## Descriptives \& Graphing <br> 

## Lecture 3

Survey Research \& Design in Psychology James Neill, 2017 Creative Commons Attribution 4.0

## Overview:

## Descriptives \& Graphing

1. Getting to know a data set
2. LOM \& types of statistics
3. Descriptive statistics
4. Normal distribution
5. Non-normal distributions
6. Effect of skew on central tendency
7. Principles of graphing
8. Univariate graphical techniques

## Getting to know <br> a data-set <br> (how to approach data)

$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$


Play with the data get to know it.

$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$


$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$


$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$

$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$


## Level of measurement \& types of statistics



## Golden rule of data analysis

A variable's level of measurement determines the type of statistics that can be used, including types of:

- descriptive statistics
- graphs
- inferential statistics


## Levels of measurement and

 non-parametric vs. parametricCategorical \& ordinal data DV
$\rightarrow$ non-parametric
(Does not assume a normal distribution)
Interval \& ratio data DV
$\rightarrow$ parametric
(Assumes a normal distribution)
$\rightarrow$ non-parametric
(If distribution is non-normal)
DVs = dependent variables
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\square$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$

## Parametric statistics

- Statistics which estimate parameters of a population, based on the normal distribution
-Univariate:
mean, standard deviation, skewness, kurtosis, $t$-tests, ANOVAs
-Bivariate:
correlation, linear regression
-Multivariate:
multiple linear regression


## Parametric statistics

- More powerful
(more sensitive)
- More assumptions (population is normally distributed)
- Vulnerable to violations of assumptions (less robust)


## Non-parametric statistics

- Statistics which do not assume sampling from a population which is normally distributed
-There are non-parametric alternatives for many parametric statistics
-e.g., sign test, chi-squared, MannWhitney U test, Wilcoxon matched-pairs signed-ranks test.
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$


## Non-parametric statistics

- Less powerful (less sensitive)
- Fewer assumptions (do not assume a normal distribution)
- Less vulnerable to assumption violation (more robust)


## Univariate descriptive statistics

## Number of variables

| Univariate | mean, median, mode, <br> histogram, bar chart |
| :--- | :--- | = one variable histogram, bar chart

## Bivariate

= two variables
correlation, $t$-test,
Multivariate scatterplot, clustered bar chart
= more than two variables
reliability analysis, factor analysis, multiple linear regression

## What do we want to describe?

The distributional properties of variables, based on:

- Central tendency(ies): e.g., frequencies, mode, median, mean
- Shape: e.g., skewness, kurtosis
- Spread (dispersion): min., max., range, IQR, percentiles, variance, standard deviation


## Measures of central tendency

Statistics which represent the
'centre' of a frequency distribution:

- Mode (most frequent)
- Median (50 th percentile)
- Mean (average)

Which ones to use depends on:
-Type of data (level of measurement)
-Shape of distribution (esp. skewness)
Reporting more than one may be appropriate.

## Measures of central tendency

|  | Mode / <br> Freq. /\%s | Median | Mean |
| :--- | :---: | :---: | :---: |
| Nominal | $\sqrt{ }$ | $x$ | $x$ |
| Ordinal | $\sqrt{ }$ | If meaningful | $x$ |
| Interval | $\sqrt{ }$ | $\sqrt{ }$ | $\sqrt{ }$ |
| Ratio | If meaningful | $\sqrt{ }$ | $\sqrt{ }$ |

## Measures of distribution

- Measures of shape, spread, dispersion, and deviation from the central tendency


## Non-parametric: Parametric:

- Min. and max.
- SD
- Range
- Skewness
- Percentiles - Kurtosis

Measures of spread / dispersion / deviation

|  | Min / Max, <br> Range | Percentile | Var/SD |
| :--- | :---: | :---: | :---: |
| Nominal | $x$ | $x$ | $x$ |
| Ordinal | $\sqrt{ }$ | If meaningful | $x$ |
| Interval | $\sqrt{ }$ | $\sqrt{ }$ | $\sqrt{ }$ |
| Ratio | $\sqrt{ }$ | $\sqrt{ }$ | $\sqrt{ }$ |

## Descriptives for nominal data

- Nominal LOM = Labelled categories
- Descriptive statistics:
-Most frequent? (Mode - e.g., females)
-Least frequent? (e.g., Males)
-Frequencies (e.g., 20 females, 10 males)
-Percentages (e.g. $67 \%$ females, $33 \%$ males)
-Cumulative percentages
-Ratios (e.g., twice as many females as males)
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$都
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$


## Descriptives for ordinal data

- Ordinal LOM = Conveys order but not distance (e.g., ranks)
- Descriptives approach is as for nominal (frequencies, mode etc.)
- Plus percentiles (including median) may be useful


## Descriptives for interval data

- Interval LOM = order and distance, but no true 0 ( 0 is arbitrary).
- Central tendency (mode, median, mean)
- Shape/Spread (min., max., range, $S D$, skewness, kurtosis)
Interval data is discrete, but is often treated as ratio/continuous (especially for > 5 intervals)


## Descriptives for ratio data

- Ratio = Numbers convey order and distance, meaningful 0 point
- As for interval, use median, mean, $S D$, skewness etc.
- Can also use ratios (e.g., Category A is twice as large as Category B)


## Mode (Mo)

- Most common score - highest point in a frequency distribution - a real score - the most common response
- Suitable for all levels of data, but may not be appropriate for ratio (continuous)
- Not affected by outliers
- Check frequencies and bar graph to see whether it is an accurate and useful statistic


## Frequencies ( $f$ ) and percentages (\%)

- \# of responses in each category
- \% of responses in each category
- Frequency table
- Visualise using a bar or pie chart


## Median (Mdn)

- Mid-point of distribution (Quartile 2, $50^{\text {th }}$ percentile)
- Not badly affected by outliers
- May not represent the central tendency in skewed data
- If the Median is useful, then consider what other percentiles may also be worth reporting
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$


## Summary: Descriptive statistics

- Level of measurement and normality determines whether data can be treated as parametric
- Describe the central tendency
-Frequencies, Percentages
-Mode, Median, Mean
- Describe the variability:
-Min., Max., Range, Quartiles
-Standard Deviation, Variance

$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$

Four moments of a normal distribution

$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$

## Four moments of a

 normal distributionFour mathematical qualities (parameters) can describe a continuous distribution which at least roughly follows a bell curve shape:

- $1^{\text {st }}=$ mean (central tendency)
- $2^{\text {nd }}=S D$ (dispersion)
- $3^{\text {rd }}=$ skewness (lean / tail)
- $4^{\text {th }}=$ kurtosis (peakedness / flattness)


## Mean <br> (1st moment)

- Average score

Mean $=\Sigma \mathrm{X} / \mathrm{N}$

- For normally distributed ratio or interval (if treating it as continuous) data.
- Influenced by extreme scores (outliers)


## Beware inappropriate averaging...

With your head in an oven
and your feet in ice

you would feel,
on average,
just fine
The majority of people have more than the average number of legs

$$
(M=1.9999) .
$$

## Standard deviation (2nd moment)

- $S D=$ square root of the variance

$$
=\Sigma \frac{(X-X)^{2}}{N-1}
$$

- For normally distributed interval or ratio data
- Affected by outliers
- Can also derive the Standard Error (SE) = SD / square root of $N$


## Skewness

(3rd moment )

- Lean of distribution
$-+v e=$ tail to right
- -ve = tail to left
- Can be caused by an outlier, or ceiling or floor effects
- Can be accurate (e.g., cars owned per person would have a skewed distribution)


## Kurtosis (4th moment)

- Flatness or peakedness of distribution

$$
\begin{aligned}
& +\mathrm{ve}=\text { peaked } \\
& \text {-ve }=\text { flattened }
\end{aligned}
$$

- By altering the X \&/or Y axis, any distribution can be made to look more peaked or flat - add a normal curve to help judge kurtosis visually.


Image source: http:///classconnection.s3.amazonaws.com/65/flashcards/2185065/ipg/kurtosis-142C1127AF2178FB244.jpg 44

## Judging severity of skewness \& kurtosis

- View histogram with normal curve
- Deal with outliers
- Rule of thumb:

Skewness and kurtosis $>-1$ or $<1$ is generally considered to sufficiently normal for meeting the assumptions of parametric inferential statistics

- Significance tests of skewness:

Tend to be overly sensitive (therefore avoid using)

## Areas under the normal curve

If distribution is normal (bell-shaped - or close):
$\sim 68 \%$ of scores within +/- $1 S D$ of $M$ $\sim 95 \%$ of scores within $+/-2 S D$ of $M$
$\sim 99.7 \%$ of scores within $+/-3 S D$ of $M$

Areas under the normal curve


## Non-normal distributions

## Types of non-normal distribution

- Modality
-Uni-modal (one peak)
-Bi-modal (two peaks)
-Multi-modal (more than two peaks)
- Skewness
-Positive (tail to right)
-Negative (tail to left)
- Kurtosis
-Platykurtic (Flat)
-Leptokurtic (Peaked)

Histogram of people's weight

$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$

Histogram of daily calorie intake


Histogram of fertility


Fertility. average number of kids
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$


Mean $=81.21$
Std. Dev. $=18.228$
$\mathrm{~N}=188$ $\qquad$
$\qquad$

$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$


Distribution for females
Distribution for males



56

## Non-normal distribution:

Use non-parametric descriptive statistics $\qquad$

- Min. \& Max.
- Range = Max. - Min.
- Percentiles
- Quartiles
-Q1
-Mdn (Q2)
-Q3
-IQR (Q3-Q1)

Effects of skew on measures of central tendency
+vely skewed distributions mode < median < mean symmetrical (normal) distributions mean $=$ median $=$ mode
-vely skewed distributions mean < median < mode

Effects of skew on measures of central tendency

positive skew

## Transformations

- Converts data using various formulae to achieve normality and allow more powerful tests
- Loses original metric
- Complicates interpretation
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$


## Review questions

1. If a survey question produces a 'floor effect', where will the mean, median and mode lie in relation to one another?

## Review questions

2. Would the mean \# of cars owned in Australia to exceed the median?

## Review questions

3. Would the mean score on an easy test exceed the median performance?
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$

$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$

$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$

## Science is beautiful <br> (Nature Video)

## (Youtube - 5:30 mins) $_{66}$

# Is Pivot a turning point for web exploration? <br> (Gary Flake) 


(TED talk - 6 min.)

## Principles of graphing

- Clear purpose
- Maximise clarity
- Minimise clutter
- Allow visual comparison


## Graphs

(Edward Tufte)

- Visualise data
- Reveal data
- Describe
- Explore
- Tabulate
- Decorate
- Communicate complex ideas with clarity, precision, and efficiency


## Graphing steps

1. Identify purpose of the graph (make large amounts of data coherent; present many \#s in small space; encourage the eye to make comparisons)
2. Select type of graph to use
3. Draw and modify graph to be clear, non-distorting, and welllabelled (maximise clarity, minimise clarity; show the data; avoid distortion; reveal data at several levels/layers)

## Software for data visualisation (graphing)

## 1. Statistical packages

- e.g., SPSS Graphs or via Analyses

2. Spreadsheet packages

- e.g., MS Excel


## 3. Word-processors

- e.g., MS Word - Insert - Object Micrograph Graph Chart



## Cleveland's hierarchy


$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$

## Univariate graphs

- Bar graph
- Pie chart
- Histogram
- Stem \& leaf plot
- Data plot / Error bar
- Box plot

$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$


## Bar chart (Bar graph)

- Allows comparison of heights of bars
- X-axis: Collapse if too many categories
- Y-axis: Count/Frequency or \% - truncation exaggerates differences
- Can add data labels (data values for each bar)

$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$


## Pie chart

- Use a bar chart instead
- Hard to read
-Difficult to show
-Small values
- Small differences
-Rotation of chart and position of slices influences perception


76

Pie chart $\rightarrow$ Use bar chart instead


77

## Histogram

- For continuous data (Likert?, Ratio)
- X-axis needs a happy medium for \# of categories
- Y-axis matters (can exaggerate)


## Histogram of male \& female heights



FIGURE 2.3.11 Histogram of heights constructed using the people. Photograph by Peter Morenus in conjunction with Prof. Linda Strausberg, University of Connecticut. Subjects are University of Connecticut genetics students, females in white tops, males in dark tops. Wild \& Seber (2000)

Stem \& leaf plot

- Use for ordinal, interval and ratio data (if rounded)
- May look confusing to unfamiliar reader

| Raw Data | Stem | Leaf |
| :---: | :---: | :---: |
| 011223444555667777 | 0 | 0112234445556677778899 |
| 8899 | 1 | 0111222333334445555556666666666777888899 |
| 1011111112121213131313 | 2 | 00112233444455667889 |
| 1314141415151515151516 | 3 | 005 |
| 1616161616161616161717 |  |  |
| 17181818181919 |  |  |
| 2020212122222323242424 |  |  |
| 242525262627282829 |  |  |
| 303035 |  |  |

## Stem \& leaf plot

- Contains actual data
- Collapses tails
- Underused alternative to histogram

| Frequency | stem \& | Leaf |
| :---: | :---: | :---: |
| 7.00 | 1. |  |
| 192.00 | 1. | 22223333333 |
| 541.00 | 1. | 444444444444444455555555555555 |
| 610.00 | 1. | 6666666666666677777777777777777777 |
| 849.00 | 1 | 88888888888888888888888888899999999999999999999 |
| 614.00 | 2 | 0000000000000000111111111111111111 |
| 602.00 | 2 | 222222222222222233333333333333333 |
| 447.00 | 2 | 4444444444444455555555555 |
| 291.00 | 2. | 66666666677777777 |
| 240.00 | 2. | 88888889999999 |
| 167.00 | 3 | 000001111 |
| 146.00 | 3. | 22223333 |
| 153.00 | 3 | 44445555 |
| 118.00 | 3 | 666777 |
| 99.00 | 3 | 888999 |
| 106.00 | 4. | 000111 |
| 54.00 | 4. | 222 |
| 339.00 | remes | (>=43) |

$\qquad$ whisker)

- Useful for interval and ratio data
- Represents min., max, median, quartiles, \& outliers
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$

Box plot (Box \& whisker)

- Alternative to histogram
- Useful for screening
- Useful for comparing variables
- Can get messy - too much info
- Confusing to unfamiliar reader


Data plot \& error bar

$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$

## Line graph

- Alternative to histogram
- Implies continuity e.g., time
- Can show multiple lines




## Graphical integrity

(part of academic integrity)
"Like good writing, good graphical displays of data communicate ideas with clarity, precision, and efficiency.

Like poor writing, bad graphical displays distort or obscure the data, make it harder to understand or compare, or otherwise thwart the communicative effect which the graph should convey."

## Tufte's graphical integrity

- Some lapses intentional, some not
- Lie Factor = size of effect in graph size of effect in data
- Misleading uses of area
- Misleading uses of perspective
- Leaving out important context
- Lack of taste and aesthetics


## Review exercise:

## Fill in the cells in this table

Level Properties Examples | Descriptive Graphs |
| :--- |
| Statistics |

Nominal
/Categorical
Ordinal /
Rank
Interval
Ratio
Answers: http://goo.gl/Ln9e1

## References

1. Chambers, J., Cleveland, B., Kleiner, B., \& Tukey, P. (1983). Graphical methods for data analysis. Boston, MA: Duxbury Press.
2. Cleveland, W. S. (1985). The elements of graphing data. Monterey, CA: Wadsworth.
3. Jones, G. E. (2006). How to lie with charts. Santa Monica, CA: LaPuerta.
4. Tufte, E. R. (1983). The visual display of quantitative information. Cheshire, CT: Graphics Press.
5. Tufte. E. R. (2001). Visualizing quantitative data. Cheshire, CT: Graphics Press.
6. Tukey J. (1977). Exploratory data analysis. Addison-Wesley.
7. Wild, C. J., \& Seber, G. A. F. (2000). Chance encounters: A first course in data analysis and inference. New York: Wiley.

## Open Office Impress

- This presentation was made using Open Office Impress.
- Free and open source software.
- http://www.openoffice.org/product/impress.html


