sociology UNIVERSITY OF LIMERICK

SO5041 Quantitative Research Methods

Brendan Halpin, Sociology, University of Limerick Brendan Halpin, Sociology
Autumn 2024/5 Autumn 2023/4

SO5041 Autumn 2023/4 – Module outline[1]

Module Code:	SO5041
Module Title:	Quantitative methods for MA research
Academic Year:	2024/5
Semester:	Autumn
Lecturer:	Dr Brendan Halpin
Lecture Locations:	Class: Mon 10-12 ER2011; Lab: Mon 13-15 A0060a
Lecturer Contact Details:	brendan.halpin@ul.ie, Phone: ext 3147
Lecturer Office Hours:	Fri 13:30-15:30

[1] The definitive version of this document is at https://teaching.sociology.ul.ie/so5041/so5041outline.pdf

Short Summary of Module:

An introduction to quantitative research methods in sociology.

Aims and Objectives of Module:

- Course focus:
  - The role of empirical reasoning in sociology, using quantitative data
  - Quantitative social science data collection, especially the survey
  - Handling quantitative data: coding it onto a computer, organising it, presenting it
  - Statistical analysis: making claims about the world using quantitative data – *sampling* and *inference*

...contd

- Practical focus:
  - Using software to analyse data and prepare findings
    - Stata: statistical software package
    - Microsoft Excel: spreadsheet
  - Carrying out questionnaire-based research
  - Becoming a critical consumer of quantitative research

Learning Outcomes:

- Apply quantitative methods to real research problems
- Critically assess published research using quantitative methods
- Choose appropriate research methods for MA research
- Use software effectively and reproducibly to manage, present and analyse social science data

Course Structure:

One two-hour lecture per week, one two-hour lab per week.

Detailed Module Plan

- Introduction to quantitative method – use number to represent information, simple descriptive statistics & presentations. Use Stata to enter & report data
- Samples, surveys and probability – the *theory* of how a sample can be used to describe a population, some elementary probability theory, basic questionnaire design and survey implementation. Manipulating data in Stata, more presentation.
- Statistical inference – the *practice* of using a sample to describe a population. Statistically informed use of Stata: testing difference of means, analysing association in tables.
- Linear regression and correlation
- Regression analysis with multiple explanatory variables?
- In parallel, reading and discussion of a small number of quantitative research reports

Lecture topics by week

Week	Topic	Lecture
1:	General introduction, (1) Introducing Quantitative Social Research	✓
2:	(2) Surveys, Questionnaires and Sampling,(3) Numbers as Information, univariate analysis	✓
3:	(4) Bivariate analysis	✓
4:	(4) Bivariate analysis continued, (5) Spread, types of variables	✓
5:	(6) Sampling, (7) Distributions	✓
6:	(8) More on Distributions, (9) Sampling Distributions and the Central Limit Theorem	✓
7:	(10) Sampling and confidence intervals, (11) Two new distributions: t and χ^2	✓
8:	Bank Holiday	✗
9:	(12) Questionnaire Design	✓
10:	(13) Hypothesis testing, (14) More t-tests	✓
11:	(15) Correlation, (16) Regression	✓
12:	(16) Regression continued	✓

Lab topics by week

Week	Topic	Lab
1:	Logging on, running Stata, general intro	✓
2:	Univariate and bivariate analysis	✓
3:	(no lab)	*X*
4:	Data entry, editing data in Stata, missing values	✓
5:	Using the Normal Distribution (by hand, spreadsheet)	✓
6:	Confidence intervals for means and proportions (by hand, Stata)	✓
7:	Understanding the standard deviation (spreadsheet exercises)	✓
8:	Bank Holiday	*X*
9:	Chi-squared test (spreadsheet and Stata)	✓
10:	Hypothesis test on a mean (spreadsheet and Stata)	✓
11:	More hypothesis tests; correlation	✓
12:	Regression analysis using Stata	✓

Texts

- Main text: Agresti, *Statistical Methods for the Social Sciences* – good introduction to formal statistical methods, very clear and accessible (will use more extensively in second semester course)
- For Stata:
  - Robson and Pevalin, *The Stata Survival Manual*
  - Kohler and Kreuter, *Data Analysis using Stata*
  - Acock, *A Gentle Introduction to Stata*
- Dip into Alan Bryman, *Social Research Methods*
- See also David de Vaus, *Surveys in Social Research*: good on survey methodology
- Other occasional readings
- Software: We will have access to Stata in the PC-Lab. You may also decide to buy a six-month student licence for Stata/BE at https://www.stata.com/order/new/edu/gradplans/student-pricing/.

Details of Module Assessment

Assessment will be by means of four assignments, worth 25% each. These will involve a range of activities, including use of Stata, online exercises on statistical concepts, and short essay-style questions. These will be due at ends of weeks 5, 9, 12 and 15. The fourth assignment will take the place of a formal exam.

Note on assignments

- Cooperation between students is encouraged but assignments must be the student's own work
- Please refer to Dept Plagiarism Policy below.
- Please use the Dept Assignment Declaration form with all assignments (except online) https://www.ul.ie/artsoc/sociology/student-resources
- Note the Dept's policy on deadlines https://www.ul.ie/artsoc/sociology/student-resources

Details of Annual Repeats

- Repeat assessment: 100% exam, August 2023

Classroom Technologies

The module will use BrightSpace for materials and submission of assignments. We will also use https://teaching.sociology.ul.ie/so5041, particularly for resources that need to be accessed directly.

FEEDBACK

Written feedback will be provided after each assignment during the semester.

Plagiarism notice

It hardly needs to be said, but all work must be your own. All material drawn from other sources must be clearly attributed. Passing off others' work as your own is considered academic dishonesty, and can be subject to substantial penalties. Please familiarise yourself with the departmental policy on plagiarism and use the coversheet declaration with all assignments (both available at http://www.ul.ie/sociology/ under Student Resources).

Deadline policy

Please also note the Department's policy on deadlines, also available at http://www.ul.ie/sociology/ under Student Resources.

Qual versus Quant?

- Social research is often divided into qualitative and quantitative:
  - Is this a real division?
  - What exactly is quantitative social research?
- In some respects the distinction is clear:
  - Quantitative research is concerned with numbers and statistical analysis, typically of large-scale surveys
  - Qualitative research is less obviously concerned with number, and typically is smaller scale and more in-depth
- 'Qualitative' actually covers a diverse range of research and method
- But sometimes this division is over-stated, e.g., "qualitative sociology" versus "quantitative sociology"
- It is only at the level of method that the division is clear

Positivism–Interpretivism?

- Some commentators see major philosophical divisions
  - quantitative ≡ positivist
  - qualitative ≡ interpretivist, anti-positivist
- While there are some parallels, this is misleading: quantitative research is not necessarily positivist, and is not incompatible with interpretivist frameworks
- Positivism is a philosophy of science that holds that only that which can be measured exists
- However, it is often used to mean a naïvely scientific approach to social research, or as an inherently negative way of referring to the quantitative method

Positivism–Interpretivism?

- In the past positivists have argued that social science should use exactly the same methods as the natural sciences,
  - suggesting that social phenomena should be studied via measurement and observation from the outside,
  - strongly downplaying the role of understanding or of *rational social action*

Critique of positivism

- Positivism has many critics, in sociology often from an interpretivist perspective: the interpretation of the meaning of action and communication is paramount for this sort of social research
- Some attacks on positivism have been extended to attacks on quantitative methods, but it is mistaken to treat them as the same
- Many interpretivists criticise the quantitative method for not being able to deal with the complexity of meaning and context in social life – this is not the same as attacking it for being positivist

Quantitative approaches and rationality

- A great deal of quantitative research works in terms of knowledgeable rational actors
  - Weber (key theorist for interpretivists) himself carried out a number of quantiative projects
  - Much neo-Weberian sociology uses quantitative method, e.g., "rational action theory" school
- Moreover, to justify quantitative sociology it is sufficient that you can say some useful things via "measurement", and conversely, all naturalistic research is in a sense dealing with measurement and observation
- Increasing use of "mixed methods": both qual and quant in the same project

Qual vs Quant

Quant	Qual
Big N	Small N
Shallow data	Rich contextual data
Highly comparable	Comparable only with complexity
Representative samples	Theoretical samples
Strong (mechanical) generalisation	Theoretical generalisation
Bureaucratic mode of production	Artisanal mode of production
Compare & contrast to detect general features	Elucidate meaning specific to context

Limitations

- Though quantitative research is not necessarily positivist, it has clear limitations
  - Rigid, almost industrial method – less opportunity for a dialogue between theory and data collection during the research process
  - Restrictive: you can only deal with the data you have
  - Weak on rare or hard-to-trace populations, or on rare events
  - Institutional context: large investment required means substantial pressure to follow the pressures of funding

Powerful

- But it is very powerful for certain types of question
  - whether certain processes or phenomena are present in a given population
  - where the *relative size* of competing effects is important
  - indeed, wherever it is necessary to generalise to a large population
- Probably true to say that the best research is methodologically ecumenical

Unit 1: Introducing Quantitative Social Research

An Anatomy of Quantitative Research

"Scientific" model?

- Though not necessarily positivist, quantitative research has strong affinities with a "scientific" model of the research process
  - the *Theory* ⇒ *Hypothesis* ⇒ *Test* cycle
  - *Falsification* in the Popperian sense
  - an interest in causal relations
- (This puts a good deal of weight on data meaning what we think it means)

The experiment

- The ideal type of this method is the `experiment` as, at least conceptually
  - subjects allocated randomly to two groups,
  - one control group
  - one "treatment" group exposed to the variable of interest
  - with the strong inference that differences in outcome between the two groups are caused by the treatment
- Experiments are, of course, almost always impossible in social research (and much other research)

Observational data

- The alternative is the observational model:
  - collect data "in the field", e.g., using surveys or "observation"
  - use *multi-variable statistical techniques* to search for *causal relations*
- Necessarily weaker than the experimental model: much harder to be sure you are not missing another reason for a given effect
- Generally harder to reason about causality with observational data: the evidence you have is simply association or correlation

Ambiguity

- That is, your survey may show that unemployed people have lower mental health than other categories (an "association" between employment status and mental health), but cannot tell you why:
  - unemployment might simply be bad for you
  - people with mental health problems may be more prone to lose jobs
  - or a third factor may affect both (e.g., geography: in one region there may be higher unemployment and higher mental ill-health for unrelated reasons)

Taking time into account

- Longitudinal observational designs help
  - if respondents are observed at two time points, and we systematically see people moving from employment to unemployment and equanimity to depression (and vice versa), we can exclude the "third factor" argument
  - if we observe people more frequently, we may be able to observe the onset on unemployment before the onset of depression at an individual (or vice versa) and argue with more confidence for a causal relationship

Surveys and Survey Research

- Social survey research is very widespread:
  - political opinion polls and market research
  - EU-wide Labour Force Survey & CSO's Quarterly National Household Survey (LFS plus)
  - EU Eurobarometer
  - European Social Survey, World Values Survey, International Social Survey Programme
  - Household Budget Surveys
  - Growing up in Ireland, TILDA
  - Slán, Irish Study of Sexual Health and Relationships
  - surveys of business opinion, of inventories etc.
  - many emanating from ESRI or government

Surveys: representativity

- The key principle of survey research is representativity: because the sample is random, summaries of the sample's characteristics can be imputed to the relevant population
- Sometimes we end up with too few cases of a subgroup to analyse – e.g., ethnic minorities; over-sampling or specially targetted surveys may help

Longitudinal surveys

- Longitudinal surveys are a special case
  - Panel surveys the same sample at regular intervals (e.g., European Community Household Panel, US Panel Study of Income Dynamics, German Socio-Economic Panel, British 'Understanding Society' Panel Study)
  - Retrospective studies ask respondents to report complete life histories retrospectively (Irish Mobility Study, UK Family and Working Lives Survey, German Life History Study, etc.)
  - Cohort studies take a group of subjects and follow them forward (e.g., the Growing Up in Ireland study, The Irish Longitudinal Study on Ageing)
- Taking time into account makes these in many ways much richer data sources

Questionnaire design 1/3

- The questionnaire is the linchpin of the survey
- Must elicit right information with minimum of ambiguity or suggestion, minimum inconvenience to the respondent
- Question design is a black art, since small changes of phrasing may cause different results

Questionnaire design 2/3

- Extensive reliance on standardised questions, or standardised forms of questions (e.g., the typical five point answer scale: strongly agree, agree, neutral, disagree, strongly disagree)
- Standard schedules exist for certain purposes, e.g., the General Health Questionnaire
- See e.g., https://www.understandingsociety.ac.uk/documentation/mainstage/variables/?s=ghq&post_type=variable_mainstage&submit=Search

sociology

Questionnaire design 3/3

- V important to minimise "open" questions: much cheaper to pre-code answers (but allow an "other, specify" answer)
- Very important to test questionnaires in a pilot survey, to trap ambiguities and other problems, and to help pre-code questions

sociology

Access

- Lots of survey data is available to the public, or to researchers
- Via data archives (e.g., the Irish Social Science Data Archive: http://www.ucd.ie/issda)
- Via government, EU, organisations like OECD
- Via website like European Social Survey: http://www.europeansocialsurvey.org/

sociology

Unit 2: Surveys, Questionnaires and Sampling

Other forms of data

Other forms of data

- Administrative data
- Online data (e.g., Twitter)
- Topic-specific data such as about Covid-19
- "Big data": commercial data, online activity trackers, mobile phone records, Fitbit, traffic records

sociology

Admin data

- Huge amounts of relevant administrative data is available
- Not survey: collected as a byproduct of the operation of the state
- Vital statistics: Births, marriages, deaths
- Censuses
- Tax, employment/unemployment, benefits, education, business
- Irish Central Statistics Office puts lots online at https://statbank.cso.ie
- See also OECD, Eurostat et alia.

sociology

"Local" social media research example

- Recent research by a former UL MA Sociology student
- How affect (positive or negative emotions) in politicians' tweets affects readers' response
- "Affect" is judged by a complex software setup created by IBM but made available to researchers
  - This sort of machine-learning or AI system is increasingly important
  - Not necessarily as accurate as human raters but can cover much more data
- https://www.nature.com/articles/s41599-021-00987-4

sociology

"Big data" increasingly important

- Matters more and more
- Requires different skills
- Sometimes threatens to replace conventional sources
  - Quicker, cheaper
  - But as accurate??
- But really big problems of representativity

sociology

Number as information

- In the Lab we will begin by coding data in numerical form, entering it on the computer and using Stata to make simple descriptive summaries of it
- This activity is an important stage of the quantitative research process: turning information into numbers and numbers into information

sociology

Steps

- There are several important steps in this process
  - Determine a clear set of information items to collect (i.e., what questions you want to ask)
  - Determine a simple way of representing this information
    - Represent it as a single “quantitative” number, e.g., Euros per annum for income, or
    - Determine a fixed set of categories grouping every possible answer, and attach a number to each category
  - Collect it (easier said than done) and enter on a computer
  - But for the analysis to make sense we need to bring in information about that the numbers mean: variable labels and value labels restore some part of the meaningful information hidden by coding as numbers

Example: Gender

- In the past, widely treated as unproblematic
- Interviewer ticked the box
- Increasing awareness that it’s not:
  - People may feel it’s none of your business (especially online)
  - Some people are non-binary

Asking the question

Let’s ask it as M/F/prefer not to say: example

Please enter your gender:

- ☐ Female
- ☐ Male
- ☐ Other/Prefer not to say

Data in the database: usually just numbers

Case no.	gndr
1	1
2	2
3	2
4	3
5	2
6	1
7	1

Raw table

```
       gndr |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |          3       42.86       42.86
          2 |          3       42.86       85.71
          3 |          1       14.29      100.00
------------+-----------------------------------
      Total |          7      100.00
```


To make it readable, we add labels

- Label gndr as "Respondent’s gender"
- Label the values:
  - 1 = "Female"
  - 2 = "Male"
  - 3 = "Declined to say"

Table with labels

```
   Respondent's |
         gender |      Freq.     Percent        Cum.
----------------+-----------------------------------
         Female |          3       42.86       42.86
           Male |          3       42.86       85.71
Declined to say |          1       14.29      100.00
----------------+-----------------------------------
          Total |          7      100.00
```


Unit 3: Numbers as Information – Data

Univariate summaries

Descriptive summaries

- When we enter data on a computer we can easily present descriptive univariate summaries (i.e., summarising one variable at a time)
- For categorical or grouped variables (i.e., where there is a “small” number of different values) we can use a frequency table: how many of each sort are there? what proportion of each sort are there?
- For variables with a “quantitative” interpretation (i.e., where the number measures something like age, income, duration, distance, or where it is a count like number of children, number of cigarettes per day) we can explore the mean or average, or perhaps the median

Frequency Table

```
. tab rjbstat
```

current economic activity	Freq.	Percent	Cum.
self-employed	929	6.82	6.82
employed	6,749	49.58	56.41
unemployed	468	3.44	59.84
retired	3,034	22.29	82.13
maternity leave	75	0.55	82.68
family care	786	5.77	88.46
ft studt, school	878	6.45	94.91
lt sick, disabld	608	4.47	99.38
gvt trng scheme	22	0.16	99.54
other	63	0.46	100.00
Total	13,612	100.00	

Summarising a “quantitative” variable

```
. su rfimn
```

Variable	Obs	Mean	Std. Dev.	Min	Max
rfimn	13,612	1454.687	1459.462	0	56916.67

Codebook

```
. codebook rfimn

rfimn                                            total income: last month

                  type:  numeric (double)
                 label:  rfimn, but 9658 nonmissing values are not labeled

                 range:  [0,56916.668]              units:  1.000e-08
         unique values:  9,658                  missing .:  0/13,612

              examples:  470.71835
                         925.07501
                         1401.9929
                         2175.7588
```


Mean, Median

- The mean is defined as the sum total of the variable, divided by the number of cases:
$$\bar{x} = \frac{\sum x_i}{n}$$
- The median is defined as the value such that 50% of cases are lower and 50% of cases are higher

Calculating the median

- To calculate the median “by hand”:
  - sort the values in order
  - if there is an odd number of cases (e.g., 101), find the middle one (e.g., 51st) – its value is the median
  - if there is an even number of cases (e.g., 100), find the middle pair (e.g., 50th and 51st) – the median is half way between their values

Median via Stata

```
. centile rfimn
```

Variable	Obs	Percentile	Centile	Binom. Interp. [95% Conf.	Interval]
rfimn	13,612	50	1141.668	1119.325	1166.667

Unit 3: Numbers as Information – Data

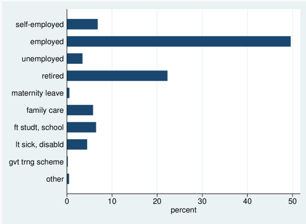
Univariate graphs

Graphs

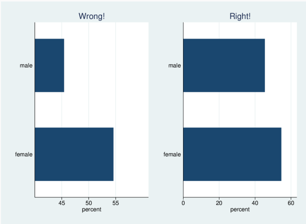
- We can also make graphical summaries
- For categorical variables the bar-chart is very good: this is like a frequency table where the height of the bars is proportional to the number/proportion in that category
- We can also use Pie-charts: these are circles with segments (pie-slices) whose angle is proportional to the size of the category (there are some arguments that bar-charts are better)
- For quantitative variables we can use histograms: these are very like bar-charts, but break the “continuous” variable into groups. The bars in a histogram touch, to show that the variable is really continuous. In a histogram, the area under the “curve” (or stepped line) for a given range is proportional to the numbers in that range. It is sometimes interesting to vary the group size (the “bin size”) to get more or less detail

Bar chart

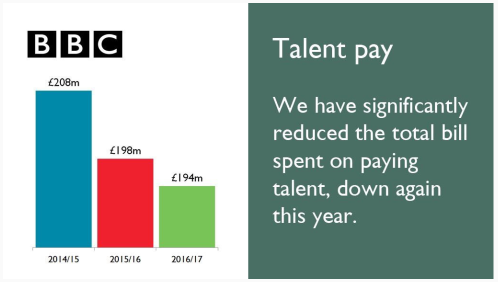
```
graph hbar, over(rjbstat)
```



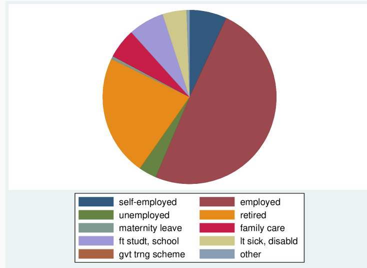
Zero is important



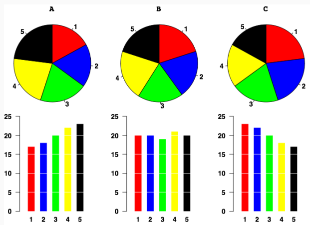
Beware: People lie with barcharts



Pie chart



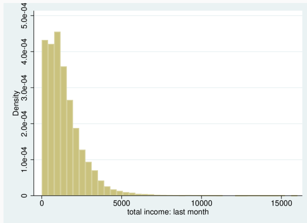
Pie is bad for you



Source: http://junkcharts.typepad.com/junk_charts/2009/08/community-outreach-pienaking.html

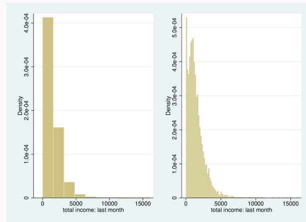
Histograms

```
hist rfimn if rfimn<=16000
```



Different bin widths

```
hist rfimn if rfimn<=16000, bin(10)
hist rfimn if rfimn<=16000, bin(100)
```



Some interesting reading on graphics

- William S Cleveland *The Elements of Graphing Data*. Rev. ed. Murray Hill, N.J.: AT&T Bell Laboratories, 1994
- Edward Tufte, *The Visual Display of Quantitative Information*. Cheshire, Conn.: Graphics Press, 1983
- Kieran Healy, *Data Visualization: A Practical Introduction* 2018. Focused on R, but Chapter 1 is a general discussion. Online at https://socviz.co/

- Agresti, *Statistical Methods for the Social Sciences*, chapter 1 (background) and chapter 3 (descriptive statistics)

- To follow: Agresti, chapter 2, Sampling and Measurement

Unit 4: Bivariate analyses

Bivariate analysis

Types of variables

- So far i have been using several terms for types of variable:
  - Categorical
  - Grouped
  - Continuous
  - Quantitative
- A more formal classification that corresponds with the sorts of analysis possible, goes:
  - Nominal
  - Ordinal
  - Interval
  - Ratio

NOIR – categorical

- Nominal variables have categories where each category is "just a name", i.e., nominal: example: religion, region of birth, political allegience
- With nominal variables we can't do much more than present frequencies
- Ordinal variables have categories where each category is something different, but where it is possible to put the categories in a meaningful high–low order. Examples include highest maths qualification, exam letter grades, attitudes (agree strongly, agree, neutral, disagree, disagree strongly) and so on. with ordinal variables, frequencies are meaningful, and the "cumulative percent" column that Stata provides is also meaningful. Additionally, with ordinal variables we can calculate the median

NOIR – quantitative

- Quantitative variables can be interval or ratio
- Interval variables are variables like temperature, where the difference between, say, 35° and 40° is the same as between 50° and 55°. That is, the meaning of an *interval* is the same no matter where it is. However, with temperature (centigrade or fahrenheit) it is not true that 40° is twice as hot as 20°, because 0° is not a true starting point. With interval variables it makes sense to calculate the mean. That is, if we have five days with temperatures of 11°, 9°, 12°, 11° and 10°, it makes sense to say the average is $\frac{11+9+12+11+10}{5} = 10.6$

NOIR – ratio variables

- Ratio variables are like interval variables, and are probably more commonly encountered. these are quantitative variables, where the zero makes sense. distance, duration, money, age are all ratio variables, because 0 kilometres, 0 seconds, €0 and 0 years are all things that make sense. ratio variables can also use mean, but unlike interval variables we can also work with proportions (twice as old, twice as rich, 10% farther and so on)

Unit 4: Bivariate analyses

Bivariate Summaries

Recap: Univariate analysis

- Frequencies: `tab foreign`
- Pie chart: `graph pie, over(foreign)`
- Bar chart: `graph bar, over(foreign)`
- Medians, etc: `centile mpg, centile(25 50 75)`
- Mean, standard deviation: `su mpg`
- Histogram: `histogram mpg`

Bivariate Analysis

- So far we have dealt with describing one variable at a time: important but limited
- Bi-variate (two-variable) summaries allow us to look at *relationships* between variables as well
- Where both variables are categorical (nominal, ordinal or grouped) we can use two-way tables, cross-tabulations

Cross-tabulation

```
. use http://teaching.sociology.ul.ie/so5041/labs/week3.dta

. tab empstat sex
```

usual employment situation	school leavers sex 1	2	Total
working for payment	388	380	768
unemployed	67	46	113
looking for 1st job	170	151	321
student	471	490	961
engaged in home dutie	0	19	19
permanent disability/	4	0	4
other	4	7	11
Total	1,104	1,093	2,197

Column Percentages

```
. use http://teaching.sociology.ul.ie/so5041/labs/week3.dta

. tab empstat sex, col
```

Key
frequency   *column percentage*

usual employment situation	school leavers sex 1	2	Total
working for payment	388 35.14	380 34.77	768 34.96
unemployed	67 6.07	46 4.21	113 5.14
looking for 1st job	170 15.40	151 13.82	321 14.61
student	471 42.66	490 44.83	961 43.74
engaged in home dutie	0 0.00	19 1.74	19 0.86
permanent disability/	4 0.36	0 0.00	4 0.18
other	4 0.36	7 0.64	11 0.50
Total	1,104 100.00	1,093 100.00	2,197 100.00

Too many percents!

```
. use http://teaching.sociology.ul.ie/so5041/labs/week3.dta
. tab empstat sex, row col
```

Key: *frequency*, *row percentage*, *column percentage*

usual employment situation	school leavers sex 1	2	Total
working for payment	388 50.52 35.14	380 49.48 34.77	768 100.00 34.96
unemployed	67 59.29 6.07	46 40.71 4.21	113 100.00 5.14
looking for 1st job	170 52.96 15.40	151 47.04 13.82	321 100.00 14.61
student	471 49.01 42.66	490 50.99 44.83	961 100.00 43.74
engaged in home dutie	0 0.00 0.00	19 100.00 1.74	19 100.00 0.86
permanent disability/	4 100.00 0.36	0 0.00 0.00	4 100.00 0.18
other	4 36.36 0.36	7 63.64 0.64	11 100.00 0.50
Total	1,104 50.25	1,093 49.75	2,197 100.00

Two ordinal variables

```
. tab mopfama mopfamb, row
```

Key: *frequency*, *row percentage*

pre-school child suffers if mother works	family suffers if mother works full-time: strongly	agree	neithr ag	disagree	strongly	Total
strongly agree	572 55.43	328 31.78	79 7.66	32 3.10	21 2.03	1,032 100.00
agree	267 6.61	2,507 62.04	876 21.68	354 8.76	37 0.92	4,041 100.00
neithr agree, disagre	55 1.09	813 16.12	3,070 60.88	1,010 20.03	95 1.88	5,043 100.00
disagree	32 0.82	295 7.59	437 11.25	2,777 71.48	344 8.85	3,885 100.00
strongly disagree	6 0.66	23 2.52	40 4.38	164 17.96	680 74.48	913 100.00
Total	932 6.25	3,966 26.59	4,502 30.19	4,337 29.08	1,177 7.89	14,914 100.00

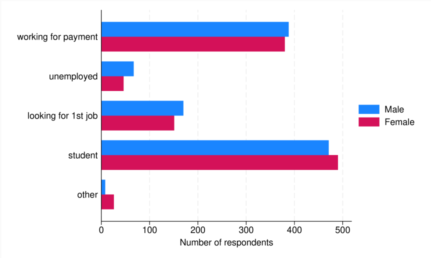
Unit 4: Bivariate analyses

Bivariate graphs

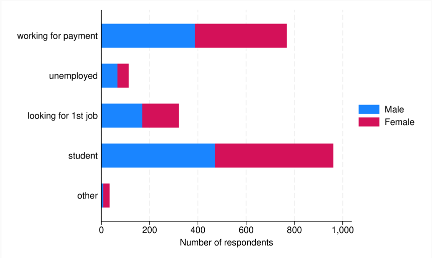
Bivariate graphs

- Graphical ways of presenting the same information as a cross-tabulation include clustered and stacked bar charts:
- Clusters are good for looking at distributions within groups
- Stacked bars are good where it is important to emphasise the sizes of the groups

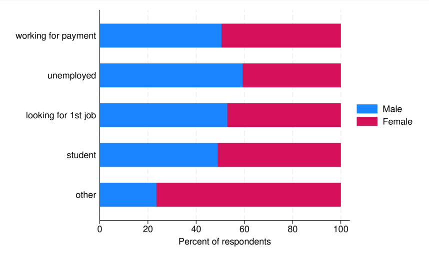
Clustered bar chart



Stacked bar chart



Stacked bar chart with percentages



Compare means – continuous within categorical

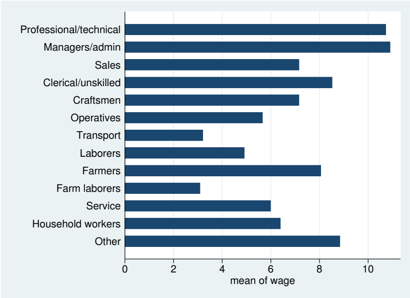
- Where one of the variables is categorical and the other is continuous (interval or ratio) we can "compare means". That is, instead of calculating mean income, we can calculate mean income for different groups

Wage by occupational group

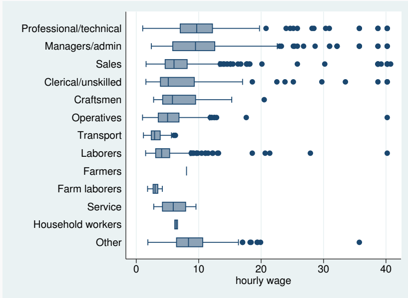
```
. sysuse nlsw88, clear
(NLSW, 1988 extract)

. tab occupation, su(wage)
```

occupation	Summary of hourly wage: Mean	Std. Dev.	Freq.
Professio	10.723624	6.3510736	317
Managers/	10.899784	7.5215875	264
Sales	7.1544893	5.0427568	726
Clerical/	8.5166856	8.5653995	102
Craftsmen	7.152988	3.7629626	53
Operative	5.6538061	3.3861932	246
Transport	3.2004937	1.3209668	28
Laborers	4.9058895	3.6083499	286
Farmers	8.0515299	0	1
Farm labo	3.0837354	.76638443	9
Service	5.9887432	2.1399333	16
Household	6.3888881	.31313052	2
Other	8.8362936	4.128914	187
Total	7.7779283	5.7613599	2,237

Another use of bar charts



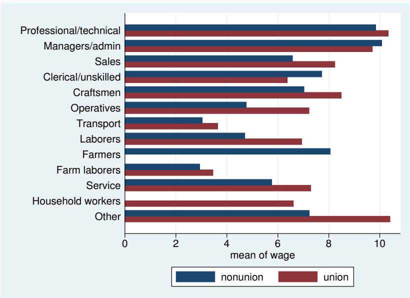
Boxplot



Boxplot

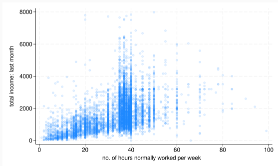
- The boxplot summarises the distribution of a variable:
  - 50% of the distribution is within the box
  - The bar in the middle is the median
  - The "whiskers" extend to the minimum and maximum, excluding "outliers"
  - Outliers are cases more than 1.5 IQR above or below the corresponding quartile (these are often marked separately)

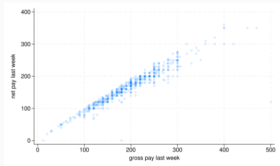
Three-way analysis



Scatterplot

- Where both variables are continuous, a very useful device is the scatterplot:





Reading to date

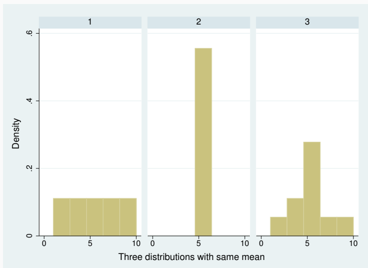
- Agresti, *Statistical Methods for the Social Sciences*:
  - Chapter 1: background
  - Chapter 3: descriptive statistics

Spread in quantitative variables

- When we have quantitative variables like age, income, we use *measures of central tendency* like mean and median to describe the "middle"
- However, the dispersion about the middle is also important: completely different distributions can have the same mean:

	A	B	C
	1	5.5	1.0
	2	5.5	4.0
	3	5.5	4.5
	4	5.5	5.2
	5	5.5	5.3
	6	5.5	5.6
	7	5.5	5.9
	8	5.5	6.0
	9	5.5	7.5
	10	5.5	10.0
Mean	5.5	5.5	5.5

Spread in quantitative variables



Summarising spread

- As well as summarising the middle, we often want to summarise the spread:
  - Range
  - Inter-quartile range
  - Standard Deviation
- The range is important, but depends too much on just the two extreme cases
- The IQR is much more stable
- The standard deviation has useful statistical properties

The "Standard Deviation"

- The standard deviation is calculated from the "deviations" from the mean, $\bar{X}$:

$$\text{Deviation} = X_i - \bar{X}$$

- The deviations add to zero so we can't just add them up: instead square them and then add them up:

$$\sum (X_i - \bar{X})^2$$

- To get a sort of average, we divide by sample size minus 1, and take the square root:

$$\sigma = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}$$

sociology

Standard Deviation: good measure of spread

- The standard deviation indicates how spread out the data is: the more spread out, the bigger the StDev
- It's a good measure because it depends on every single case, not just the extreme pairs (range) or the quartiles
- It has useful statistical properties which help when we come to statistical inference

sociology

Online app

Use this online-app to explore what happens to the mean and standard deviation as you move the data values around

https://teaching.sociology.ul.ie/apps/deviation/

sociology

Unit 5: Spread

More on types of variables

More on types of variables

- We have already divided variables into nominal/ordinal/interval/ratio
- We can also differentiate between "qualitative" and "quantitative":
  - Nominal variables are clearly qualitative: observations with different values have different qualities
  - Interval/ratio variables directly represent quantities: amount of money, number of children, distance

sociology

What about ordinal variables?

- Where do ordinal variables (*e.g.*, level of education) fit in?
  - In between?
  - Often treated as qualitative – categories are clearly "different"
  - Sometimes treated as quantitative: *e.g.*, letter-grades are given scores for calculating QCA; "Likert" scale answers (strongly agree to strongly disagree) are given scores of -2, -1, 0, 1, 2
  - Applying scores to an ordinal variable implies there is an underlying interval/ratio variable which we do not measure
  - In the QCA example this is true: the individual exam marks are on a ratio scale
  - In the likert-scale example, we may feel there is a continuum of agreement–disagreement which is approximately measured by the five options

sociology

Sensitivity

- There, however, we may feel that the steps between neutral and (dis)agree are bigger than between (dis)agree and strongly (dis)agree so we might want the scores to go -3, -2, 0, 2, 3 (or -3, -1, 0, 1, 3)
- When we base an analysis on applied scores like these, it is good to do a "sensitivity analysis": compare the results with different scoring systems to see how sensitive your analysis is to the scoring method

sociology

Summary of 5-point ordinal scale

Question	VP	P	Neutral	G	VG	N	Score
Overall satisfaction with Department of Sociology	1.4	1.7	23.2	52.6	21.1	289	0.781
National reputation of University of Limerick	0.7	2.5	16.5	36.6	43.7	279	0.840
Overall quality of academic programme within UL	0.0	5.1	13.6	54.2	27.1	295	0.807
Overall quality of services provided by UL	1.0	7.7	20.1	47.8	23.4	299	0.770
Overall quality of the social aspects at UL services	0.7	3.2	23.6	36.6	35.9	284	0.808
Quality of Student Academic Administration office	3.9	12.5	33.9	37.1	12.5	280	0.684
Quality of Fees office	2.9	7.5	47.3	33.5	8.8	239	0.675

sociology

Yet another distinction!

- A new distinction: **discrete** versus **continuous**
- Discrete variables have a finite number of possible values
  - Nominal and ordinal variables are by definition discrete
  - Count variables are discrete: 0, 1, 2, 3 . . . – you can't actually have 2.4 children
  - Grouped interval variables are discrete
- Continuous variables can have an infinite number of values in principle: 2, 2.4, 2.43, 2.435, 2.4358 and so on.
- Ungrouped interval/ratio variables are continuous in principle: an infinite number of values are possible
- In practice, we treat income as continuous though it can't vary in amounts less than 1 cent (and age, though we often measure it in integer years, or to the nearest month, etc.)

sociology

Outline

- Samples representing populations
- Random sampling
- Non-random sampling
- Representativity and bias

Reading

Chapter 2 of Agresti, Chapter 8 of Bryman

Unit 6: Sampling

Understanding sampling

Sample not Census

- Most quantitative research proceeds by using samples: rather than ask the entire electorate every month or so who they would vote for, ask a random sample
- If these individuals are chosen at random, their responses will approximate those of the whole electorate
- Random here does not mean completely without control, rather that each individual in the reference population has a known chance of selection

Simple Random Sampling

- The basic sampling strategy is the simple random sample
  - every subject in the population has the same chance of being selected
  - every possible sample of that size has the same chance of being selected

How to select a random sample

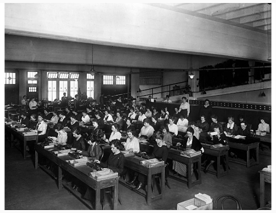
- Acquire a sampling frame: a list of all individuals in the reference population
- Number the individuals from 1 to N, the size of the population
- Draw n (sample size) numbers from a random number table or a computer, in the range 1 to N
- Interview the corresponding individuals

Random numbers

- In the past random number tables were widely used:

Line/Col	(1)	(2)	(3)	(4)
01	96252	48687	59641	99460
02	71334	10218	07459	20339
03	35932	10229	83514	76461
04	33568	36397	92080	98430
05	31535	29990	89596	77529
06	16483	30849	18676	63225
07	26374	60736	14522	66096
08	12040	90130	91860	08280

- Calculated meticulously by hand or using simple mechanical devices



Now computerised

- Now computers do the job very fast and easilly
- For instance in Excel typing `=rand()` in a cell gives you a random number between 0 and 1
- Or the `runiform()` function in Stata (try `display runiform()` repeatedly)
- If we have a list of names and addresses in a spreadsheet or a database the computer can generate a properly random sample in a blink of an eye

Probability sampling

- Simple random samples are an example of probability sampling
- Probability sample has the enormously important theoretical advantage of representativity
- If properly conducted, the sample is guaranteed to be *representative*
- That is, all sample characteristics *approximate* those of the population – even characteristics you haven't thought of
- Non-probability sampling does not have these advantages but is sometimes used for other reasons

Unit 6: Sampling

Non-random sampling

Non-probability sampling

- There are many forms of non-probability sampling:
  - volunteer sampling (self-selection)
  - "streetcorner interview"
  - quota sampling

Volunteer sampling

- Volunteer sampling is popular in the media: questionnaires in magazines, phone-in lines (vote for/against TV programme issue)
- Sometimes the sample is very large
- However, because respondents are self-selecting serious bias can enter

Volunteer sample – dodgy inference!

- Agresti/Finlay quote several examples (pp 20–21): e.g., TV phone vote on "should UN stay in US?" had a response of 186,000 with 67% voting for "No"
- A simultaneous random sample of 500 had 28% voting no: *much* smaller but far more reliable
- TV phone-in subject to bias:
  - Who is watching the program? Time of day, interest, etc.
  - Who is motivated to phone? Those with a strong opinion on the subject
- Twitter polls a particularly bad example of volunteer samples ("retweet for a bigger sample!")

"Opportunistic" samples

- "Streetcorner interview" sampling simply interviews whoever comes along: random in that sense
- However, where and when the interviewing is done will affect who is likely to be interviewed: under-represent workers if during the working day, under-represent country people if done in the city, over-represent shoppers if done in shopping area, etc.

Quota sampling

- Quota sampling is a form of sampling that avoids these bias problems by using quotas (of age group, sex, employment status, income group, region etc.) – people passing by at "random" are interviewed only if they fit the quota
- Not really probability sampling, but often a very effective and efficient way of "faking" a true random sample, especially for well-understood purposes like opinion polling

Unit 6: Sampling

Characteristics of samples

Samples are *representative*

- The characteristics of a random sample approximate those of the reference population by virtue of randomisation
- For instance, mean income in a sample (the sample statistic) will be more-or-less close to the mean income for the population from which it is drawn (the population parameter)
- The sample statistic is a more-or-less good estimator of the population parameter
- This is in general true for any statistic: proportion answering yes, distribution of qualifications, average working hours, joint distributions (e.g., crosstabs) etc.

Samples are *approximate*

- But because it comes from a sample, the sample statistic will not be an exact estimate of the population parameter
- The sample statistic depends on chance: the exact contents of the sample
- Repeating the exercise with a different sample will give a different sample statistic – still approximating the population parameter, of course
- In principle we can think of there being a distribution of possible sample statistic values, all falling more or less close to the true value of the population parameter

Recap

- Research often uses samples, not censuses
- Samples are cheaper, and very useful if *representative*
- Random sampling gives representativity
- Non-random sampling builds in bias (but is often more practical)
- Characteristics of representative samples approximate those of the reference population
- This approximation is with error: we will deal with sampling error next

Unit 6: Sampling

Error in sampling

Sampling error

- This difference between the sample statistic and the true value is called sampling error
- If the true proportion of UL students with part-time jobs is 65% and a sample reports 72% we have a sampling error of 72 – 65 = +7%
- How useful a sample is depends on the size of sampling error
  - The bigger the sample the smaller the sampling error
  - The more variability in the population the larger the sampling error

How big is sampling error?

- How do we know how big the sampling error is?
- We don't know the true value, ever, but we can reason about how big it is likely to be
- With certain assumptions about the range of the true value and information about the dispersion of the sample we can estimate it – e.g., with a sample of 1,000 sampling error for proportions (%age working etc.) is close to ±%3

A simple simulation

- This computer simulation draws a sample from a population with a known mean value, and compares the sample mean to the true mean
- Use it repeatedly to see how sample values jump around
- See that larger samples have less volatility
- See that if the population variability (standard deviation) is higher, volatility is higher
- `https://teaching.sociology.ul.ie/apps/so4046/sampling/`

Systematic error

- Sampling error is due to sampling variability: always present
- Other sources of error can arise – systematic error is a problem

Error from sampling problems

- Undercoverage – e.g., homeless people, geographically mobile people
- Bad sampling frame – e.g., out-of-date electoral register

Error from response problems

- Outright refusal – "non-response"
- Refusal to answer certain questions – "item non-response"
- Bias or deception in answers

Survey design and error

- Sampling error is part of the design
- Non-sampling error is a problem and can invalidate the conclusions we wish to make – bias the sample away from representativity of the reference population
- Undercoverage or non-response introduces the possibility that the sample we get differs systematically from the population (by differing from the part we don't get)

Undercoverage and non-response

- For instance, if we are interested in financial wellbeing and our survey underrepresents homeless people or people who move very often the people in the sample may be systematically better off than those we fail to trace
- If UL students working part-time are too busy to respond to a questionnaire our sample will under-estimate the average hours spent working part-time

Unit 6: Sampling

More complex sampling strategies

More sampling strategies

- We have discussed several types of sampling strategy
  - Simple random sampling
  - Volunteer sampling and "street-corner interviewing"
  - Quota sampling
- Random sampling is probability sampling; the others are not
- Other forms of probability sampling also exist:
  - Systematic sampling
  - Stratified sampling
  - Cluster sampling
  - Multistage sampling

Systematic sampling

- Systematic sampling is used where the sampling frame has a more-or-less random order
- Pick a name at random near the start of the sampling frame
- Skip k cases, pick the next name, and continue until finished
- k, the size of the skip, is calculated as N/n, so that you arrive at the full size of the desired sample
- This works as a simple random sample, on the assumption that there is no relationship between where you are in the list and the relevant characteristics you have
- If there is any order to the sampling frame (like people being grouped together, for instance, or periodicity) this will not yield a random sample

Stratified sampling

- Stratified sampling works by dividing the population into strata and collecting random samples within the strata
- If the groups or strata are defined according to variables important to the study (e.g., age, gender, occupational group) this method yields statistically more efficient samples (i.e., sampling error is less than with a simple random sample)

Stratified sampling: efficient, flexible

- Where we know the population distribution of age or gender (e.g., from a census) our sample will automatically have the right proportions
- This method also allows over-sampling of small groups: ethnic minorities, people on government training schemes, the unemployed – this allows comparisons small groups and others that would not be possible in a simple random sample

Cluster sampling

- Cluster sampling divides the population into groups called clusters
- These are often geographic areas, or organisationally based
- A random sample of clusters is drawn
- Within each cluster, simple random samples of individuals are drawn

Cluster sampling – pros and cons

- Advantages:
  - Cost – much less interviewer travel time
  - Can draw better random samples within cluster than from national sampling frame (e.g., identify all households in cluster first, then sample: better than out-of-date list of addresses)
- Disadvantage: statistically less efficient – larger sample needed for same accuracy

Multistage sampling

- Multistage samples use several of these methods in combination
- For instance, using electoral wards as clusters, take a simple random sample of clusters; within each ward, treat streets as clusters and sample again; within each street identify every separate address and do a systematic sample of every nth household

Reading

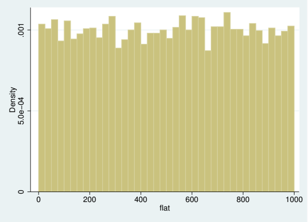
- Agresti Chapter 2
- Bryman Ch 8 (see also Ch 7)

Distributions

- We have seen how to display and summarise the distribution of variables:
  - Categorical: frequency distribution, percentage distribution, bar and pie charts
  - Quantitative (interval/ratio): mean, median, IQR, standard deviation, histogram, box-plot

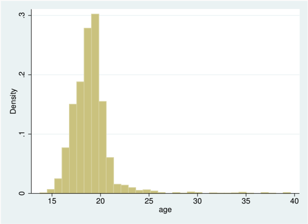
Distributions have shapes

- The shape of the histogram tells us about the distribution of the variable
- If a variable is “uniformly” distributed we see a flat distribution between the extremes:



Heaped distributions

- More often we see “heaped distributions” where more of the observations cluster around the centre, like this example (age) from the 1999 school-leavers’ survey:

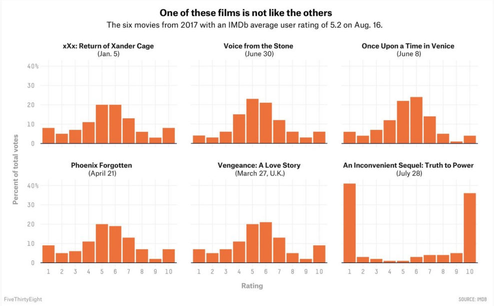


Many patterns

There are many patterns we might see:

- Uniform
- Extremes
- Bimodal
- Uni-modal

Polarisation

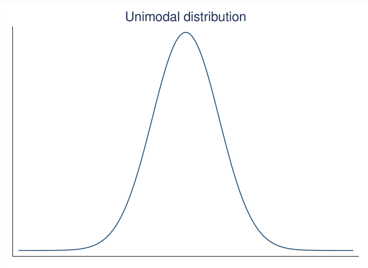


http://www.fivethirtyeight.com

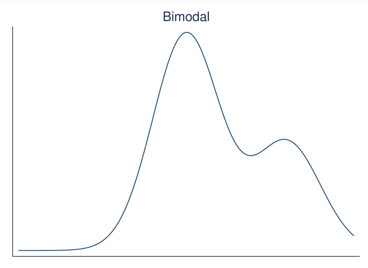
Uni-modal differences

- Symmetric (with different levels of kurtosis)
  - platykurtic – flatter
  - mesokurtic – average
  - leptokurtic – very concentrated around centre
- Asymmetric
  - Positively skewed (to right)
  - Negatively skewed (to left)

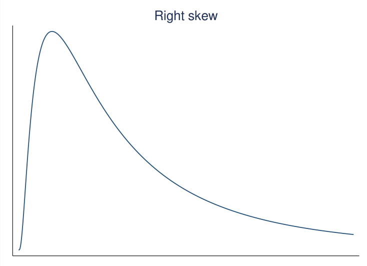
Symmetric unimodal



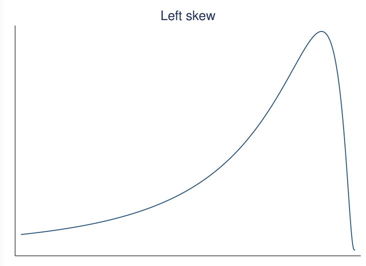
Asymmetric bimodal



Asymmetric: right skew



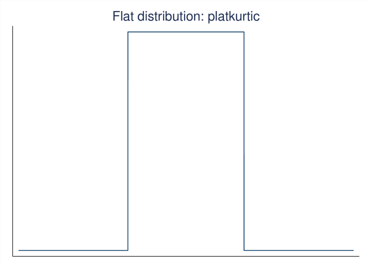
Asymmetric: left skew



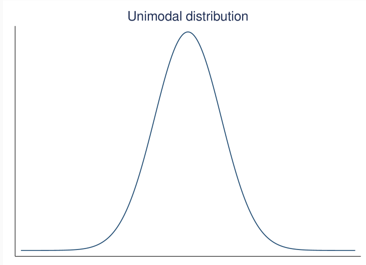
Different symmetric shapes: kurtosis

- Different distributions with the same mean and standard deviation can have different shapes
- Kurtosis: balance between peak, shoulders and tails

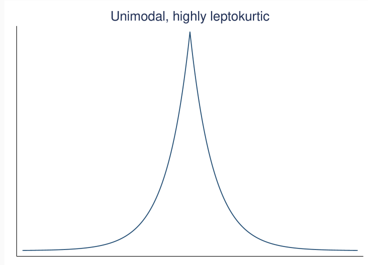
Flat: low kurtosis



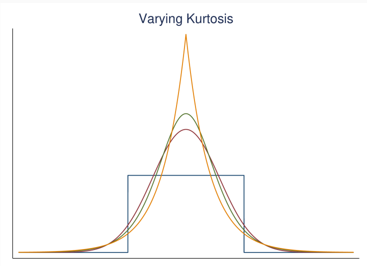
Normal: mid-kurtosis



Very peaky: high-kurtosis



Varying kurtosis

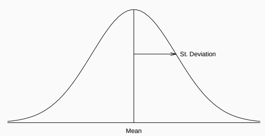


Unit 7: Distributions

The Normal distribution

Mathematically defined distributions

- There are some patterns that are defined “theoretically” or mathematically
- Among these the most important is the Normal Distribution:



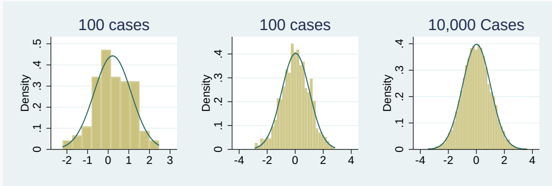
- This is the famous “bell-shaped curve”

The Normal Distribution

- The normal distribution is
  - symmetric (no skew)
  - mesokurtic (between flatter and pointier)
- The mean, mode and median are the same
- The farther you go from the mean, the lower the proportion of cases, in each direction symmetrically

Histograms display distributions

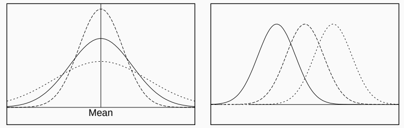
- We can see what a histogram of a variable drawn from a normally distributed population looks like:



- As the sample gets bigger, the histogram approximates the theoretical distribution better

Normal: mean and standard deviation

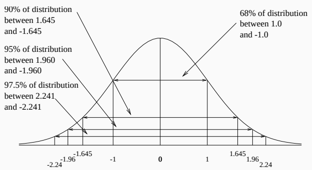
- What makes the normal distribution useful is that its form is well understood:
  - It is completely characterised by its mean and its standard deviation



Same mean, different SD — Same SD, different mean

Online app: https://teaching.sociology.ul.ie/apps/normsd

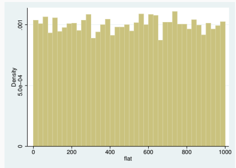
Normal distribution: well-understood



- About 68% of the cases in a normal distribution are within 1 SD on either side of its mean
- 95% are within $\pm$ 1.96 standard deviations of the mean
- 97.5% are within $\pm$ 2.24 standard deviations of the mean

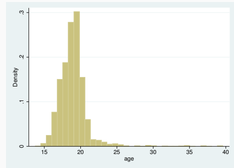
Histograms and distributions

- Histograms display distributions
- Probability distributions describe them formally
- The set of ticket numbers in a raffle has a uniform distribution
  - flat histogram
  - equal numbers of tickets in all ranges, e.g., 1–100, 101–200, 201–300
  - Thus winning ticket is equally likely to fall in any range: number is not related to chance of selection



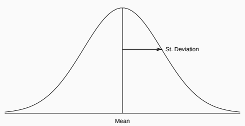
Heaped histograms

- However, if we were to pick a school-leaver at random from the school-leaver's survey, there is a relationship with date of birth:
  - ages near 19 more likely
  - ages much younger or much older much less likely
- . . . a clustered distribution



The Normal Distribution

- The extent of clustering depends on shape, standard deviation
- Smaller standard deviation means individual chosen at random more likely to fall near the mean
- The Normal Distribution a kind of "ideal type" of clustered symmetric distribution (arises naturally in many contexts)

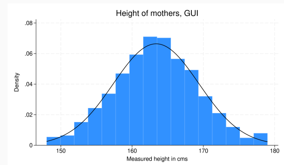


Normal

- The normal distribution is
  - symmetric
  - uni-modal
  - meso-kurtic
  - range $-\infty$ to $+\infty$
  - continuous, not discrete
- Completely defined by its mean and standard deviation:
  https://teaching.sociology.ul.ie/apps/normsd/

Examples of normally distributed variables

- IQ scores and other standardised tests – designed that way
- Time for a Deliveroo – possibly normal, e.g., mean of 30 mins, standard deviation of 5 mins
- Adult human height (separately for males and females)



Why does the Normal Distribution crop up so often?

- Where there is a core value, but lots of small things pushing it either way
- First observed in physical measurements, e.g., height of mountain, speed of light
  - Unknown correct answer
  - Each measurement full of small factors (errors) pushing it up and down
  - Some errors cancel each other, some compound
- Measurements will tend to have a normal distribution (hopefully) centred on the true value
- Normally distributed if many small factors, pushing up equally to down
  - additive
  - independent

sociology

Visualisations

https://commons.wikimedia.org/wiki/File:Galton_box.webm

https://teaching.sociology.ul.ie/so4046/quincunx.mp4

sociology

It crops up in sampling

- Each case in a sample pulls the sample mean up and down
  - For calculating things like means, each case has an additive effect
  - In simple random samples each case is independent of all others
- Therefore the set of all sample means has a normal distribution: https://teaching.sociology.ul.ie/apps/so4046/sampling/
- We will come back to this in the next lecture

sociology

Unit 8: The Normal Distribution

The Standard Normal Distribution

Standard Normal Distribution

- The Standard Normal Distribution is a special case of the normal distribution:
  - Mean = 0
  - Standard deviation = 1
- We can map any given ND onto it by
  - subtracting the mean
  - dividing by the standard deviation
- We can thus use the SND to estimate probabilities/proportions for any normal distribution, once we know the mean and standard deviation

sociology

Standard normal app

https://teaching.sociology.ul.ie/apps/snd

sociology

Reading the SND in the app

- The app tells us the
  - proportion of the distribution, or equivalently,
  - the chance of picking a case at random
- above or below a certain value
- or inside or outside +/- times that value
- we can also use it to calculate proportions/probabilities between (or outside) any two values

sociology

Proportion above or below a given level

- For example, given mean 100 and standard deviation 20, what's the chance of observing a value above 130?

$$\mu = 100, \sigma = 20, X = 130$$

$$\Rightarrow z = \frac{X - \mu}{\sigma} = \frac{130 - 100}{20} = 1.5$$

- From the app, we see that z=1.5 corresponds to p=0.0668 or about 6.7%: 6.7% of the distribution is above 130
- Clearly, 100% - 6.7% = 93.3% of the distribution is below 130

sociology

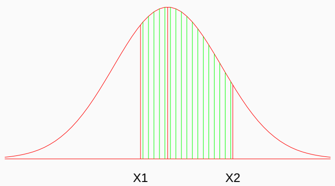
Height of GUI mothers

```
. su hwomen

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
      hwomen |      6,300     163.373    6.01054        148        179
. count if hwomen >= 175
  229

. count if hwomen >= 170
  949

. count if hwomen <= 160
  2,018

. count if hwomen <= 155
  626

. centile hwomen, centile(75 90)

                                                         Binom. interp.
    Variable |       Obs  Percentile    Centile        [95% conf. interval]
-------------+-------------------------------------------------------------
      hwomen |     6,300         75         167             167        168
             |                   90         171             171        171
```

sociology

Proportion between two values

- Once we know how to calculate the proportion of the distribution above or below any value, we can calculate the proportion between any pair of values



Proportion between two values

- To calculate the proportion between X1 and X2, calculate
  - The proportion below X1: $P(X < X1)$
  - The proportion above X2: $P(X > X2)$
  - $P(X1 < X < X2) = 1 - P(X < X1) - P(X > X2)$

Working backwards: given p find z

- We may also wish to work in the opposite direction
- Instead of asking what proportion of the distribution is above X, we may ask what is the level such that proportion p of the distribution is above it?
- For example, given the same distribution, what is the level such that only 5% of the distribution is above it?

Working backwards: given p find z

- Given the same distribution, what is the level such that only 5% of the distribution is above it?
- We work backwards, starting in the body of the table by searching for the value nearest to 5% or 0.050
- This corresponds with z = 1.645 (falls between 1.64 and 1.65)
- Reverse the formula: $X = \sigma \times z + \mu = 20 \times 1.645 + 100 = 132.9$
- Therefore, 5% of this distribution is above 132.9

Unit 8: The Normal Distribution

Online apps

App: P<X

Find the proportion below X, (above or below the mean)

- Link: Proportion below X

App: P>X

Find the proportion above X, (above or below the mean)

- Link: Proportion above X

App: P between X1 and X2

X1 and X2 may both be on either side of, or straddle the mean.

- Link: Proportion between X1 and X2

Outline

- In this unit we will examine
  - Population parameters and sample estimates
  - The Central Limit Theorem: this is why the normal distribution is important
  - How to deal with the imprecision of sample estimates: confidence intervals
  - Reading: Agresti, Chapter 4

Unit 9: The Normal Distribution

Reasoning about sampling error

Sampling error and sampling distributions

- How do we reason about sampling and sampling error?
- Sampling Error = true value – sample value; but true value unknown!
- We can reason about the likely distribution of sampling error
- If the true value is μ, how far is a sample value likely to fall from it?
- $\longrightarrow$ a *distribution* of possible sample values

Referendum sample

- Example: referendum with yes/no answer
- The probability a randomly selected voter goes yes vs no is unknown
- That is equivalent to saying the yes/no proportion is unknown
- What if we want to know overall proportion, and poll 1500 random voters?
- If we get a result of 54/46: how do we know we have any accuracy?

Computer simulation

- Agresti does a computer simulation to test
- Computer selects 1500 random numbers: $0 - 0.5 \Rightarrow$ yes; $0.5 - 1 \Rightarrow$ no
- First run gets 51.6% yes, next 49.1%
- They run it a million times, and make a histogram of the results: vast majority between 0.46 and 0.54, ie +- 4%

Simulation by hand: $N = 4$

- We can alternatively work through a “toy” example
  - Yes and No equally likely
  - Sample size of 4
  - 16 different possible combinations, each equally likely
- Result is a “binomial” distribution
- Note number of possible combinations is 2^N so this calculation is practical only with tiny samples

Simulating with N=4

N			
Y			

Simulating with N=4

N	N		
N	Y		
Y	N		
Y	Y		

Simulating with N=4

N	N	N	
N	N	Y	
N	Y	N	
N	Y	Y	
Y	N	N	
Y	N	Y	
Y	Y	N	
Y	Y	Y	

Simulating with N=4

N	N	N	N
N	N	N	Y
N	N	Y	N
N	N	Y	Y
N	Y	N	N
N	Y	N	Y
N	Y	Y	N
N	Y	Y	Y
Y	N	N	N
Y	N	N	Y
Y	N	Y	N
Y	N	Y	Y
Y	Y	N	N
Y	Y	N	Y
Y	Y	Y	N
Y	Y	Y	Y

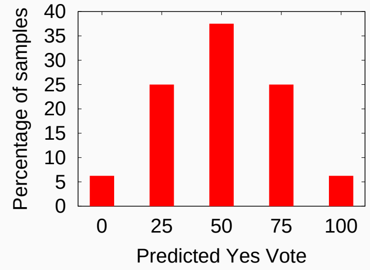
Simulating with N=4

N	N	N	N	0 : 4
N	N	N	Y	1 : 3
N	N	Y	N	1 : 3
N	N	Y	Y	2 : 2
N	Y	N	N	1 : 3
N	Y	N	Y	2 : 2
N	Y	Y	N	2 : 2
N	Y	Y	Y	3 : 1
Y	N	N	N	1 : 3
Y	N	N	Y	2 : 2
Y	N	Y	N	2 : 2
Y	N	Y	Y	3 : 1
Y	Y	N	N	2 : 2
Y	Y	N	Y	3 : 1
Y	Y	Y	N	3 : 1
Y	Y	Y	Y	4 : 0

Simulating with N=4

N	N	N	N	0 : 4
N	N	N	Y	1 : 3
N	N	Y	N	1 : 3
N	Y	N	N	1 : 3
Y	N	N	N	1 : 3
N	N	Y	Y	2 : 2
N	Y	N	Y	2 : 2
N	Y	Y	N	2 : 2
Y	N	N	Y	2 : 2
Y	N	Y	N	2 : 2
Y	Y	N	N	2 : 2
N	Y	Y	Y	3 : 1
Y	N	Y	Y	3 : 1
Y	Y	N	Y	3 : 1
Y	Y	Y	N	3 : 1
Y	Y	Y	Y	4 : 0

$N = 4$



Sampling distribution for N=4

Online Apps:

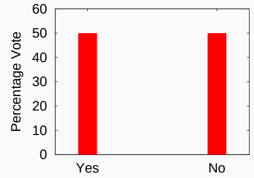
- Heads and Tails
- Simulating binomial sampling

Central Limit Theorem

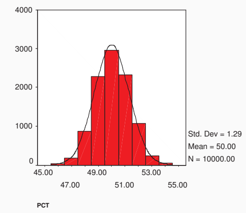
- The Central Limit Theorem: for a sufficiently large sample size, the *sampling distribution* of a statistic such as the sample mean will be approximately normal
- This is true, whatever the distribution of the population of interest
- We have seen this with the simulation of sampling vote where the true population proportions are 50:50

Binomial distribution

Population Distribution



Sampling Distribution



Uniform and normal populations

See `https://teaching.sociology.ul.ie/apps/so4046/sampling`

Sampling distribution

- This means that for sufficiently large samples, our sample statistic can be regarded as being drawn from a random distribution with mean μ and standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
- This is a very important theorem because it allows us to reason about the precision of sample estimates
- Note: $\sigma_{\bar{X}}$, the standard deviation of the sample mean, is known as the *standard error*

This unit

- Understanding imprecision due to sampling
- Estimating imprecision: confidence intervals
  - For means
  - For proportions
- Reading: Agresti Ch 5, sections 1 to 3.

Unit 10:

Estimating imprecision: confidence intervals

Samples are uncertain

- The characteristics of representative samples approximate those of the reference population
- But with uncertainty
- How do we characterise this uncertainty?
- With margins of error such as "confidence intervals"

Point estimates

- Agresti: A point estimator of a parameter is a sample statistic that predicts the value of that parameter
- For instance, our sample mean $\bar{X}$ is a point estimator of the population mean μ
- Good point estimators require two things:
  - To be centred around the true value (unbiased)
  - To have as small a sampling error as possible (efficient)

Unbiased, efficient

- Unbiasedness means that the estimate will fall around the true value, "on average"
- Efficient means that it will fall close to the true value
- Simple Random Sample estimates are efficient and unbiased, e.g.,

Unbiased, efficient

- Unbiasedness means that the estimate will fall around the true value, "on average"
- Efficient means that it will fall close to the true value
- Simple Random Sample estimates are efficient and unbiased, e.g.,

- Sample mean:

$$\bar{X} = \frac{\sum X_i}{n}$$

Unbiased, efficient

- Unbiasedness means that the estimate will fall around the true value, "on average"
- Efficient means that it will fall close to the true value
- Simple Random Sample estimates are efficient and unbiased, e.g.,
- Sample mean:

$$\bar{X} = \frac{\sum X_i}{n}$$

- Sample standard deviation:

$$\hat{\sigma} = s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$$

Unbiased, efficient

- Unbiasedness means that the estimate will fall around the true value, "on average"
- Efficient means that it will fall close to the true value
- Simple Random Sample estimates are efficient and unbiased, e.g.,
- Sample mean:

$$\bar{X} = \frac{\sum X_i}{n}$$

- Sample standard deviation:

$$\hat{\sigma} = s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$$

- Sample proportion:

$$\hat{\pi}_1 = \frac{n_1}{n_1 + n_2}$$

Sampling distributions and sampling error

- We have seen that sample estimates have sampling error, and now understand something of its characteristics, by exploring sampling distributions
- Sample estimates can be considered as being drawn from an imaginary random distribution, which (if the estimator is unbiased) will centre on the true population parameter, with a level of imprecision measured by the standard error (which will be as low as possible if the estimator is efficient)

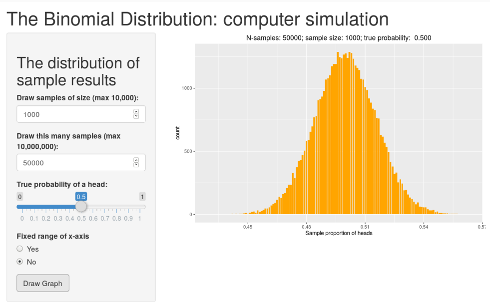
Sampling distributions: Central Limit Theorem

- Where the sample is sufficiently large, the Central Limit Theorem tells us that the sampling distribution is normal, with mean μ and standard deviation $\sigma_{\bar{X}}$
- The standard deviation of the sampling distribution is called the standard error

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{N}}$$

Simulation: Yes/No, 1000 cases, 50000 repeats

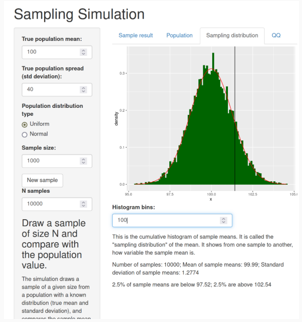
https://teaching.sociology.ul.ie/apps/binsim



Simulation: Uniform population distribution

- Population: mean 100, Std Dev 40
- Sampling distribution
  - mean 99.99
  - Std Dev = Standard Error 1.2774
- Theoretical SE: $\frac{40}{\sqrt{1000}} = 1.2650$

https://teaching.sociology.ul.ie/apps/so4046/sampling



Changing sample size and population SD

- With the same simulation, we can see what happens to sampling variability
  - for different sample sizes
  - for different population standard deviations
  - for uniform and normal population distributions
- The previous simulation shows the same for binomial variables

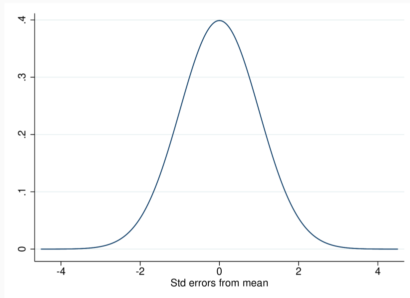
Confidence intervals

- We can use the CLT to add to our point estimate a measure of its precision: the *Confidence Interval*
- Agresti: "A confidence interval for a parameter is a range of numbers within which the parameter is believed to fall"
- "The probability that the confidence interval contains the parameter is called the confidence coefficient" which is a number close to 1, like 95% or 99%

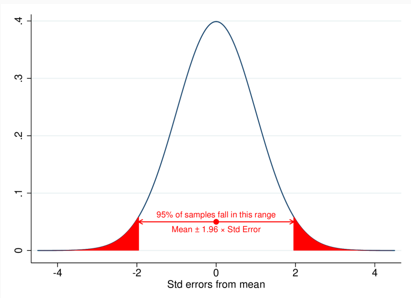
Estimating intervals

- A confidence interval is a band around our point estimate within which we can claim that there is a, for instance, 95% chance that the true population value lies – how do we calculate this?
- We work from the sampling distribution – if we are estimating a mean value, we know that 95% of estimates will fall within ± 1.96 standard errors of the mean
- The 1.96 comes from the normal distribution: 95% of the distribution is between 1.96 standard deviations above and below the mean

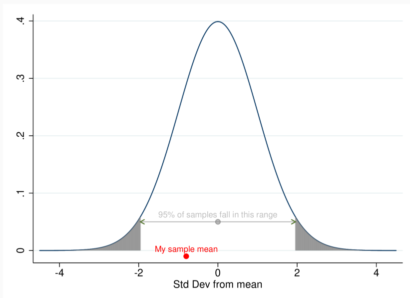
Central limit theorem: sample statistics normal



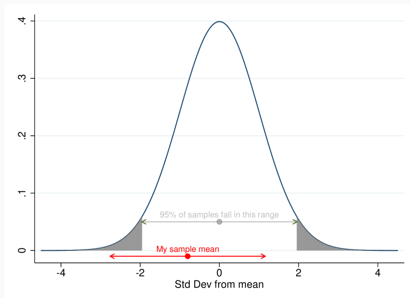
95% within ± 1.96 SE of mean



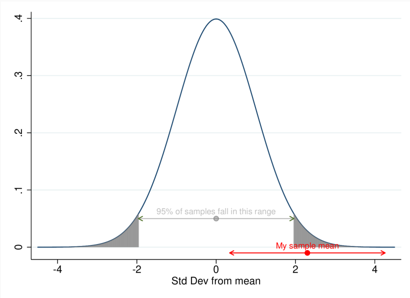
A sample: one of 95% within ± 1.96



Sample mean ± 1.96 SEs contains true mean



But 5% of sample means do not



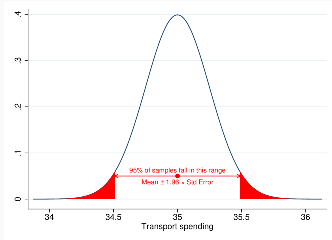
Example: transport spending

- Let's say spending on transport has a true mean of €35 per week, standard deviation €10. With a sample of 1600, 95% of all possible sample estimates will fall between:

$$35 \pm 1.96 \times \frac{10}{\sqrt{1600}}$$

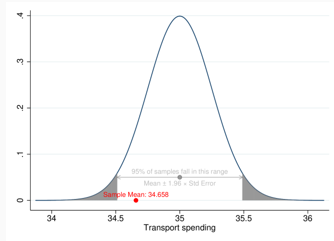
which is

$$\mu \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$$



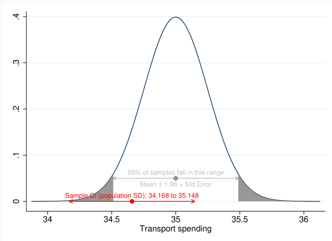
Sample results

- Let's say our sample gives a mean of €34.658, with a standard deviation of €10.123



Reverse the reasoning

- We don't know $\mu = 35$, only $\bar{X} = 34.658$
- We can reverse the reasoning and say that the true value has a 95% chance of falling in the range $\bar{X} \pm 1.96 \times \sigma_{\bar{X}}$



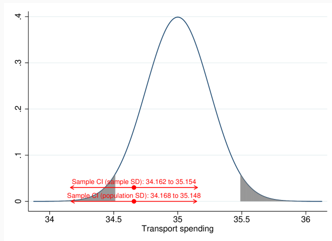
Sample SD

- But we don't know the true standard error
- We can use the sample estimate instead:

$$34.658 \pm 1.96 \times \frac{10.123}{\sqrt{1600}}$$

- In this case, a very slightly wider interval

$$\bar{X} \pm z_{0.95} \times \hat{\sigma}_{\bar{X}}$$



Calculating the interval

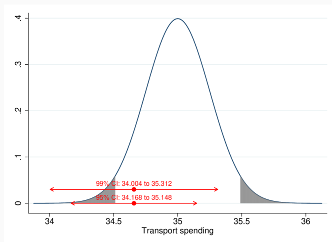
- Thus with sample mean is €34.658 and sample standard deviation €10.123 we calculate:
  - Standard Error is $\frac{10.123}{\sqrt{1600}} = \frac{10.123}{40} = 0.2531$
  - The lower bound is $34.658 - 1.96 \times 0.2531 = 34.162$
  - The upper bound is $34.658 + 1.96 \times 0.2531 = 35.154$
- We can interpret this as saying we are 95% confident that the true value is in this range
- This is because with 95% of all possible samples, such a confidence interval will include the true value

Levels of confidence

- The 95% confidence level is often used but sometimes the chance of being wrong one time in twenty is not acceptable
- We can use the normal distribution to choose other levels of confidence: for instance, 99% of the normal distribution lies within ± 2.575 standard deviations
- Check for yourself on `https://teaching.sociology.ul.ie/apps/snd`

99% Confidence

- 99% CI:
  - Lower: $34.658 - 2.575 \times 0.2531 = 34.006$
  - Upper: $34.658 + 2.575 \times 0.2531 = 35.310$
- To be more sure, we are necessarily less precise



Unit 10:

CIs for proportions

Standard deviation/error for proportions

- The example above is for sample means: the CI for sample proportions is similar
- The sampling distribution for a proportion (percent unemployed, percent voting Republican, percent satisfied etc.) is also normal for large samples
- The standard deviation of a 0/1 or yes/no variable depends on the proportions in the population however:

$$\sigma_\pi = \sqrt{\pi \times (1 - \pi)}$$

- We can estimate the standard error then as

$$\hat{\sigma}_{\hat{\pi}} = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}$$

See `https://teaching.sociology.ul.ie/apps/binsim`

CI for proportions

- The 95% confidence interval for a proportion uses this, the same way as for the mean:

$$\hat{\pi} \pm z_{0.95} \times \hat{\sigma}_{\hat{\pi}}$$

- If we have a sample 1000 and find 45% of voters expressing an intention to vote yes we calculate as follows:

$$\hat{\sigma}_{\hat{\pi}} = \sqrt{\frac{0.45(1 - 0.45)}{1000}} = 0.0157$$

and thus the CI is

$$0.45 \pm 1.96 \times 0.0157$$

- That is $\hat{\pi} \pm 0.031$, i.e., plus or minus 3%

Example: Proportion unionised

```
. sysuse nlsw88
(NLSW, 1988 extract)

. tab union

      union |
     worker |      Freq.     Percent        Cum.
------------+-----------------------------------
   nonunion |      1,417       75.45       75.45
      union |        461       24.55      100.00
------------+-----------------------------------
      Total |      1,878      100.00
```

- Proportion = 461/1878 = 0.2455

Example: Proportion unionised

```
. sysuse nlsw88
(NLSW, 1988 extract)

. tab union

      union |
     worker |      Freq.     Percent        Cum.
------------+-----------------------------------
   nonunion |      1,417       75.45       75.45
      union |        461       24.55      100.00
------------+-----------------------------------
      Total |      1,878      100.00
```

- Proportion = 461/1878 = 0.2455
- Std Dev = $\sqrt{0.2455 * (1 - 0.2455)} = 0.4304$

Example: Proportion unionised

```
. sysuse nlsw88
(NLSW, 1988 extract)

. tab union

      union |
     worker |      Freq.     Percent        Cum.
------------+-----------------------------------
   nonunion |      1,417       75.45       75.45
      union |        461       24.55      100.00
------------+-----------------------------------
      Total |      1,878      100.00
```

- Proportion = 461/1878 = 0.2455
- Std Dev = $\sqrt{0.2455 * (1 - 0.2455)} = 0.4304$
- Std Err = $\frac{0.4304}{\sqrt{1878}} = 0.0099$

Example: Proportion unionised

```
. sysuse nlsw88
(NLSW, 1988 extract)

. tab union

      union |
     worker |      Freq.     Percent        Cum.
------------+-----------------------------------
   nonunion |      1,417       75.45       75.45
      union |        461       24.55      100.00
------------+-----------------------------------
      Total |      1,878      100.00
```

- Proportion = 461/1878 = 0.2455
- Std Dev = $\sqrt{0.2455 * (1 - 0.2455)} = 0.4304$
- Std Err = $\frac{0.4304}{\sqrt{1878}} = 0.0099$
- Half width = 1.96*0.0099 = 0.0195

Example: Proportion unionised

```
. sysuse nlsw88
(NLSW, 1988 extract)

. tab union

      union |
     worker |      Freq.     Percent        Cum.
------------+-----------------------------------
   nonunion |      1,417       75.45       75.45
      union |        461       24.55      100.00
------------+-----------------------------------
      Total |      1,878      100.00
```

- Proportion = 461/1878 = 0.2455
- Std Dev = $\sqrt{0.2455 * (1 - 0.2455)} = 0.4304$
- Std Err = $\frac{0.4304}{\sqrt{1878}} = 0.0099$
- Half width = 1.96*0.0099 = 0.0195
- Low: 0.2455 - 0.0195 = 0.2260
- High: 0.2455 + 0.0195 =0.2650

Summary

- CLT: Sampling distribution of a sample statistic is normally distributed
  - Centred on true value, standard deviation is SE = $\frac{\sigma}{\sqrt{n}}$
- 95% of samples fall in range $\mu \pm SE \times 1.96$
- 95% Confidence interval: $\bar{X} \pm \hat{SE} \times 1.96$
  - "95% confident the true value lies in this interval"
- 99% Confidence interval: $\bar{X} \pm \hat{SE} \times 2.575$
- Similar process from proportions where SE is $\sigma_\pi = \sqrt{\pi \times (1 - \pi)}$

$$\hat{\pi} \pm z_{0.95} \times \hat{SE}_\pi$$

- A way of presenting a sample statistic with a measure of its precision

Today

- The χ^2 test for association in tables (chi-square)
- The t-distribution: replaces the normal for small samples

Unit 11:

Samples and tables

Association in tables

- When we make a table, we read it for association by looking at percentages
- Either the row percentages, up and down the columns
- Or the column percentages across the rows
- (If one variable happens before or causes the other, percentages within that variable are easier to interpret)

ESS: selected countries and attitude, row %

```
. tab cntry freehms, row nokey
```

Country	Gays and lesbians free to live life as they wish: Agree str	Agree	Neither a	Disagree	Disagree	Total
DE	1,135	926	144	78	56	2,339
	48.53	39.59	6.16	3.33	2.39	100.00
FR	1,329	450	119	47	49	1,994
	66.65	22.57	5.97	2.36	2.46	100.00
GB	1,057	884	176	49	28	2,194
	48.18	40.29	8.02	2.23	1.28	100.00
IE	1,001	970	126	53	34	2,184
	45.83	44.41	5.77	2.43	1.56	100.00
Total	4,522	3,230	565	227	167	8,711
	51.91	37.08	6.49	2.61	1.92	100.00

Source: European Social Survey, Wave 9 (2018/9)

ESS: selected countries and attitude, column %

```
. tab cntry freehms, col nokey
```

Country	Gays and lesbians free to live life as they wish: Agree str	Agree	Neither a	Disagree	Disagree	Total
DE	1,135	926	144	78	56	2,339
	25.10	28.67	25.49	34.36	33.53	26.85
FR	1,329	450	119	47	49	1,994
	29.39	13.93	21.06	20.70	29.34	22.89
GB	1,057	884	176	49	28	2,194
	23.37	27.37	31.15	21.59	16.77	25.19
IE	1,001	970	126	53	34	2,184
	22.14	30.03	22.30	23.35	20.36	25.07
Total	4,522	3,230	565	227	167	8,711
	100.00	100.00	100.00	100.00	100.00	100.00

Detecting association

- If these percentages differ from each other (or from the row/col total percentages) we say there is association
- The distribution of one variable depends on the other value of the other
- But how big do differences need to be before we say it's a meaningful level of association?
- NB: A sample from a population with no association will have small percentage differences just by chance!

Fake data with no association, row %

```
. tab c1 freehms, row chi nokey
```

c1	Gays and lesbians free to live life as they wish: Agree str	Agree	Neither a	Disagree	Disagree	Total
DE	1,206	889	147	57	40	2,339
	51.56	38.01	6.28	2.44	1.71	100.00
FR	992	755	145	55	47	1,994
	49.75	37.86	7.27	2.76	2.36	100.00
GB	1,146	797	146	63	42	2,194
	52.23	36.33	6.65	2.87	1.91	100.00
IE	1,178	789	127	52	38	2,184
	53.94	36.13	5.82	2.38	1.74	100.00
Total	4,522	3,230	565	227	167	8,711
	51.91	37.08	6.49	2.61	1.92	100.00

```
Pearson chi2(12) =  13.3692   Pr = 0.343
```


Problem: how small is small

- A table that logically cannot have a link between nationality and attitude has small percentage differences just by chance
- We need a rule to decide
  - how small is small enough to say there is no association?
  - how big is big enough to count as evidence of association?
- Answer: the χ^2 test (chi-square)

Unit 11:

The χ^2 test

The χ^2 test for independence in tables

- The χ^2 test for association in tables
- Independence: no association between two variables
  - pattern of row percentages the same in all rows
  - pattern of column percentages the same in all columns
- Even if independence holds in the population, sampling variability leads to differences in percentages
- How big can the differences be before we can be convinced that there is really association in the population?

Compare "observed" with "expected"

- Method: compare the real table ("observed") with hypothetical table under independence ("expected")
- Summarise the difference into a single figure (χ^2 statistic, chi-sq)
- Compare χ^2 statistic with known distribution
- . . . What is the probability of getting a sample statistic "at least this big" by simple sampling variability *if independence holds in the population*?

"Expected" ⇒"independence"

- The "expected" table has the same row and column totals, but the cell values are such that the percentages are the same as in the total row and column:

$$n_{ij} = \frac{R_i C_j}{T}$$

- For each cell we summarise the difference between observed (O) and expected (E) values as

$$\frac{(O-E)^2}{E}$$

- The summary for the table as a whole is the sum of this quantity across all cells:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Illustration

Demonstration with spreadsheet

X^2 has a χ^2 distribution

- This statistic has a known distribution, the χ^2 distribution
- That is, if we take a large number of samples from a population where there is no association, and calculate the statistic, they will have a distribution in a known form, and we can calculate the probability of finding a value "at least as large as" any given number
- Depends only on the "degrees of freedom": number of rows minus one times the number of columns minus one:

$$df = (r-1)(c-1)$$

χ^2 distribution

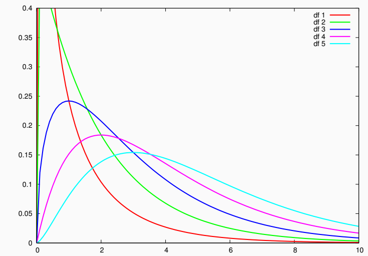


Figure 1: The χ^2 Distribution for various degrees of freedom

App

This app draws the χ^2 distribution for different degrees of freedom, and calculates the proportion of the distribution above a given χ^2 statistic.

`https://teaching.sociology.ul.ie/apps/chidist`

With Stata

```
. tab cntry freehms, row nokey chi
```

Country	Agree str	Agree	Neither a	Disagree	Disagree	Total
	Gays and lesbians free to live life as they wish					
DE	1,135 48.53	926 39.59	144 6.16	78 3.33	56 2.39	2,339 100.00
FR	1,329 66.65	450 22.57	119 5.97	47 2.36	49 2.46	1,994 100.00
GB	1,057 48.18	884 40.29	176 8.02	49 2.23	28 1.28	2,194 100.00
IE	1,001 45.83	970 44.41	126 5.77	53 2.43	34 1.56	2,184 100.00
Total	4,522 51.91	3,230 37.08	565 6.49	227 2.61	167 1.92	8,711 100.00

Pearson chi2(12) = 294.6550   Pr = 0.000

Unit 11:

The t-distribution

The t-distribution

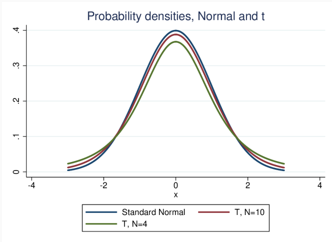
- We have used the Standard Normal Distribution to make confidence intervals around sample means and sample proportions
- The depends on the Central Limit Theorem:

  A sufficiently large sample drawn from any population will have a normal distribution, even when the population distribution is not normal

- In practice, for small samples the approximation to the normal distribution is not good, so we use a related distribution called the t-distribution

Student's *t*

- The t-distribution is like the normal distribution but wider
- The bigger the sample the closer it is to the normal distribution
- Small samples give much fatter tails
- It makes the CI a little wider for smaller sample sizes



Student's identity

- It is known as "Student's t", after the Guinness statistician who invented it
- William Sealey Gosset (1876–1937)
- Guinness were wary of losing trade secrets so publication of research was not allowed
- Gosset published anonymously, as "Student"

Sample size matters: degrees of freedom

- If the sample is small and we are calculating, say, a mean, the normal distribution underestimates the uncertainty
- That is, repeated samples from the same population will vary more widely
- The true sampling distribution of the mean is wider
  - depends on degrees of freedom which is N-1

Online web applets

- See web applets:
  - Standard Normal Distribution: `https://teaching.sociology.ul.ie/apps/snd`
  - t-Distribution: `https://teaching.sociology.ul.ie/apps/tdist`
- More spread at small N, more like normal distribution at high N
- Use the app to see P above/below a t value, or inside/outside ± t

Worked example

- Let's use the same example as in lecture 12, transport spending:
  - sample mean: €34.658
  - sample standard deviation: €10.123
- But this time, let's say the sample was 20
- This makes the standard error bigger:

$$\frac{10.123}{\sqrt{20}} = 2.264$$

Calculating the interval

- CI: $\bar{X} \pm t_{0.95,19} \times SE$
- We can use the web-app to find the t-value for 95% confidence
- DF = 20 - 1 = 19

Calculating the interval

- CI: $\bar{X} \pm t_{0.95,19} \times SE$
- We can use the web-app to find the t-value for 95% confidence
- DF = 20 - 1 = 19

- t = 2.093
- Interval:

$$34.658 \pm 2.093 \times 2.264$$

$$29.920 \leftarrow 34.658 \rightarrow 39.396$$

A wider interval

- A much wider interval than in Lecture 12, but that is mostly because of the small sample size making the standard error much bigger
- However, it is also a wider interval than if we used the SND
  - Correct:

$$29.920 \leftarrow 34.658 \rightarrow 39.396$$

  - Using 1.96 instead of 2.093 (incorrect)

$$30.221 \leftarrow 34.658 \rightarrow 39.095$$

Why it matters

- The interpretation of a 95% confidence interval is:
  - For 95% of samples, the true value will fall in the interval
  - We say we are 95% confident the true value falls in the interval
- If we use z = 1.96, our interval is too narrow and the true value will fall outside the interval more than 5% of the time

When to use the t-Distribution

- If working by hand (e.g., doing an exam question):
  - If N is above about 60-100, use the Normal Distribution
  - Otherwise use the t-distribution
- If using a computer, always use t
  - If N is small, it is more correct
  - If N is large, it makes no difference
- Stats programs (Stata, R, SPSS, etc) will always use t
- Note t is relevant only for quantities like means
  - Don't use for proportions

Unit 11:

Spreadsheet exercise

Spreadsheet exercise

- We now use a spreadsheet to generate a confidence interval step by step
- We will calculate the mean and the standard deviation for 16 observations
- Then calculate the standard error and create a confidence interval

Data

16 observations have been collected as follows:

36	54	25	54
84	44	98	19
57	27	97	51
60	13	68	81

Formulas

- Mean: $\bar{X} = \frac{\sum X_i}{N}$
- Standard deviation: $s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{N-1}}$
- Standard error: $se = \frac{s}{\sqrt{N}}$
- CI: $\bar{X} \pm t \times se$

- Three important topics today:
  - Hypothesis testing
  - Significance
  - *t*-test for paired samples

Unit 12: Hypothesis testing

Hypothesis Testing

Confidence intervals assess imprecision

- Confidence intervals allow us to present sample information appropriately
  - Point estimate, e.g., mean or other sample statistic: our "best guess" of the true value
  - Confidence interval: range in which we are "confident" true value (the population parameter) lies
- In combination, the point estimate and CI give us the answer with a measure of its precision

Hypothesis testing and CIs

- Statistical inference proceeds by hypothesis testing: a formalisation of the Confidence Interval approach
- If we wish to test whether a variable has an effect on another (e.g., does a switch to a 4-day week change productivity) we set up a hypothesis, for instance
- If we wish to test whether a variable has an effect on another (e.g., does a switch to a 4-day week change productivity) we set up a hypothesis, for instance
  H_1: *Productivity after is not the same as productivity before,* $X_a \neq X_b$
- We then assess whether the data support the hypothesis

Recent example

- Microsoft recently experimented with a 4-day week in Tokyo:

https://www.theguardian.com/technology/2019/nov/04/microsoft-japan-four-day-work-week-productivity

- They found productivity increased, as well as worker satisfaction

Null hypothesis

- To test this turn it around, and set up a null hypothesis that says the opposite: H_0: *Productivity after is equal to productivity before,* $X_a = X_b$
- If we can *reject the null hypothesis* on the basis of our sample data, we can say the data support the main hypothesis

Claims about population

- Hypothesis testing is a way of using the reasoning behind CIs to make specific claims about the population
- Say we want to know if there is a relationship between two variables, e.g., whether switching to a 4-day week has an effect on productivity (positive or negative)
- We look at a sample of workers to make inferences about the population
- We begin with a hypothesis: $X_a \neq X_b$ or $X_a > X_b$
- We negate the hypothesis to form a null hypothesis, called H_0: $H_0 : X_a = X_b$ which is equivalent to $H_0 : D_x = X_a - X_b = 0$
- That is, on average, the productivity difference is zero

Testing the null hypothesis

- We then test the null hypothesis:
  - First we calculate a sample mean productivity difference, $\hat{D}_x$
  - Then we construct a confidence interval at our chosen level of confidence (e.g., 95%): $\hat{D}_x \pm z_{0.95} \times SE$
  - The CI gives us a band around the point estimate within which we are 95% sure the true value lies
  - If zero lies outside the interval, we are at least 95% sure the true population value is not zero, and we can reject the null hypothesis
  - If zero lies within the interval, then zero is in the range of plausible values, so we cannot reject the null hypothesis
  - We don't say we "accept the null hypothesis" because zero is just in the range of plausible values, and other values in this range are approximately as likely

Reject or fail to reject

- Rejecting the null hypothesis constitutes support for the initial or "alternative" hypothesis
- Failing to reject the null hypothesis means the data fail to support the initial hypothesis: "there is no evidence that the switch to a 4-day week affects productivity"
- Failure to support the initial hypothesis may be because
  - It is actually false, i.e., $D_x = 0$
  - The effect is small and/or very variable, and thus the sample is too small to detect it

Example: Productivity before and after an intervention

ID	Before	After
1	3.5	4.6
2	1.8	2.3
3	2.4	2.5
4	3.3	4.4
5	1.7	2.1
6	3.7	5.1
7	4.4	5.6
8	3.4	4.1
9	1.8	1.4
10	2.3	1.7

http://teaching.sociology.ul.ie/so5041/unit12.csv

How do we "test" the null hypothesis?

- In this example we calculate the difference in productivity before and after, in the sample data
- Some may be negative, some may be positive, but we are interested in the average: is it systematically different from zero?
- Strategy: calculate the mean of the differences, and construct a CI around it (say at 95%)
- If zero lies outside the CI, then we are at least 95% sure the true value lies in a range that does not include zero
- If zero within the CI, then the range within which we think the true value lies does include zero

Reject or not?

- In the former case (zero outside interval), we can reject the null hypothesis: we are at least 95% sure that zero is not the true value
- We can therefore say the data support the initial hypothesis that the switch to a 4-day week affects productivity
- In the latter case (interval contains zero), we cannot reject the null hypothesis: zero is in the range where we feel the true value lies
- In this case there is no evidence in the sample data of an effect of the 4-day week on productivity
- This is not the same as evidence of no effect!
- That zero lies within the CI is not the same as zero being the true value!

Summary

- Extension of Confidence Intervals to answer questions: hypothesis testing
- Negate the initial hypothesis to create a "null" hypothesis
- Look at the data: would it be likely if the null hypothesis were true?
- Make a CI around the sample statistic: does it include the null value?
- If no, the null is unlikely to be true: reject, support initial hypothesis
- If yes, the null may be true: fail to reject, fail to support initial hypothesis

Unit 12: Hypothesis testing

Statistical significance

Statistical significance

- Significance: let's say we do a hypothesis test with a 95% confidence level, and we find the zero is way outside the CI
- We can try again with a 99% confidence level:
  - If it is still outside the interval we are not "at least 95%" but "at least 99%" sure that zero is not the true value

Keep looking

- If we keep trying with CIs with higher confidence levels we will eventually find one where zero is just outside the interval
- If that is at confidence level *C* we can say that we are *C*% sure (not "at least" any more) that zero is not the true value

App: confidence intervals and significance

https://teaching.sociology.ul.ie/apps/cislide

The chance of being wrong?

- There is then a C% chance that the true value is in the range that doesn't include zero, or a 100% - C% chance that the true value is outside the CI, and therefore could include zero
- This p = 100%-C% value is the probability that we get a sample statistic as different from zero as we did, even though the true value was zero
- This is known as the significance of the sample estimate, or its p-value

The p-value

- We want this p-value to be as small as possible, typically under 5% (0.05)
- Using p<5% as a threshold is the same as using 95% confidence
- p-values are widely used – stats programs report them in many places
- In general the interpretation is "what's the probability of getting this result by chance if the null hypothesis was true?"
- Looking at the exact p-value can be more interesting than yes/no on whether zero is inside the CI

t-test

- Rather than use the CI we can set this up as a "t-test"
- We can find the *t* corresponding to the CI just touching zero thus:

$$t = \frac{\bar{X}}{SE}$$

- If that *t* is larger than the critical value, then the CI using the critical value is smaller and doesn't overlap zero
- The significance is the exact p-value of that *t*
- This example is a "paired sample *t*-test"

t-test example in Stata

```
. gen diff = after-before

. ttest diff == 0

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    diff |      10    .5499999    .2166667    .6851601    .0598659    1.040134
------------------------------------------------------------------------------
    mean = mean(diff)                                             t =   2.5385
Ho: mean = 0                                     degrees of freedom =        9

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 0.9841         Pr(|T| > |t|) = 0.0318          Pr(T > t) = 0.0159
```


t-test – paired

```
. ttest before==after

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  before |      10        2.83     .299648      .94757    2.152149    3.507851
   after |      10        3.38    .4859812    1.536808    2.280634    4.479366
---------+--------------------------------------------------------------------
    diff |      10   -.5499999    .2166667    .6851601   -1.040134   -.0598659
------------------------------------------------------------------------------
     mean(diff) = mean(before - after)                            t =  -2.5385
 Ho: mean(diff) = 0                              degrees of freedom =        9

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.0159         Pr(|T| > |t|) = 0.0318          Pr(T > t) = 0.9841
```

χ^2 test and significance

- Another example of significance occurs in the χ^2 (chi-sq) test for association in a table
- Here the initial hypothesis is that the two variables are associated
- Thus the null hypothesis is that they are not associated (the "independence hypothesis")
- When we calculate the χ^2 statistic ($\sum \frac{(O-E)^2}{E}$) we compare its value with the range of possible values we would get if H_0 were true
- This is what we read from the table of the χ^2 distribution

χ^2 example in Stata

```
. tab empstat sex, chi

                      |  school leavers sex
   usual employment   |        {s11}
   situation {s81}    |      male     female |     Total
----------------------+----------------------+----------
 working for payment  |       388        380 |       768
          unemployed  |        67         46 |       113
 looking for 1st job  |       170        151 |       321
             student  |       471        490 |       961
               other  |         8         26 |        34
----------------------+----------------------+----------
               Total  |     1,104      1,093 |     2,197

          Pearson chi2(4) =  14.9610   Pr = 0.005
```


Signficance and error

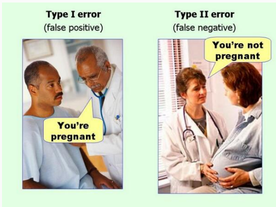
- Another way of looking at significance is "the chance of seeing a pattern as strong as this if the null hypothesis is true"
- For instance, if there is one chance in twenty ($p = 0.05$) that the true value is outside the CI, then by basing our decision on the CI we will be wrong one time in twenty (if the null is true)
- The actual chance of being wrong also depends on the chance the null hypothesis is true

Type I error

- This is known as Type I Error: rejecting the null hypothesis when it is true
  - e.g., the true value might be zero but a small number of possible samples generate CIs that don't include zero
  - e.g., there may be no association but a small number of possible samples yield high χ^2 statistics
- The risk of such False Positive error is 5% times the (unknown) probability of the null being true

Type I and Type II error

- If very important to avoid Type I error, use high confidence levels (e.g., 99.5% instead of 95%) or insist on low p-values (e.g., 0.005 instead of 0.05)
- However, there is a second type of error, Type II
  - Failing to reject the null hypothesis when it is false
- That is, failing to support the initial hypothesis even though it is true



Type I and Type II

- If we raise the confidence level we reduce the risk of Type I error but raise the risk of Type II error
- That is, if we make a special effort not to accept an initial hypothesis unless there is very clear evidence, we necessarily fail to accept it where there is only fairly clear evidence
- For a given p-value, we can only reduce the Type II error by increasing the sample size

Summary

- From confidence intervals to test statistics
- From test-statistics to p-values : the probability of observing an effect this strong if the null hypothesis is true
- General approach, for testing means, association in tables, and lots of other measures

The basic t-test: sample versus reference-point

- The simplest t-test compares a sample mean against a fixed number:

$$H_0 : \mu = X_r$$

$$t = \frac{\bar{X} - X_r}{SE}$$

- If t is bigger than the critical value for 95%, or its p-value is below 0.05, we reject the null hypothesis

Paired-sample t-test

- Comparing "paired samples" is the same, with the difference being compared with zero:

$$D = X_{after} - X_{before}$$

$$H_0 : \delta = 0$$

$$t = \frac{\bar{D} - 0}{SE}$$

“t-test” for a proportion

- Comparing a sample proportion against a fixed number such as 50% has a similar logic
- It doesn’t use the t-distribution, but can use standard normal for “large” samples

$$H_0 : \pi = \pi_r$$

$$z = \frac{p - \pi_r}{SE}$$

Example: referendum

- For an upcoming referendum, 1000 voters are polled, and 542 say they will vote yes (no don’t knows)
- p = 0.542
- SD = $\sqrt{p * (1 - p)} = 0.498$
- SE = 0.0158
- Test statistic:

$$\frac{0.542 - 0.500}{0.0158} = 2.67$$

Unit 13: More t-tests

Directional hypotheses

Directional hypotheses

- A complication: some hypotheses are directional
- e.g., holidays make you *happier*, training *raises* your earning power

$$H_1 : W_{after} > W_{before}$$

$$H_o : W_{after} \leq W_{before}$$

$$H_o : D \leq 0$$

Direction ⇒ one-way t-test

- If zero is *below* the CI, reject the null hypothesis
- If zero is within or above the CI, cannot reject
- Net result: higher confidence for the same CI – 1.96 $\times$ *SE* for 97.5% not 95%

Unit 13: More t-tests

Comparing means across groups

Comparing means across groups

- We don’t always have situations where we want to test something as simple as whether the true answer is zero
- However, we very often want to test whether a mean (e.g., income) is different according to values of another variable (e.g., sex)
- If sex affects wage, we would expect to see the mean wage for men (X_m) to be different from the mean wage for women (X_w)
- We can consider the sample difference ($\bar{X}_m - \bar{X}_w$) to be a point estimate of the population difference
- The null hypothesis is that $X_m = X_w$ or $X_m - X_w = 0$

Independent-samples t-test

- To construct a CI we need the SE, which is very like the normal one if both groups have the same population variance (or standard deviation):

$$\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- If this cannot be assumed, the SE is more complex, and depends on the separate standard deviations

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Two-sample t-test in Stata

```
. ttest grsearn, by(sex)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    male |     953    25.89192    1.584105    48.90243    22.78318    29.00066
  female |     978    27.92025    1.541657    48.21223     24.8949    30.94559
---------+--------------------------------------------------------------------
combined |   1,931    26.91921    1.104885     48.5521    24.75232    29.08611
---------+--------------------------------------------------------------------
    diff |           -2.028325    2.210045               -6.362652    2.306002
------------------------------------------------------------------------------
    diff = mean(male) - mean(female)                              t =  -0.9178
Ho: diff = 0                                     degrees of freedom =     1929

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.1794         Pr(|T| > |t|) = 0.3589          Pr(T > t) = 0.8206
```

Two-sample t-test, unequal variance

```
. ttest grsearn, by(sex) unequal

Two-sample t test with unequal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    male |     953    25.89192    1.584105    48.90243    22.78318    29.00066
  female |     978    27.92025    1.541657    48.21223     24.8949    30.94559
---------+--------------------------------------------------------------------
combined |   1,931    26.91921    1.104885     48.5521    24.75232    29.08611
---------+--------------------------------------------------------------------
    diff |           -2.028325    2.210451               -6.363455    2.306805
------------------------------------------------------------------------------
    diff = mean(male) - mean(female)                              t =  -0.9176
Ho: diff = 0                     Satterthwaite´s degrees of freedom =   1925.9

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.1795         Pr(|T| > |t|) = 0.3589          Pr(T > t) = 0.8205
```


Summary: Key concepts

- Hypothesis test
  - Null hypothesis
  - Initial or alternative hypothesis
- Types I and II error
- Statistical significance and p-values
- t-tests:
  - Single value compared with reference
  - Paired values compared with implicit zero reference
  - Independent samples t-test: comparing two groups
    - Equal vs unequal variance

Unit 13: More t-tests

Summarising inference

Summarising inference

- For large samples, we use the normal distribution to construct confidence intervals around means of "quantitative" (interval/ratio) variables
- For small samples we use the t-distribution to construct the confidence interval
- For interval/ratio variables we usually estimate the mean, and use

$$SE = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{\sqrt{\frac{\sum(X-\bar{X})^2}{n-1}}}{\sqrt{n}}$$

Summarising inference: proportions

- For nominal variables like vote, sex, etc. we calculate proportions, not means (where we split in two)
- With large samples (at least 20 in each category) we can construct confidence intervals using the normal distribution and the formula $\sigma_{\hat{\pi}} = \sqrt{\frac{p(1-p)}{n}}$ for the standard error
- With this we can conduct hypothesis tests in exactly the same way as with interval/ratio data
- However, with small samples the approximation to the normal distribution no longer holds and we have to use another distribution, the binomial distribution

Summarising inference: χ^2 test

- For nominal variables with more categories, and for tables made from nominal variables, we can use the χ^2 test
- Again, this has "large sample" requirements – few of the expected values should be < 5
- If some rows & columns have low values, combined expected values will be very low: collapse these rows and columns into other categories

Surveys and Questionnaires

- The main way of collecting survey data is by means of questionnaires and structured interviews
- There are several important issues relating to the design and implementation of questionnaire-based surveys
- Reading:
  - Bryman Chs 9 and 10
  - DA de Vaus, *Surveys in Social Research*, chapter 6 ("Constructing Questionnaires")

Questionnaires

- The questionnaire is the backbone – the script for the interview
- Two main types
  - In structured interviewing it is sometimes referred to as the interview schedule
  - Questionnaires filled out by the respondent him/herself are referred

  to as self-completion questionnaires
- The practical purpose of a questionnaire survey is to get standard info from a large number of people relatively cheaply and reliably

Purpose of questionnaire

- Must minimise of ambiguity in the meaning of the recorded answers – need to be clear and precise
- Also minimise non-response
  - Item non-response: refusal to answer or "don't know" to specific questions
  - Respondent refusal: flat refusal to participate
- Hence good design is important:
  - Clarity of questions and structure
  - No unnecessary burden on the respondent

Unit 14: Surveys and Questionnaires

Questionnaire Design

Questionnaire Design

- There are two aspects to questionnaire design
  - The overall structure of the document
  - The individual questions

Composing questions

- Questions must be clear and unambiguous
  - Express the questions in clear and simple language
  - Don't use leading questions (Avoid "Isn't it the case that X is a good idea?"; prefer "Do you think X is a good idea or a bad idea?")
  - Ask a single thing at a time (Avoid "Do you have a job, and if so how much do you earn?")

Composing questions: direct information

- Avoid hypothetical questions: often don't give useful information (e.g., "If you had a grant how much more money would you spend on drink?")
- If you need to ask for information about others, restrict it to factual information ("What is your partner's job?", but not "What does your partner think about X?")

Composing questions: easy to answer

- Make the questions easy to answer:
  - provide a set of options (perhaps on a prompt card)
  - with amounts (time, money etc.) provide options help to precision (e.g., times per week, ranges)

Document structure

- The document must be structured simply and logically
  - Group questions in a way that will seem reasonable to the respondent
  - Use clear routing:

    5: Do you have a job?
    (if yes, ask qn 6, else skip to qn 7)
    6: What is your job?
    7: . . .

- Reserve sensitive questions (e.g., income, drug use) to the end of the questionnaire: less likely to scare off respondents there

Question structure

- The structure of questions is also important
- Each question must have its answer space: a tick-box, a space for writing, a set of categories
- Closed questions are preferred: a fixed set of options makes it easier to answer and to analyse
- Always allow the category "Other" with closed questions, with space to write ("If other, please specify:")

Types of closed question

- Closed responses can take several forms:
  - One only: "Tick the category that best describes your job"
  - One or more: "Tick all of the following reasons that are relevant"
  - Ranking: "Rank the importance for you of each of the following reasons (1, 2, 3 etc.)"

Likert scales

- Measurement of attitudes often uses "Likert" scales:
  - a set of statements relevant to the attitude being measured
  - with options indicating agreement:
    - Strongly disagree
    - Disagree
    - Neither agree nor disagree
    - Agree
    - Strongly agree

Practical organisation

- The practical organisation of the questionnaire must take question structure into account: easy to answer and easy to process
- Processing involves several steps
  - Recording the information as it is being collected
  - Checking it is consistent and the right questions are answered
  - Coding it onto a computer
  - Checking it is consistent once on computer
  - Adding variable and value labels to help the computer data set to make sense

Design principles

- The design of the questionnaire should keep the first three of these in mind
  - It should be easy to record the information
  - It should be easy to read the recorded information to see that it makes sense, to see that the correct routing has been followed
  - The layout of the questionnaire should anticipate the structure of the computer data set

Example questionnaire

Example Questionnaire Extract Office use only

1: Sex: Male ☐1 Female ☐2 ☐q1

2: Age: ☐q2

18 or under	☐	1
19 – 23	☐	2
24 – 30	☐	3
31 – 40	☐	4
41 – 50	☐	5
51 – 64	☐	6
65 or more	☐	7

Many options, tick one

3: Which of the following options best describes your current situation? (show card): ☐q3

Self-employed	☐	1
Employed	☐	2
Unemployed	☐	3
Retired	☐	4
Family care	☐	6
Full time student, school	☐	7
Long-term sick, disabled	☐	8
Training scheme	☐	9
Other	☐	10

If "Other" please specify:____________________

Ordinal options (with a flaw)

4: How often do you read newspapers? Tick the category that is closest: ☐q4

Never	☐	1	(Skip to question 6)
More than 1 per day	☐	2	
1 every day	☐	3	
2–4 per week	☐	4	
1 per week	☐	5	
Less than 1 per week	☐	6	

Picking zero or more from a list

5: Which of the following newspapers do you read? (show card) Tick each one that applies:

Irish Times	☐ a		☐ q5a
Irish Independent	☐ b		☐ q5b
Irish Examiner	☐ c		☐ q5c
Limerick Leader	☐ d		☐ q5d
Sunday Independent	☐ e		☐ q5e
etc. etc.	☐ f		☐ q5f
Other	☐ g		☐ q5g

If "Other" please specify:____________________

Attitude questions with Likert answers

6: People have different views about women's role in society. Please indicate whether you agree or disagree with each of the following things people might say, using the categories provided:

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree	
A woman's place is in the home	☐1	☐2	☐3	☐4	☐5	☐q6a
Young children suffer if the mother works outside the home	☐1	☐2	☐3	☐4	☐5	☐q6b
Women are entitled to the same pay for the same work as men	☐1	☐2	☐3	☐4	☐5	☐q6c
When a wife works it is important for the husband to do his share of the housework	☐1	☐2	☐3	☐4	☐5	☐q6d

Pilot your questionnaire

- It is very important that a questionnaire work well once in the field
- Test it beforehand,
  - on friends/colleagues
  - on a "pilot" subsample drawn from the reference population
- Are the questions comprehensible?
- Are the categories in closed questions adequate? Do they cover everything? Make the right distinctions? No very unlikely ones? Will they make sense to the respondent?

Piloting

- Is the structure okay? Not too complicated or illogical?
- Is there anything likely to confuse the respondent or interviewer?
- Exactly how long will it take? If too long, shorten it now!
- Piloting will also allow you to create closed categories for questions, using real-world feedback rather than your imagination

Example survey

https://teaching.sociology.ul.ie/apps/so4046/survey/

Remaining topics

- From today we look at two related techniques for interval and ratio variables:
  - Correlation: a single measure of association for interval/ratio variables
  - Linear Regression: a very powerful technique for describing the relationship between one interval/ratio variable and one or more others

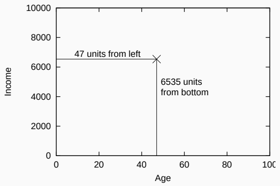
Unit 15: Correlation

Measures of association for interval/ratio variables

Scatterplots

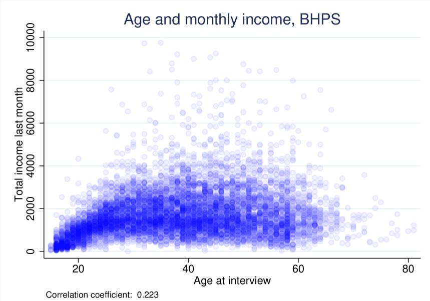
- Scatterplots can be considered as interval/ratio analogue of cross-tabs: arbitrarily many values mapped out in 2-dimensions

Figure 1: Age 47, income €6,535.



Age vs Income

Figure 2: Age and income, Wave 15 BHPS.

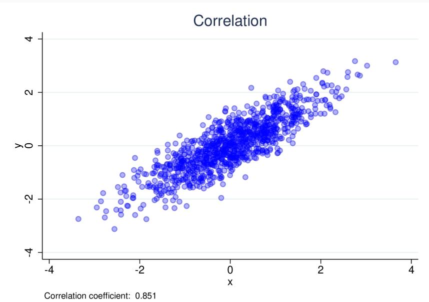


Summarising association simply

- We can see a lot of detail in a scatterplot, but sometimes we can summarise it in simple ways
- For instance the two variables may have a positive association: when one is high the other tends to be high, and vice versa
- Or a negative association: when one is high the other tends to be low

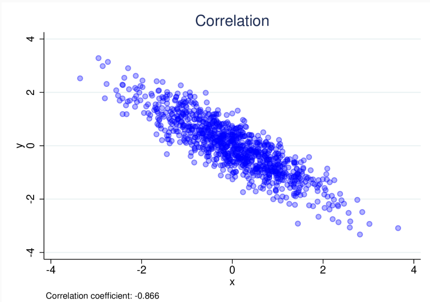
Strong positive

Figure 3: Fictional data displaying a strong positive linear relationship.



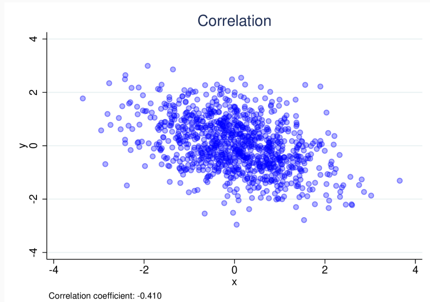
Strong negative

Figure 4: Fictional data displaying a strong negative linear relationship.



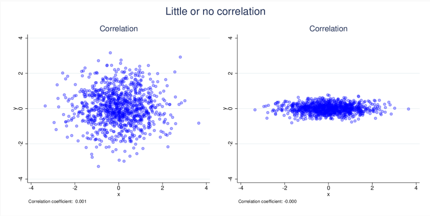
Weak negative

Figure 5: Fictional data displaying a weak negative linear relationship.



No relationship

Figure 6: Fictional data displaying absence of a relationship.



App

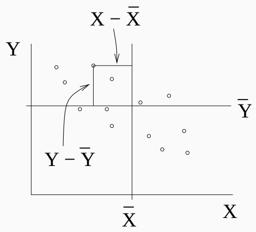
https://teaching.sociology.ul.ie/apps/corrgame

The correlation coefficient

- How does it work?
- Combines X deviations $(X_i - \bar{X})$ and Y deviations $(Y_i - \bar{Y})$ – i.e., compares each point with the mean for X and the mean for Y
- With positive association, cases below average on X tend to be below average on Y (and above average on X tend to be above average on Y)
- With negative association, cases below average on X tend to be above average on Y and vice versa

Deviations

Figure 7: X and Y deviations for correlation



Combining deviations

- With positive association below the mean, both $X_i - \bar{X}$ and $Y_i - \bar{Y}$ are negative, so $(X_i - \bar{X})(Y_i - \bar{Y})$ is positive
- With negative association, $X_i - \bar{X}$ and $Y_i - \bar{Y}$ tend to have opposite signs, so $(X_i - \bar{X})(Y_i - \bar{Y})$ is negative

Pearson Product-Moment Correlation

Coefficient

- Pearson Product-Moment Correlation Coefficient (r)

$$r = \frac{SXY}{\sqrt{SXX.SYY}}$$

$$SXX = \Sigma(X - \bar{X})^2$$

$$SYY = \Sigma(Y - \bar{Y})^2$$

$$SXY = \Sigma(X - \bar{X})(Y - \bar{Y})$$

- Range: $-1 \leq r \leq +1$
- r is a symmetric measure: $r_{xy} = r_{yx}$

Robust to transformations

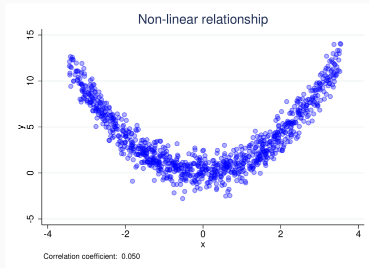
- Changing the scale of one variable (additively or multiplicatively) doesn't change the results
  - Corr(x, y) = Corr(x + c, y)
  - Corr(x, y) = Corr(x * c, y)
- "Scale invariant"
- Suits both interval and ratio variables

Pitfalls

- Correlation is not causality
- Absence of correlation is not absence of relationship: non-linearity

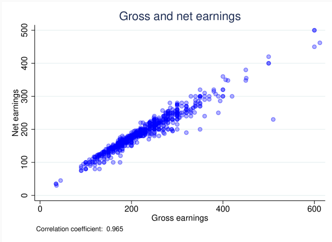
Non-linear!

Figure 8: A strong non-linear relationship and a near-zero correlation.



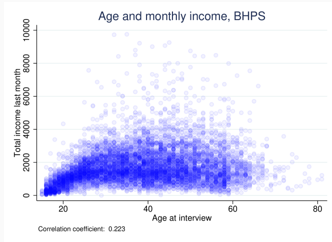
Correlations on data: School Leaver's Earnings

```
. corr grsearn netearn
(obs=756)

             | grsearn  netearn
-------------+------------------
     grsearn |  1.0000
     netearn |  0.9650   1.0000
```



Correlations on data: BHPS

```
. corr ofimn oage
(obs=7,934)

             |   ofimn     oage
-------------+------------------
       ofimn |  1.0000
        oage |  0.2228   1.0000
```



Strong and weak

- There is a strong simple mechanism linking gross and net earnings, leading to a very high correlation of 0.965
- The relationship between age and income is much more complex, but is still real: thus a much lower correlation of 0.223

Hypothesis testing

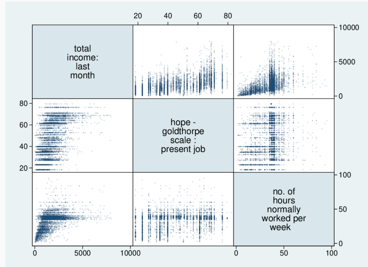
- Null hypothesis: no association, correlation = 0
- Test statistic: $\frac{r}{SE}$, normally distributed (SE not in output)
- P-value: Chance of getting a correlation this far from zero if null is true

```
. pwcorr ofimn oage, sig

             |   ofimn     oage
-------------+------------------
       ofimn |  1.0000
             |
             |
        oage |  0.2228   1.0000
             |  0.0000
             |
```


Estimating correlations in Stata

```
. pwcorr ofimn ojbhgs ojbhrs, sig

             |   ofimn   ojbhgs   ojbhrs
-------------+---------------------------
       ofimn |  1.0000
             |
             |
      ojbhgs |  0.4851   1.0000
             |  0.0000
             |
      ojbhrs |  0.4245   0.2489   1.0000
             |  0.0000   0.0000
             |
```


Viewing the same correlations



Summary

- Correlation: summarising straight-line association between two interval/ratio variables
- Range: -1 (perfect negative) through 0 (no association) to 1 (perfect positive)
- Beware pitfalls
  - "correlation isn't causation"
  - absence of linear association isn't absence of association

Outline

- Moving on from the idea of correlation: relating two (or more) interval/ratio variables
- Identifying the line that best summarises the scatterplot
- Directional: X predicts Y
- Predictive: Given the relationship between X and Y, knowing X helps us predict Y better
- Bivariate regression: one X variable predicting one Y
- Multiple regression: multiple X variables predicting one Y
- Reading: Agresti Ch 9

Unit 16: Regression

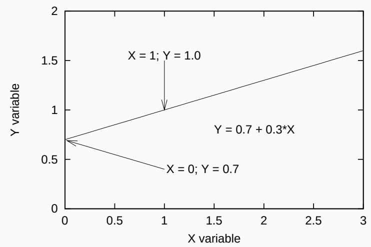
Bivariate Linear Regression

Bivariate regression analysis

- Regression Analysis: Fitting the “best” line through the scatter
- Very closely related to correlation, but treats one variable as dependent and the other(s) as explanatory, while correlation is asymmetric

Some geometry: equation of a line

$$Y = a + bX$$



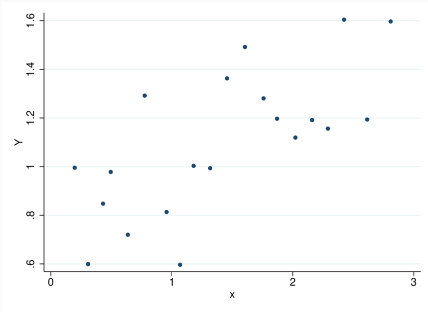
Applet

https://teaching.sociology.ul.ie/apps/abx/

Predictive in intent

- Asymmetric: use X to predict Y
- Find the best *a* and *b* to summarise the data scatter:
  - ‘Best’ is defined as minimising the squared deviations between the observed data-points and the fitted line, hence often called ‘least-squares’ regression
  - Deviations are the vertical distance between the line and the observed data points.
  - Very similar logic to the mean (minimise variance).

A simple example: scatterplot



Regression in Stata

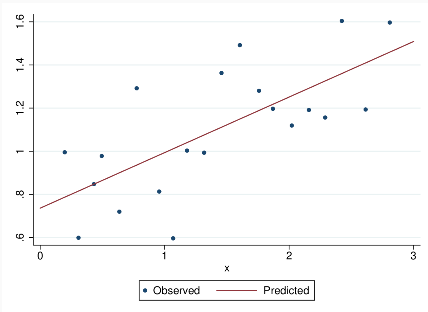
```
. reg y x

      Source |       SS           df       MS      Number of obs   =        20
-------------+----------------------------------   F(1, 18)        =     17.51
       Model |  .820567701         1  .820567701   Prob > F        =    0.0006
    Residual |  .843474028        18  .046859668   R-squared       =    0.4931
-------------+----------------------------------   Adj R-squared   =    0.4650
       Total |  1.66404173        19  .087581144   Root MSE        =    .21647

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .2574678   .0615269     4.18   0.001     .1282045    .3867311
       _cons |   .7363586   .0998061     7.38   0.000     .5266737    .9460435
------------------------------------------------------------------------------
```


Regression equation

$$\hat{Y} = 0.7364 + X \times 0.2575$$

- To draw the line by hand, calculate two predicted values for Y at opposite sides of the graph
  - e.g., for X = 0, Y = a = 0.7364
  - for X = 3, Y = a + 3*b = 0.7364 + 3*0.2575 = 1.509
- Join them with a ruler!

The line



Predicted values

- The line gives a predicted value of Y for each value of X:

$$\hat{Y} = a + bX$$

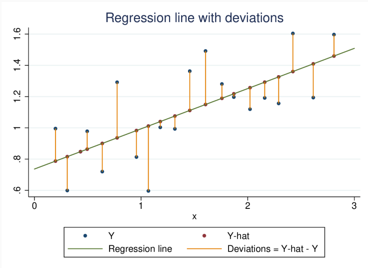
$$Y = \hat{Y} + e$$

$$Y = a + bX + e$$

e is the ‘residual’ or deviation.

- That is, knowing X we “predict” or guess Y as $a + bX$

Deviations from the line



Regression equation

- Regression equation: the estimate of Y, called $\hat{Y}$, depends on X:

$$\hat{Y} = a + bX$$

- The regression slope b depends on SXY and SXX, the intercept a is calculated from b and the mean values of Y and X:

$$b = \frac{SXY}{SXX}$$

$$a = \bar{Y} - b\bar{X}$$

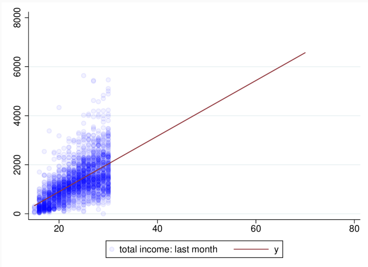
$$SXX = \sum (X_i - \bar{X})^2$$

$$SXY = \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

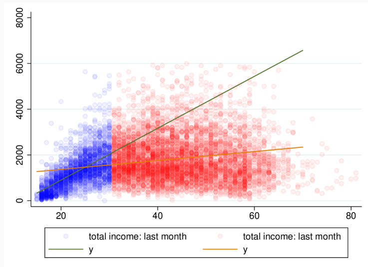
Pitfalls

- Spurious relationships will fit just as well as real ones (e.g., if A affects B and A affects C, B and C will seem to be related and a regression line might fit well)
- Predicting outside the range of the data: the relationship we see only holds for the data we use, and it may well not hold for higher (or lower) values of X and Y
- Like correlation, non-linear relationships may be missed

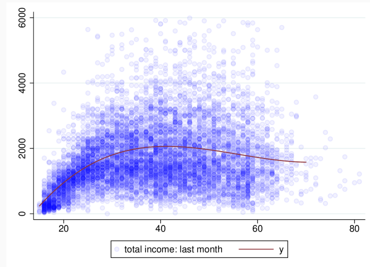
Predicting outside the range of the data: income and age for <30 years



Predicting outside the range of the data: full age range



Income and age: true relationship is non-linear



Fit

- How well does it "fit"? We use R^2 to tell:
  - ranges from 0: no relationship at all
  - to 1: perfect relationship, all Ys are exactly equal to $a + bX$
  - values from 0.7 up indicate quite a good relationship
  - smaller values may indicate an interesting relationship
- In the case of bivariate regression (one independent variable), R^2 is the same as $r \times r$ (squared correlation coefficient).

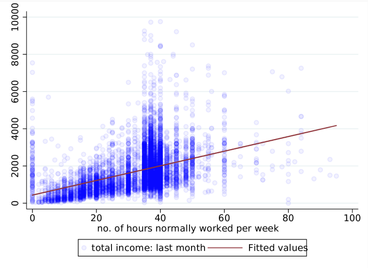
Hypothesis testing

- If we regress Y on X we are asking whether X has a (statistical) effect on Y
- The null is therefore that X has no "effect" on Y
- This is equivalent to a slope, β, of zero: $\hat{Y} = \alpha + 0 \times X$
- In any sample from a population where this is true, b, the estimate of β is likely to be close to, but not exactly zero
- We use its SE to create a t-stat: $t = \frac{\hat{\beta}}{\mathrm{SE}}$
- DF = n - k + 1 where k in number of X variables

Regression in Stata:

```
. reg ofimn ojbhrs

      Source |       SS           df       MS      Number of obs   =     7,945
-------------+----------------------------------   F(1, 7943)      =   1398.95
       Model |  1.7000e+09         1  1.7000e+09   Prob > F        =    0.0000
    Residual |  9.6522e+09     7,943   1215179.2   R-squared       =    0.1497
-------------+----------------------------------   Adj R-squared   =    0.1496
       Total |  1.1352e+10     7,944  1429021.17   Root MSE        =    1102.4

------------------------------------------------------------------------------
       ofimn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ojbhrs |   39.34202   1.051854    37.40   0.000     37.28011    41.40393
       _cons |   434.7389    36.8029    11.81   0.000     362.5955    506.8822
------------------------------------------------------------------------------
```

Predicted regression line



Unit 16: Regression

Multiple Regression

Multiple explanatory variables

- Regression analysis can be extended to the case where there is more than one explanatory variable – multiple regression
- This allows us to estimate the net simultaneous effect of many variables, and thus to begin to disentangle more complex relationships
- Interpretation is relatively easy: each variable gets its own slope coefficient, standard error and significance
- The slope coefficient is the effect on the dependent variable of a 1 unit change in the explanatory variable, *while taking account of the other variables*

Example

- Example: domestic work time may be affected by gender, and also by paid work time: competing explanations – one or the other, or both could have effects
- We can fit bivariate regressions:

$$DWT = a + b \times PaidWork$$

or

$$DWT = a + b \times Female$$

- We can also fit a single multiple regression

$$DWT = a + b \times PaidWork + c \times Female$$

Dichotomous variables

- We deal with gender in a special way: this is a *binary* or *dichotomous* variable – has two values
- We turn it into a yes/no or 0/1 variable – e.g., female or not
- If we put this in as an explanatory variable a *one unit change in the explanatory variable* is the difference between being male and female
- Thus the *c* coefficient we get in the $DWT = a + b \times PaidWork + c \times Female$ regression is the net change in predicted domestic work time for females, once you take account of paid work time.
- The *b* coefficient is then the net effect of a unit change in paid work time, once you take gender into account.

Sex only predicting income

```
. reg ofimn i.osex

      Source |       SS           df       MS      Number of obs   =     7,945
-------------+----------------------------------   F(1, 7943)      =    606.72
       Model |   805586626         1   805586626   Prob > F        =    0.0000
    Residual |  1.0547e+10     7,943  1327780.12   R-squared       =    0.0710
-------------+----------------------------------   Adj R-squared   =    0.0708
       Total |  1.1352e+10     7,944  1429021.17   Root MSE        =    1152.3

------------------------------------------------------------------------------
       ofimn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        osex |
     female  |  -637.3352   25.87467   -24.63   0.000    -688.0563    -586.614
       _cons |   2062.275   18.64855   110.59   0.000     2025.719    2098.831
------------------------------------------------------------------------------
```


Sex and job hours predicting income

```
. reg ofimn ojbhrs i.osex

      Source |       SS           df       MS      Number of obs   =     7,945
-------------+----------------------------------   F(2, 7942)      =    794.96
       Model |  1.8935e+09         2   946761687   Prob > F        =    0.0000
    Residual |  9.4586e+09     7,942  1190962.07   R-squared       =    0.1668
-------------+----------------------------------   Adj R-squared   =    0.1666
       Total |  1.1352e+10     7,944  1429021.17   Root MSE        =    1091.3

------------------------------------------------------------------------------
       ofimn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ojbhrs |   33.96065   1.123629    30.22   0.000     31.75804    36.16326
             |
        osex |
     female  |  -337.0889   26.44232   -12.75   0.000    -388.9228    -285.255
       _cons |   787.1759   45.73595    17.21   0.000     697.5214    876.8304
------------------------------------------------------------------------------
```


Sex and hours