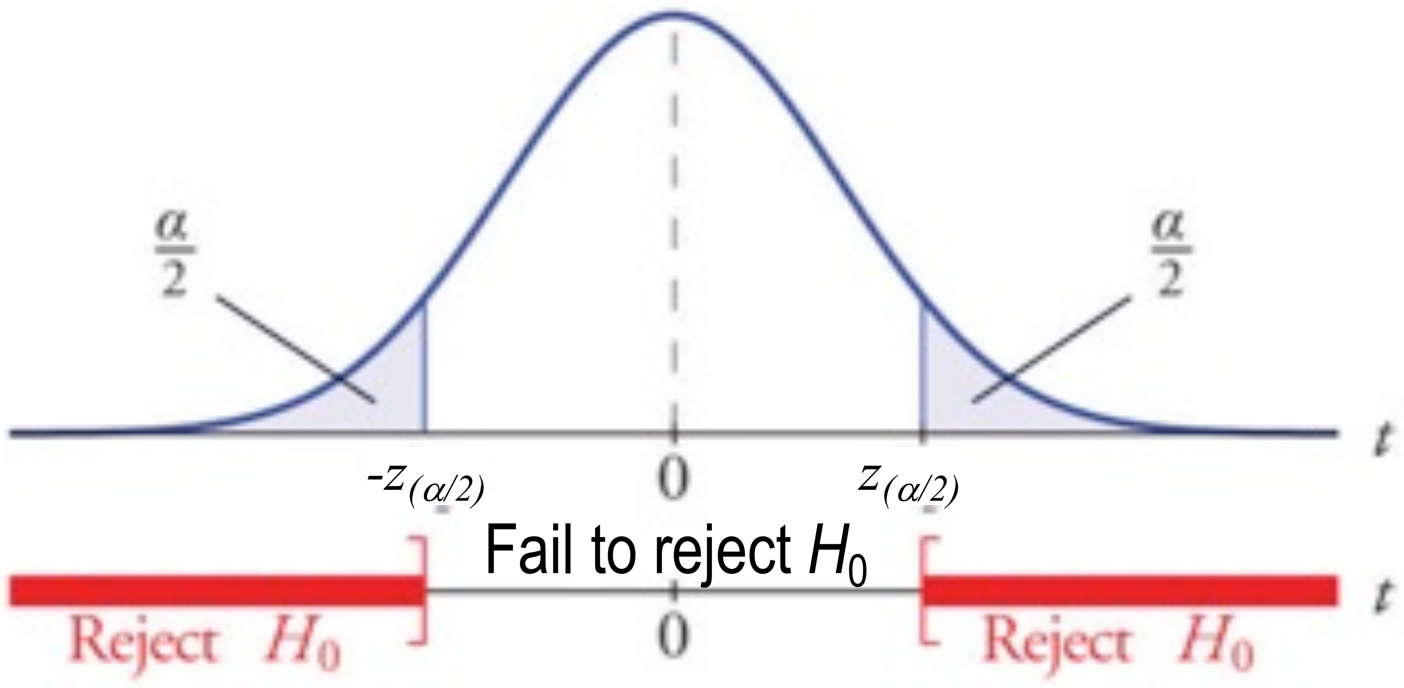
where, $\chi^2_{\alpha/2}$=*129,56* and $\chi^2_{1-\alpha/2}$=*74,22, for* $n-1=99$ Degress of Freedom.

Hypothesis Tests (I)

- **Hypothesis tests** are used to assess the evidence on whether a **difference in a population parameter** (a mean, variance, proportion, etc.) between two or more groups is likely to have **arisen by chance** or whether some other **factor is responsible for the difference**.
- Two competing statistical hypotheses:
  - The **null hypothesis (H_0)** *is an assertion about one or more population parameters* assumed to be **true**
  - The **alternative hypothesis, (H_a)**, is the assertion of all situations **not covered by the null hypothesis** (i.e., wrong).
- *They* constitute a set of hypotheses that covers all possible values of the parameter or parameters in question.

Visualization of Hypothesis testing

$$H_0 : \mu = \mu_0$$
$$H_a : \mu \neq \mu_0$$



Hypothesis Tests (II)

- An hypothesis test aim to determine if is appropriate **to reject** or **not the null hypothesis**.
- The nature of the hypothesis test is determined by the question being asked
  - E.g., if speed signals are expected to change the mean of vehicle speeds, then a null hypothesis of no difference in means is appropriate.
- The process is the following:
  - the empirical evidence is assessed
  - The results of the test will either **refute** or **fail to refute** the null hypothesis based on a pre-specified level of confidence $(1-\alpha)$.

Hypothesis Tests (III)

- It can never be proved that a **statistical hypothesis is true using the results of a statistical test**.
- We simply **admit that H_0 cannot be ruled out** by the **observed data**.
- However, errors do occur among possible results of a test of hypothesis, including type I and II errors.

		Reality	
	Test Result	**H_0 is true**	**H_0 is false**
Decision	**Reject**	Type I error P(Type I error) = α	Correct decision
	Do not reject	Correct decision	Type II error P(Type II error) = β

Visualizing Type I errors

95% of all sample means ($\bar{x}_i$) are hypothesized to be in this region

Fail to reject the null hypothesis

Fail to reject the null hypothesis

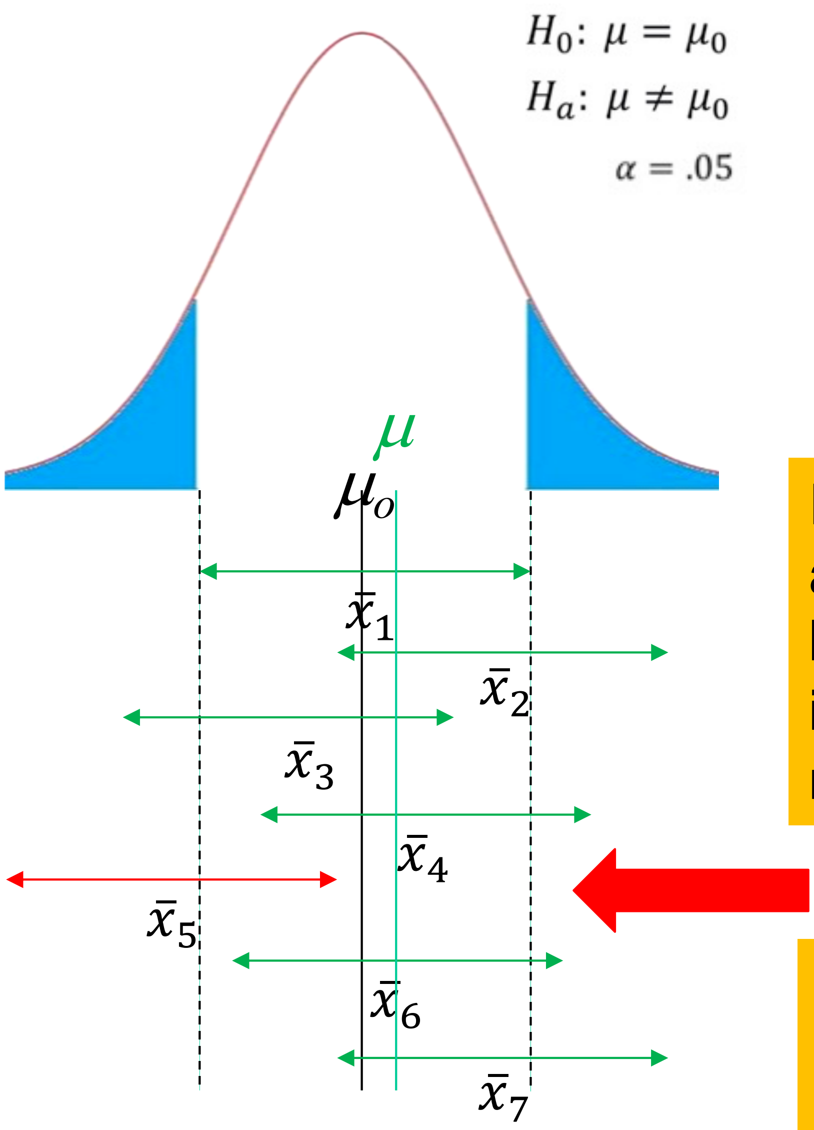
Fail to reject the null hypothesis

Fail to reject the null hypothesis

Reject the null hypothesis

Fail to reject the null hypothesis

Fail to reject the null hypothesis



If we took a sample and it was by chance like x_5, we would incorrectly reject the null hypothesis.

Type I error

α is the "level of tolerance" or our tolerance for making a Type I error.

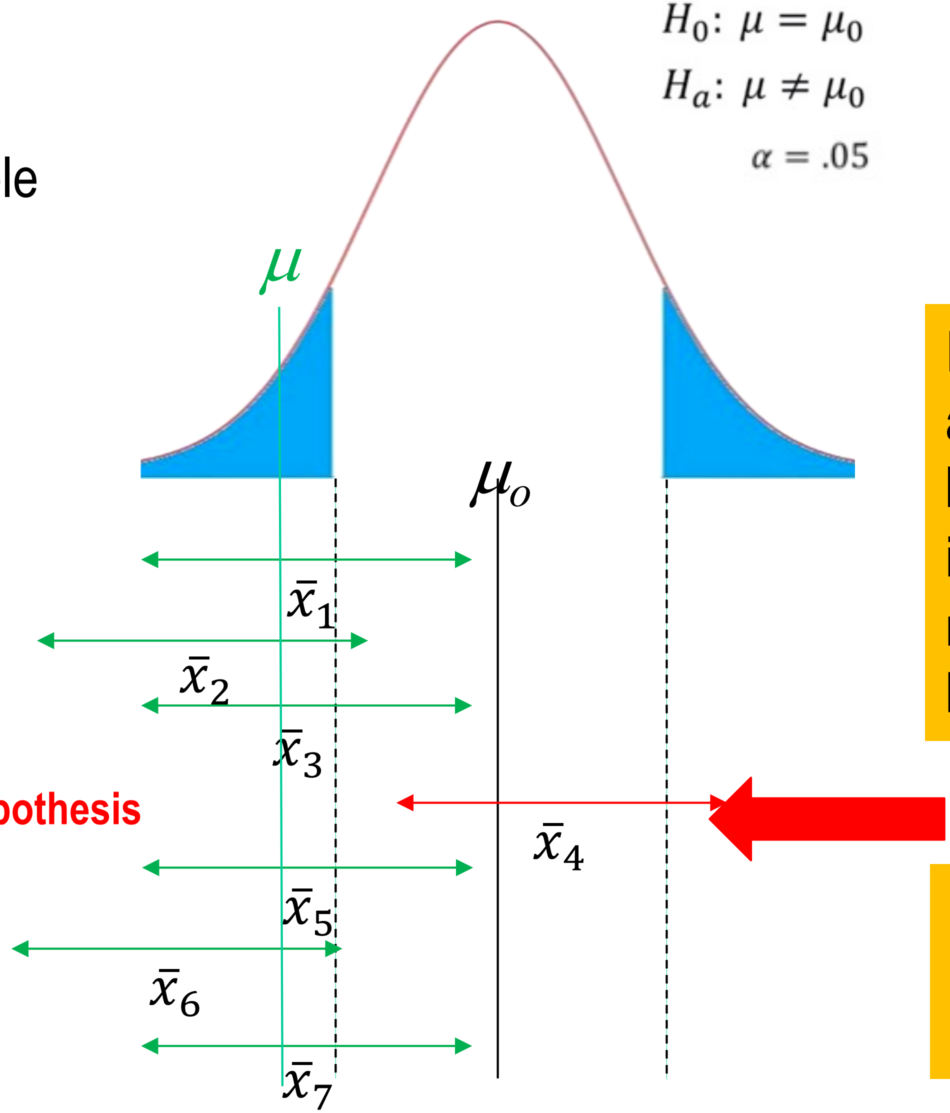
Visualizing Type II errors

H_0: $\mu = \mu_0$

H_a: $\mu \neq \mu_0$

$\alpha = .05$

95% of all sample means ($\bar{x}_i$) are hypothesized to be in this region



Reject the null hypothesis

Reject the null hypothesis

Reject the null hypothesis

Fail to reject the null hypothesis

Reject the null hypothesis

Reject the null hypothesis

Reject the null hypothesis

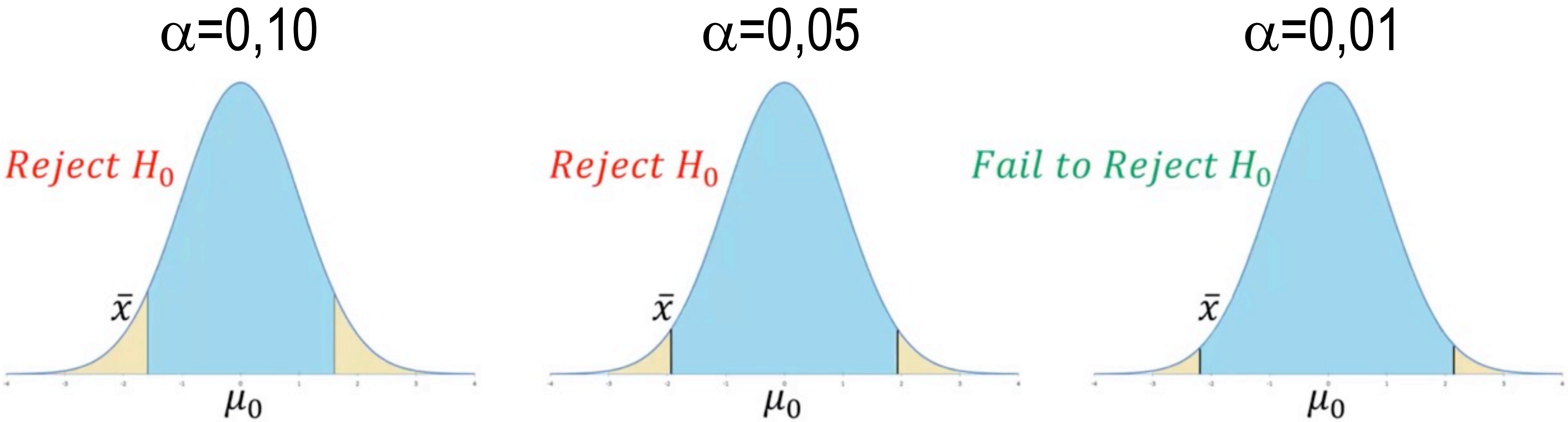
If we took a sample and it was by chance like x_4, we would incorrectly fail to reject the null hypothesis.

Type II error

β is the probability of committing Type II error. The value of β varies with experimental factors.

Type I and II errors and Level of significance

- As α decreases so does the Type I error. The critical value to reject the null hypothesis moves outwards thus "capturing" more sample means.



- However the move outward of the critical values may also "capture" a mean from a different population off to the side. We would fail to reject the null H when indeed we should. Thus the chance of Type II error increases as α decreases.

Main causes of Type I and II errors

- When selecting samples we are always subject to the randomness of data and the chance of getting “wrong” samples
- We may, by random chance alone, select a sample that is not representative of the population
  - Sample of one “type” of data not ranging the full range of possible types (for example, by chance only, interview young white collars)
  - Sample being in the far out tails of the sampling distribution
- Sampling techniques may be flawed / biased
  - Wrong sample frame
  - Wrong sampling approach
  - Systematic error in the collection procedure

FEUP

Type I and type II errors, in statistics (III)

- Since both probabilities **α and β reflect probabilities of making errors**, they should be kept as **small as possible**.
  - There is a trade-off between the two.
  - Usually, the probability of making a Type II error is often ignored.
- **The smaller the α, the larger the β**.
  - Making α really small increases the probability of making a Type II error, all else being equal.
- The **consequences** of making Type I and Type II errors, as well as the research question, should guide the decision on which **statistical error is least desirable**.

Hypothesis Tests (V)

- As discussed previously, the decision of whether the null hypothesis is rejected (or not) is based on the **rejection region.**
- **Two tailed test**:

$$H_0 : \mu = c \qquad Z^* = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \qquad P\left[Z^* \geq Z_c\right] = P\left[Z^* \leq -Z_c\right] = \alpha/2$$
$$H_a : \mu \neq c$$

  - If $|Z^*| \geq Z_c$, then the **probability** of observing this value (or **larger**) **is** α. **H_0 is rejected in favor of H_a**.
  - If $|Z^*| < Z_c$, then the **probability** of observing this value (or **smaller**) is $(1-\alpha)$. **H_0 fails to be rejected**.

Example 5

- Assuming the data of example 1, test the following hypothesis:

$$H_0 : \mu = 60$$
$$H_a : \mu \neq 60$$

Example 5 – Answer (I)

- Relevant formulas:
  - Confidence interval: $\overline{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
  - Standardized test statistic: $Z^* = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}}$
- Test of hypothesis:

$$H_0 : \mu = 60$$
$$H_a : \mu \neq 60$$

- Let X be the continuous variable of “*vehicular speed on a specific road*”.
- It is said that: μ = 58,86km/h; σ = 5,5km/h; and n =1296.

Example 5 – Answer (II)

- Interval of confidence

$$\overline{X} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} \Leftrightarrow 58,86 \pm 1,96 \times \frac{5,5}{\sqrt{1296}} \Leftrightarrow 58,86 \pm 0,30 \Leftrightarrow [58,56;59,16]$$

  - Since the value of 60km/h is within the rejection area, then we reject the null hypothesis, that the mean speed in that road is 60km/h.

- Standardized test statistic

$$Z^* = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = \frac{58,86 - 60,00}{5,5 / \sqrt{1296}} = -7,46$$

  - Since the test statistic |–7,46| = 7,47 is greater than 1,96, the critical value for a two-tailed test at the 5% level of significance, the null hypothesis is rejected.
  - As expected, a confidence interval and the standardized test statistic lead to identical conclusions.

Hypothesis tests (VI)

- Testing the **Population Mean** with Unknown Variance

$$t^* = \frac{\overline{X} - \mu}{s / \sqrt{n}}$$

, where t^* has t distribution with n-1 degrees of freedom

- Testing the **Population Variance**

$$X^2 = \frac{(n-1)s^2}{\sigma^2}$$

, where X^{2*} has χ^2 distribution with n-1 degrees of freedom, when the population variance is normally distributed with variance equal to s^2.

- Testing for a **Population Proportion**

$$Z^* = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

, where the estimated sample proportion p^ is equal to the number of "successes" observed in the sample divided by the sample size, *n*, and *q* = *1* – p.

Example 6

- A test of whether the variance of speeds on Indiana roads is larger than 20 is calculated at the 5% level of significance, assuming a sample size of 100, the sample variance is 19,51km/h.
- The parameter of interest is the population variance, and the hypothesis to be tested is:

$$H_0 : \sigma^2 \leq 20$$

$$H_a : \sigma^2 > 20$$

Example 6 – Answer (I)

- Relevant formulas:
  - Standardized test statistic: $X^2 = \frac{(n-1)s^2}{\sigma^2}$
- Test of hypothesis:

$$H_0 : \sigma^2 \leq 20$$

$$H_a : \sigma^2 > 20$$

- Let X be the continuous variable of "*vehicular speed variance on a specific road*".
- It is said that: s^2 = 19,51km/h; and n=100.

Example 6 – Answer (II)

- The standardized test statistic is:

$$X^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{99(19.51)}{20} = 96.57$$

- The critical value for a chi-squared random variable with 99 degrees of freedom, α = 0.05 and a right-tailed test is 123.??? =chisqr.inv.rt(0,05;99)
- As such, the null hypothesis cannot be rejected at the 0.05 level of significance.

FEUP

Hypothesis tests
Comparing two populations

- **Comparing parameters of two different populations** is extremely useful in transport studies
  - Example: compare quantities such as speeds, accident rates, pavement performance, etc.
- These tests could be about:
  - **Differences in means**
  - **Differences in proportions**
  - **Differences in variances**

Testing the difference between two means: Independent samples (I)

❒ The test of hypothesis and standardized test statistics are:

$$H_0: \mu_a - \mu_b = 0$$

$$H_a: \mu_a - \mu_b \neq 0$$

$$Z^* = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- For small populations a *t* distribution is used with the following number of degrees of freedom

$$df = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}$$

Testing the difference between two means: Independent samples (II)

- When both **universe variances are equal** there is an alternative test for the difference between two population means, using the *t* distribution

$$t^{*} = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

  - This test uses a pooled variance, s^2_p

$$s_p^2 = \frac{\left(n_1 - 1\right)s_1^2 + \left(n_2 - 1\right)s_2^2}{n_1 + n_2 - 2}$$

  - The degrees of freedom in this equation are n_1+n_2-2

Example 6

- Interest is focused on whether the cancellation of the NMSL (National Maximum Speed Law) had an effect on the mean speeds on Indiana roads.
- To test this hypothesis, 744 observations in the period before and 552 observations in the period after cancellation are used. A 5% significance level is used.
- Descriptive statistics show that average speeds in the before and after periods are $\overline{X}_a$ = 57,65 and $\overline{X}_b$ = 60,48, respectively. Further, the variances for the before and after the cancellation periods are s_a^2 = 16,4 and s_b^2 = 19,1, respectively.
- Test the competing hypotheses : $H_0 : \mu_a - \mu_b = 0$

$$H_a : \mu_a - \mu_b \neq 0$$

Example 6 – Answer

- Relevant formula and calculation:

$$Z^* = \frac{\left(\overline{X}_a - \overline{X}_b\right) - \left(\mu_a - \mu_b\right)}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}} = \frac{\left(60,48 - 57,65\right) - 0}{\sqrt{\frac{19,1}{552} + \frac{16,4}{744}}} = 11,89$$

  - Since the test statistic 11,89 is much larger that 1,96, the critical value for a two-tailed test at the 5% significance level, and so the null hypothesis is rejected.
  - This result indicates that the mean speed increased after the repeal of the NMSL and that this increase is not likely to have arisen by random chance.

Testing the difference between two means: Paired Observations

- **Paired observations** exist when the change in one condition is tested with the same individuals.
- This results in a improved experiment, because it **removes variations in the measurements** due to different characteristics of the individuals.
  - For example testing different types of tires on different sets of vehicles or in the same set.
- The **test of hypothesis** is:

$$H_0: \mu_d = 0 \qquad H_a: \mu_d \neq 0$$

with $\mu_d = \mu_1 - \mu_2$ and $t^* = \dfrac{\overline{X}_d - \mu_d}{s_d / \sqrt{n_d}}$

, where: μ_d the average difference between each pair of observations;

s_d standard deviation of the differences

n_d number of paired observations

Testing the difference between two population proportions

- The method pertains to **data measured on a qualitative (nominal)**, rather than a quantitative, scale.
- With samples sufficiently large the difference between proportions is approximately normally distributed
- The **test of hypothesis** is:

$$H_0: p_1 - p_2 = 0$$

$$H_a: p_1 - p_2 \neq 0$$

with $Z^* = \dfrac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$, where $\hat{p}$ is the combined proportions of both samples

and $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$ $\quad \hat{p}_1 = x_1 / n_1 \qquad \hat{p}_2 = x_2 / n_2$

- If the **difference between proportions is some constant c**

$$H_0: p_1 - p_2 \leq 0$$

$$H_a: p_1 - p_2 > 0$$

with $Z^* = \dfrac{(\hat{p}_1 - \hat{p}_2) - c}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$

Non-Parametric tests

- Non- parametric methods are used in situations where only **fewer stringent assumptions could be met** (less information contained in the data)
- Non-parametric methods should be used when:
  - Sample data are **frequency counts**
  - The sample data are measured using an **ordinal scale**
  - The research hypothesis are **not concerned with specific parameters** (e.g. μ and σ^2)
  - Requirements like approximate **normality, large sample size and continuous variables** are **violated**

The purpose of Sampling

- Transport demand analysis often requires estimates of the characteristics of **large populations**:
  - Levels of usage of public transportation;
  - Number of trips per individual;
  - Car ownership levels.
- Since **we cannot survey the entire population**, we should resort to survey a part of it.
- Sampling makes it possible to estimate these characteristics with **adequate accuracy** while:
  - Saving money and time;
  - Reducing survey administration problems;
  - Minimizing intrusion.

Definition of sampling terms (I)

- **Population** - The set of all things (people, objects, firms, etc.) for which we wish to estimate its characteristics.
- **Population Element** - An individual unit within the population.
- **Sampling Unit** - An element that makes up a sample such as people, dwelling units, stores and products. It could comprise a number of population elements such as individual persons in a household.
- **Sampling Frame** - A list of sampling units (or a source of information) used to draw a sample. For mobility surveys data from the census is very useful.
- **Sampling Strategy** -The rule of selecting sampling units from the sampling frame.

Definition of sampling terms (II)

- **Sampling error** -The error in an estimate of a population characteristic which is based on a sample rather than a census.
- **Non-response bias** - The error due to the inability to collect information from some respondents in a sample (usually refusals to answer the survey).
- **Response bias** - The error due to systematic distortion of survey responses. Several reasons:
  - social desirability;
  - prestige seeking;
  - post purchase or behavior justification.

The process

1. Define population
2. Identify sampling frame (List of sampling units)
3. Select sampling strategy (How to select sampling units from the sampling frame)
4. Determine sample size
5. Draw sample/collect data

Sampling strategies

- **Probability Sampling** - Any sampling method in which the chance of any population element's inclusion in the sample is known and greater than zero.
  - Can be used to obtain statistically valid estimates of population characteristics.
  - Allows calculation of the magnitudes of sampling errors.

- **Non-Probability Sampling** - Any sampling method in which the probability of any population element's inclusion in the sample is unknown; e.g., convenience and judgmental sampling.

Sampling methods (I)

- **Simple Random Sampling**
  - Each element has an equal chance of being chosen.
- **Systematic Random Sampling**
  - Randomly select a value between 1 and $k=N/n$. *Choose randomly* $1<=j<=k$ and then select all the following elements $j, j+k \ldots j+(n-1)k$.
- **Stratified Random Sampling**
  - The population is subdivided (stratified) into mutually exclusive groups;
  - A simple random sample is then chosen independently from each group (stratum).
  - It has a lower variance than a random sample.
  - Best when variance within strata is very low.

Total variance = variance between strata + variance within strata

FEUP

Sampling methods (II)

- **Cluster Sampling**
  - A random sample of groups is selected and all members of the groups become part of the sample.
- **Multi-Stage Sampling**
  - Consists of several sampling methods used sequentially to select groups of sampling units.
- **Sequential Sampling**
  - An initial small sample is taken and analyzed. Based on the results, a decision is made on subsequent sampling.

Factors that determine sample size

- Number of groups and subgroups in the sample that need to be analyzed
- Required accuracy / effect size
- Cost
- Variability within the population
- Level of confidence and power

Simple random sample
Basic formulas

- Everyone has the same probability of being interviewed
- The inclusion of someone in the sample doesn´t influence the possible inclusion of other

- Estimation of scalar values (average value)

Absolute error for an infinite population :

$$\varepsilon = t_{\alpha/2}\frac{s_x}{\sqrt{n}}$$

,where $t_{\alpha/2}$ is the Student law for a level of significance of α and sample size of *n*

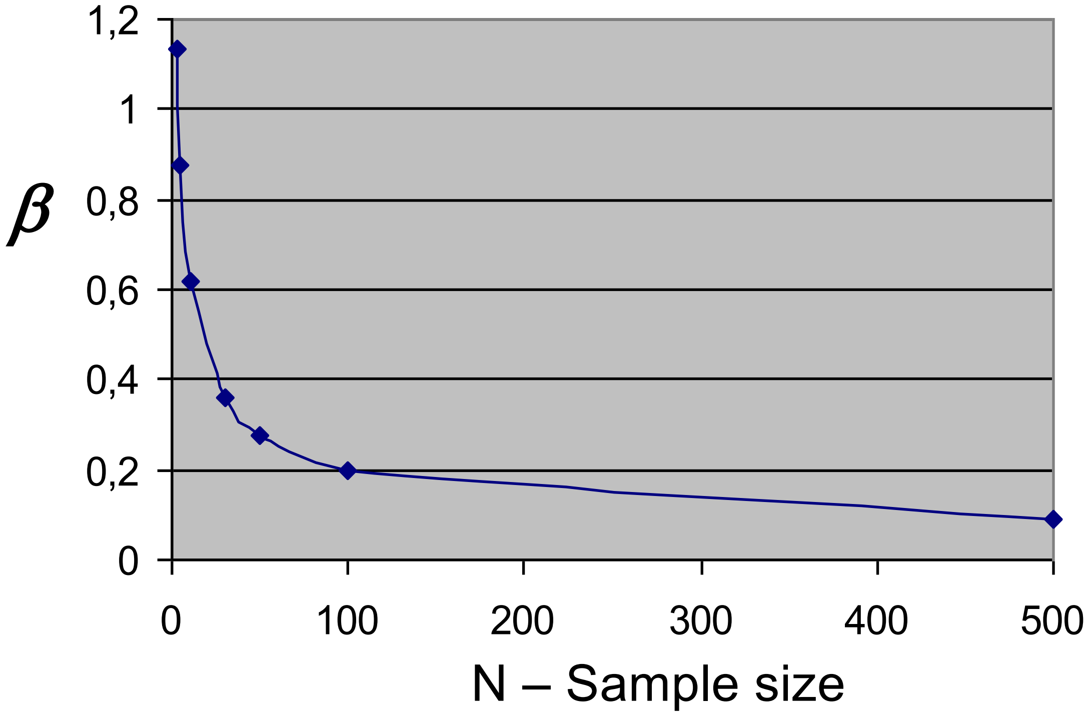
Correcting for a finite population

$$\varepsilon = t_{\alpha/2}\frac{s_x}{\sqrt{n}}\sqrt{\frac{N-n}{N}}$$

Simple random sample
Basic formulas

Relative error (in number of standard deviations) it becomes independent from the variance of the variable that we want to estimate

$$\beta = \frac{\varepsilon}{s_x} = t_{\alpha/2}\frac{1}{\sqrt{n}}\sqrt{\frac{N-n}{N}}$$



Sample dimension as a function of relative error

$$n = \frac{N t_{\alpha/2}^2}{N\beta^2 + t_{\alpha/2}^2} \quad N \to \infty \; n = \frac{t_{\alpha/2}^2}{\beta^2}$$

Proportions
Sample size

For *p* the proportion of a certain cell the confidence interval semi lenght is:

$$\varepsilon = Z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

From the previous expression we have

$$n = \frac{z^2_{\alpha/2}}{\left(\frac{\varepsilon}{p}\right)^2}\frac{(1-p)}{p}$$

When we fix the significance level and the relative error expected the sample varies with the ratio *(1-p)/p*

Stratified Sample (I)

- The population is divided into strata and a sample is taken from each.
- Stratified sampling is worthwhile if
  - The population variance differs by strata, and/or
  - The cost of data collection differs by strata.
- Proportionate Allocation

$$N_{sg} = p_g N_s$$

$$\sum_{g=1}^{G} p_g = 1$$

p_g the proportion of group *g in the population*

N_{sg} *sample size in stratum g*

N_s *total sample size*

Stratified Sample (II)

- Optimal Allocation:
  - Minimizes variance of the estimator $\overline{X}$ subject to budget constraint (Ben Akiva and Lerman (1985), chapter 8)

$$N_{sg} = \frac{p_g \sigma_g / \sqrt{C_g}}{\sum_{g=1}^{G} p_g \, \sigma_g / \sqrt{C_g}} N_s$$

, where σ_g is the standard deviation of stratum *g*

C_g *is* the unit cost of data collection in stratum g

- Often it is a 2-step process
  - Small, simple random sample to learn about strata
  - Optimal sample

Recommended readings

- Washington, Simon P., Karlaftis, Mathew G. e Mannering (2003) “Statistical and econometric Methods for Transportation Data Analysis”, CRC – Chapter 2
- Ben-Akiva, Moshe and Lerman, Steven R (1985) “Discrete Choice Analysis: Theory and Applications to Travel Demand”, MIT Press – Chapter 8
- Juan de Dios Ortúzar, Luis G. Willumsen (2001) “Modeling Transport (3rd edition)”, Wiley and Sons - Chapters 2 and 3