

Applied Statistics

Lecture 3

Prof.ssa Chiara Seghieri, Dott.ssa Costanza Tortù

Laboratorio di Management e Sanità, Istituto di Management, L'EMbeDS
Scuola Superiore Sant'Anna, Pisa

c.seghieri@santannapisa.it

c.tortu@santannapisa.it

Outline

1. Methodological Framework
2. Hypothesis test for single values
3. Hypothesis test to compare two samples
4. The ANOVA test
5. Tests to measure the association between categorical variables
6. Hypothesis test for correlation

1) Methodological Framework

Two Types of Inference

1. Confidence Intervals:

- Confidence Intervals give us a range in which the population parameter is likely to fall.
- We use confidence intervals whenever the research question calls for an estimation of a population parameter.

Example: What is the mean age of trees in the Black forest?

Which is the proportion of US adults who would vote for candidate A.

2. Hypothesis Testing:

- Hypothesis tests are tests of population parameters.

Example: Is the proportion of US adult women who would vote for candidate A $> 50\%$?

Is the mean age of trees in the forest > 50 years?

Is average income greater in men than in women?

Hypothesis Testing:

- **Hypothesis:** statement regarding the range of values of the unknown population parameter of interest.

Example of hypothesis regarding the mean of a population μ : *“Is the true mean different from 100?”*

- **Hypothesis testing:** method to make decisions or **inferences on hypotheses** using sample data (evidence). Sample data used to choose between two choices i.e. **hypotheses** or statements about a population parameter. We typically do this by comparing what we have observed to what we expected if one of the statements (**Null Hypothesis**) was true.

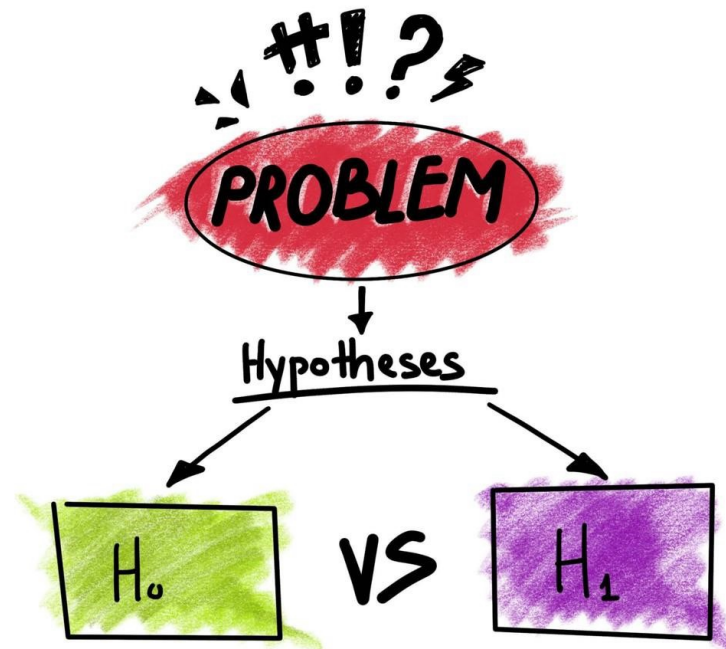
Since CIs provide a set of plausible values for the true value of the population parameter they could be used for testing hypotheses.

Hypothesis Testing: concepts

There are **two** hypotheses called the *null hypothesis* and the *alternative* or *research hypothesis*. The usual notation is:

H_0 : — the 'null' hypothesis

H_1 (or H_a): — the 'alternative' or 'research' hypothesis



Null Hypothesis:

H_0

- ❖ **Null hypothesis:** statement indicating no change, no difference (the status quo); it is usually expressed such as that the value of a population parameter (proportion, mean, standard deviation,...) is **equal to** some claimed value.
- ❖ We test the null hypothesis directly.
- ❖ We can either **reject** H_0 or **fail to reject** H_0 .

Alternative Hypothesis:

$$H_1$$

- ❖ The **alternative hypothesis** (denoted by H_1 or H_a or H_A) is the statement that the parameter has a value that somehow differs from the null hypothesis.
- ❖ The symbolic form of the alternative hypothesis must use one of these symbols: \neq , $<$, $>$.

To sum up:

H_A : Research (Alternative) Hypothesis

- What we aim to gather evidence of
- Typically that there **is** a difference/effect/relationship etc.
- The research hypothesis should be set up by the investigator before any data are collected.

H_0 : Null Hypothesis

- What we assume is true to begin with
- Typically that there is **no** difference/effect/relationship etc.

STRATEGY

The standard procedure is to assume **H_0 is true** - just as we presume innocent until proven guilty. Using probability theory, we try to determine whether there is sufficient evidence to declare H_0 false.

Example: hypothesis testing for the population mean

Null hypothesis – Mean age on the population is equal to 50

$$H_0 : \mu = 50$$

Alternative hypothesis - Mean age is different from 50

$$H_A : \mu \neq 50 \quad \text{two-tailed test}$$

or $H_A : \mu < 50 \quad \text{lower-tailed test}$

$$H_A : \mu > 50 \quad \text{upper-tailed test}$$

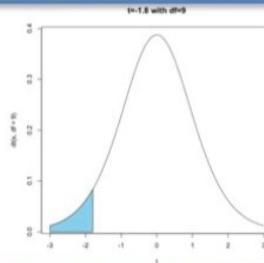
Example: hypothesis testing for the population mean

H_a Description

Figure

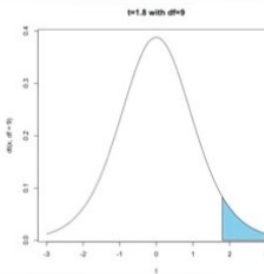
One-tailed

Population mean is below the standard



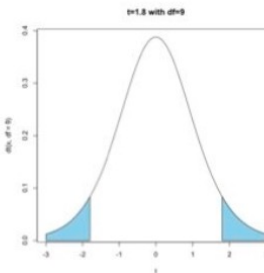
One-tailed

Population mean is above the standard



Two-tailed

Population mean is different from the standard



	2-Tailed Test	Right-Tailed	Left Tailed
Null hypothesis	$H_0: \mu = \mu_0$	$H_0: \mu \leq \mu_0$	$H_0: \mu \geq \mu_0$
Alternative hypothesis	$H_a: \mu \neq \mu_0$	$H_a: \mu > \mu_0$	$H_a: \mu < \mu_0$

Illustrative Example: “Body Weight”

- **The problem:** In the 1970s, 20–29 years old men in the U.S. had a mean μ body weight of 170 pounds. Standard deviation σ was 40 pounds. We test whether mean body weight in the population now differs.
- **Null hypothesis** $H_0: \mu = 170$ (“no difference”)
- The **alternative hypothesis** can be either $H_a: \mu > 170$ (**one-sided test**) body weight in this group has increased since 1970 or $H_a: \mu \neq 170$ (**two-sided test**)

Test Statistics

To check the null hypothesis we calculate a figure known as a **test statistic**, which is based on data from our sample.

Note: Different types of problems require different test statistics.

IDEA: If the test statistic shows you “observed an unlikely result”, you reject the null hypothesis in favour of the alternative hypothesis.

Test statistic: measure employed to make a decision about the null hypothesis. It is found by converting the sample statistic to a score with the assumption that the null hypothesis is true.

Examples of Test Statistic - Formulas

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

Test statistic for
proportions

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

Test statistic for
mean

Decision Criterion

A

Traditional method: the rejection region (typically used when computing statistics manually):

Reject H_0 if the test statistic falls within the critical region of the sampling distribution.

Fail to reject H_0 if the test statistic does not fall within the critical region of the sampling distribution.

B

Alternative Method: the p-value approach (generally used with a computer and statistical software).

Note: failing to reject the null hypothesis does not mean this is true!

Hypothesis Testing: type of errors

		Decision	
		Accept H_0	Reject H_0
Null Hypothesis (H_0)	True	Correct "Confidence Level" Probability = $1 - \alpha$	Type I Error "False Positive" Probability = α
	False	Type II Error "False Negative" Probability = β	Correct "Statistical Power" Probability = $1 - \beta$

$$\alpha = P(\text{Type I Error}) \quad \beta = P(\text{Type II Error})$$

Hypothesis Testing: type of errors

Type I and type II errors are **inversely related**.

By convention it is given more importance to **$P(\text{Type I error}) = \alpha$**
which is usually fixed= 0.05, 0.01 or 0.1

STRATEGY: Fix α , reach the smallest b

NOTE These errors have different consequences in the scientific research!

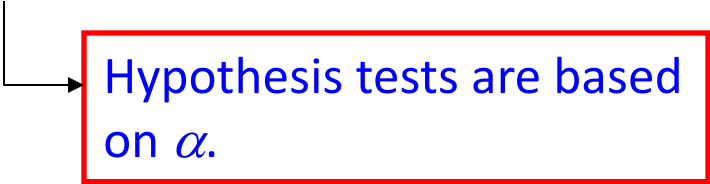
A type I error (reject the null hypothesis when the null is true) introduces a false conclusion into the scientific community and can lead to a tremendous waste of resources before further research invalidates the original finding.

Type II errors can be costly as well, but generally go unnoticed

A type II error – failing to recognize a scientific breakthrough – represents a missed opportunity for scientific progress.

Level of Significance

In a hypothesis test, the **level of significance** is your maximum allowable probability of making a type I error. It is denoted by α .



Hypothesis tests are based on α .

Defines the unlikely values of the sample statistic if the null hypothesis is true.
Defines rejection region of the sampling distribution

By setting the level of significance at a small value, you are saying that you want the probability of rejecting a true null hypothesis to be small.

Commonly used levels of significance:

$$\alpha = 0.10$$

$$\alpha = 0.05$$

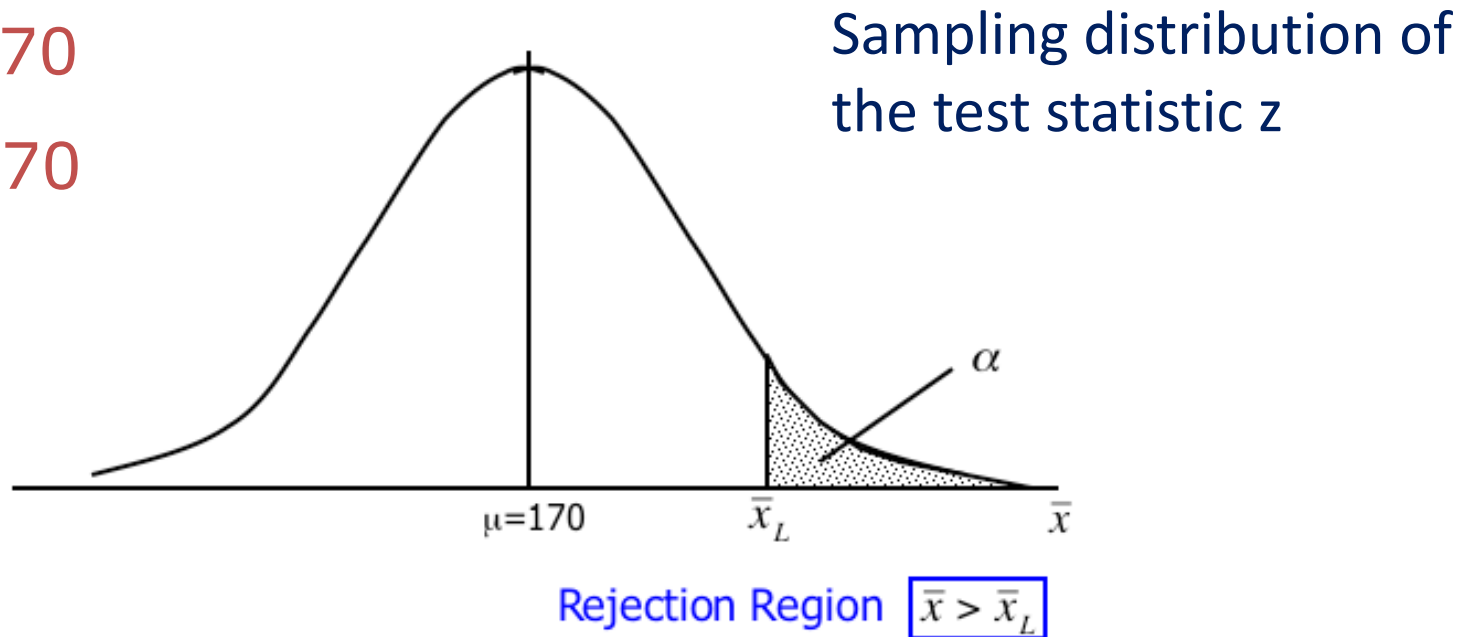
$$\alpha = 0.01$$

The Rejection Region

A **rejection region** (or **critical region**) of the sampling distribution is the range of values for which the null hypothesis is not probable. **If the test statistic falls into that range, we decide to reject the null hypothesis.**

$$H_0: \mu = 170$$

$$H_A: \mu > 170$$

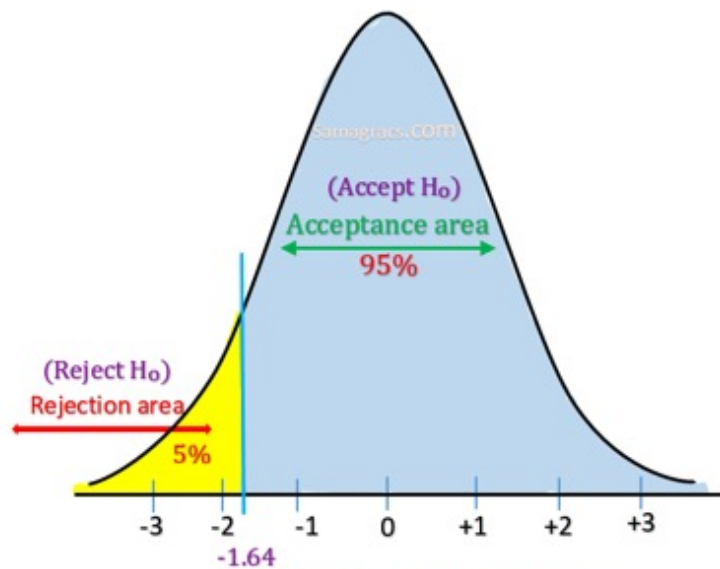
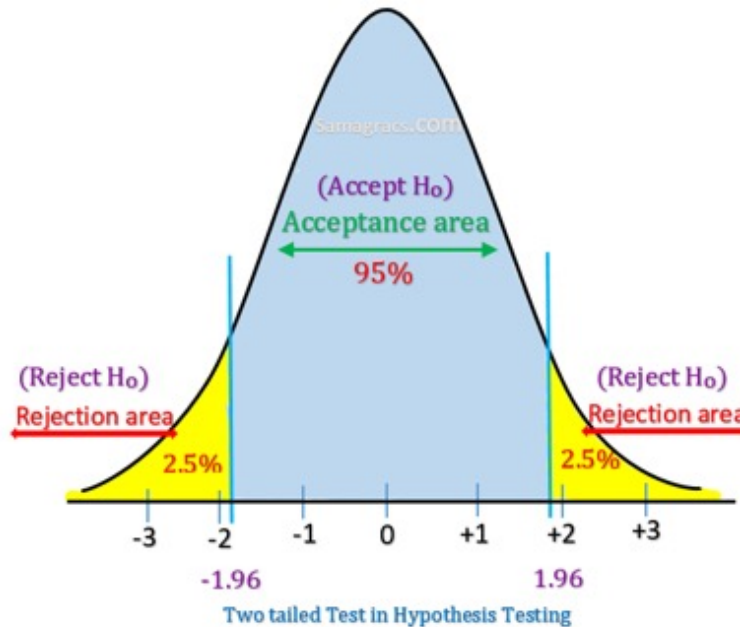


\bar{x}_L is the critical value of \bar{x} to reject H_0 .

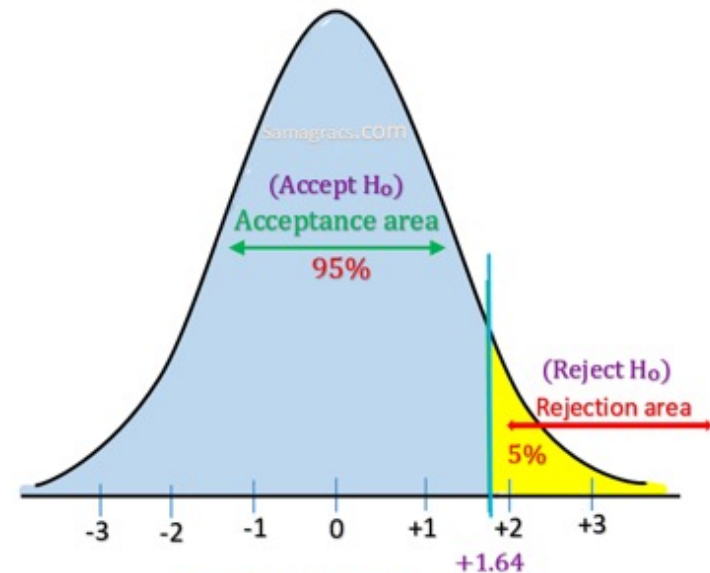
Rejection Region (dependent on H_A)

Alternative Hypothesis	Critical Region
$H_A : \mu \neq \mu_0$	$z < -z_{\alpha/2} \text{ or } z > z_{\alpha/2}$
$H_A : \mu > \mu_0$	$z > z_{\alpha}$
$H_A : \mu < \mu_0$	$z < -z_{\alpha}$

Decision Criterion: traditional method

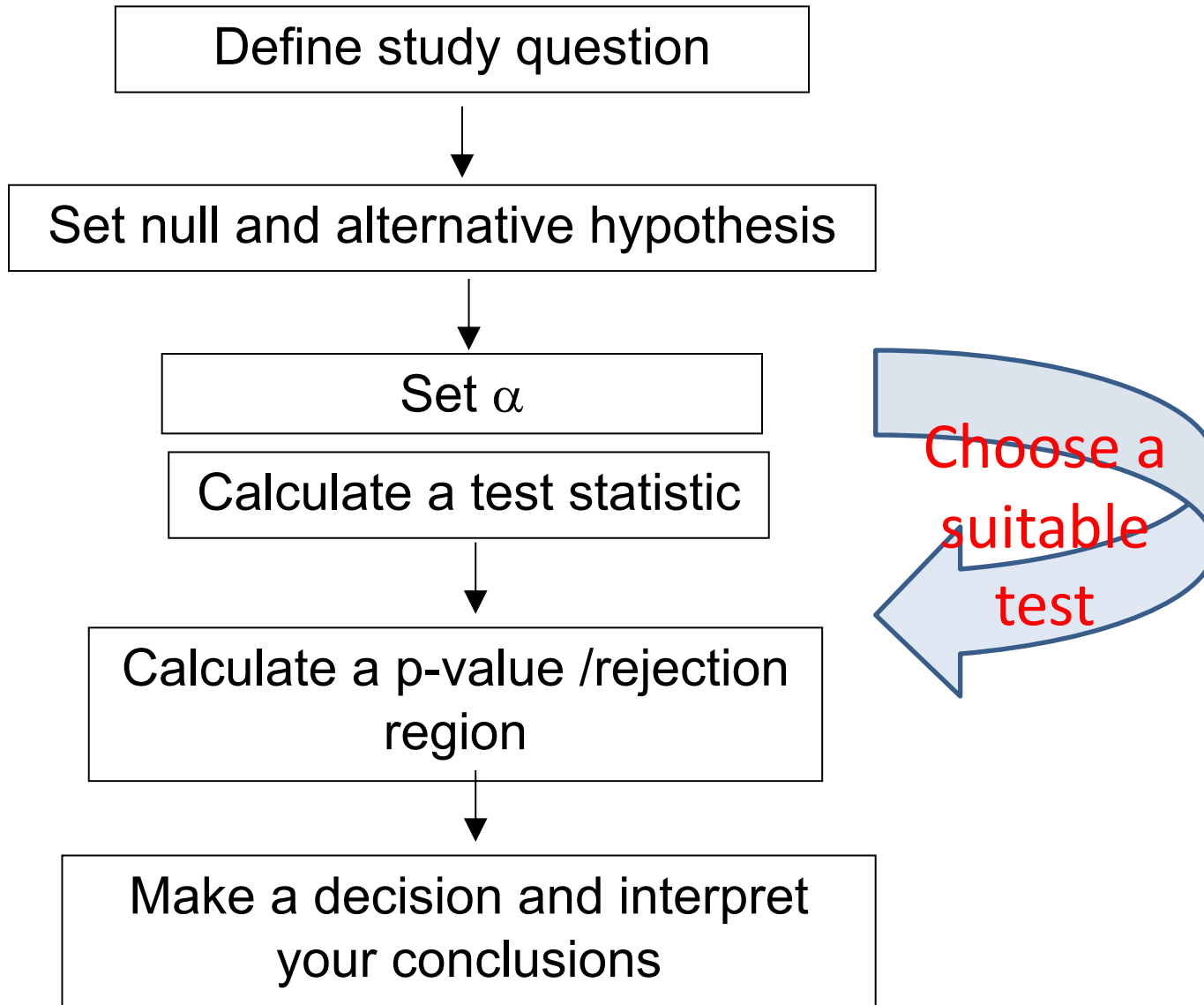


Left tailed Test in Hypothesis Testing



Right tailed Test in Hypothesis Testing

Decision Criterion: traditional method - STEPS



Decision Criterion: alternative method

P-value based method: The p-value is the probability of observing a so extreme value of the test statistics when H_0 is true.

Reject H_0 if the $P\text{-value} \leq \alpha$ (where α is the significance level, such as 0.05).

Fail to reject H_0 if the $P\text{-value} > \alpha$.

IDEA

A small p-value indicates that the realization of a given test statistic would be unlikely under the null hypothesis.

The difference of the sample statistic with respect to the population parameter is not by chance.

The lower the p-value, the more evidence there is in favour of rejecting the null hypothesis.

2) Hypothesis Test for single values

Hypothesis Test for single values

A - Test for mean difference:

- Null Hypothesis $\mathbf{H_0: \mu = \mu_0}$
- Alternative $\mathbf{H_1: \mu \neq \mu_0}$

Test Statistic when σ known

$$z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}} \right)}$$

\bar{x} : sample mean

μ : population mean

σ : population standard deviation

n : sample size

Under H_0 $\mu = \mu_0$. So, the test concludes whether the mean is equal to a given value (in the bilateral case)

Examples

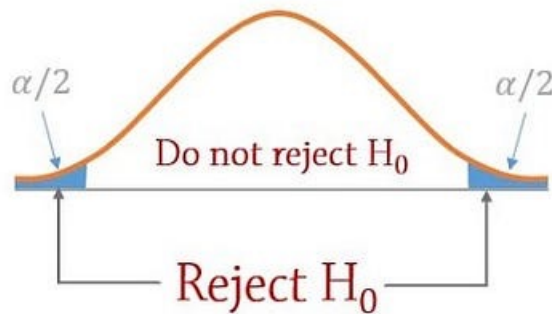
- You want to test whether the mean height in a class is 150cm
- You want to test that the mean test score is above 60/100

Hypothesis Test for single values

Two-tailed

$$H_0: \mu = 23$$

$$H_1: \mu \neq 23$$

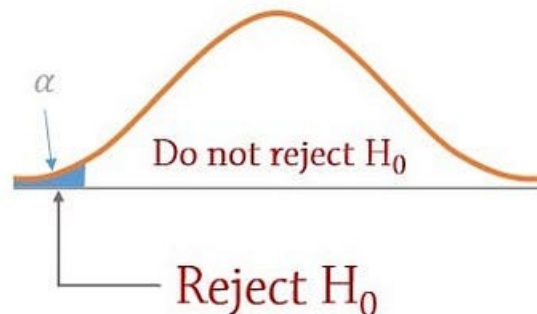


One-tailed

Left-tailed

$$H_0: \mu \geq 23$$

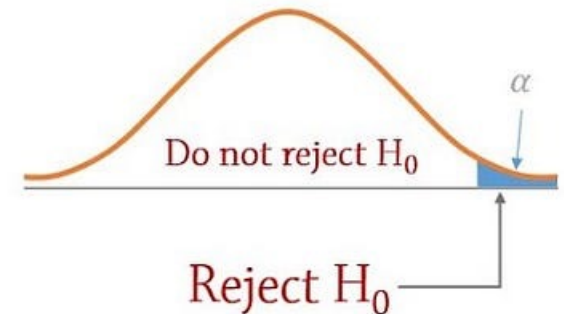
$$H_1: \mu < 23$$



Right-tailed

$$H_0: \mu \leq 23$$

$$H_1: \mu > 23$$



Hypothesis Test for single values

A - Test for mean difference:

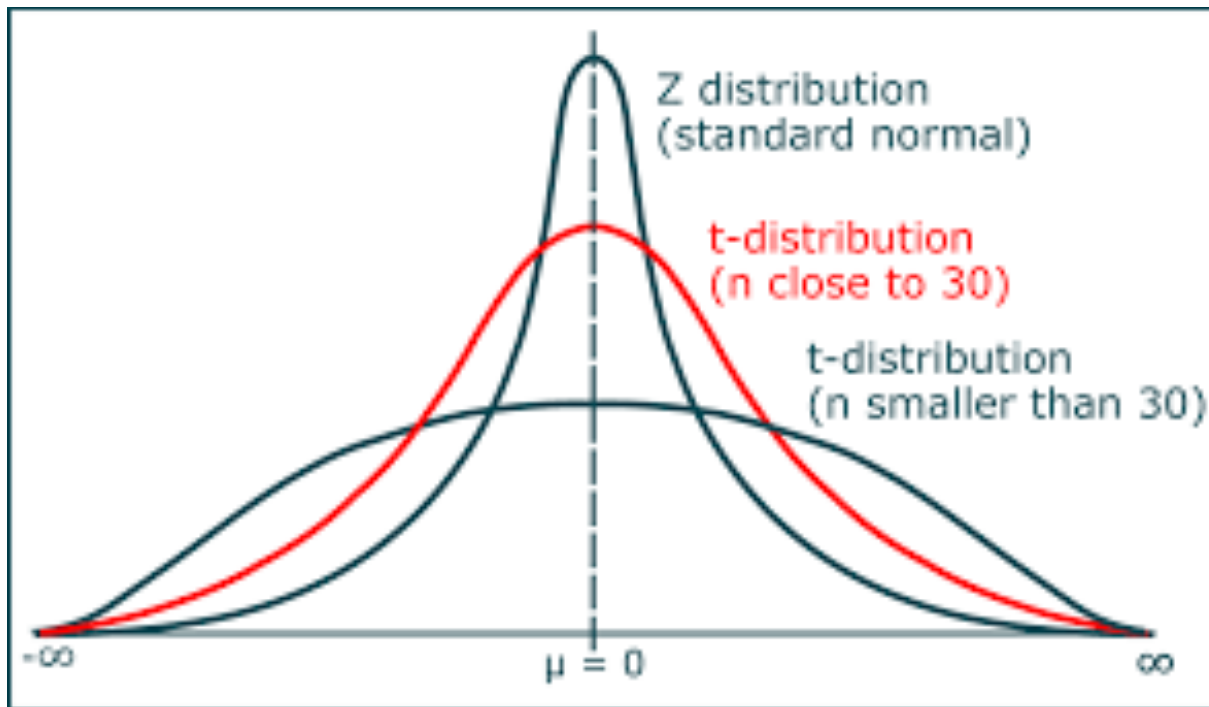
– Null Hypothesis $H_0: \mu = \mu_0$

– Alternative $H_1: \mu \neq \mu_0$

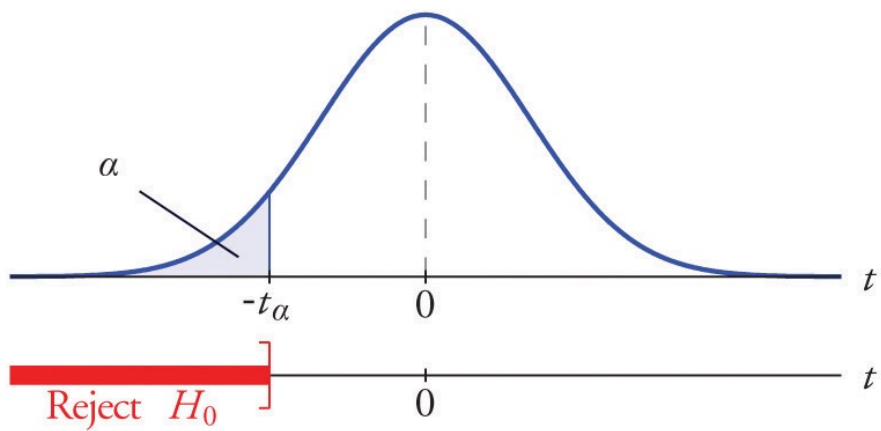
*What if variances
are unknown? →*

T-Student

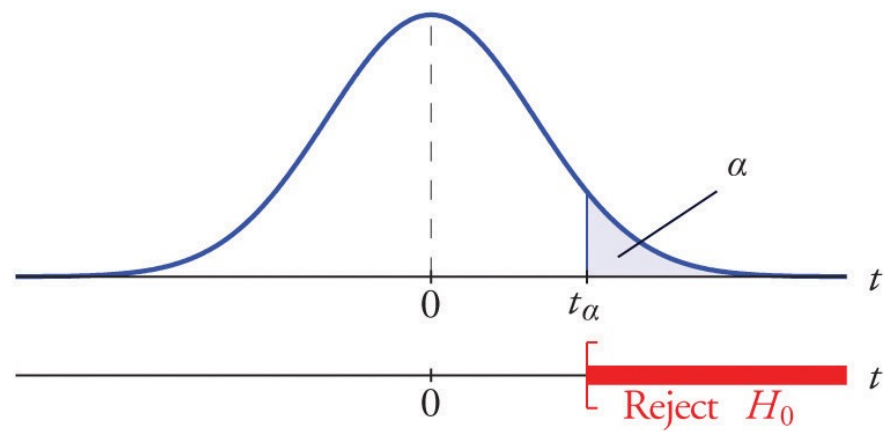
$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$



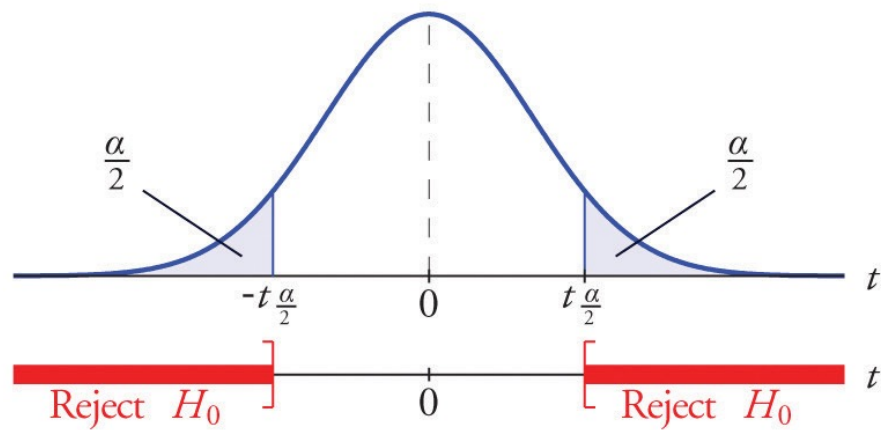
$$H_a : \mu < \mu_0$$



$$H_a : \mu > \mu_0$$



$$H_a : \mu \neq \mu_0$$



Hypothesis Test for single values

B - Test for difference in proportions

- Null Hypothesis $H_0: p = p_0$
- Alternative $H_1: p \neq p_0$

$$\text{test statistic: } z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

\hat{p} = sample proportion
 p = population proportion
 n = sample size

Under H_0 $p = p_0$. So, the test concludes whether the estimated proportion is equal to a given value (in the bilateral case)

Examples

- You want to test whether the proportion of citizens who vote for Democrats in a given district is above 0.50
- You want to test whether the proportion of citizens who regularly pay taxes is above 0.80

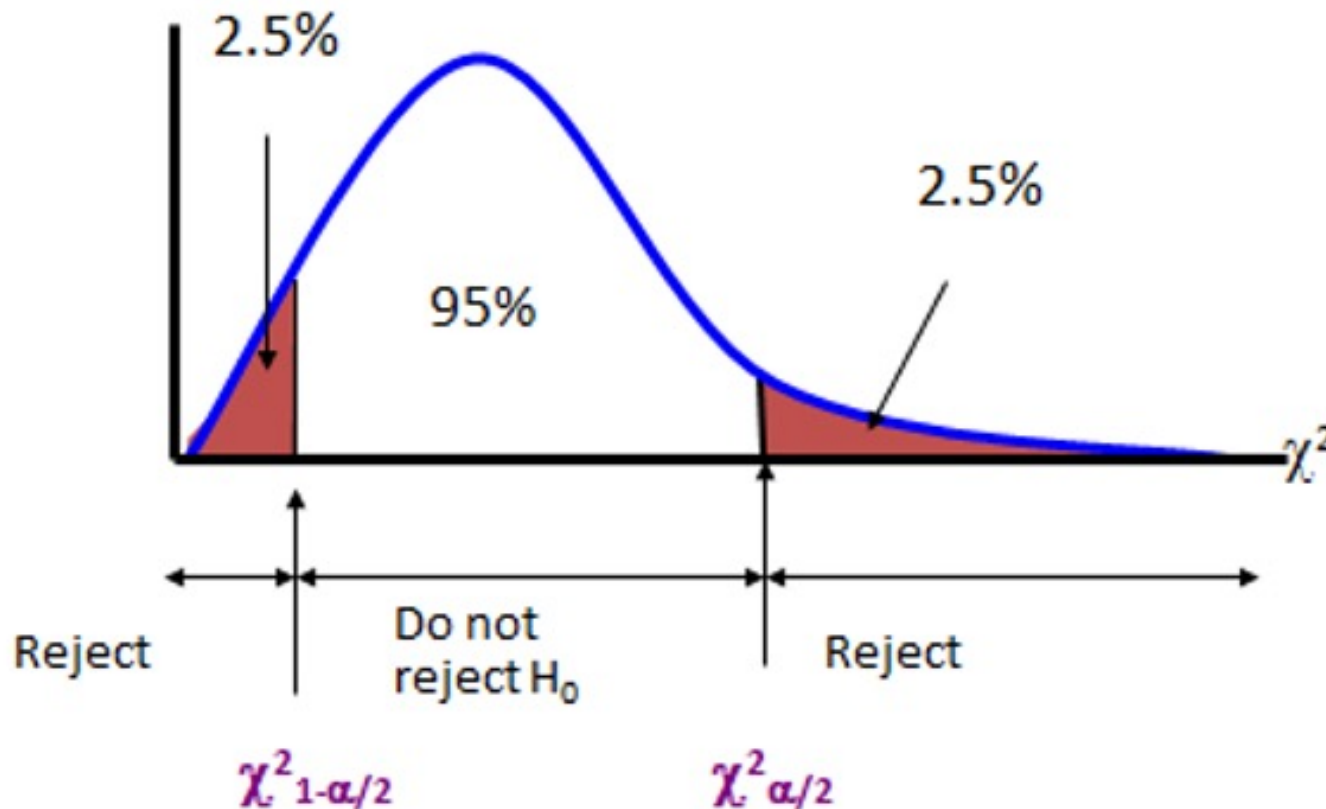
Hypothesis Test for single values

C - Test for the population variance:

– Null Hypothesis $\mathbf{H_0: \sigma = \sigma_0}$

– Alternative $\mathbf{H_1: \sigma \neq \sigma_0}$

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



3) Hypothesis Test to compare two samples

Hypothesis Test for Two Independent Samples

A - Test for mean difference:

– Null Hypothesis $\mathbf{H_0: \mu_1 = \mu_2}$

– Alternative $\mathbf{H_1: \mu_1 \neq \mu_2}$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Under H_0 $\mu_1 - \mu_2 = 0$. So, the test concludes whether there is a difference between the means or not.

Examples

- high income consumers spend more on the product than low income consumers
- Male scholars receive lower test scores than females

Hypothesis Test for Two Independent Samples

A - Test for mean difference:

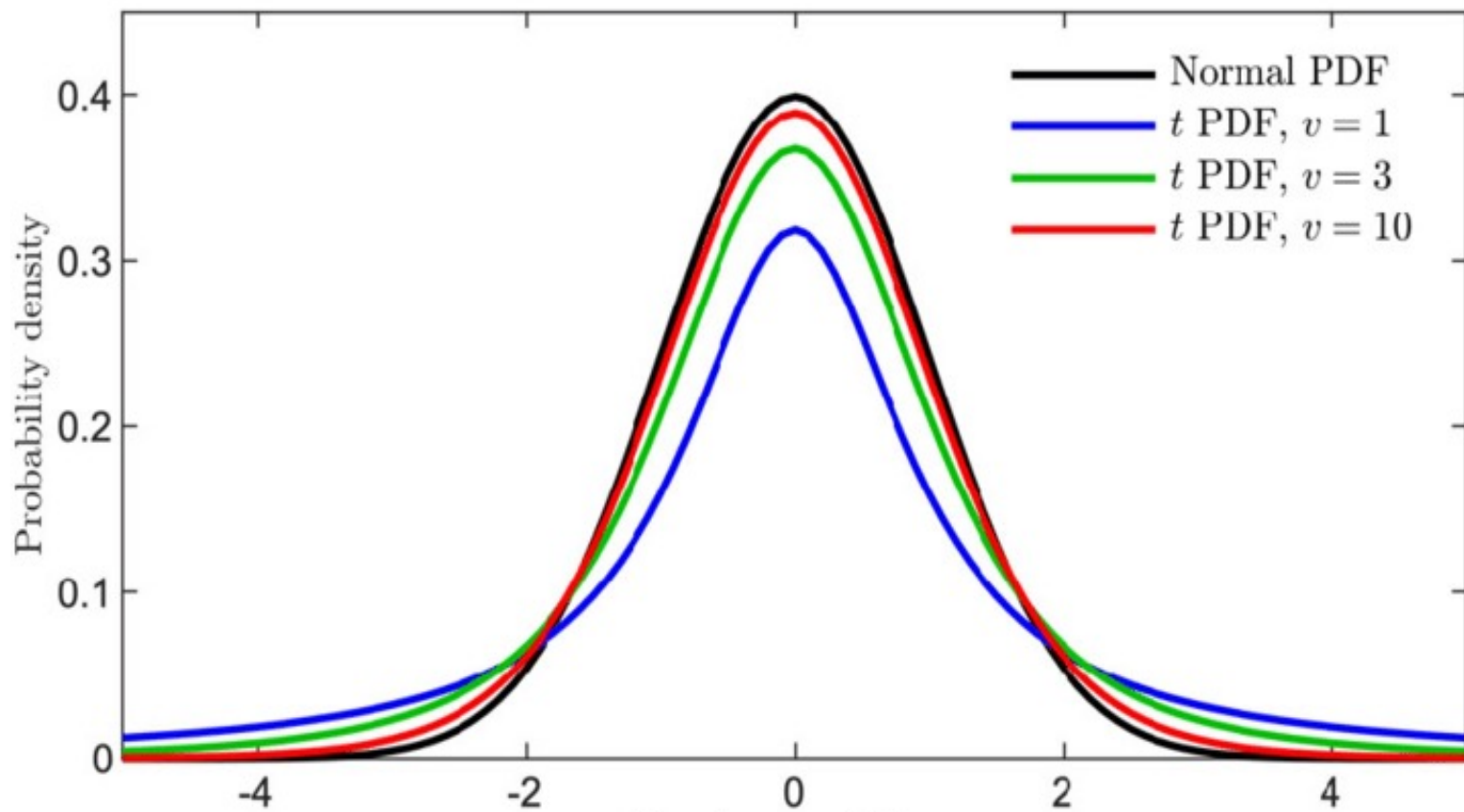
– Null Hypothesis $H_0: \mu_1 = \mu_2$

– Alternative $H_1: \mu_1 \neq \mu_2$

*What if variances
are unknown? →*

T-Student

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$



Hypothesis Test for Two Independent Samples

B - Test for difference in proportions

– Null Hypothesis $\mathbf{H_0: p_1 = p_2}$

– Alternative $\mathbf{H_1: p_1 \neq p_2}$

$$Z = \frac{\widehat{p}_1 - \widehat{p}_2 - 0}{\sqrt{p_0(1 - p_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Under H_0 $p_1 - p_2 = 0$. So, the test concludes whether there is a significant difference between the proportion of a given event in two independent samples

Examples

- The proportion of brand-loyal users in Segment 1 (eg males) is more than the proportion in segment II (e.g. females)
- The proportion of households with Internet in Canada exceeds that in USA

Hypothesis Test for Two Independent Samples

C - Test to compare variances:

– Null Hypothesis $\mathbf{H_0: \sigma_1 = \sigma_2}$

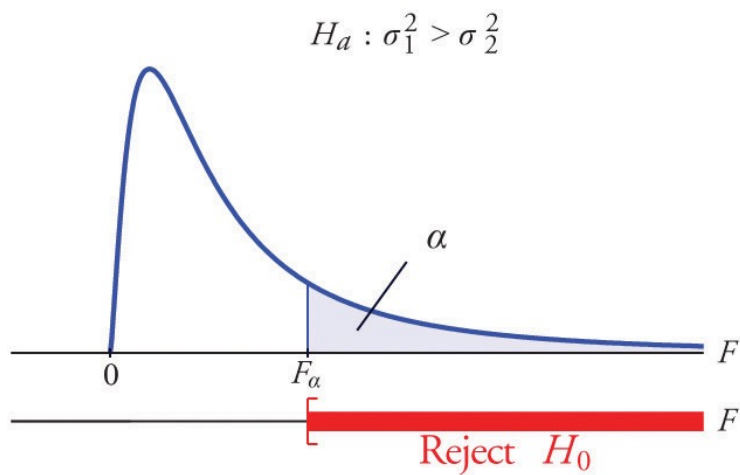
– Alternative $\mathbf{H_1: \sigma_1 \neq \sigma_2}$

$$F = \frac{s_1^2}{s_2^2}$$

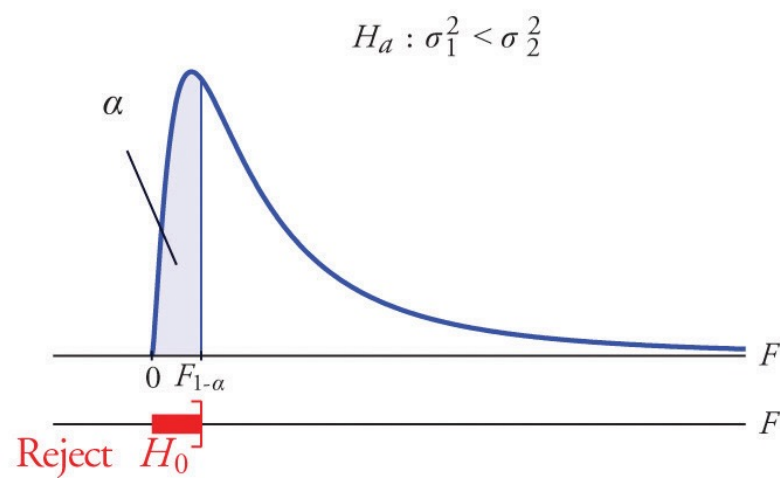


In some practical situations the difference between the population standard deviations and is also of interest.

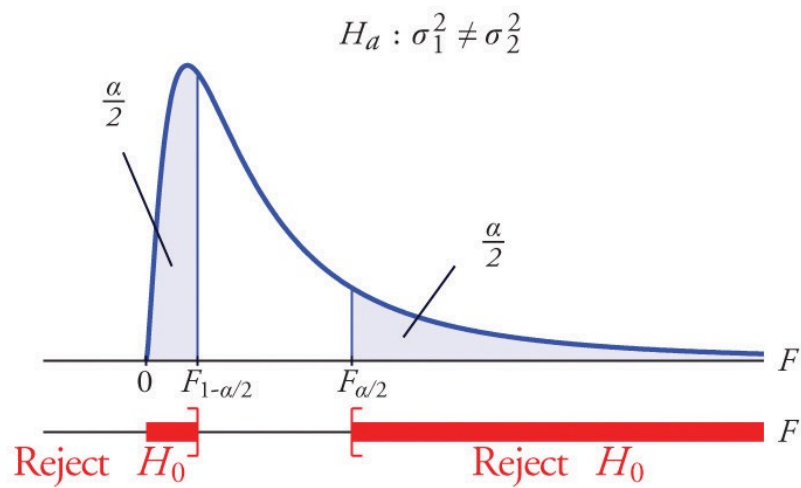
(a)



(b)



(c)



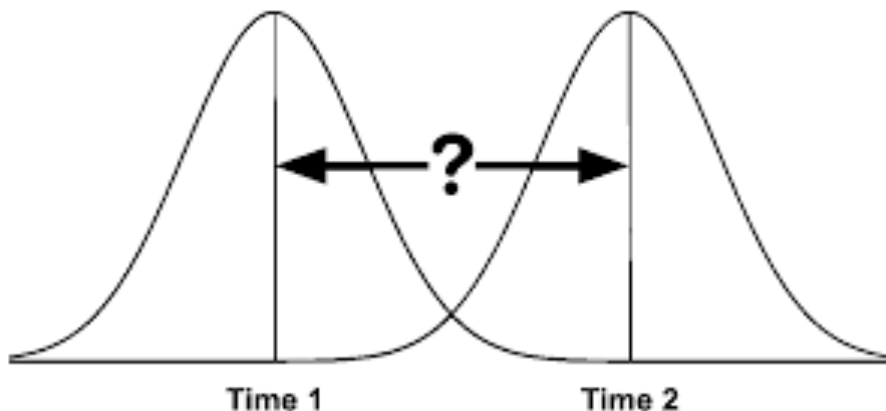
Hypothesis Test for Paired data

- Null Hypothesis $H_0: \mu_{t1} = \mu_{t2}$
- Alternative $H_1: \mu_{t1} \neq \mu_{t2}$

Under H_0 $\mu_{t1} - \mu_{t1} = 0$. So, the test concludes whether there is a difference between the mean at $t1$ and the one at $t2$

Dataset 1		Dataset 2
Observation 1	← Pair #1 →	Observation 1
Observation 2	← Pair #2 →	Observation 2
Observation 3	← Pair #3 →	Observation 3
Observation 4	...	Observation 4
Observation 5		Observation 5
Observation 6		Observation 6
Observation 7		Observation 7
Observation 8		Observation 8
Observation 9		Observation 9
...		...

Paired Samples T-Test



Hypothesis Test for Paired data

$$t = \frac{\sum d}{\sqrt{\frac{n(\sum d^2) - (\sum d)^2}{n-1}}}$$

where d: difference per paired value
n: number of samples

One sample t-test



Is there a **difference** between a **group** and the **population**

Independent samples t-test



Is there a **difference** between **two groups**

Paired samples t-test



Is there a **difference** in a **group** between **two points in time**

4) The ANOVA Test

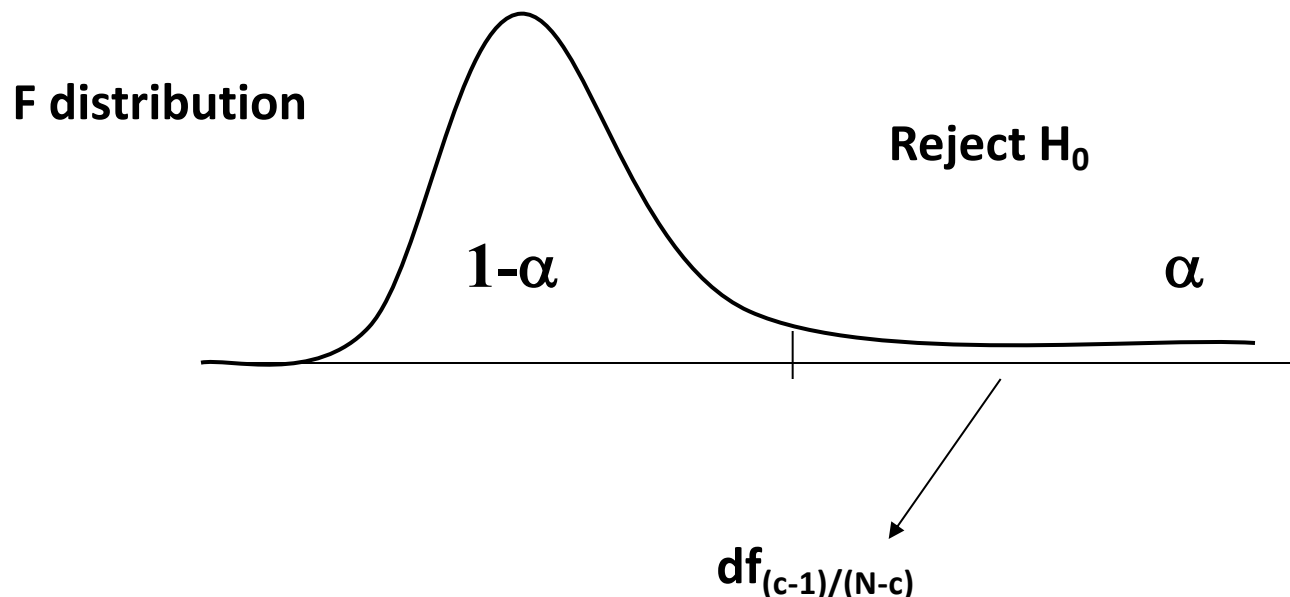
Comparing more than two means: the ANOVA test

The null hypothesis tests whether the mean of all the independent samples is equal

$$H_0 \mu_1 = \mu_2 = \mu_3 \dots = \mu_n$$

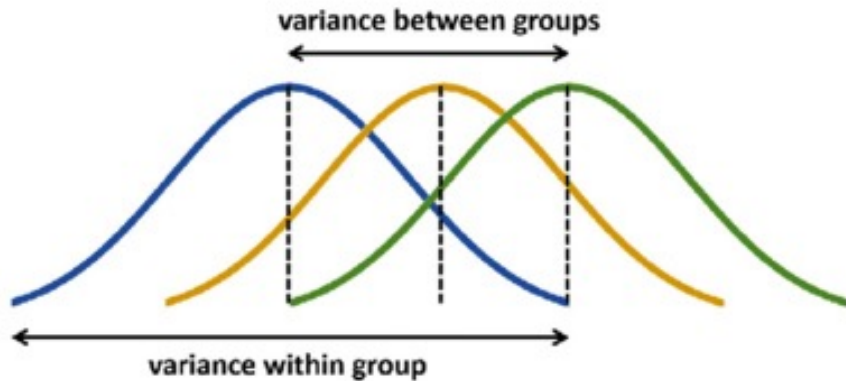
$$H_1 \mu_1 \neq \mu_2 \neq \mu_3 \dots \neq \mu_n$$

- The null hypothesis would be tested with the F distribution

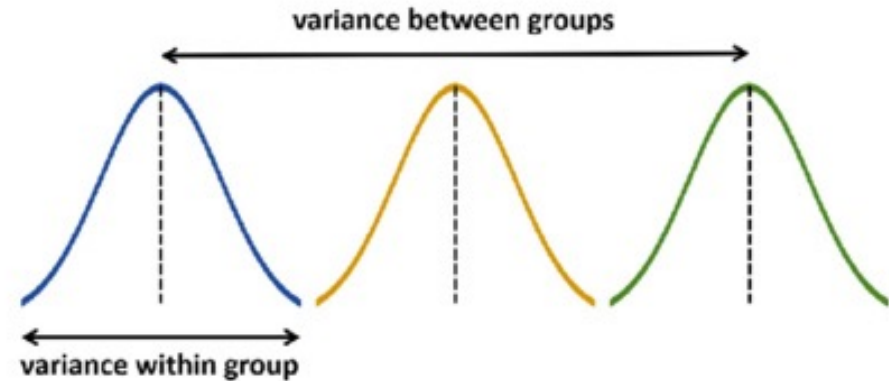


The “F-test”

A



B



Is the difference in the means of the groups more than background noise (=variability within groups)?

Summarizes the mean differences between all groups at once.

$$F = \frac{\text{Variability between groups}}{\text{Variability within groups}}$$

Analogous to pooled variance from a ttest.

5) Tests to measure the association between categorical variables

Chi squared Test

- **Null:** There is **NO** association between class and survival
- **Alternative:** There **IS** an association between class and survival

3 x 2
contingency table

Class * Survived? Crosstabulation				
Count				
		Survived?		Total
		Died	Survived	
Class	1st	123	200	323
	2nd	158	119	277
	3rd	528	181	709
Total		809	500	1309

Chi-squared test statistic

- The chi-squared test is used when we want to see if two categorical variables are related
- The test statistic for the Chi-squared test uses the sum of the squared differences between each pair of observed (O) and expected values – in case H0 is true - (E)

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

Key
<i>frequency</i>
<i>row percentage</i>
<i>column percentage</i>

Interpretation

pclass	survived		Total
	0	1	
1	123	200	323
	38.08	61.92	100.00
	15.20	40.00	24.68
2	158	119	277
	57.04	42.96	100.00
	19.53	23.80	21.16
3	528	181	709
	74.47	25.53	100.00
	65.27	36.20	54.16
Total	809	500	1,309
	61.80	38.20	100.00
	100.00	100.00	100.00

Pearson chi2(2) = 127.8592 Pr = 0.000

Since $p < 0.05$ we reject the null

There is evidence to suggest that there is an association between class and survival

value of the chi square test statistic

p- value
p < 0.005

Two-way –Contingency- tables

Which percentages between row and col are better for investigating whether class had an effect on survival?

Column

Row

pclass	survived		Total
	0	1	
1	123 38.08 15.20	200 61.92 40.00	323 100.00 24.68
2	158 57.04 19.53	119 42.96 23.80	277 100.00 21.16
3	528 74.47 65.27	181 25.53 36.20	709 100.00 54.16
Total	809 61.80 100.00	500 38.20 100.00	1,309 100.00 100.00

65.3% of
those who
died were in
3rd class

74.5% of
those in 3rd
class died

Did class affect survival? **Solution**

%’s within each class are preferable due to different class frequencies

pclass	survived		Total
	0	1	
1	123 38.08	200 61.92	323 100.00
2	158 57.04	119 42.96	277 100.00
3	528 74.47	181 25.53	709 100.00
Total	809 61.80	500 38.20	1,309 100.00

The question of interest is whether the class of an individual affected their chance of survival.

As there are different numbers in the classes, the percentages within those who died (col freq) are misleading

Did class affect survival? **Solution**

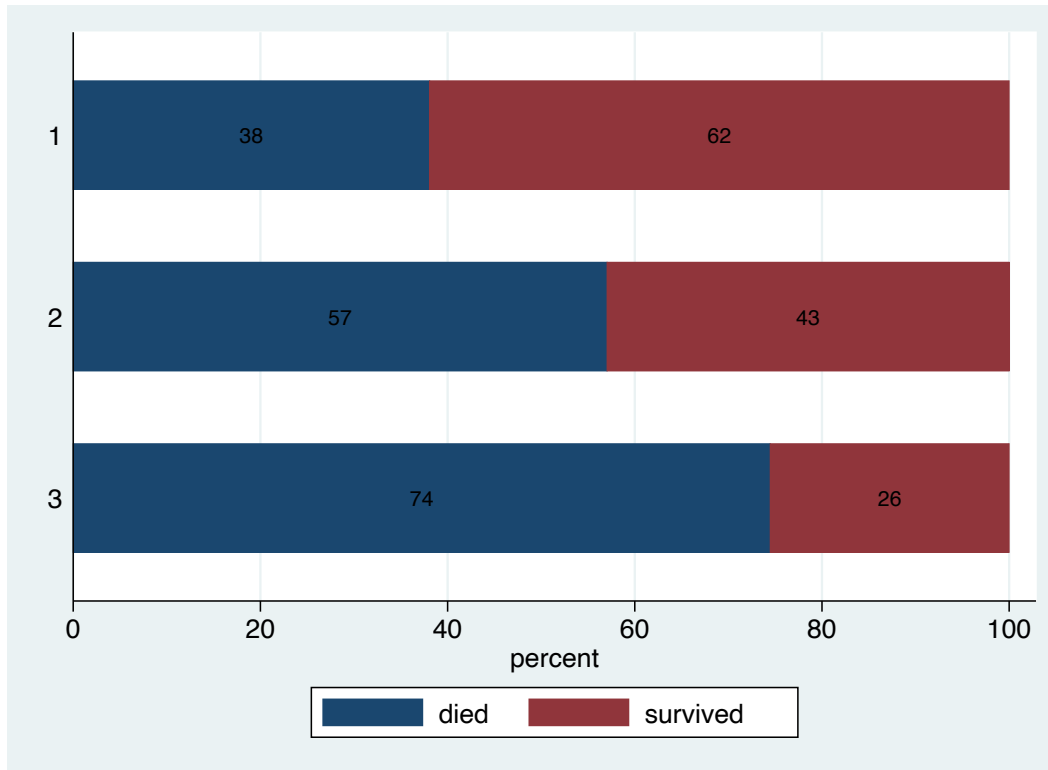


Figure 1: Bar chart showing % of passengers surviving within each class

Data collected on 1309 passengers aboard the Titanic was used to investigate whether class had an effect on chances of survival. There was evidence ($p < 0.005$) to suggest that there is an association between class and survival.

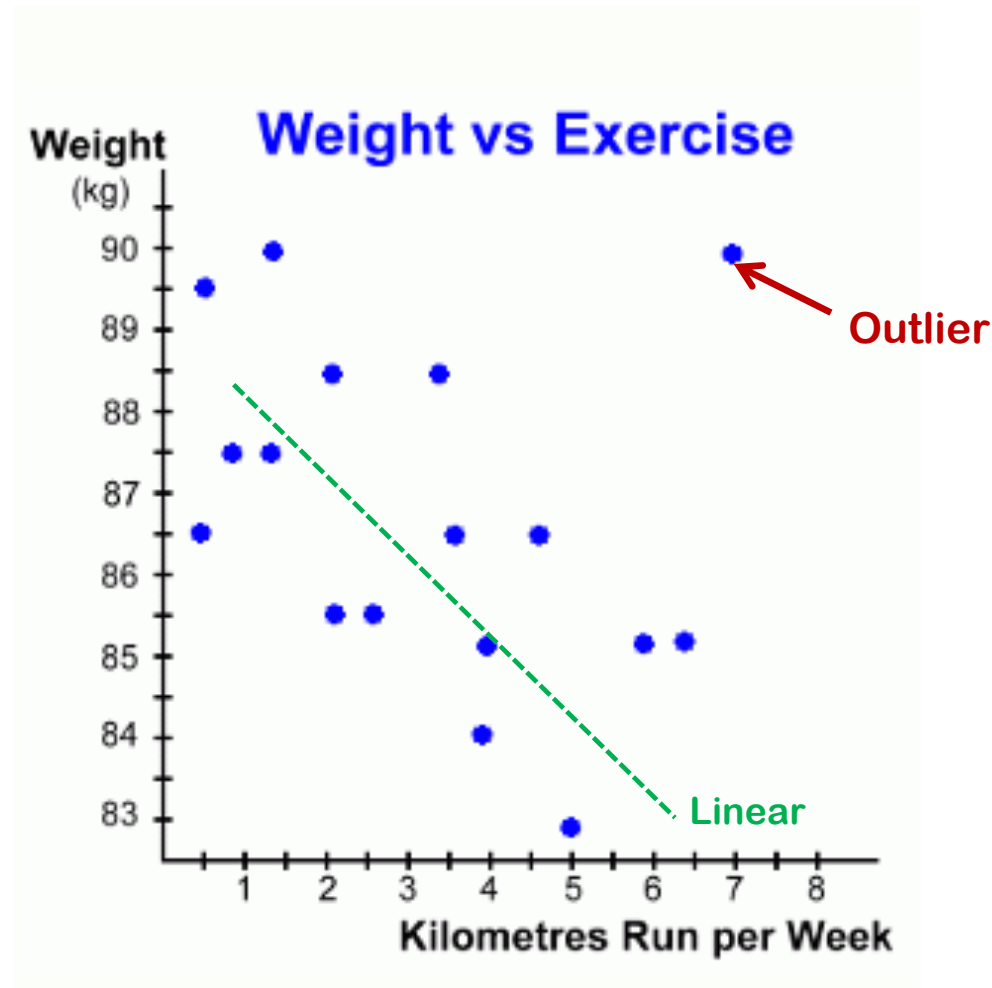
Figure 1 shows that class and chances of survival were related. As class decreases, the percentage of those surviving also decreases from 62% in 1st Class to 26% in 3rd Class.

6) Hypothesis Test for correlation

Scatterplot

Relationship between two quantitative variables:

- Explores the way the two co-vary: (correlate)
 - Positive / negative
 - Linear / non-linear
 - Strong / weak
- Presence of outliers
- Statistic used:
 r = correlation coefficient

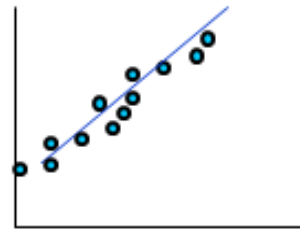


Correlation Coefficient r

- ▶ Measures strength of a relationship between two continuous variables

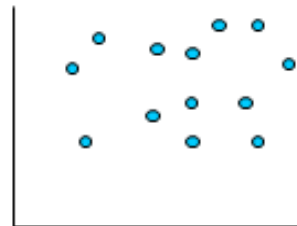
$$-1 \leq r \leq 1$$

Strong positive linear relationship



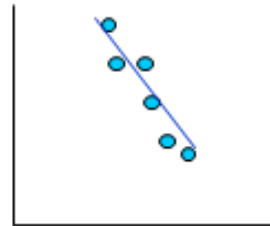
r close to 1

No linear relationship



r close to zero

Strong negative linear relationship



r close to -1

Correlation Interpretation

Correlation quantifies this relationship. Correlation coefficients range from -1 and +1 with 0 meaning there is no relationship at all.

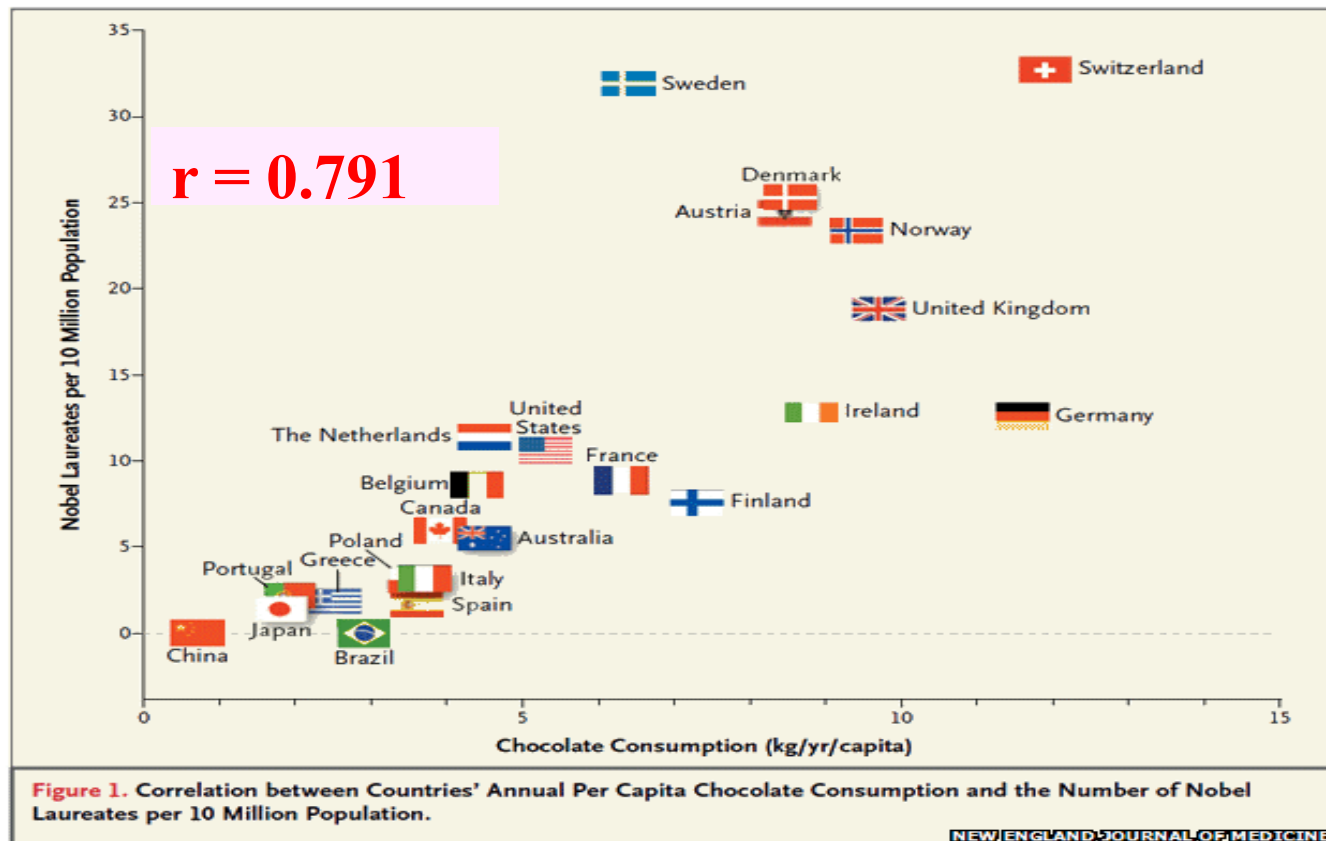
The further away from 0 the coefficient is, the stronger the relationship. A positive number means that as x increases, so does y and negative coefficients that y decreases as x increases.

Correlation coefficient value			Relationship
-0.3 to +0.3			Weak
-0.5 to -0.3	or	0.3 to 0.5	Moderate
-0.9 to -0.5	or	0.5 to 0.9	Strong
-1.0 to -0.9	or	0.9 to 1.0	Very strong

Cohen, L. (1992). Power Primer. Psychological Bulletin, 112(1) 155-159

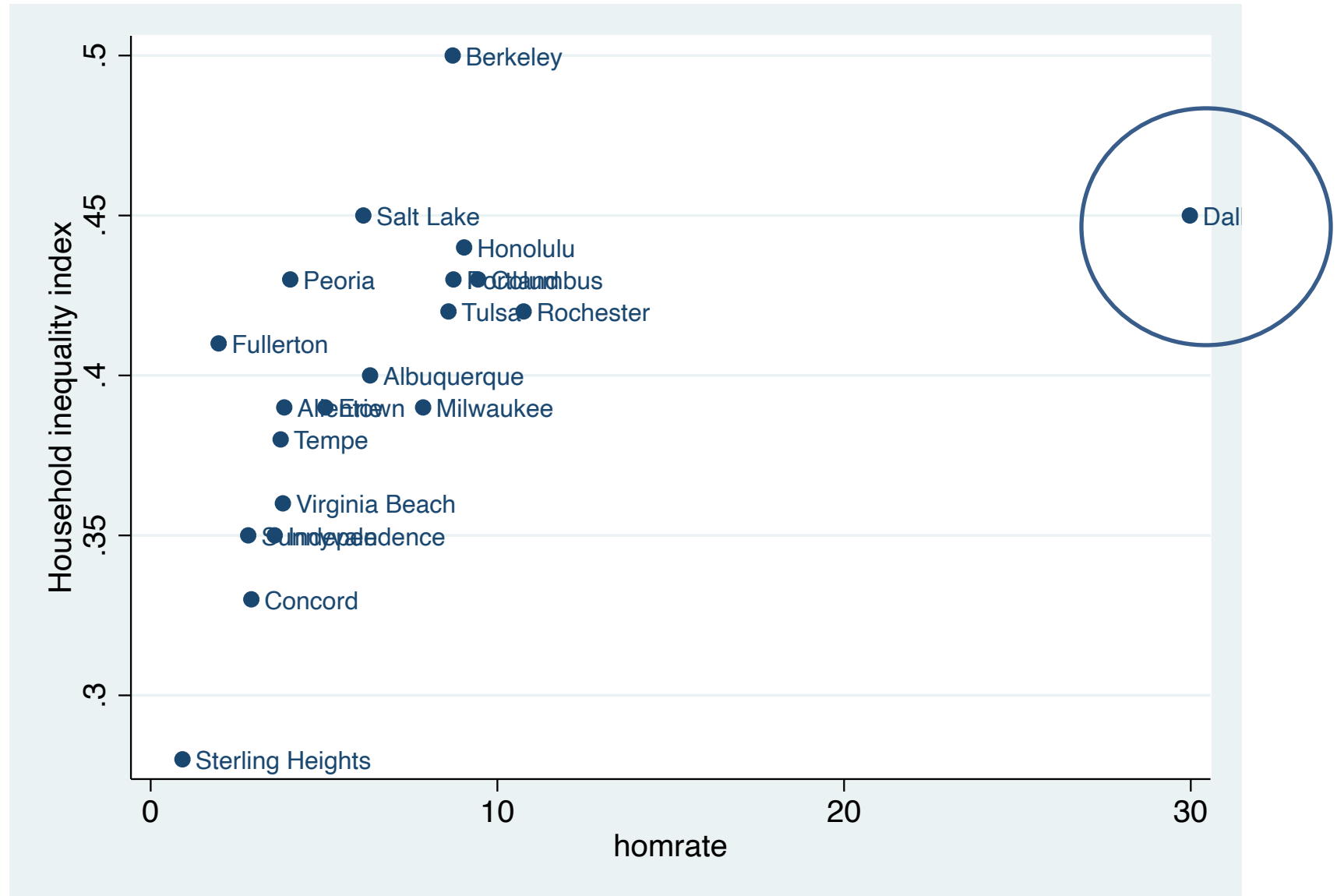
Does chocolate make you clever or crazy?

- ▶ A paper in the New England Journal of Medicine claimed a relationship between chocolate and Nobel Prize winners



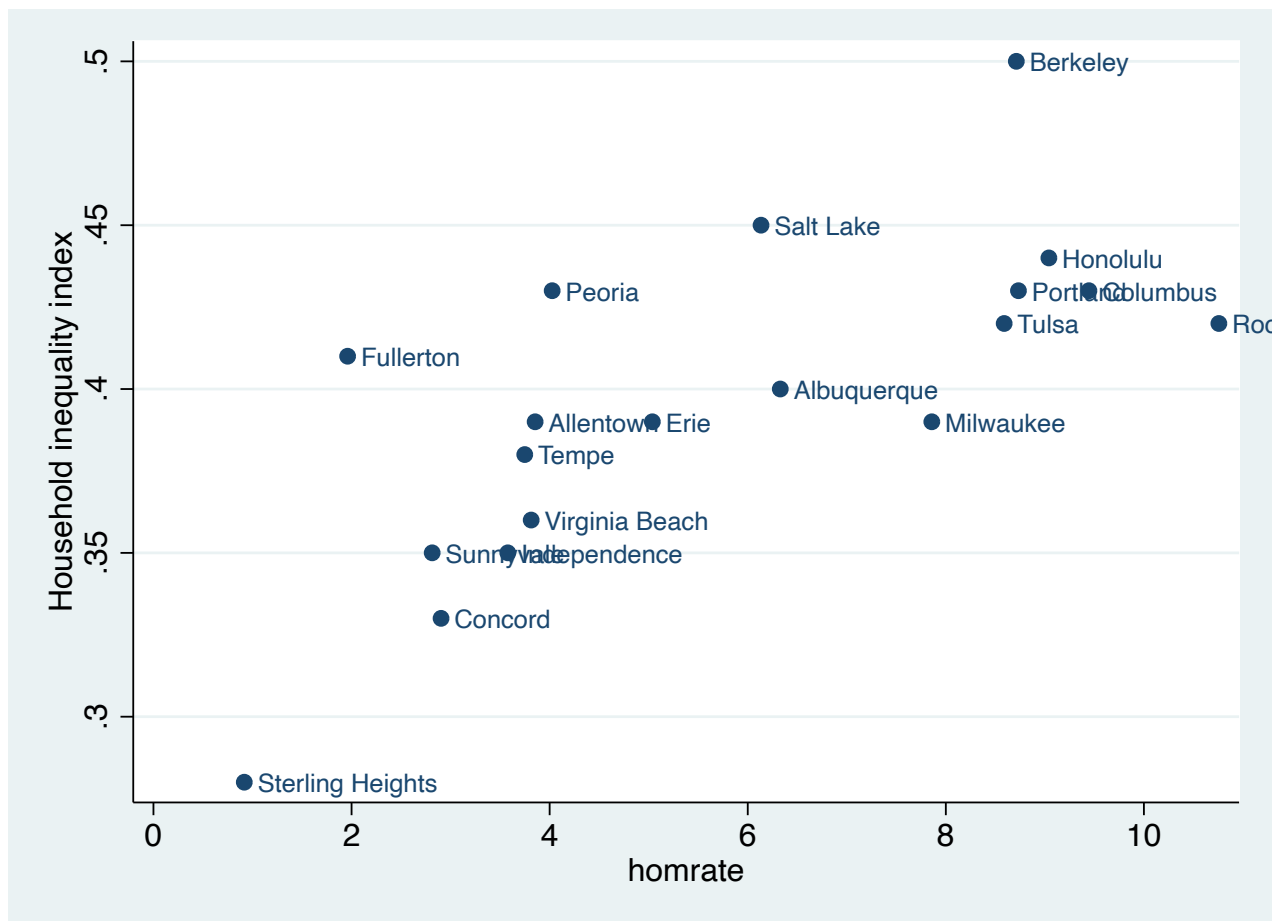
<http://www.nejm.org/doi/full/10.1056/NEJMon1211064>

Is higher inequality associated with higher homicide rate?



Is higher inequality associated with higher homicide rate?

Note: same graph but without Dallas



Hypothesis tests for r

Tests the null hypothesis that the population correlation:

$$H_0 r = 0$$

$$H_1 r \neq 0$$

`pwcorr inequal homrate if city!="Dallas", sig`

	inequal	homrate
inequal	1.0000	
homrate	0.7163 0.0006	1.0000

$r=0.71$

$p\text{-value}<0.005$

A significant result just means that there is evidence to suggest that r is not 0

Exercise - Solution

$$r=0.71$$

There is a significant and strong positive relationship between inequalities and homicides and the relationship looks like a linear relationship (it can be approximated by a line)

