

Quantitative Research Methods

Week 2: Descriptive Statistics Measurements and Scales

Overview

- Recap from previous session
- Descriptive statistics
- Frequencies
- Averages
- Normal distribution
- Standard deviation
- Cross tabulations
- Checking significance: chi square
- Coding scales to make sure they match

Recap

- What do we use data analysis for?
- How do we translate everything into numbers?
- What is the difference between a dependent and independent variable?
- What do we know about sampling?
- What is statistical significance? How do we measure it?
- What is the difference between a nominal, ordinal and interval variable?

Descriptive statistics

- Descriptive statistics **illustrate information** provided by the **data**
- They are not inferential ie **do not predict or explain** more than is evident from the data
- They use existing data to **identify patterns and relationships**
- Presenting descriptive statistics:
 - Bar: plot summaries for groups of cases or separate variables
 - Pie charts: plot frequencies and distributions
 - Histograms: frequencies (distribution) of different scores
 - Frequency tables: frequency (no. of cases) of each value within a variable
 - Crosstabulations: show whether there is a relationship between 2 variables (bivariate analysis)
 - Chi- square: measures degree of association between 2 variables (ie whether there is a significant relationship between the 2 variables)

Frequencies

- Frequencies can be displayed in a number of ways and can be very useful
- Most statistics that we see are frequencies (**often expressed as percentages**): “11% of the electorate voted for this party”; “8 out of 10 cats” etc
- Looking at **frequencies can also allow us to see the distribution** of a group
 - **Eg** if we are looking at the age of respondents, we can see if there is an even **spread** or whether most of the respondents are concentrated within a particular age group
- This can be **important for establishing how reliable the sample is**; how reflective it is of the population; how other findings may be affected by the spread of cases within a particular variable; etc

Frequencies – example

Which party did u vote for in G.ElecQ216

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Conservative	735	20.3	26.4	26.4
	Labour	1367	37.8	49.1	75.5
	Liberal Democrat	459	12.7	16.5	92.0
	SNP	132	3.7	4.7	96.7
	Plaid Cymru	11	.3	.4	97.1
	Green Party	7	.2	.3	97.4
	Other Party	19	.5	.7	98.1
	Referendum Party	54	1.5	1.9	100.0
	Total	2784	77.0	100.0	
Missing	Skp,didnt vote	766	21.2		
	Refused to disclose voting	62	1.7		
	Not answered	3	.1		
	Total	831	23.0		
Total		3615	100.0		

Averages

- Mean, median or mode?
- Mean: add all the values together, divide by number of cases
- Mode: the value which occurs most frequently
- Median: rank cases in order of values. Median is the midpoint

Numerical descriptions

Let y denote a quantitative variable, with observations $y_1, y_2, y_3, \dots, y_n$

a. Describing the *centre*

Median: Middle measurement of ordered sample

Mean:
$$\bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n} = \frac{\sum y_i}{n}$$

Example: Annual per capita carbon dioxide emissions (metric tons) for $n = 8$ largest nations in population size

Bangladesh 0.3, Brazil 1.8, China 2.3, India 1.2, Indonesia 1.4, Pakistan 0.7, Russia 9.9, U.S. 20.1

Ordered sample:

Median =

$\bar{y}$

Mean =

Example: Annual per capita carbon dioxide emissions (metric tons) for $n = 8$ largest nations in population size

Bangladesh 0.3, Brazil 1.8, China 2.3, India 1.2, Indonesia 1.4, Pakistan 0.7, Russia 9.9, U.S. 20.1

Ordered sample: 0.3, 0.7, 1.2, 1.4, 1.8, 2.3, 9.9, 20.1

Median =

$\bar{y}$

Mean =

Example: Annual per capita carbon dioxide emissions (metric tons) for $n = 8$ largest nations in population size

Bangladesh 0.3, Brazil 1.8, China 2.3, India 1.2, Indonesia 1.4, Pakistan 0.7, Russia 9.9, U.S. 20.1

Ordered sample: 0.3, 0.7, 1.2, 1.4, 1.8, 2.3, 9.9, 20.1

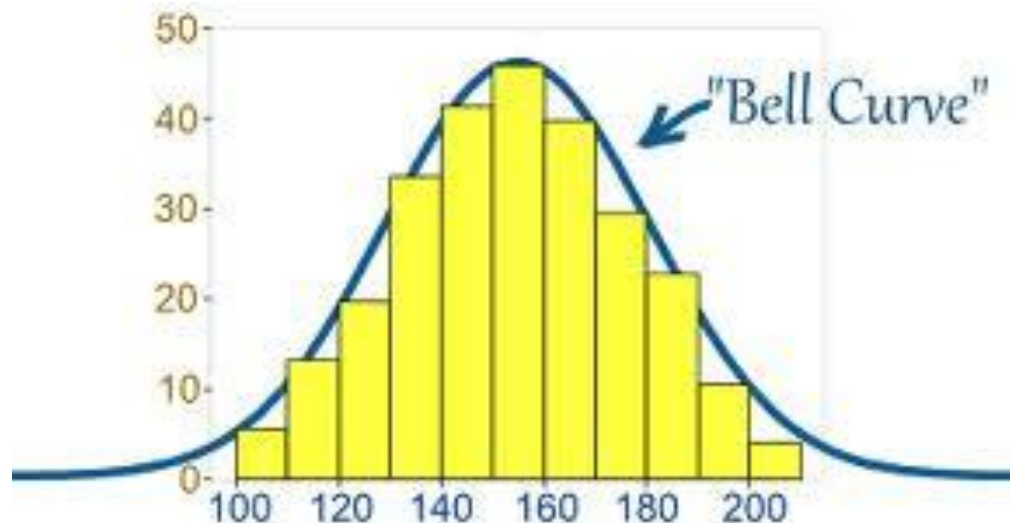
Median = $(1.4 + 1.8)/2 = 1.6$

$\bar{y}$

Mean = $(0.3 + 0.7 + 1.2 + \dots + 20.1)/8 = 4.7$

Normal distribution

- A normal distribution is when cases are distributed symmetrically around the mean, tailing off on either side – sometimes known as a “bell-shaped curve”.



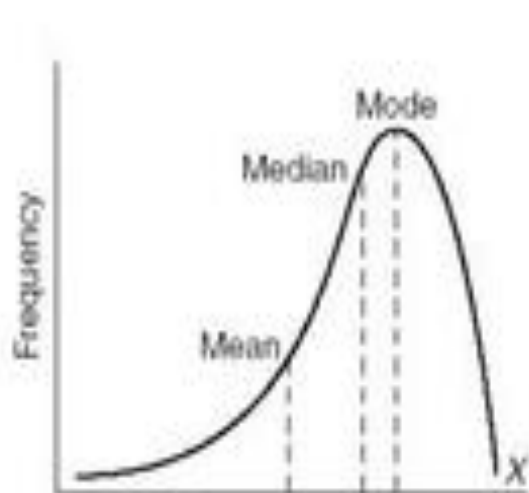
Normal distribution (cont)

- Sometimes the **mean can be misleading** if the cases don't follow this "normal" pattern
- **Eg** cases might be **clustered around the extremes**, with only a few cases near the midpoint
- Or there may be quite a few "**outliers**" – cases that do not fit the general trend
- Or the data may be **skewed on one side or the other** (positively or negatively skewed)
 - **Eg** on a value ranging from 1-20, the mean might be 8 and the median might be 7 but the mode might be 2, if cases were clustered in the lower range with a few outliers in the higher range to boost the mean

Normal distribution (cont)

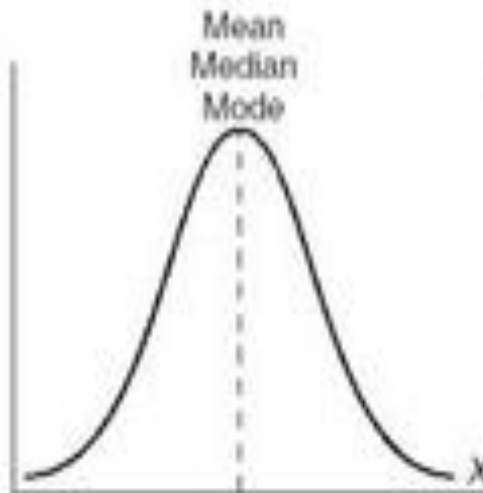
- Skewness is a measure of symmetry (or the lack there of). A distribution, or data set, is symmetrical if it looks the same to the left and right of the centre point.
- Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. Data sets with high kurtosis have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak. A uniform distribution would be the extreme case
- The easiest way to spot these trends is by using a histogram

(a) Negatively skewed



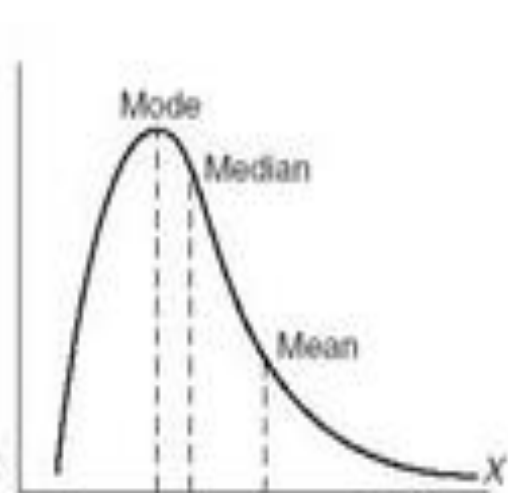
← Negative direction

(b) Normal (no skew)



The normal curve represents a perfectly symmetrical distribution

(c) Positively skewed



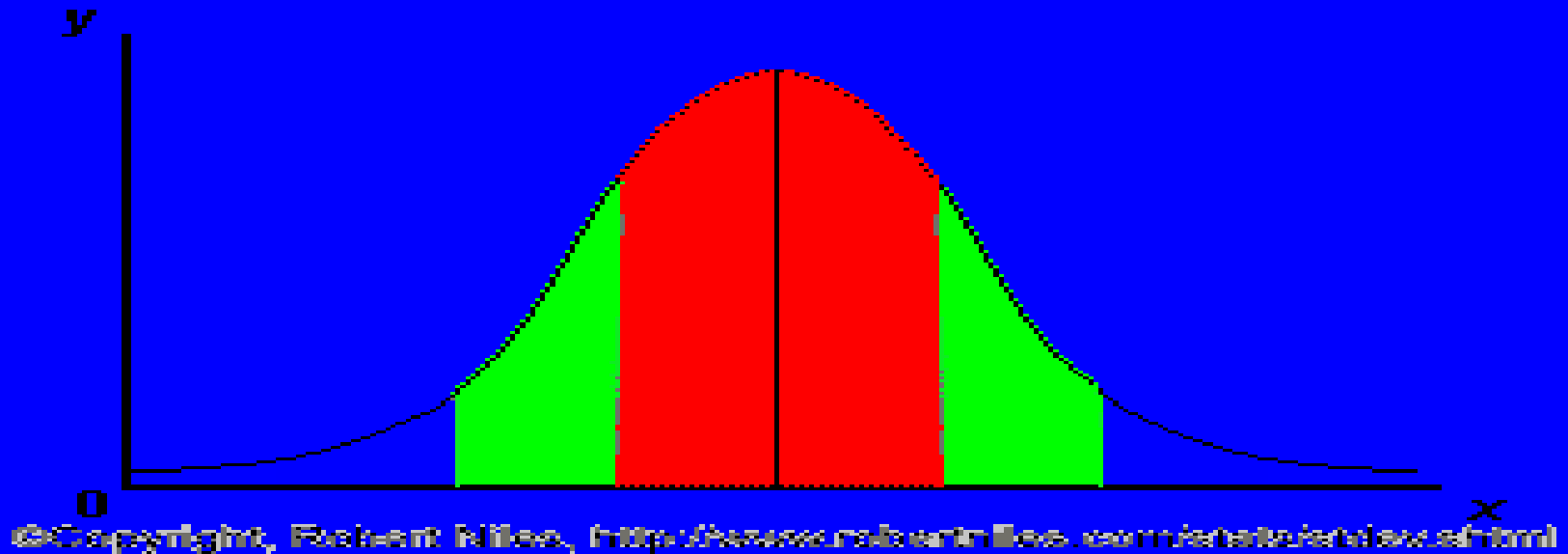
→ Positive direction

Examples of normal and skewed distributions

Standard deviation

- **Standard deviation** (SD) measures the **spread** of data about the mean value
- It is useful in comparing variables which have the same mean but different ranges
 - **Eg** the mean of these sets is the same: {15, 15, 15, 14, 16} and {2, 7, 14, 22, 30}. But the second set is much more spread out.
- **SD** is the **square root** of the variance from the mean
- If a variable has a **low SD**, the values **are not spread out too much**
- An **important** attribute of SD is that **if the mean and SD of a normal distribution are known**, we can measure the percentile rank for any given score. In a normal distribution, c. **68% of the scores are within one SD** of the mean and c. 95% of the scores are within two SDs of the mean.

Standard deviation (cont)



- The sections highlighted in **red** are within **one standard deviation** from the mean (c.68% of cases).
- The sections highlighted in **green** are within **two standard deviations** (95% of cases)
- The sections highlighted in **blue** are within **three standard deviations** (99% of cases)
- The wider the spread of cases, the further out you have to go to reach three standard deviations

Standard deviation: A “typical” distance from the mean

The *deviation* of observation i from the mean is

$$y_i - \bar{y}$$

The *variance* of the n observations is

$$s^2 = \frac{\Sigma(y_i - \bar{y})^2}{n-1} = \frac{(y_1 - \bar{y})^2 + \dots + (y_n - \bar{y})^2}{n-1}$$

The standard deviation s is the square root of the variance,

$$s = \sqrt{s^2}$$

Example: Political ideology

- For those in the student sample who attend religious services at least once a week ($n = 9$ of the 60),
- $y = 2, 3, 7, 5, 6, 7, 5, 6, 4$

$$\bar{y} = 5.0,$$

$$s^2 = \frac{(2-5)^2 + (3-5)^2 + \dots + (4-5)^2}{9-1} = \frac{24}{8} = 3.0$$

$$s = \sqrt{3.0} = 1.7$$

For entire sample ($n = 60$), mean = 3.0, standard deviation = 1.6, tends to have similar variability but be more liberal

Cross-tabulations

- **Cross-tabulations**

allow you to compare and analyse two variables at the same time to spot relationships and interactions

- **Example**

from my research: looking at results from the 1997 French elections. Which party had the highest proportion of women MPs?

Party which won in 1997 * Sex of winning candidate in 1997 Crosstabulation

Count		Sex of winning candidate in 1997		Total
		Male	Female	
Party which won in 1997	RPR	130	6	136
	UDF	101	6	107
	DVD	12	0	12
	PS	204	39	243
	PCF	32	4	36
	Verts	4	2	6
	DVG	33	3	36
	FN	1	0	1
Total		517	60	577

Cross-tabulations

- **Cross-tabs** can also be used to look at **more than two variables** at the same time (you can experiment with this)
- You can **split the data** to compare groups eg by sex so that every other test you do shows results for both groups
 - **Eg** if you were interested in the interaction between income and vote, you could also see whether this varies by sex
- You can ask SPSS to produce statistics such as percentages – these are reported in the output.

Cross-tabulations

- As you become more advanced you will be able to combine information gathered from descriptive statistics to illustrate more complex examples
- It is important when writing up your results to include the following:
 - A title (what are you showing?)
 - The total n
 - Any tests you have done to show whether the findings are significant
 - Percentages to facilitate interpretation of the results. Pay attention
 - do you need percentages for rows, columns or both?

Presentable Crosstabs

Party of vote by sex in the 1997 BES

Sex	Party of Vote			Total
	Conservative	Labour	Lib-Dem	
Male	28%	52%	20%	100%
Female	32%	50%	18%	100%

N=1677

When constructing two way tables the independent variable is typically arranged on the vertical axis and the dependent variable on the horizontal axis. In this case party of vote is the dependent variable. You must always include a title and the total percentages.

Could this table be improved? How?

Checking significance: chi square

- **The statistic most commonly used** when looking at **nominal data** is the chi square
-  This is the chi square symbol. Chi square is a non-parametric test. It **relies on weaker assumptions about the distribution** of the data than many other tests.
- **In order to calculate a Chi-square (which you don't need to do for this course) you work out the expected frequencies for each category if no relationship existed and compare these figures to the actual observations**

Checking significance: chi square

- Thus in our example of sex and vote we would expect the same proportion of men and women to vote for each party.

- **if there is no relationship.**

The **Chi square** for the example of sex and party of vote (Conservative, Labour, Lib Dem & other) **is 0.226**. This shows us that **there is a 22.6% chance** that there is no association between our variables in the population

- **Chi-square helps to determine whether there is a statistically significant relationship between two nominal variables**
- Good news: it is easy to calculate on SPSS and easy to interpret! 😊

Scales: do they match?

- Sometimes questionnaires may use a series of questions to understand a person's position on an issue
- **Eg** instead of (or as well as) asking someone directly if they are left or right wing, it might ask their positions on a range of economic and social issues
- If these issues are used to form an overall impression, it is important to make sure they are all coded the same way
 - **Eg** On a scale of 1-10 where 1 is agree strongly and 10 is disagree strongly, people might evaluate these two statements:
 - “Governments should tax the rich to help the poor”
 - “All government assets should be privatised”
 - If you agree with the first statement, you are probably left-wing; if you agree with the second statement, you are probably right-wing
 - So it may be necessary to recode the variables so that 1 is the most left-wing position on all questions and 10 is the most right-wing position

Summary

- **Descriptive** statistics give us basic (but very useful) information about our data
- It is important to present descriptive statistics correctly
- We can use **frequencies** to see if our results are **normally distributed**
- **The standard deviation** gives us an indication of how widely the results are spread out
- **Cross-tabulations** allow us to measure relationships between variables
- We can use the **chi-square test** to see if our results are significant
- Scales should be coded in the same direction if we wish to make meaningful comparisons between them

Measurements and Scales

Measurement

- ***Variable*** – a characteristic that can vary in value among subjects in a sample or a population.

Types of variables

- ***Categorical*** (also called *qualitative*)
- ***Quantitative***

(There are different statistical methods for each type)

Categorical variable – scale for measurement is a set of categories

Examples:

Racial-ethnic group (white, black, Hispanic)

Political party identification (Dem., Repub., Indep.)

Vegetarian? (yes, no)

Mental health evaluation (well, mild symptom formation, moderate symptom formation, impaired)

Happiness (very happy, pretty happy, not too happy)

Religious affiliation

Occupation

Quantitative variable – possible values differ in magnitude

Examples:

Age, height, weight, BMI = $\text{weight(kg)}/[\text{height(m)}]^2$

Annual income

GDP

Time spent studying for PhD

Number of times you presented your paper in conferences

Scales of measurement

For *categorical variables*, two types:

- ***Nominal scale***
 - unordered categories
 - Preference for President, Race, Gender,
 - Religious affiliation, Opinion items (favor vs. oppose, yes vs. no)
- ***Ordinal scale***
 - ordered categories
 - Political ideology (very liberal, liberal, moderate, conservative, very conservative)
 - Country income groups (high, middle, low)
 - Mental impairment (none, mild, moderate, severe)
 - Government spending on edu. (up, same, down)

For *quantitative variables*, set of possible values is called an ***interval scale***. (i.e., numerical interval between each possible pair of values)

Note: In practice, *ordinal* categorical variables often treated as *interval* by assigning scores

(e.g., Grades A,B,C,D,E an ordinal scale, but treated as interval if assign scores 4,3,2,1,0 to construct a GPA)

Another classification:

Discrete/Continuous

Discrete variable

– possible values a set of separate numbers, such as 0, 1, 2, ...

Example: Number of ...
e-mail messages sent in previous day

Continuous variable

– infinite continuum of possible values

Example: Amount of time spent on Internet in previous day

(In practice, distinction often blurry - discrete variables taking *many* possible values (e.g., age) are treated as continuous)

What type of variable is

1. No. of movies seen this summer
2. Favorite music, type of
3. Happiness
4. Economic growth rate (in percentage)
5. Job satisfaction
6. Salary
7. Highest degree attained
8. Student ID no.