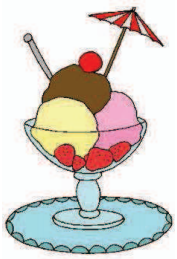
Lecture 5

Hypothesis testing; analysing continuous and categorical data

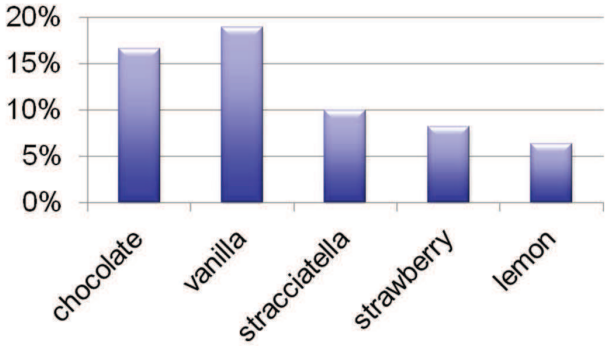
Outline

- Principles of statistical testing
- Comparing continuous variables (t-test)
- Comparing categorical variables (Chi^2 test)
- Non-parametric tests
- Stratification

Students prefer stawberry ice cream



- Favourite types of ice cream in Ghana:



- In contrast, a survey showed 40% of students from Accra prefer strawberry ice cream
- …5 students were interviewed …

Why statistical testing?

- "Problem" when interpreting results of research studies:

  Study population ≠ target population

- Aim: draw conclusions about unknown target population based on the known study population

→ three basic principles of statistical inference

Three basic principles of statistical inference

- Step 1: Estimate the variable of interest

  → Calculation based on study population, result applied to target population

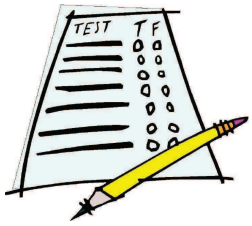
- Step 2: Calculate confidence intervals

  → Confidence intervals since study population ≠ target population

- Step 3: Performing **statistical tests**

  → Yes/No question to evaluate the variable of interest

Statistical tests I



Definition

A statistical test is a method to investigate an assumption (hypothesis) about a particular parameter.

Example

I suspect that (*or:* I want to check whether) the average size of male students is different (taller) from the general population (μ = 179 cm).

A sample of n = 21 is available. It was found: $\bar{x} = 182 \quad s = 5.6$

Is the data consistent with the hypothesis?

Research question: $\bar{x} > \mu$ or $\bar{x} > 179$ cm

→ two hypotheses are formed

Statistical Tests II

H_0:	**Null hypothesis**	$\bar{x}$ = 179 cm (or smaller)
H_1:	**Alternative hypothesis**	$\bar{x}$ > 179 cm

The null hypothesis is always the logical opposite of the research question

Possible results of the statistical test are:

i. data support the null hypothesis (H_0 is not rejected)

or

ii. data do not support the null hypothesis (H_0 is rejected, H_1 is accepted)

Statistical tests III

Statistical tests decide on two opposing hypotheses (H_0, H_1) concerning the target population based on observations in a subsample (study population).

Results of a statistical test are:

"The hypothesis H_0 is accepted"
"The observations provide a statistically non-significant result"
or
"The hypothesis H_0 is rejected"
"The observations provide a statistically significant result"

Errors in statistical tests I

The following errors are possible:

Type I Error: occurs when the null hypothesis is rejected, although it is true

Type II Error: occurs when the null hypothesis is not rejected even though the alternative hypothesis is true

		Decision	
		H_0	H_1
Truth	H_0 true	Correct	False (Type I Error)
	H_1 true	False (Type II Error)	Correct

Errors in statistical tests II

- Both errors cannot be ruled out completely!
- An epidemiological study is a kind of random process
  So error tolerances can be adopted as probabilities:
  Type I Error: Error probability α
  Type II Error: Error probability β
- In statistical tests the error probabilities α and β depend on each other:
  → if α is chosen to be small, then β increases

Statistical tests IV

The essential step in the deduction of a statistical test is to derive a **test statistic (T)** from the data, which has a known distribution under the null hypothesis.

Test statistic (T)

The value (T) is obtained from the sample data with a certain formula. This formula can be simple, but may also be complex.

Accept or reject the null hypothesis

The null hypothesis is rejected if **T** exceeds a certain critical value.

The null hypothesis is not rejected if T does not exceed a certain critical value

Comparing continuous variables (t-test)

- Based on the mean of a sample the **one-sample t-test** tests whether an estimate of a population is equal to, smaller or greater than an predetermined value
- Based on the means of two samples the **two-sample t-test** tests whether the estimates of two populations are equal, smaller or greater

  - paired samples e.g. blood pressure before and after a drug intake (same patient)

  - independent samples e.g. blood pressure with and without medication (two groups of patients)

Example: one-sample t-test

Given a random sample $x_1, x_2, ..., x_{21}$
The available data are the body sizes of 21 male students.

The following results are obtained from the sample:

ESTIMATE OF THE MEAN: $\bar{x}$ =182.24

ESTIMATE OF THE VARIANCE: s^2 =31.36

Value for comparison (hypothesis): $\mu > 179$

Example: one-sample t-test

H_0: Null hypothesis $\bar{x} = \mu$

H_1: Alternative hypothesis $\bar{x} > \mu$

test statistic is calculated by the following formula

$$T = \sqrt{n}\,\frac{\bar{x} - \mu}{s} \qquad T=$$

Extract from t-distribution

Degrees of freedom	1 – α (one sided test)				
	0.85	0.90	0.95	0.99	0.995
1	1.963	3.078	6.314	31.821	63.656
5	1.156	1.476	2.015	3.365	4.032
10	1.093	1.372	1.812	2.764	3.169
15	1.074	1.341	1.753	2.602	2.947
20	1.064	1.325	1.725	2.528	2.845

What are degrees of freedom?

- The number of independent observed values in a sample
- Example: An interview of 30 women (sample) results in the annual number of shoes bought by the women:
  - → Mean: 6 pairs of shoes
  - → Degrees of freedom (df) = n-1 = (30 - 1) = 29
- The first 29 observations are independent, but the 30th observation is fixed by the equation to calculate the mean of 6 pairs of shoes
- A contingency table (categorical data) gives the degrees of freedom (df) by: **df = (s-1)*(z-1)**

  where: **s** = number of columns, **z** = number of lines

Example: one-sample t-test

- Test at significance level α= 0.05: Reject H_0 if $T > T_{crit}$ (*here* T_{crit} = 1.725)
- "Significance level": A false rejection of the null hypothesis shall happen with a small probability α. This probability is usually set to 0.05, i.e. 5%.
- 1.725 is the so-called "critical level" (T_{crit}) and results from the t-distribution (for 20 degrees of freedom and α= 0.05)

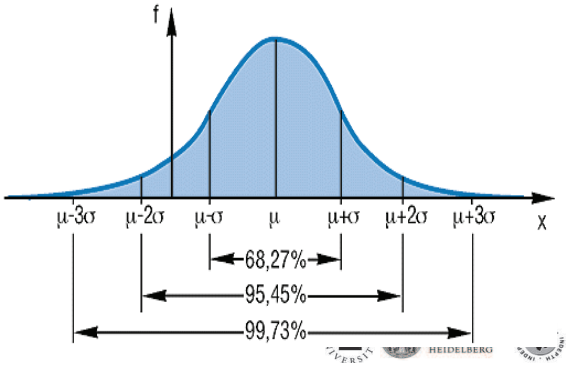
- Since 2.65 > 1.725, the null hypothesis is rejected
  - male students are significantly taller than the general population (with α = 5%)

p-value

- The p-value is the probability of obtaining a test statistic at least as extreme as the observed, given the null hypothesis is true.

- If the p-value < α (usually 0.05) then the null hypothesis is rejected and the result is said to be statistically significant

- Statistical software such as SAS, STATA, SPSS, R, ... calculate the test statistic (T for the t-test) as well as the exact p-value



Hypothesis testing in general

1. Formulate H_0 and H_1
   Usually H_0: there is no difference…..
   results are due to pure chance …….
   H_1: hypothesis of interest
2. Find an appropriate test statistic
3. Decide on α
4. Collect data

<u>without computer</u>

5. Find a critical value test from table
6. Calculate test statistic
7. Compare test statistic with critical value:
   If test statistic > than critical value → reject H_0 otherwise accept H_0

<u>with computer</u>

5. Let the computer calculate statistic and p-value
6. Compare p-value with α:
   p=<α → reject H_0
   p> α → accept H_0

Significance ≠ Relevance

Results of five trials of drugs to lower serum cholesterol

Trial	Drug	Cost	No of patients per group	Difference in mean cholesterol (mg/decilitre)	S.E of difference	95% CI for difference	P-Value
1	A	Cheap	30	-40	40	-118.4 to 38.4	0.32
2	A	Cheap	3000	-40	4	-47.8 to -32.2	<0.001
3	B	Cheap	40	-20	33	-84.7 to 44.7	0.54
4	B	Cheap	4000	-2	3.3	-8.5 to 4.5	0.54
5	C	Expensive	5000	-5	2	-8.9 to -1.1	0.012

Multiplicity of tests

- Given H_0 is true there is a 5% chance that we get a significant result at 5% level
- If we perform 100 tests and all 100 null hypotheses are true we expect 5 significant results (at 5% level).
- When performing many tests it may be necessary to control p-values to adjust for multiplicity of testing

Comparing categorical variables

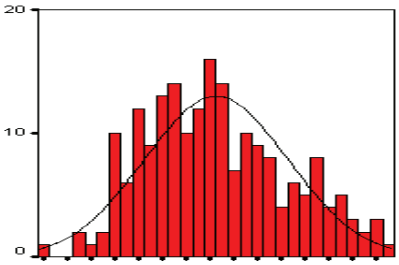
- For each individual in the sample the outcome is one of two (or more) alternatives
  - subject may have experienced a particular disease or remained healthy
  - Gender
  - Categories of education or socioeconomic status (low, middle, high)

Chi^2 – Test (χ^2- Test)

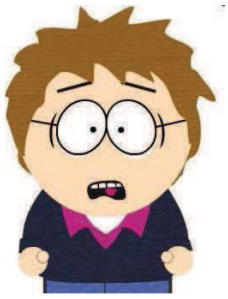
χ^2 - test tests distribution properties of a statistical population

- **Are the two variables independent of each other?**
- Are the frequencies distributed in a certain way?
- Example: Is the population normally distributed?

  (required for statistical tests)



Example Chi² - test (χ^2-test)



- Are men more likely than women to wear glasses (with a significance level $\alpha = 5\%$)?
- Every person has two features: wearing glasses and the sex. Chi²-test tests whether the two traits are independent.
- Out of 100 men, 50 wear glasses and only 30 women out of 100

Contingency table:

	Men	Women	
Wear Glasses	50	30	80
No Glasses	50	70	120
	100	100	200

Example Chi² - test (χ^2-test)

	Men	Women	
Wear Glasses	50 (40)	30 (40)	80
No Glasses	50 (60)	70 (60)	120
	100	100	200

O = observed value

E = expected value

$$E = \frac{\text{(row total x column total)}}{\text{grand total}}$$

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$= 8.33$$

Example Chi^2 - test (χ^2-test)

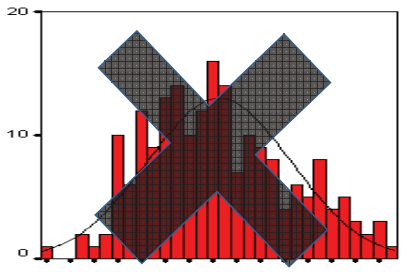
Example: Are men more likely than women to wear glasses (α= 5%)?

$\chi^2 = 8.33$

- The value of χ^2 is looked up in a table of the chi-square distribution:
- Degrees of freedom depends on the number of fields in the table. Here there are 4 fields = 1df
- The χ^2 value of the sample, i.e. 8.33 is greater than 3.84.
- The study shows that men are significantly more likely to wear glasses (for α = 5%).

	1-α					
f	0,900	0,950	0,975	0,990	0,995	0,999
1	2,71	3,84	5,02	6,63	7,88	10,83
2	4,61	5,99	7,38	9,21	10,60	13,82
3	6,25	7,81	9,35	11,34	12,84	16,27

Non-parametric tests



When to use non-parametric tests

- Data do not come from normal distribution
- Small sample size
- If transformation of data to normal distribution is hard to interpret

- Wilcoxon signed rank test
- Kruskal-Wallis test

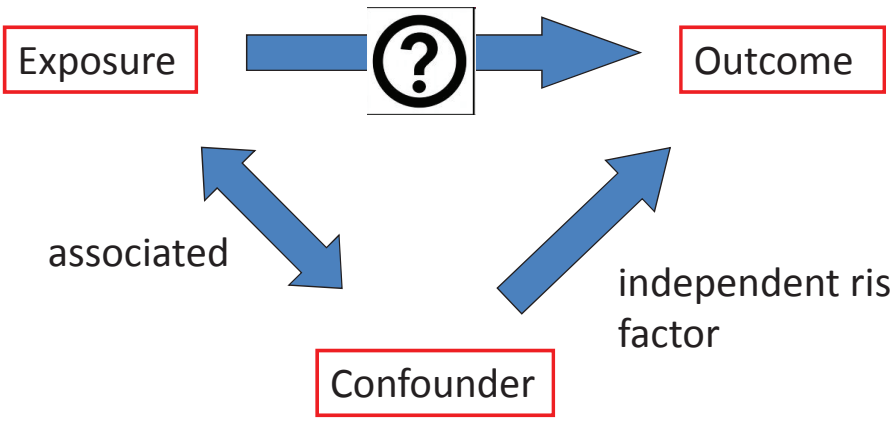
Overview of test procedures

Type of variable

Quantitative (Normal distribution) *(Non-parametric)*	**Qualitative** (Binomial distribution)
1 sample:	
One-sample t-test	test for a single proportion
Sign test, Wilcoxon signed rank test	
2 paired samples:	
paired t-test	McNemar χ^2-test
Sign test, Wilcoxon signed rank test	
2 independent samples:	
t-test Welch-test	χ^2-test
Wilcoxon rank sum test	
≥ 3 independent samples:	
ANOVA	χ^2-test
Kruskal-Wallis-test	

What is confounding?

- A *confounder* is an independent risk factor of the outcome
- A *confounder* must be associated with the exposure

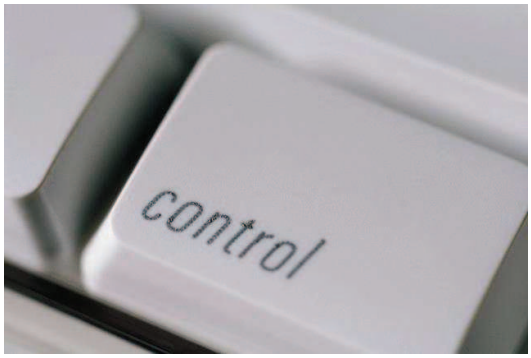


Controlling for confounding

- To control for confounding you must take the confounding variable out of the picture

- Control at the design stage
  - Randomization
  - Restriction
  - Matching

- Control at the analysis stage
  - Stratified analyses
  - Multivariate analyses



Stratification

- Stratification allows for assessment of confounding and effect modification
- "Fix" the level of the confounding variable and produce groups within which the confounder does not vary
- Then evaluate the association within each stratum of the confounder
- Within each stratum, the confounder cannot confound because it does not vary

- Question: Who is an epidemiologist?
- Answer: A physician broken down by age and sex!

An example

	Smoker	Nonsmoker	Total
No heart attack	340 (85%)	510 (85%)	850 (85%)
Heart attack	60 (15%)	90 (15%)	150 (15%)
Total	400 (100%)	600 (100%)	1000 (100%)

- What is the risk ratio for heart attack comparing smokers with non-smokers?
- RR = 15% / 15% = 1.0
- What do we conclude?

Age stratification

Age < 65 years	Smoker	Nonsmoker	Total
No heart attack	282	194	476
Heart attack	18	6	24
Total	300 (60%)	200 (40%)	500 (100%)

Age > 65 years	Smoker	Nonsmoker	Total
No heart attack	58	316	374
Heart attack	42	84	126
Total	100 (20%)	400 (80%)	500 (100%)

Is smoking associated with age?

Prevalence ratio for smoking (old vs young) = 20% / 60% = 0.33

Age stratification

Age < 65 years	Smoker	Nonsmoker	Total
No heart attack	282	194	476 (95%)
Heart attack	18	6	24 (5%)
Total	300	200	500 (100%)

Age > 65 years	Smoker	Nonsmoker	Total
No heart attack	58	316	374 (75%)
Heart attack	42	84	126 (25%)
Total	100	400	500 (100%)

Is risk of heart attack associated with age?

Risk ratio for heart attack (old vs young) = 25% / 5% = 5.0

Age stratification

Age < 65 years	Smoker	Nonsmoker	Total
No heart attack	282 (94%)	194 (97%)	476 (95%)
Heart attack	18 (6%)	6 (3%)	24 (5%)
Total	300 (100%)	200 (100%)	500 (100%)

Age > 65 years	Smoker	Nonsmoker	Total
No heart attack	58 (58%)	316 (79%)	374 (75%)
Heart attack	42 (42%)	84 (21%)	126 (25%)
Total	100 (100%)	400 (100%)	500 (100%)

Is risk of heart attack associated with smoking inside the age groups?

Young: Risk ratio for heart attack (smoking vs not) = 6% / 3% = 2.0

Old: Risk ratio for heart attack (smoking vs not) = 42% / 21% = 2.0

What happened?

Age < 65 years	Smoker	Nonsmoker	Total
No heart attack	282 (94%)	194 (97%)	476 (95%)
Heart attack	18 (6%)	6 (3%)	24 (5%)
Total	300 (100%)	200 (100%)	500 (100%)

Age > 65 years	Smoker	Nonsmoker	Total
No heart attack	58 (58%)	316 (79%)	374 (75%)
Heart attack	42 (42%)	84 (21%)	126 (25%)
Total	100 (100%)	400 (100%)	500 (100%)

Total	Smoker	Nonsmoker	Total
No heart attack	340 (85%)	510 (85%)	850 (85%)
Heart attack	60 (15%)	90 (15%)	150 (15%)
Total	400 (100%)	600 (100%)	1000 (100%)

Stratified Analysis

Crude

OR_{Crude}

Stratum 1 Stratum 2

OR_1 OR_2

Crude 2 x 2 table
Calculate Crude OR (or RR)

Stratify by Confounder

Calculate OR's
for each stratum

If stratum-specific OR's are similar,
calculate adjusted OR (e.g. M-H)

If Crude OR =/= Adjusted OR,
confounding is likely

If Crude OR = Adjusted
OR, confounding is
unlikely

Direction of confounding

- Confounding "pulls" the observed association away from the true association
- It can either exaggerate/over-estimate the association (positive confounding)
  - RR crude = 3
  - RR adjusted = 1
- It can hide/underestimate the true association (negative confounding)
  - RR crude = 1
  - RR adjusted = 3

Crude vs. Adjusted Effects

- Crude:
  - does not take into account the effect of the confounding variable
- Adjusted:
  - accounts for the confounding variable(s)
  - generated using multivariate analyses (e.g. logistic regression)
- Confounding is likely when:
  - RR crude =/= RR adjusted
  - OR crude =/= OR adjusted

Limitations of stratification

- Cannot be used to adjust for several covariates simultaneously
  - adjustment is only for the association between one independent variable and one outcome at a time
- Can adjust for categorical covariates only
- When data is sparse the methods are not useful (i.e. can not calculate stratum-specific rates if the sample size is 0)

Multivariate analysis

- Stratified analysis works best when there are few strata (i.e. if only 1 or 2 confounders have to be controlled)
- If the number of potential confounders is large, multivariate analysis offers the only real solution
  - Can handle large numbers of confounders (e.g. could control for smoking, alcohol, physical activity, diet, in the same analysis)
  - Based on statistical regression models
  - Always done with statistical software packages