Applied Statistics

Feb 9-11 2021

Chiara Seghieri,
Laboratorio MeS
Istituto di Management
Scuola Superiore Sant'Anna
c.seghieri@santannapisa.it

Two Types of Inference

1. <u>**Confidence Intervals:**</u>
   - **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. <u>**Hypothesis Testing:**</u>
   - **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 is a statement regarding the range of values of the unknown population parameter of interest. As, for example the mean μ: "Is the true mean different from 100?

•**Hypothesis testing:** an objective method of making decisions or **inferences** from 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 other 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

- The null hypothesis is a 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.
- Either reject H_0 or fail to reject H_0.

Always two hypotheses:

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.

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.

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: $\neq$, $<$, $>$.

Example:

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

$$H_0 : \mu = 50$$

Alternative hypothesis - Mean age is different from 50

$$H_A : \mu \neq 50$$ **two-tailed test**

or $$H_A : \mu < 50$$ **lower-tailed test**

$$H_A : \mu > 50$$ **upper-tailed test**

Illustrative Example: "Body Weight"

- **The problem:** In the 1970s, 20–29 year 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** $H_{0:}$ μ = 170 ("no difference")
- The **alternative hypothesis** can be either $H_{a:}$ μ > 170 (**one-sided test**) body weight in this group has increased since 1970 or
  $H_{a:}$ μ ≠ 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 samples.

Different types of problems require different test statistics.
If the test statistic shows you "observed an unlikely result", you reject the null hypothesis and accept the alternative hypothesis.

Test Statistic

The test statistic is a value used in making a decision about the null hypothesis, and it is found by converting the sample statistic to a score with the assumption that the null hypothesis is true.

It tells us, if H_0 is true, how likely it is that we would obtain the given sample result.

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

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.

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

Types of Errors

No matter which hypothesis represents the claim, always begin the hypothesis test **assuming that the null hypothesis is true**.

At the end of the test, one of two decisions will be made:

1. reject the null hypothesis, or
2. fail to reject the null hypothesis.

There are **two** possible decisions that can be made:

A. Conclude that there ***is enough evidence*** to support the alternative hypothesis (rejecting the null hypothesis in favor of the alternative)

B. Conclude that there ***is not enough evidence*** to support the alternative hypothesis (**not** rejecting the null hypothesis in favor of the alternative)

We reject H_0 only when the chance is small that H_0 is true. Since our decisions are based on probability rather than certainty, we can make errors.

Hypothesis Testing: type of errors

True State	Fail to reject H_0	reject H_0
H_0 True	Correct Decision	Type I Error
H_0 False	Type II Error	Correct Decision

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

Goal: Keep α, β reasonably small

Type I and type II errors are inversely related. Decreasing one increases the other.

They are different sorts of mistakes and have different consequences. For instance 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.

By convention it is given more importance to **P(Type I error) = α which is usually fixed= 0.05, 0.01 or 0.1**

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.

The probability of making a type II error is denoted by β.

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:

$\alpha = 0.10$ $\alpha = 0.05$ $\alpha = 0.01$

Level of Significance, α

- **Defines the unlikely values of the sample statistic if the null hypothesis is true**
  - Defines rejection region of the sampling distribution
  - Typical values are 0.01, 0.05, or 0.10
- Is selected by the researcher at the beginning

You should be aware that Type II error is also important. A small sample size, for example, might lead to frequent Type II errors, i.e. it could be that your (alternative) hypotheses are right, but because your sample is so small, you fail to reject the null even though you should.

Steps to undertaking a Hypothesis test

Define study question

↓

Set null and alternative hypothesis

↓

Set α

Calculate a test statistic

↓

Calculate a p-value /rejection region

↓

Make a decision and interpret your conclusions

Choose a suitable test

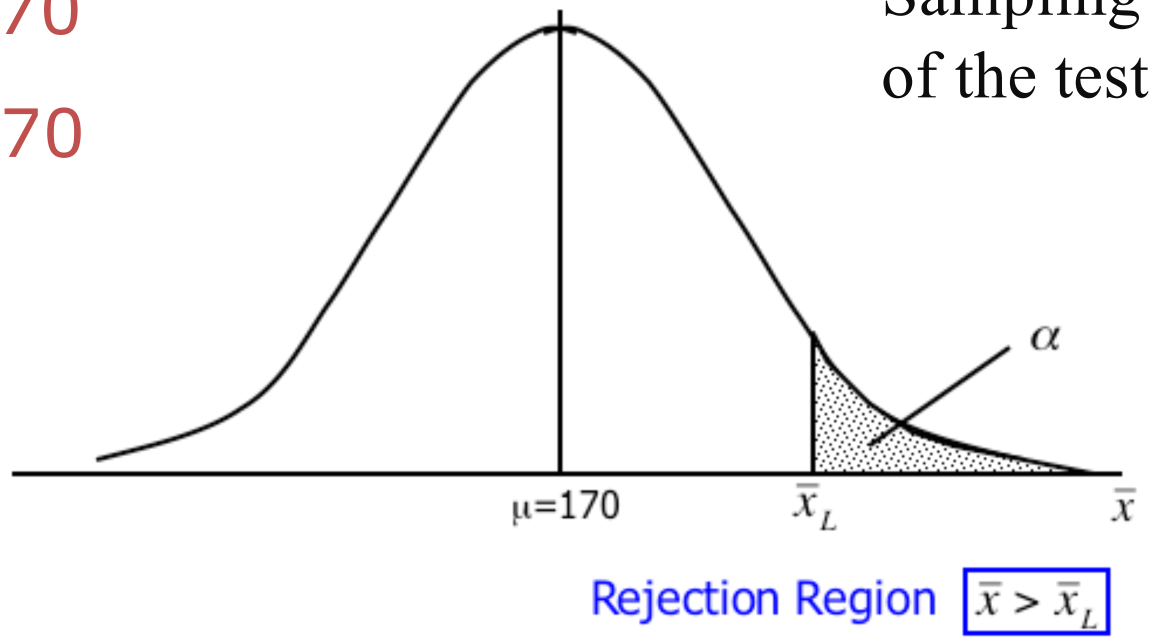
The Rejection Region

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.**

H0: $\mu=170$

HA: $\mu>170$

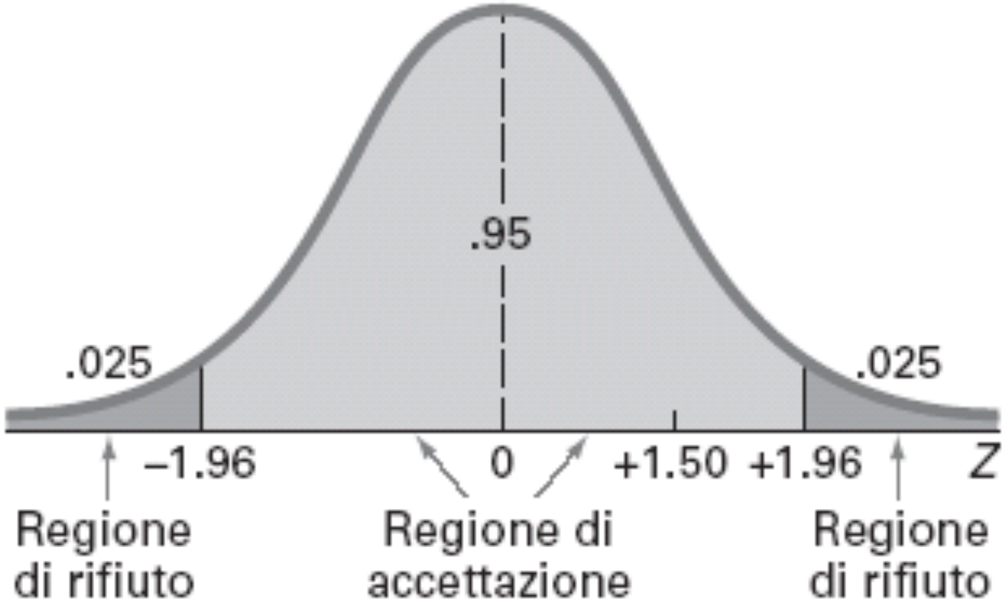
Sampling distribution of the test statistic z



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

Rejection Region (dependent on H_A)

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



.95
.025
.025
–1.96
0
+1.50
+1.96
Z
Regione di rifiuto
Regione di accettazione
Regione di rifiuto

Decision Criterion - cont

P-value method:

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

Fail to reject H_0 if the *P*-value $> \alpha$.

The p-value approach:

The p-value is the probability that the data could deviate from the null hypothesis as much as they did or more. It si not the probability that the null hypothesis is correct.

A small p-value indicates that observation of the test statistic would be unlikely if the null hypothesis is true→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.

Hypothesis Test for Two Independent Samples

Test for mean difference:

- Null Hypothesis H_0: $\mu_1 = \mu_2$
- Alternative H_1: $\mu_1 \neq \mu_2$

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

Some examples

examples

- populations of users and non-users of a brand differ in perceptions of the brand
- high income consumers spend more on the product than low income consumers
- 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

Can be used for examining differences between means and proportions

```
. ttest incomeppp, by(sex)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       1 |  14,339    30459.53    228.5309    27365.57    30011.58    30907.48
       2 |  15,473     15746.7    148.0887    18420.83    15456.43    16036.97
---------+--------------------------------------------------------------------
combined |  29,812    22823.29    140.7191    24296.77    22547.47    23099.11
---------+--------------------------------------------------------------------
    diff |            14712.83    268.4461                14186.66    15238.99
------------------------------------------------------------------------------
    diff = mean(1) - mean(2)                                      t =  54.8074
Ho: diff = 0                                     degrees of freedom =    29810

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Comparing more than two means: the ANOVA test

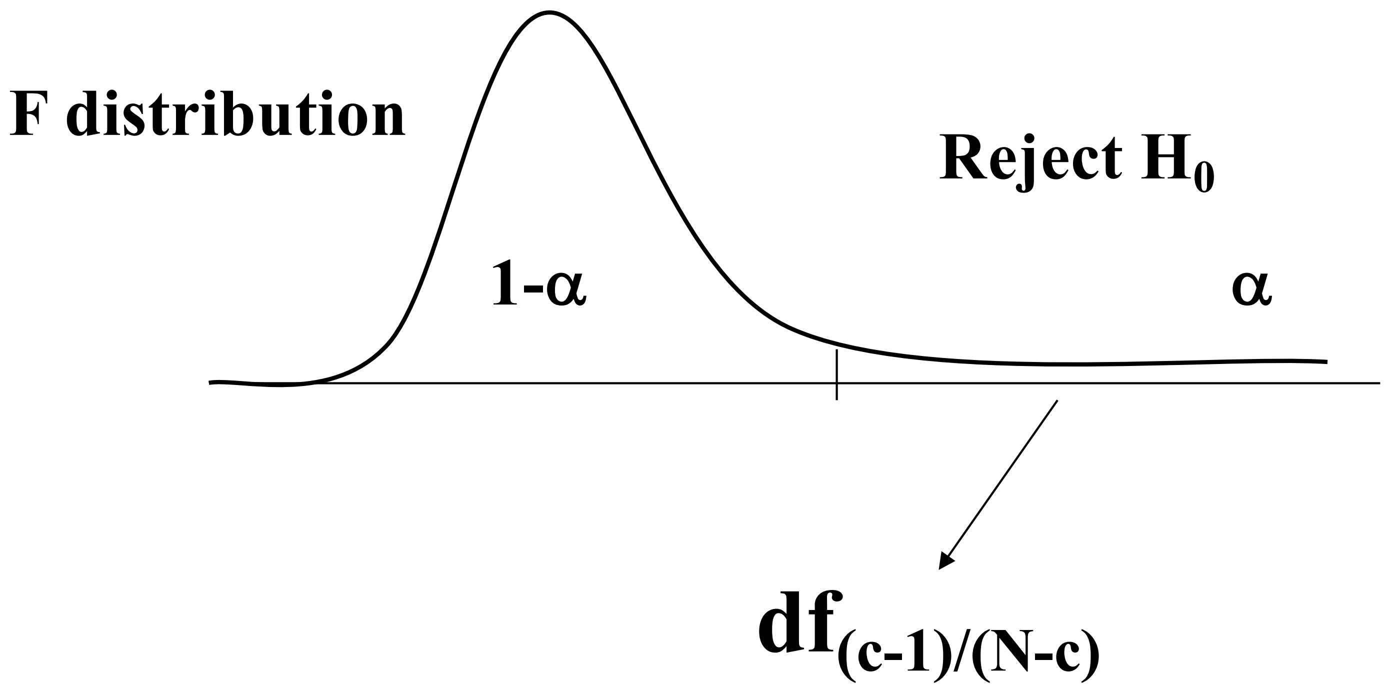
The null hypothesis tests whether the mean of all the independent samples is equal

$$H_0 \;\; \mu_1 = \mu_2 = \mu_3 \; ..... = \mu_n$$

$$H_1 \;\; \mu_1 \neq \mu_2 \neq \mu_3 \; ..... \neq \mu_n$$

ANOVA Test

- **The null hypothesis would be tested with the F distribution**



The "F-test"

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.

oneway incomeppp country

Analysis of Variance

Source	SS	df	MS	F	Prob >
Between groups	**1.9988e+12**	**3**	**6.6625e+11**	**1273.08**	**0.00**
Within groups	**1.5600e+13**	**29808**	**523338228**		
Total	**1.7598e+13**	**29811**	**590333165**		

Example: Titanic

Research question: Did class (of travel) affect survival?

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)

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

tab pclass survived, row col chi2

Key
frequency
row percentage
column percentage

pclass	survived 0	1	Total
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**

value of the chi square test statistic

p- value

$p < 0.005$

Interpretation

Since $p < 0.05$ we reject the null

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

But… what is the nature of this association/relationship?

I expect that ‘wealthy’ people more likely to survive on board the Titanic

- Choose the right percentages from the two-way table to investigate

Two-way –Contingency- tables

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

pclass	survived 0	1	Total
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 0	1	Total
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

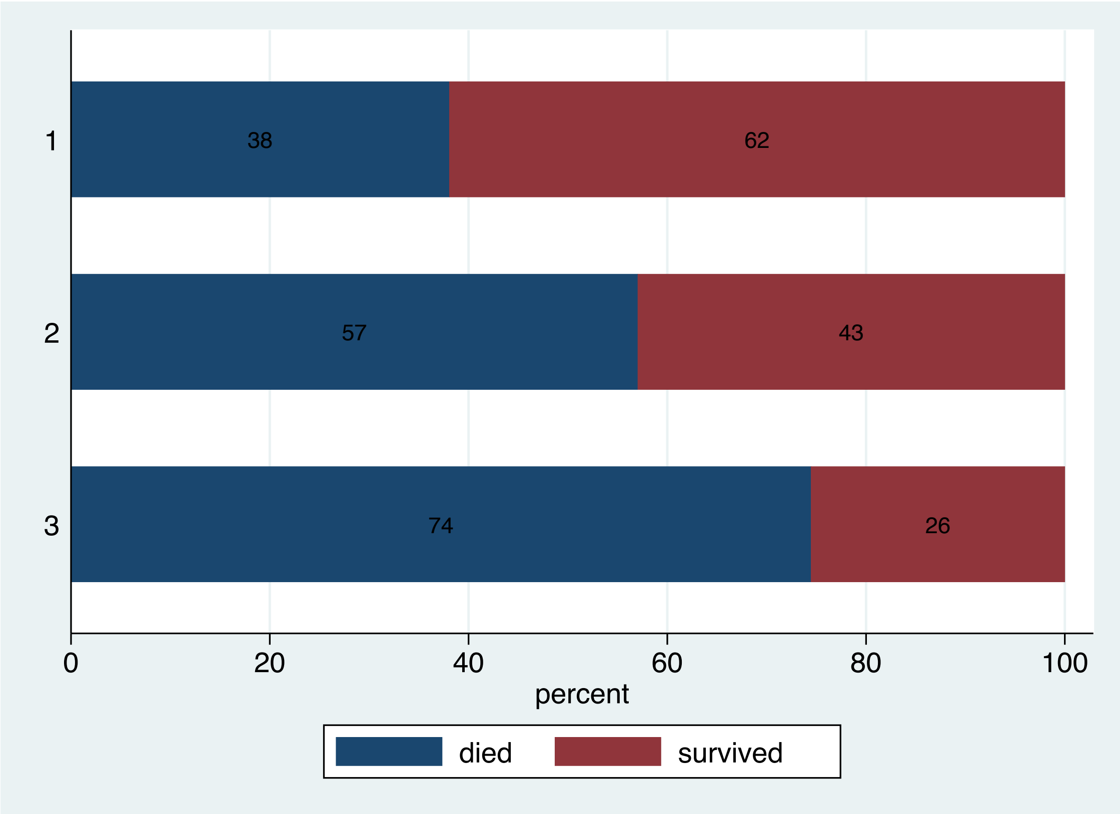


Figure 1: Bar chart showing % of passengers surviving within each class

catplot survived pclass, percent(pclass) asyvars stack ///
blabel(bar, pos(center) format(%2.0f))
legend(pos(bottom) col(5))

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.

Scatterplot

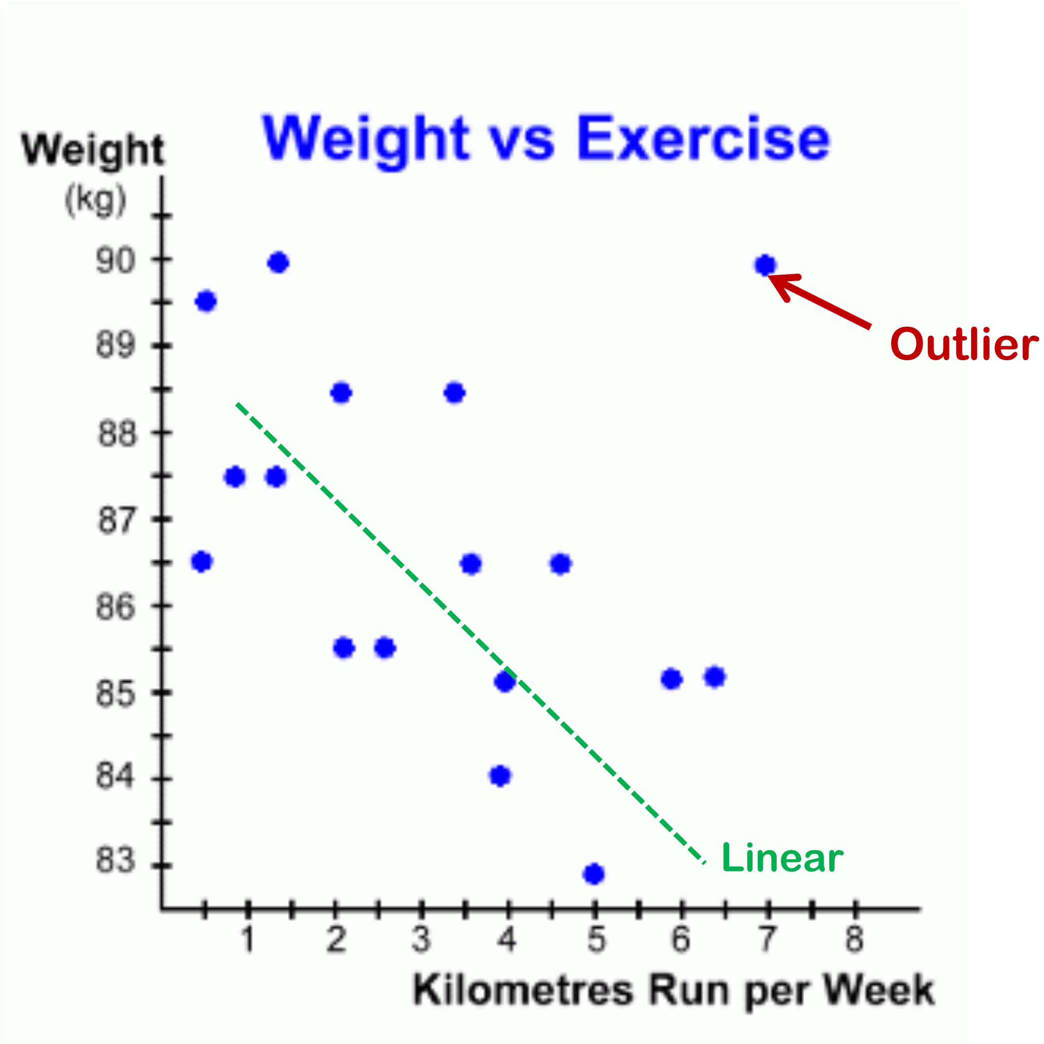
Relationship between two quantitative variables:

- ➢ Explores the way the two co-vary: (correlate)
  - Positive / negative
  - Linear / non-linear
  - Strong / weak

- ➢ Presence of outliers

- ➢ Statistic used:

r = correlation coefficient

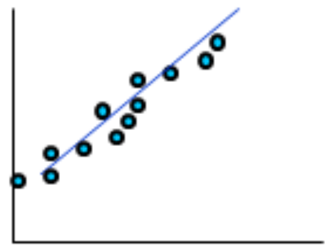


Correlation Coefficient r

- **Measures strength of a relationship between two continuous variables**

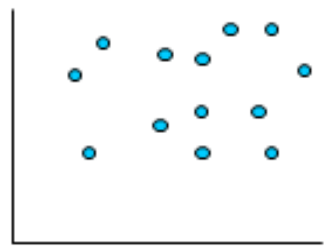
$-1 \leq r \leq 1$

Strong positive linear relationship



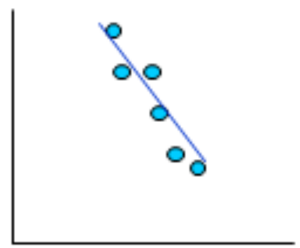
r close to 1

No linear relationship



r close to zero

Strong negative linear relationship



r close to -1

Correlation 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 Interpretation

An interpretation of the size of the coefficient has been described by Cohen (1992) as:

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

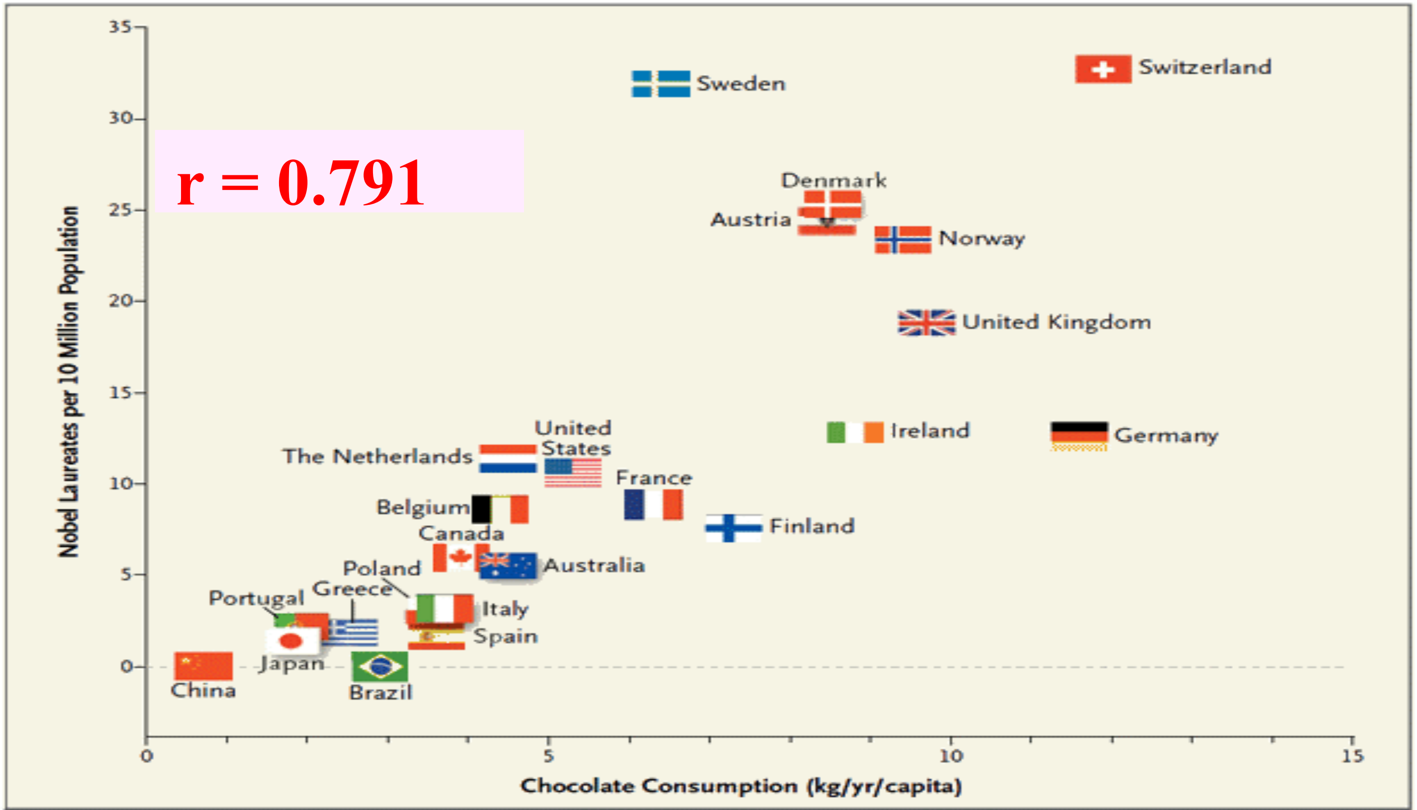
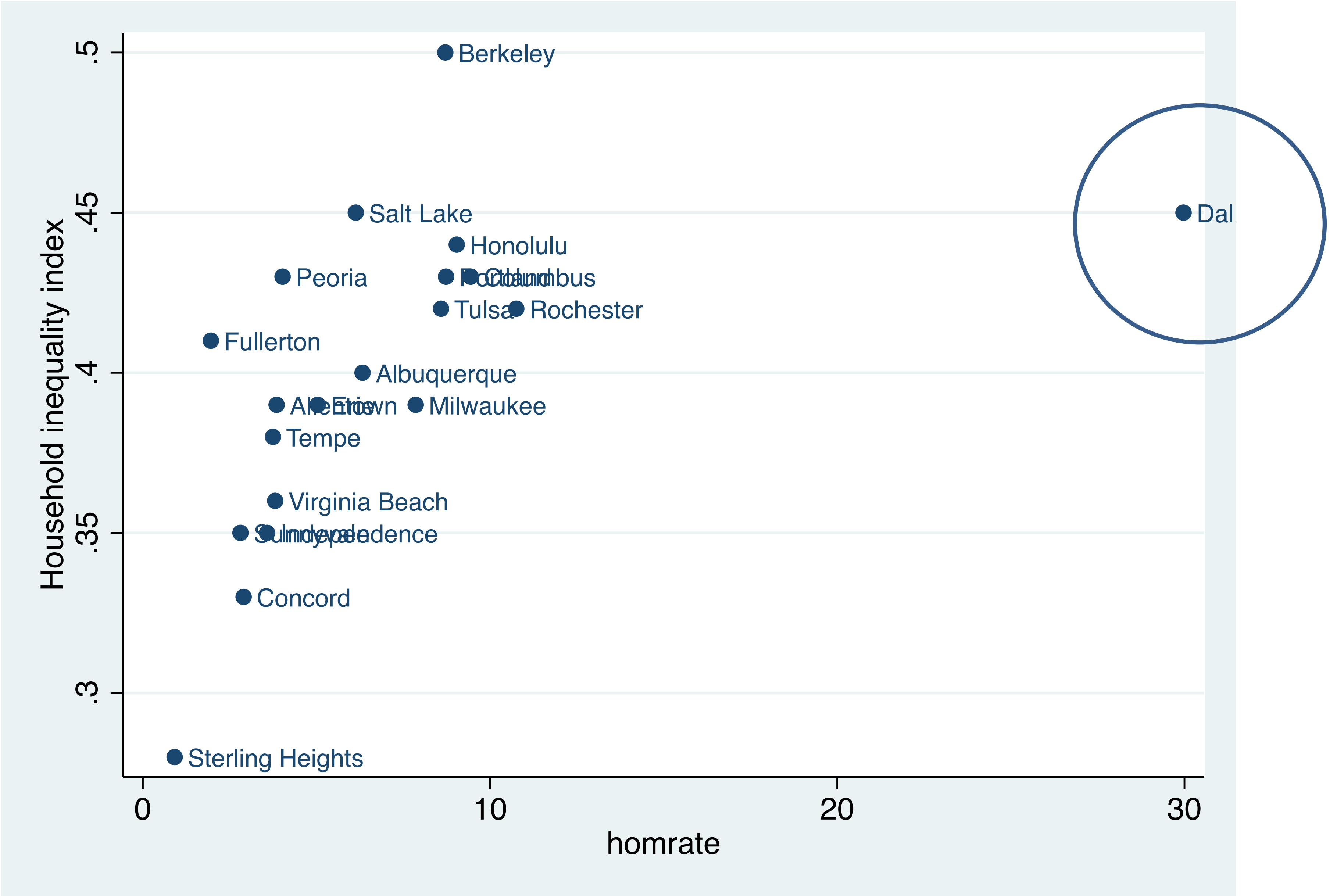


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

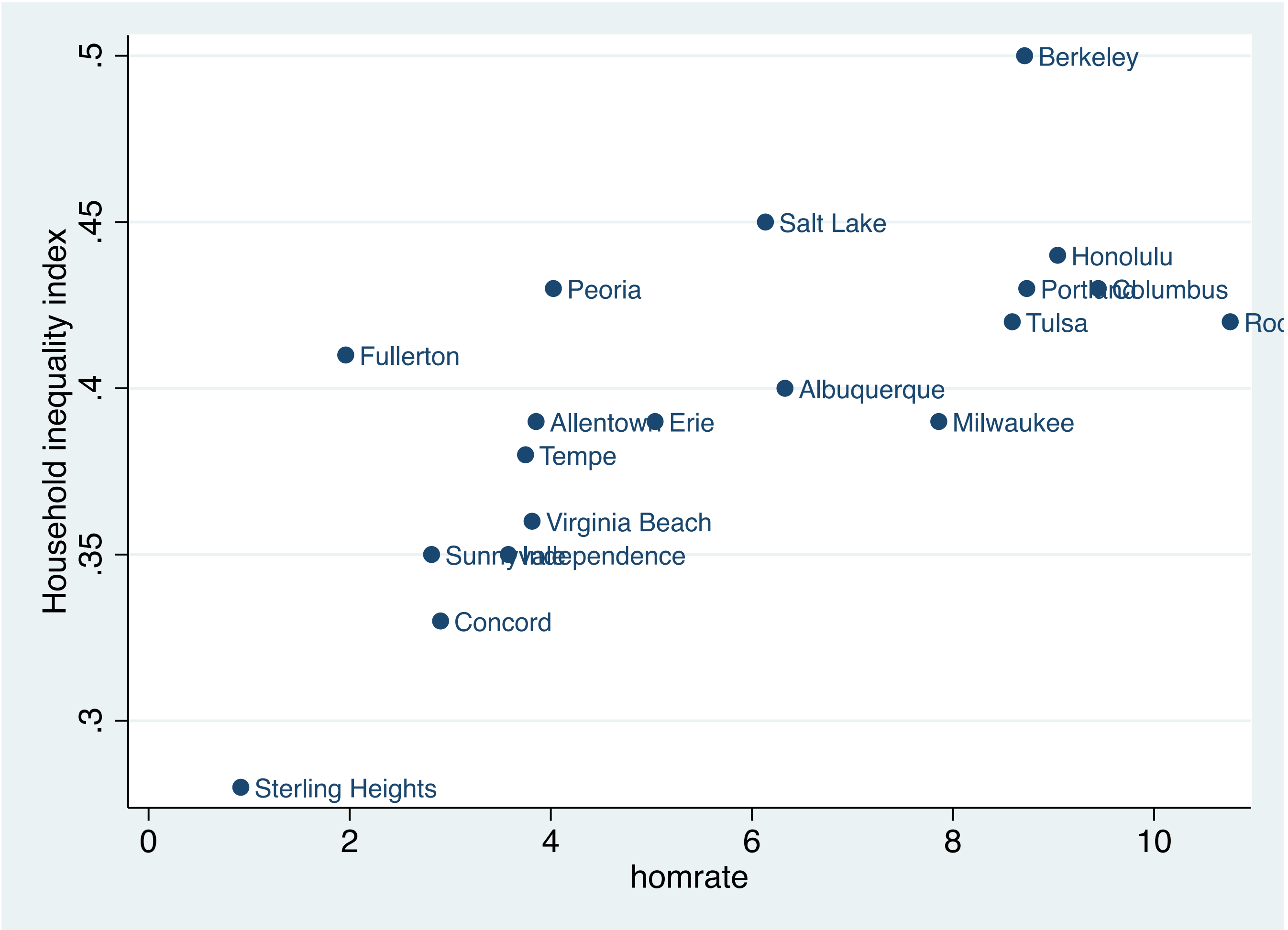
http://www.nejm.org/doi/full/10.1056/NEJMon1211064

Is higher inequality associated with higher homicide rate?



Is higher inequality associated with higher homicide rate?

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-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)