Sampling

Probability sample

Non probability sample

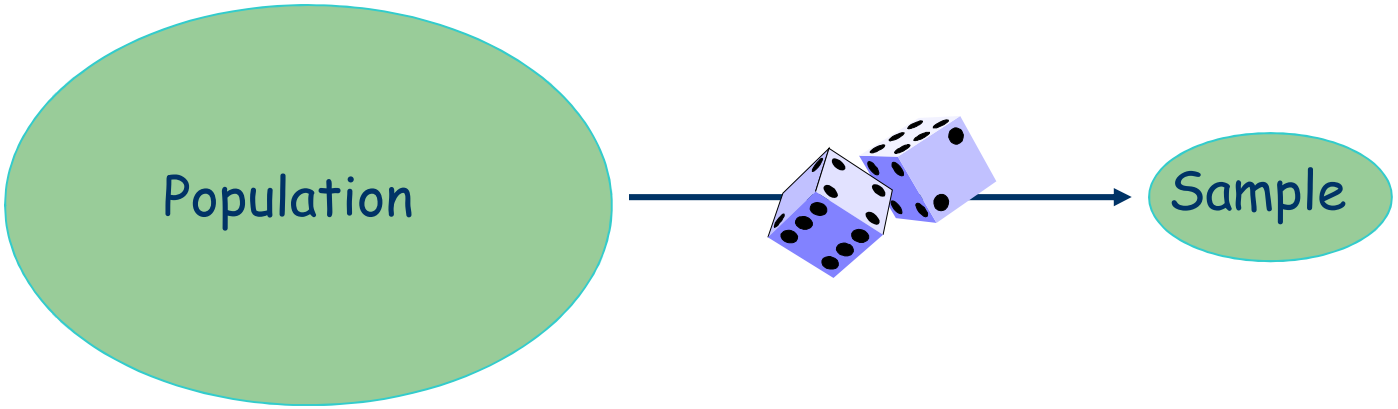
Statistical inference

Sampling error

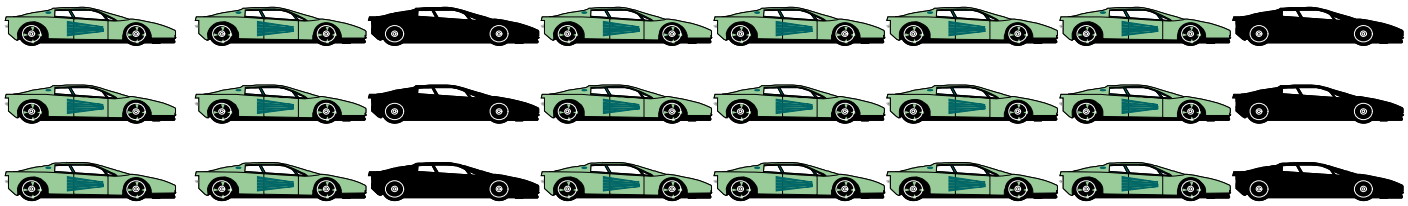
Probability sample

- Goal: A representative sample = miniature of the population
- You can use simple random sampling, systematic sampling, stratified sampling, clustered sampling or combination of these methods to get a probability sample
- Probability sample $\Rightarrow$ You can draw conclusions about the whole population

Simple Random

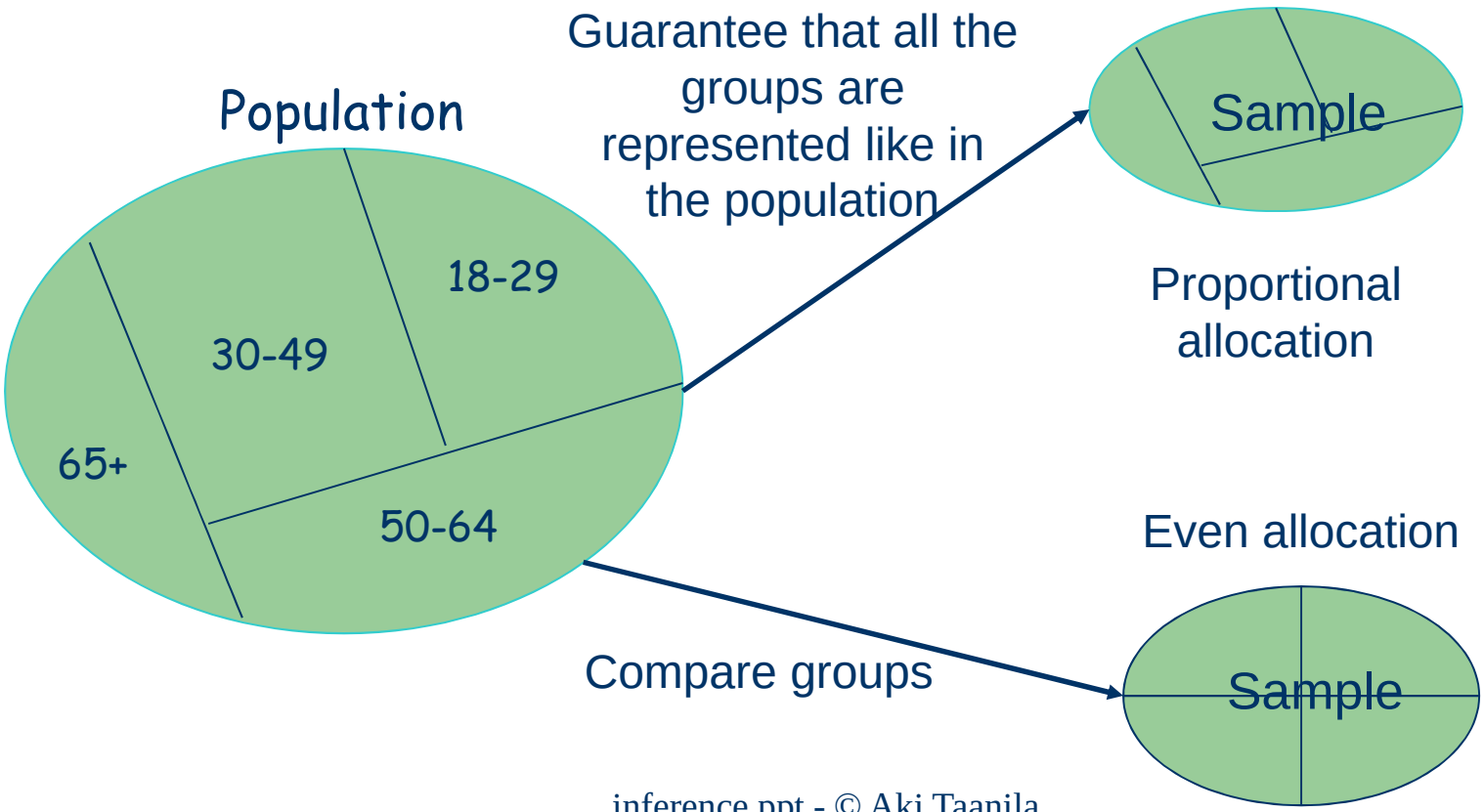


Systematic

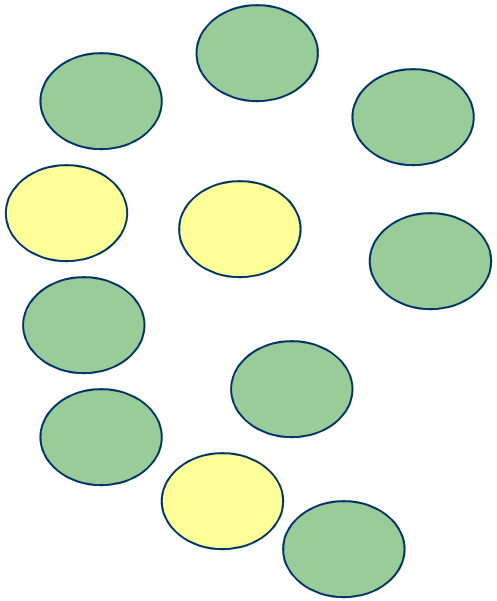


- Select picking interval e.g. every fifth
- Choose randomly one among the first five (or whatever the picking interval is)
- Pick out every fifth (or whatever the picking interval is) beginning from the chosen one

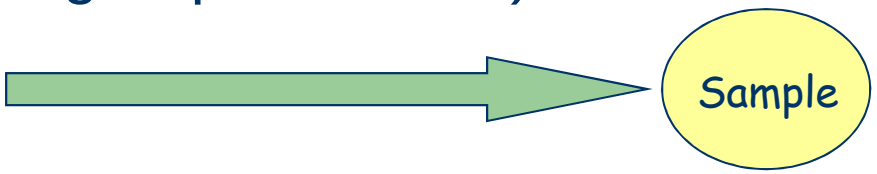
Stratified


Population
18-29
30-49
65+
50-64
Guarantee that all the groups are represented like in the population
Sample
Proportional allocation
Even allocation
Compare groups
Sample

Cluster



- Divide population into the clusters (schools, districts,…)
- Choose randomly some of the clusters
- Draw sample from the chosen clusters using appropriate sampling method (or investigate chosen groups in whole)



Non probability Sample

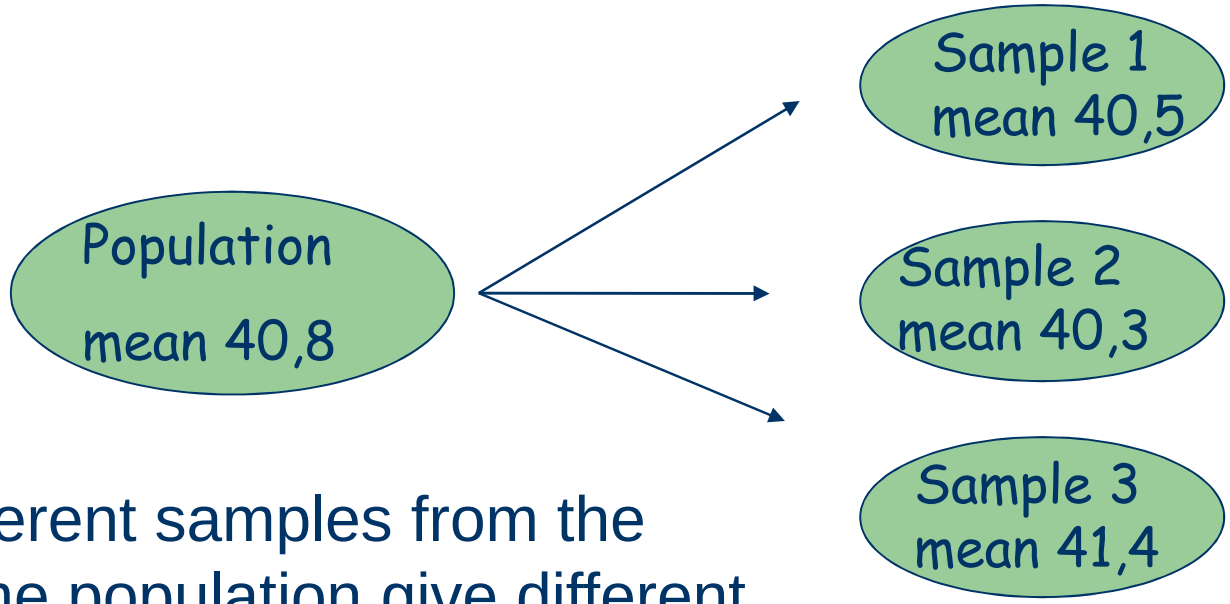
- When a sample is not drawn randomly it is called a non probability sample
- For example, when you use elements most available, like in self-selecting surveys or street interviews
- In the case of a non probability sample you should not draw conclusions about the whole population

Statistical inference

- Statistical inference: Drawing conclusions about the whole population on the basis of a sample

- Precondition for statistical inference: A sample is randomly selected from the population (=probability sample)

Sampling Error



- Different samples from the same population give different results
- Due to chance

Sampling distributions

Mean
- Normal distribution
- T-distribution

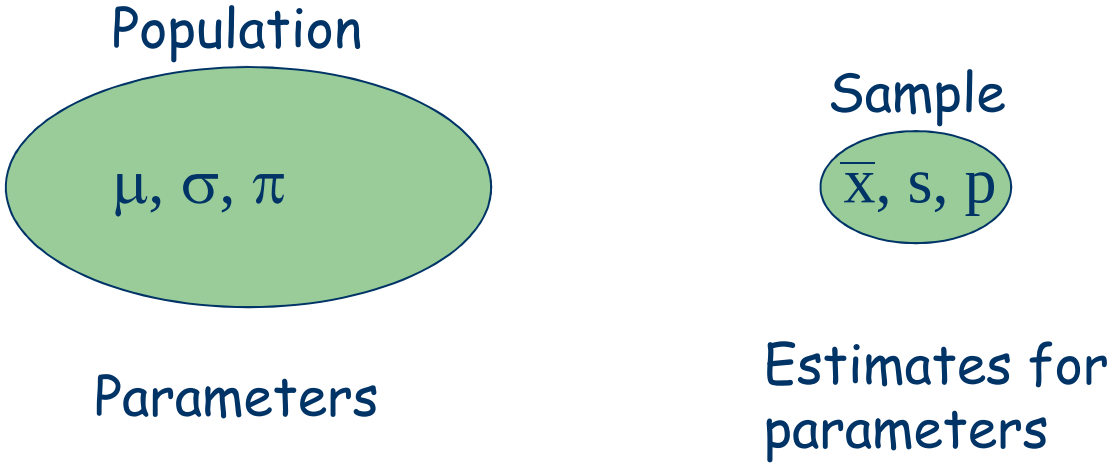
Proportion
- Normal distribution

Sampling distribution

- Most of the statistical inference methods are based on sampling distributions
- You can apply statistical inference without knowing sampling distributions
- Still, it is useful to know, at least the basic idea of sampling distribution
- In the following the sampling distributions of mean and proportion are presented as examples of sampling distributions

Denotations

- Population parameters are denoted using Greek letters μ (mean), σ (standard deviation), π (proportion)
- Sample values are denoted $\overline{x}$ (mean), s (standard deviation), p (proportion)



Sampling Distribution of Mean 1

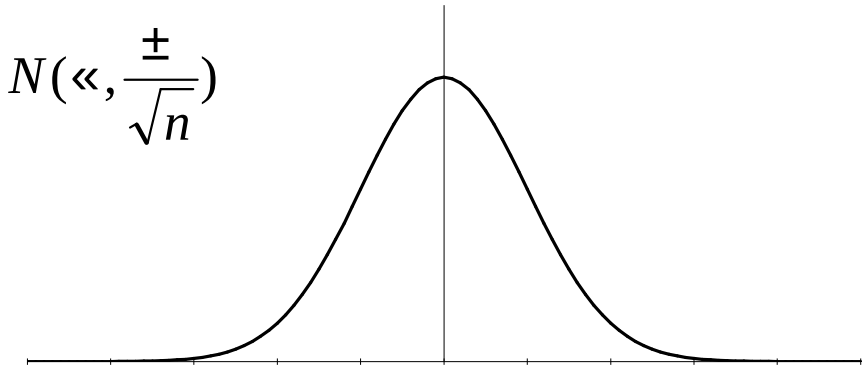
Mean calculated from a sample is usually the best guess for population mean. But different samples give different sample means!

It can be shown that sample means from samples of size n are normally distributed:

$$N(\mu, \frac{\sigma}{\sqrt{n}})$$

Term $\frac{\sigma}{\sqrt{n}}$ is called standard error (standard deviation of sample means).

Sampling Distribution of Mean 2



Sample mean comes from the normal distribution above.

Knowing normal distribution properties, we can be 95% sure that sample mean is in the range:

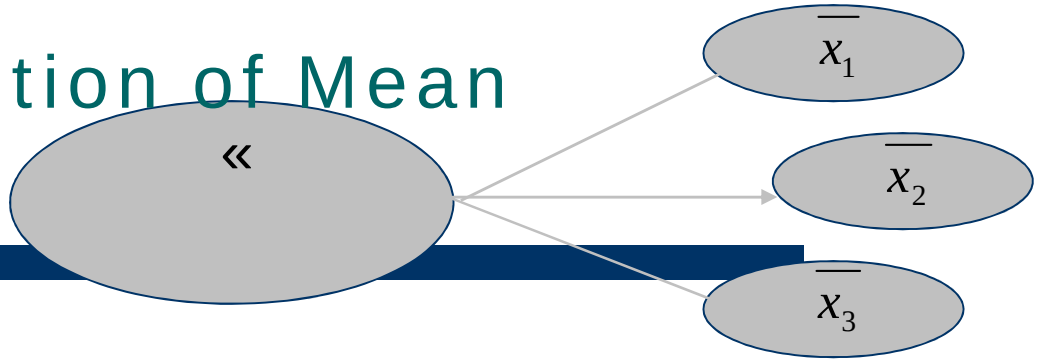
$$\mu - 1{,}96 \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 1{,}96 \cdot \frac{\sigma}{\sqrt{n}}$$

Confidence interval for mean

Based on the previous slide, we can be 95% sure that population mean is in the range:

$$\bar{x} - 1{,}96 \cdot \frac{\pm}{\sqrt{n}} \leq \text{«} \leq \bar{x} + 1{,}96 \cdot \frac{\pm}{\sqrt{n}}$$

Sampling Distribution of Mean
σ unknown



If population standard deviation is unknown then it can be shown that sample means from samples of size n are t-distributed with n-1 degrees of freedom

As an estimate for standard error we can use $\frac{s}{\sqrt{n}}$

Confidence interval for mean
σ unkown

Based on the previous slide, we can be 95% sure that population mean is in the range:

$$\bar{x} - t_{critical} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{critical} \cdot \frac{s}{\sqrt{n}}$$

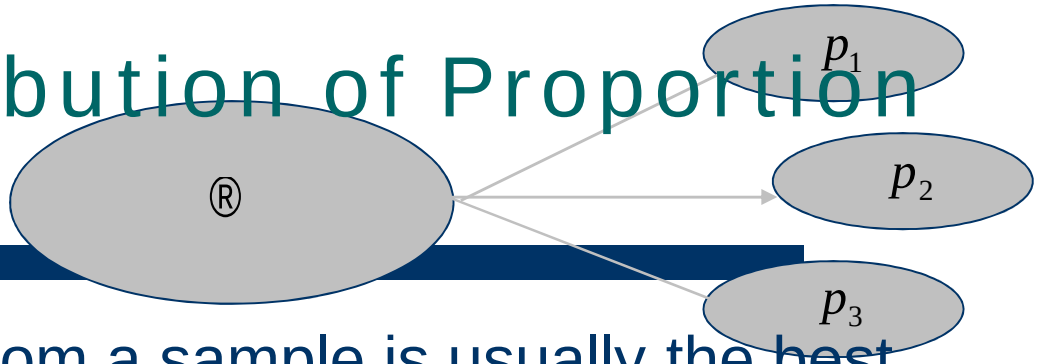
T-distribution

- T-distribution is quite similar to normal distribution, but the exact shape of t-distribution depends on sample size
- When sample size increases then t-distribution approaches normal distribution
- T-distribution's critical values can be calculated with Excel

  =TINV(probability;degrees of freedom)
- In the case of error margin for mean degrees of freedom equals n – 1 (n=sample size)
- Ex. Critical value for 95% confidence level when sample size is 50:

  =TINV(0,05;49) results 2,00957

Sampling Distribution of Proportion



- Proportion calculated from a sample is usually the best guess for population proportion. But different samples give different sample proportions!
- It can be shown that proportions from samples of size n are normally distributed $N(®, \sqrt{\frac{®(1-®)}{n}})$
- Standard error (standard deviation of sample proportions) is $\sqrt{\frac{®(1-®)}{n}}$
- As an estimate for standard error we use $\sqrt{\frac{p(1-p)}{n}}$

Error margin for proportion

- Based on the sampling distribution of proportion we can be 95% sure that population proportion is (95% confidence interval)

$$p - 1{,}96\sqrt{\frac{p(1-p)}{n}} \leq ® \leq p + 1{,}96\sqrt{\frac{p(1-p)}{n}}$$

Parameter Estimation

Parameter and its estimate

Error margin

Parameter estimation

- Objective is to estimate the unknown population parameter using the value calculated from the sample
- The parameter may be for example mean or proportion

Error margin

- A value calculated from the sample is the best guess when estimating corresponding population value
- Estimate is still uncertain due to sampling error
- Error margin is a measure of uncertainty
- Using error margin you can state confidence interval: estimate $\pm$ error margin

Error margin for mean - σ known

- If population standard deviation σ is known then error margin for population mean is

$$1{,}96 \cdot \frac{\pm}{\sqrt{n}}$$

- We can be 95% sure that population mean is (95% confidence interval):

$$\bar{x} - 1{,}96 \cdot \frac{\pm}{\sqrt{n}} \leq \text{«} \leq \bar{x} + 1{,}96 \cdot \frac{\pm}{\sqrt{n}}$$

Error margin for mean - σ unknown

- If population standard deviation is unknown then error margin for population mean is

$$t_{critical} \cdot \frac{s}{\sqrt{n}}$$

- We can be 95% sure that population mean is (95% confidence interval):

$$\bar{x} - t_{critical} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{critical} \cdot \frac{s}{\sqrt{n}}$$

Confidence level

- Confidence level can be selected to be different from 95%
- If population standard deviation σ is known then critical value can be calculated from normal distribution
- Ex. In Excel =-NORMSINV(0,005) gives the critical value for 99% confidence level (0,005 is half of 0,01)
- If population standard deviation σ is unknown then critical value can be calculated from t-distribution
- Ex. In Excel =TINV(0,01;79) gives critical value when sample size is 80 and confidence level is 99%

Error margin for proportion

- Error margin for proportion is

$$1{,}96\sqrt{\frac{p(1-p)}{n}}$$

- We can be 95% sure that population proportion is (95% confidence interval)

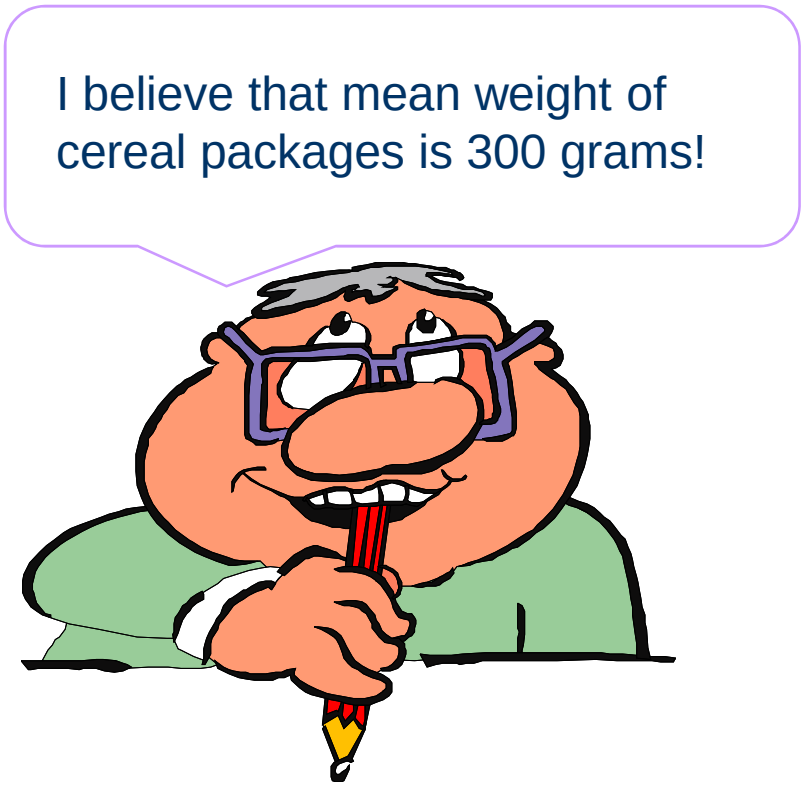
$$p - 1{,}96\sqrt{\frac{p(1-p)}{n}} \leq ® \leq p + 1{,}96\sqrt{\frac{p(1-p)}{n}}$$

Hypothesis testing

Null hypothesis
Alternative hypothesis
2-tailed or 1 –tailed
P-value

Hypothesis 1

- Hypothesis is a belief concerning a parameter
- Parameter may be population mean, proportion, correlation coefficient,...



Hypothesis 2

- Null hypothesis is prevalent opinion, previous knowledge, basic assumption, prevailing theory,...
- Alternative hypothesis is rival opinion
- Null hypothesis is assumed to be true as long as we find evidence against it
- If a sample gives strong enough evidence against null hypothesis then alternative hypothesis comes into force.

Hypothesis examples

H0: Mean height of males equals 174.
H1: Mean height is bigger than 174.

H0: Half of the population is in favour of nuclear power plant.
H1: More than half of the population is in favour of nuclear power plant.

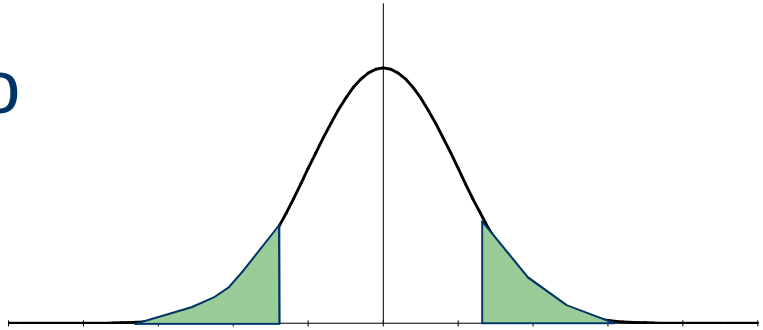
H0: The amount of overtime work is equal for males and females.
H1: The amount of overtime work is not equal for males and females.

H0: There is no correlation between interest rate and gold price.
H1: There is correlation between interest rate and gold price.

2-tailed Test

Use 2-tailed if there is no reason for 1-tailed.

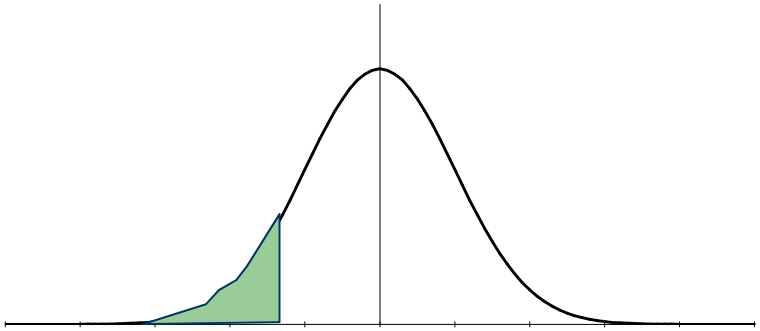
In 2-tailed test deviations (from the null hypothesis) to the both directions are interesting.



Alternative hypothesis takes the form ”different than”.

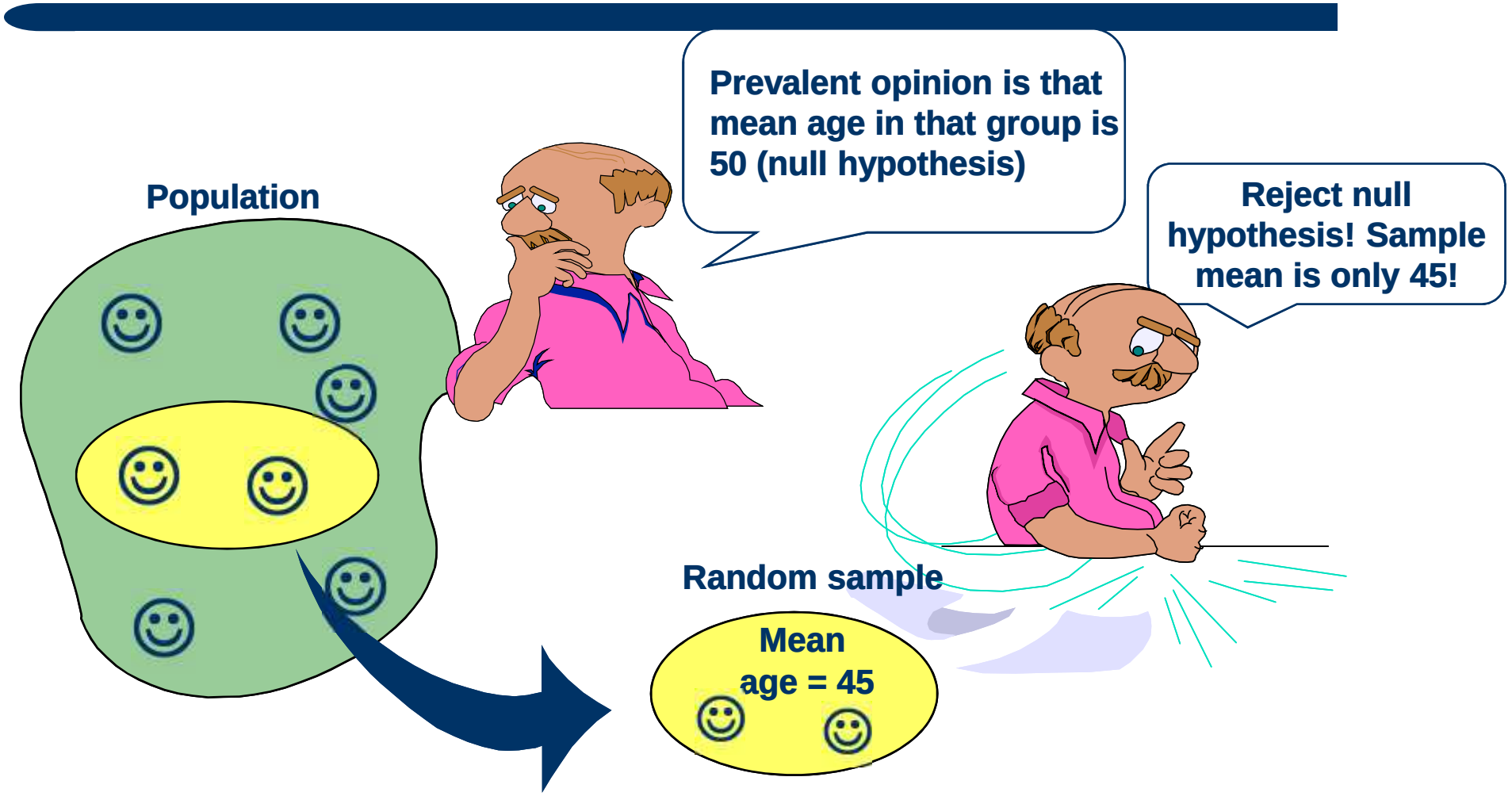
1-tailed Test

In 1-tailed test we know beforehand that only deviations to one direction are possible or interesting.



Alternative hypothesis takes the form ”less than” or ”greater than”.

Logic behind hypothesis testing



Risk of being wrong

Not Guilty until proved otherwise!

Null hypothesis remains valid until proved otherwise!

Sometimes it happens that innocent person is proved guilty.

Same may happen in hypothesis testing: We may reject null hypothesis although it is true. (**there is always a risk of being wrong when we reject null hypothesis; risk is due to sampling error**).

Significance Level

- When we reject the null hypothesis there is a risk of drawing a wrong conclusion
- Risk of drawing a wrong conclusion (called p-value or observed significance level) can be calculated
- Researcher decides the maximum risk (called significance level) he is ready to take
- Usual significance level is 5%

P-value

- We start from the basic assumption: The null hypothesis is true
- P-value is the probability of getting a value equal to or more extreme than the sample result, given that the null hypothesis is true
- Decision rule: If p-value is less than 5% then reject the null hypothesis; if p-value is 5% or more then the null hypothesis remains valid
- In any case, you must give the p-value as a justification for your decision.

Steps in hypothesis testing!

1. Set the null hypothesis and the alternative hypothesis.
2. Calculate the p-value.
3. Decision rule: If the p-value is less than 5% then reject the null hypothesis otherwise the null hypothesis remains valid. In any case, you must give the p-value as a justification for your decision.

Testing mean

- Null hypothesis: Mean equals x_0
- Alternative hypothesis (2-tailed): Mean is different from x_0
- Alternative hypothesis (1-tailed): Mean is less than x_0
- Alternative hypothesis (1-tailed): Mean is bigger than x_0

Testing mean - σ known
p-value

- Calculate standardized sample mean

$$z = \frac{\bar{x} - «}{\pm / \sqrt{n}}$$

- Calculate the p-value that indicates, how likely it is to get this kind of value if we assume that null hypothesis is true
- In Excel you can calculate the p-value: =NORMSDIST(-ABS(z))

Testing mean - σ unknown
p-value

- Calculate standardized sample mean

$$t = \frac{\bar{x} - «}{s/\sqrt{n}}$$

- Calculate the p-value that indicates, how likely it is to get this kind of value if we assume that null hypothesis is true
- In Excel you can calculate the p-value: =TDIST(ABS(t),degrees of freedom,tails); in this case degrees of freedom equals n-1 and tails defines whether you use one-tailed (1) or two-tailed (2) test

Testing Proportion

- In the following p_0 is a value between 0 and 1
- Null hypothesis: Proportion equals p_0*100%
- Alternative hypothesis (2-tailed): Proportion is different from p_0*100%
- Alternative hypothesis (1-tailed): Proportion is less than p_0*100%
- Alternative hypothesis (1-tailed): Proportion is bigger than p_0*100%

Testing proportion
p-value

- Calculate standardized sample proportion

$$z = \frac{p - ®}{\sqrt{\frac{p(1-p)}{n}}}$$

- Calculate the p-value that indicates, how likely it is to get this kind of value if we assume that null hypothesis is true
- In Excel you can calculate the p-value: =NORMSDIST(-ABS(z))

Comparing two group means

- Null hypothesis: Group means are equal
- Alternative hypothesis (2-tailed): Group means are not equal
- Alternative hypothesis (1-tailed): Mean in a group is bigger than in another group

Comparing two group means - Selecting appropriate t-test

- If we have an experiment, in which observations are paired (e.g. group1: salesmen's monthly sales before training and group2: same salesmen's monthly sales after training), then we should use paired sample t-test.
- If we compare two independent groups with equal variances then we should use independent samples t-test for equal variances.
- If we compare two independent groups with unequal variances then we should use independent samples t-test for unequal variances.

Comparing two group means
t-test p-value

- Calculate the p-value using function
  =TTEST(group1;group2;tail;type)

- Group1 refers to cells containing data for group1 and group2 refers to cells containing data for group2
- Tail may be 1 (1-tailed test) or 2 (2-tailed test).
- Type may be 1 (paired t-test), 2 (independent samples t-test for equal variances) or 3 (independent samples t-test for unequal variances).

Equal or unequal variances?

- Independent samples t-test is calculated differently depending on whether we assume population variances equal or unequal
- If sample standard deviations are near each other then you can use equal variances test
- In most cases both ways give almost the same p-value
- If you are unsure about which one to use then you can test whether the variances are equal or not by using F-test
- You should use 2-tailed test with the following hypothesis
  H0: Variances are equal

  H1: Variances are unequal

Equal or unequal variances
p-value

- F-test is included in Tools-menu's Data Analysis – tools
- As an output you get among other things p-value for 1-tailed test
- You have to multiply p-value by two to get p-value for 2-tailed test
- If 2-tailed p-value is less than 0,05 (5%) then you should reject H0 and use t-test for unequal variances

Testing cross tabulation

- Null hypothesis: No relationship in the population
- Alternative hypothesis: Relationship in the population

Testing cross tabulation
p-value

- See http://myy.helia.fi/~taaak/q/inference6.htm
- See SPSS instructions http://myy.helia.fi/~taaak/r/spinference6.htm

Testing correlation

- Null hypothesis: Correlation coefficient equals 0 (no correlation)
- Alternative hypothesis (2-tailed): Correlation coefficient is different from 0
- Alternative hypothesis (1-tailed): Correlation coefficient is less than 0
- Alternative hypothesis (1-tailed): Correlation coefficient is bigger than 0

Testing correlation
p-value

- See http://myy.helia.fi/~taaak/q/inference5.htm
- See SPSS instructions http://myy.helia.fi/~taaak/r/spinference5.htm