Wrapping up

Course: Methoden und Anwendungen

Instructor: Steven Moran, University of Zurich

Started off with some ambitious goals

19.2	Overview
26.2	Fieldwork & data elicitation
5.3	Linguistic analysis, data formats
12.3	Research methods
19.3	Introduction to R
26.3	Data transformation and query I
2.4	Data transformation and query II
16.4	Data analysis I
23.4	Data analysis II
30.4	Tutorial session
7.5	Probability and statistical inference
21.5	Inferential statistics: Chi-square test
28.5	Inferential statistics: T-test

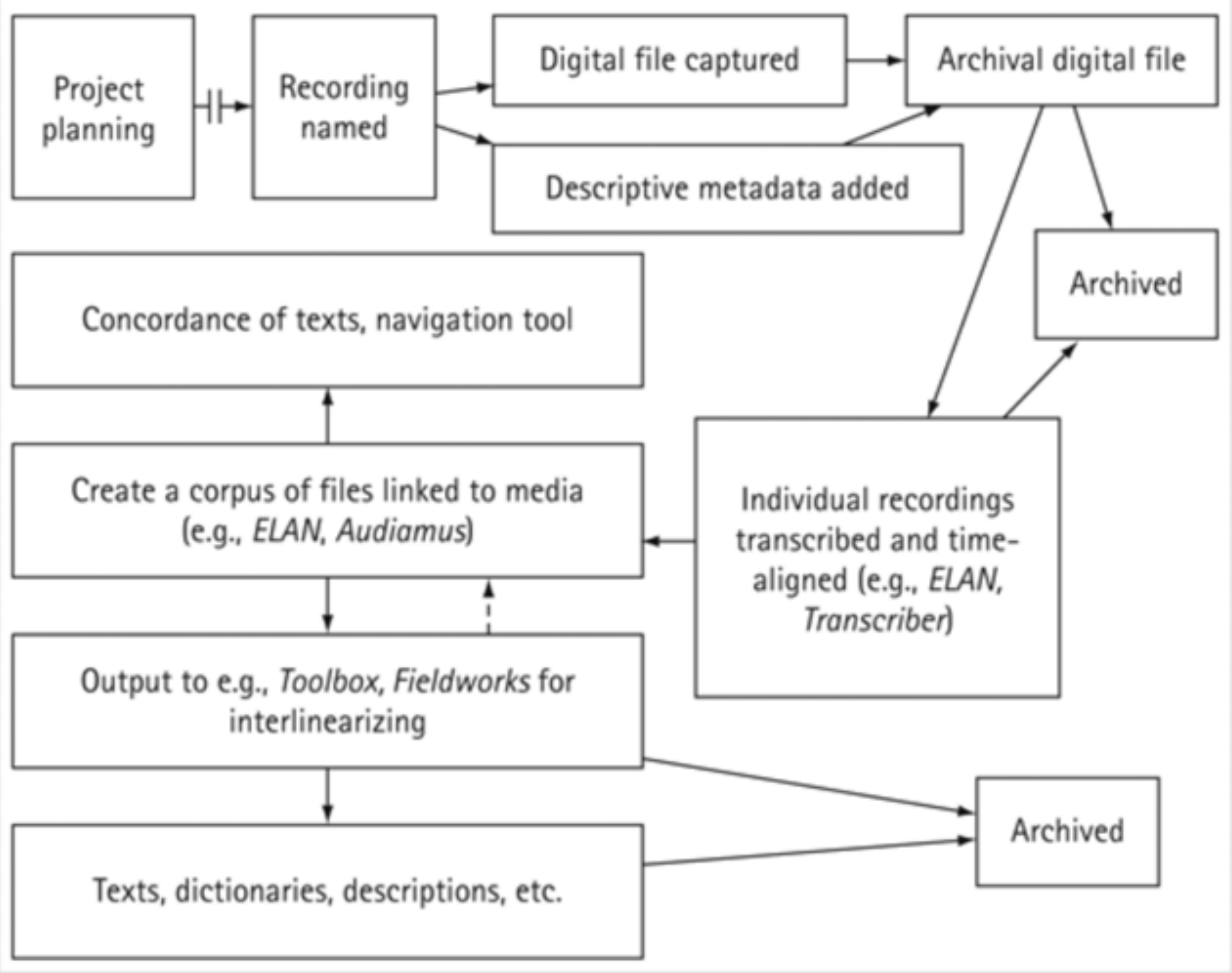
Fieldwork & data elicitation

- Why?
- Language documentation vs description
- Concerns
- How?

Linguistic analysis, data formats

- Phonological analysis
    - How do we figure out which sounds are contrastive in a language?
    - Phonemes vs allophones
    - Phonological rules
- Morphological analysis
    - identify phonemes and their types
    - general solutions towards elegant description cognitively realistic description
        - NOUN + “s” == plural in English
- What phono/morpho facts are preferred, dispreferred? How do we ask?

Linguistic analysis, data formats


Project planning
Recording named
Digital file captured
Archival digital file
Descriptive metadata added
Archived
Concordance of texts, navigation tool
Create a corpus of files linked to media (e.g., *ELAN*, *Audiamus*)
Individual recordings transcribed and time-aligned (e.g., *ELAN*, *Transcriber*)
Output to e.g., *Toolbox*, *Fieldworks* for interlinearizing
Archived
Texts, dictionaries, descriptions, etc.

Linguistic analysis, data formats

- What is data?
  - Fieldwork and analysis result in data
  - Metadata (or how do we describe that data?)
- Form and content
- Linguistic data types
  - paradigms, dictionaries, wordlists, (annotated) texts, corpora, IGT...
- Electronic data formats
  - http://www.itchyfeetcomic.com/2015/02/expressive-vowels.html?m=1
- Data modeling in tables, databases, etc. becomes input for software (R)

Research methods

- Birthday paradox
- Differences between qualitative and quantitative research
  - Qualitative
    - how something is as opposed to how much/many
    - inductive (theory driven from the results)
    - used in preliminary studies
  - Quantitative
    - deductive
    - based on a theory we develop an explicit hypothesis; try to prove / disprove the hypothesis
- Labov 1972

Qualitative vs quantitative research

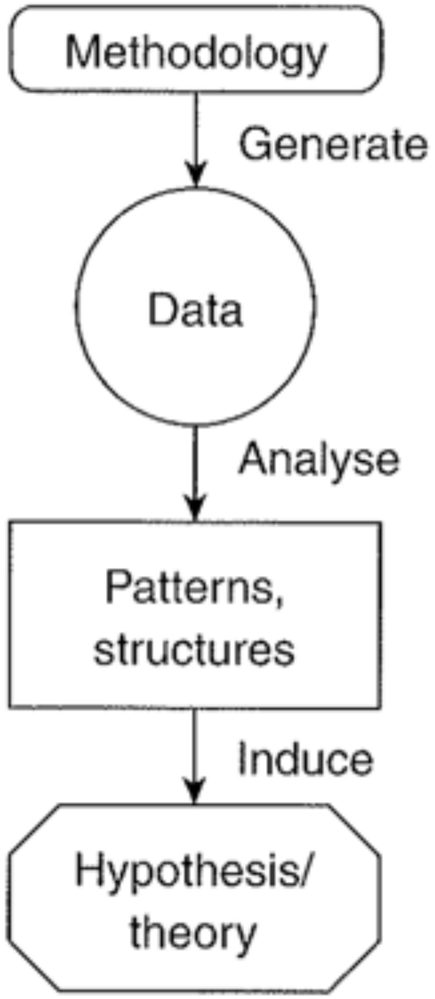


FIGURE 2.2 *Qualitative-inductive approach.*

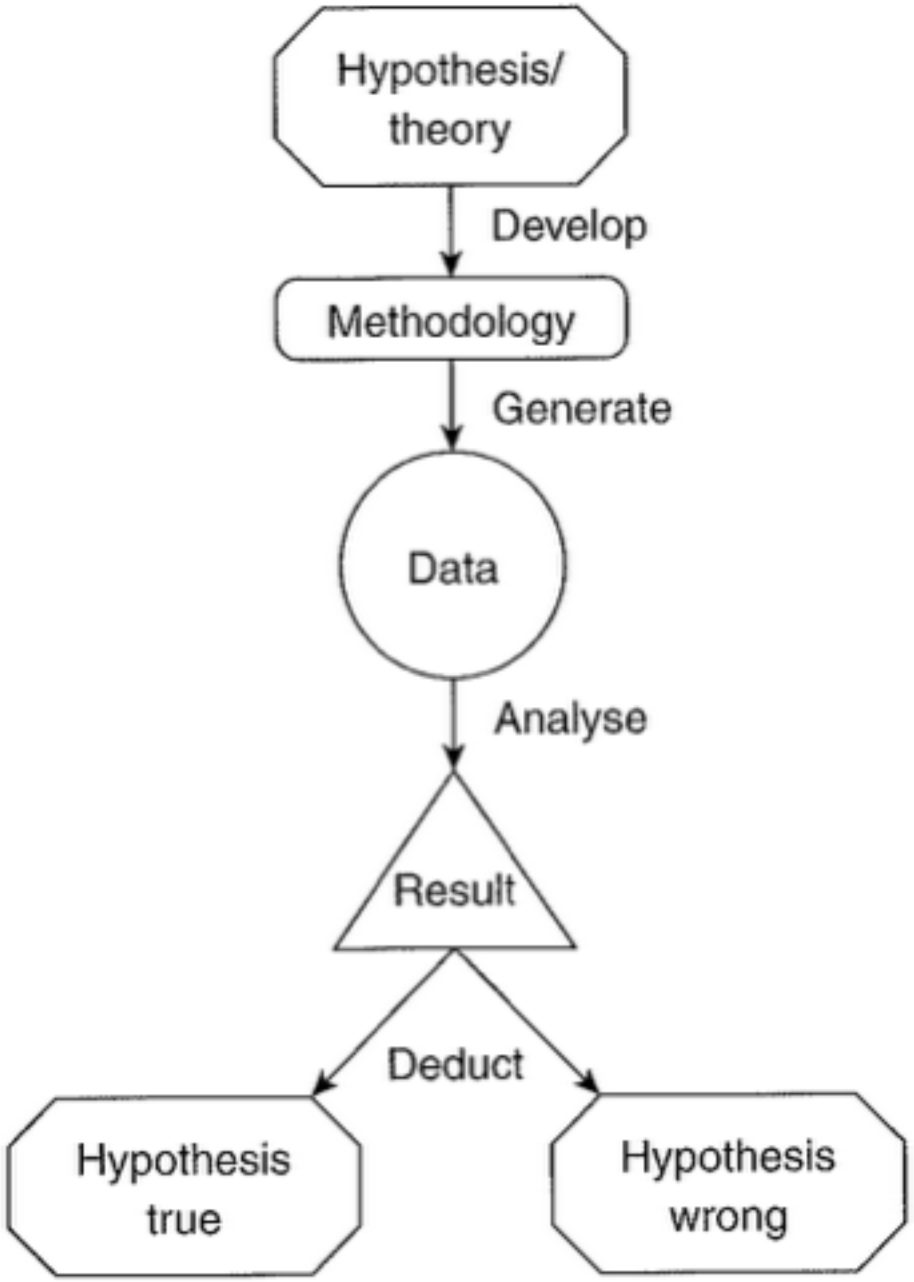


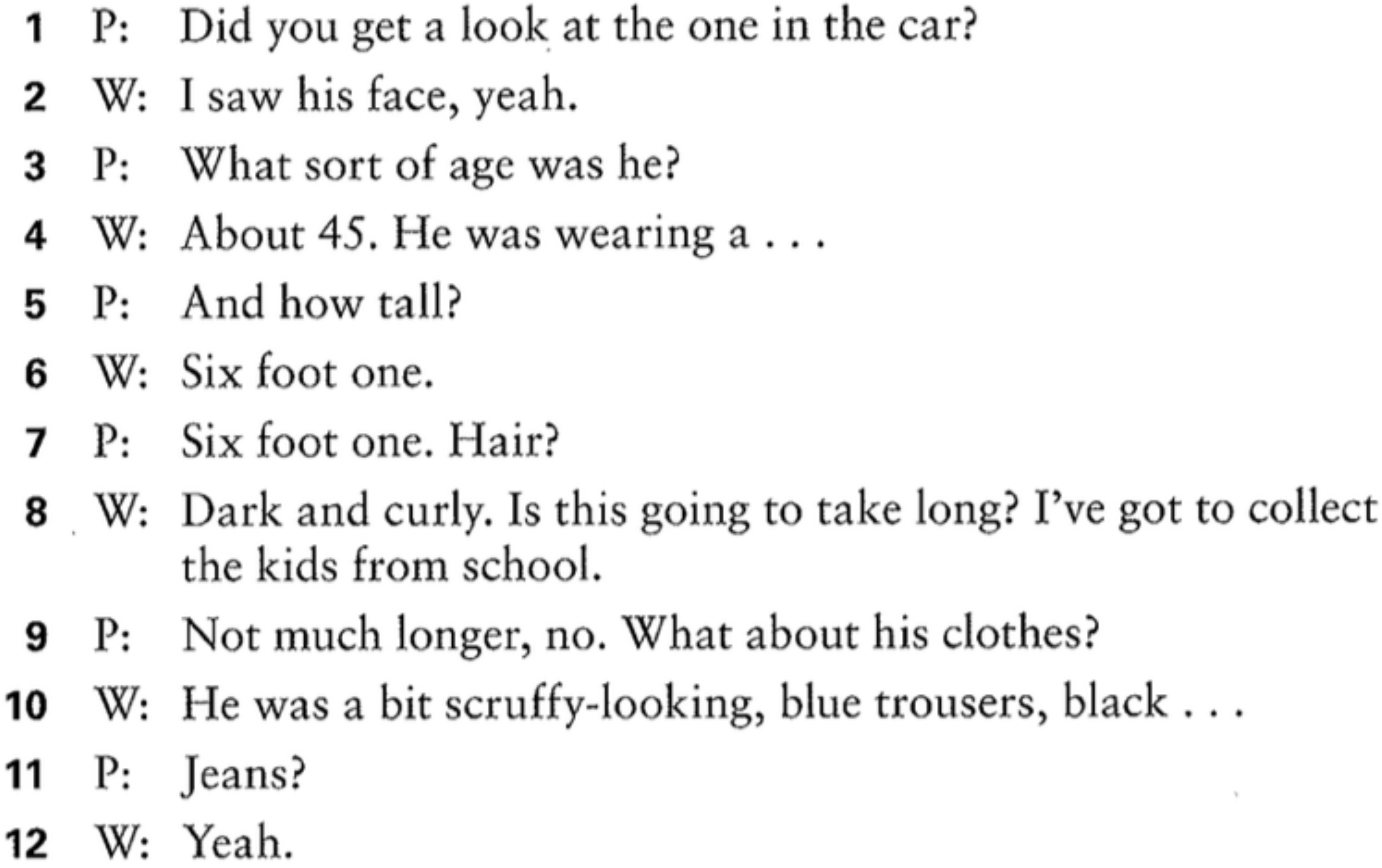
FIGURE 2.1 *Quantitative-deductive approach.*

Quantitative research

- Requires a methodology: hypothesis testing
  - hypothesis is a statement about a particular aspect of reality
  - findings are based on previous research (or observations) and the aim is to prove or disprove it
- Based on a precisely formulated hypothesis, develop a methodology
  - i.e. a set of instruments that allow us to measure reality in such a way that the results allow us to prove the hypothesis right or wrong
  - includes using or developing analytical tools to analyze the data
    - so we learned to use the statistical software package R
  - data collection methods must enable us to collect data which fits our research question and data analysis tools

An example

- interview between police officer (P) and witness (W)



1 P: Did you get a look at the one in the car?
2 W: I saw his face, yeah.
3 P: What sort of age was he?
4 W: About 45. He was wearing a . . .
5 P: And how tall?
6 W: Six foot one.
7 P: Six foot one. Hair?
8 W: Dark and curly. Is this going to take long? I've got to collect the kids from school.
9 P: Not much longer, no. What about his clothes?
10 W: He was a bit scruffy-looking, blue trousers, black . . .
11 P: Jeans?
12 W: Yeah.

Fairclough (1998:18): Language and Power.

An example

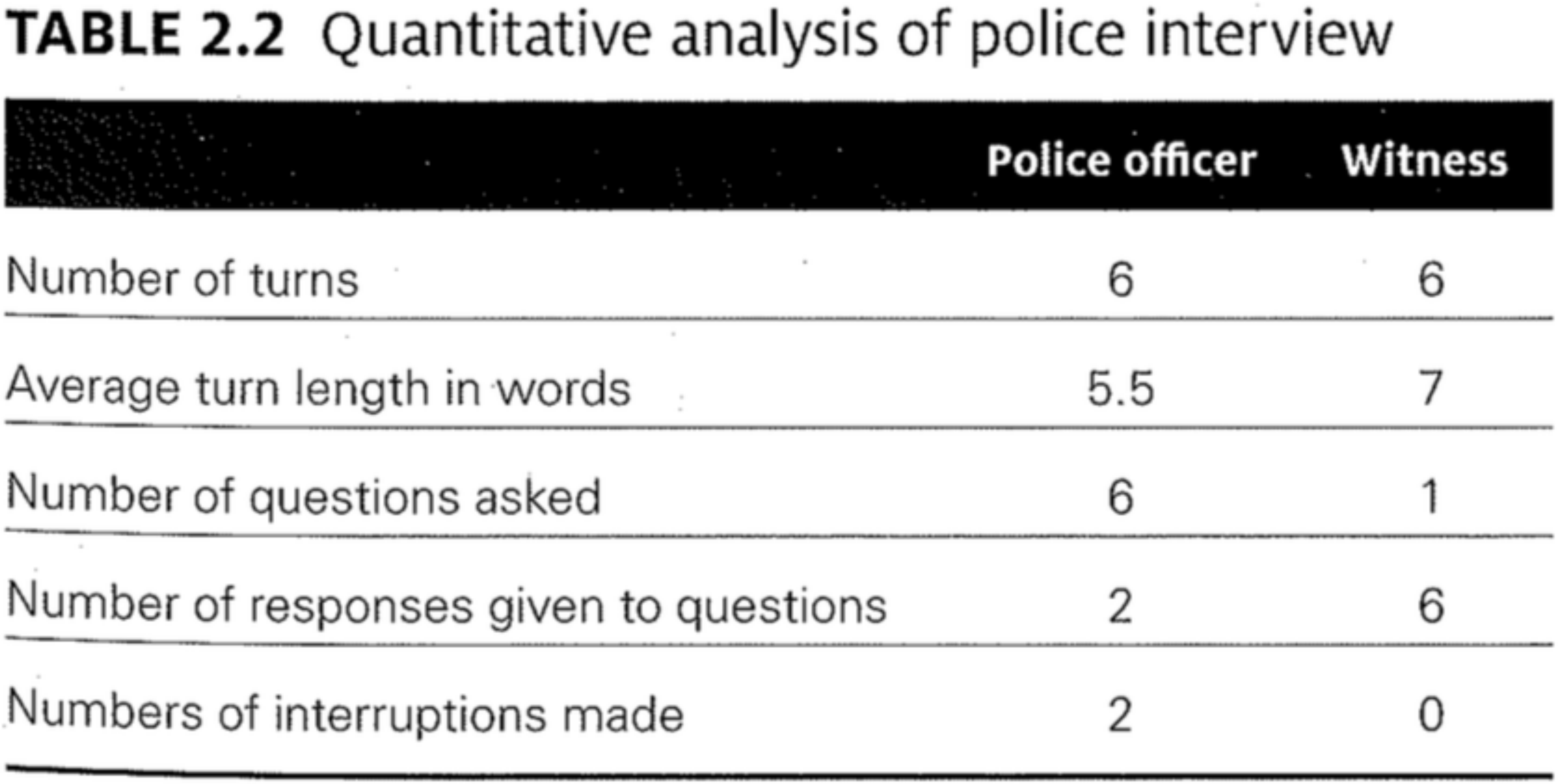


TABLE 2.2 Quantitative analysis of police interview

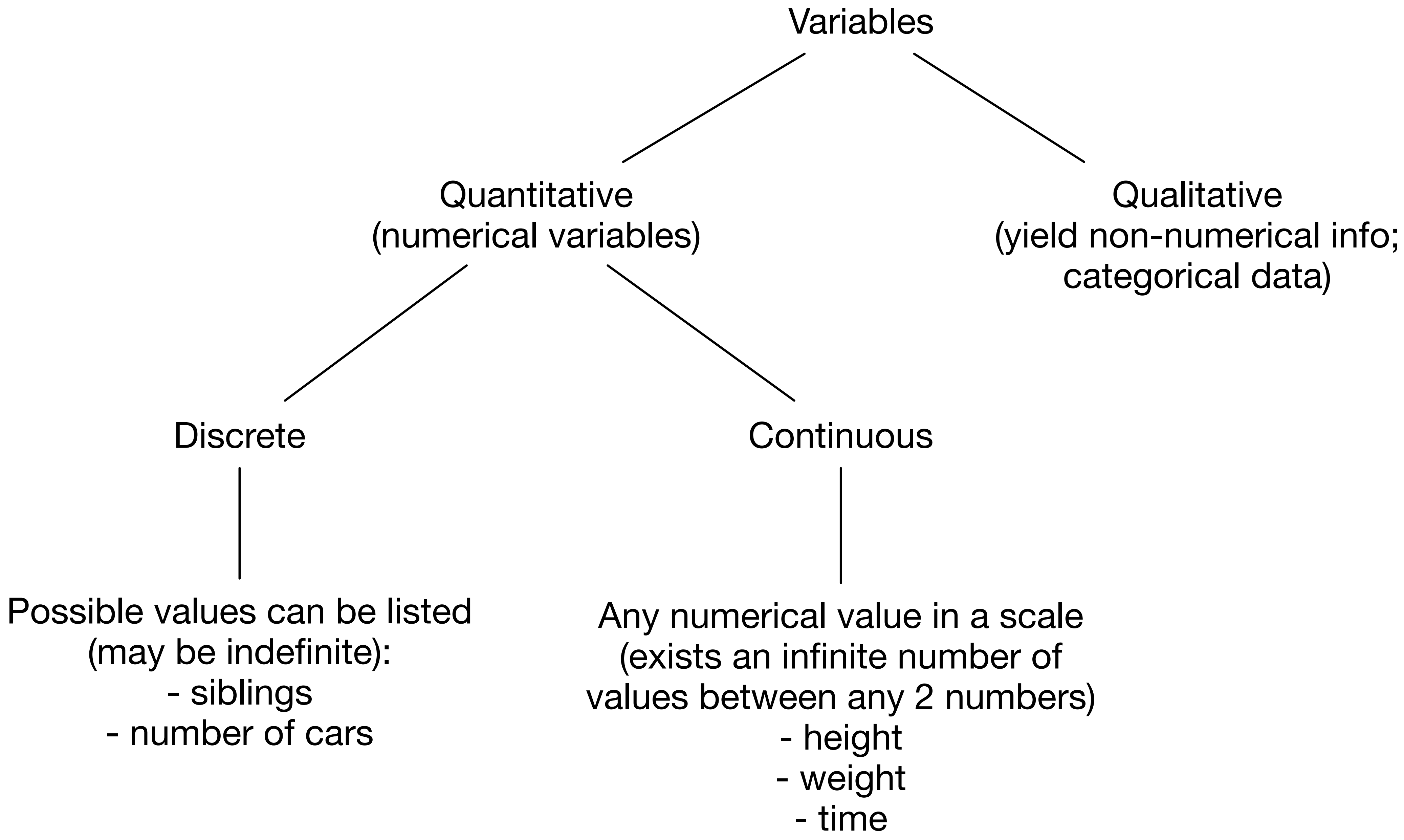
	Police officer	Witness
Number of turns	6	6
Average turn length in words	5.5	7
Number of questions asked	6	1
Number of responses given to questions	2	6
Numbers of interruptions made	2	0

When is a quantitative analysis inevitable?

- any study whose main argument is based on counting things
- any study that aims at proving two or more groups of people (or objects) are distinctively different, e.g. counting and the comparison
- any study that aims at showing that two variables are related, i.e. they cooccur in a particular pattern, e.g. the higher/lower X, the higher/lower Y (one variable results in the change of another)
- Remember: quantitative research is deductive
    - questions and hypotheses are based on already existing theories
    - we need to know how we measure/count things

Differences between qualitative and quantitative variables

- Different types of variables



Variables

Quantitative
(numerical variables)

Qualitative
(yield non-numerical info;
categorical data)

Discrete

Continuous

Possible values can be listed
(may be indefinite):
- siblings
- number of cars

Any numerical value in a scale
(exists an infinite number of
values between any 2 numbers)
- height
- weight
- time

Continuous and discrete data

- discrete data = finite or countable values
  - number of students in class (no 6.5, etc.)
- continuous data = infinitely many possible values on a continuous scale without gaps or interruptions
  - a student's age (years, days, hours, seconds, etc.)
- distinction between continuous and discrete applies to raw data and not to the results of any kind of statistical analysis (unlikely that on average 12.2 students per class)

Spatz, Chris. Basic statistics: Tales of distributions. Cengage Learning, 2007.

Scales of measurement

- What number were you wearing in the race? “5”
- What place did you finish in? “5”
- How many minutes did it take you to finish? “5”
  - The three 5s all look the same. However, the three variables (identification number, finish place, and time) are quite different. Because of the differences, each 5 has a different interpretation.
  - Consider another person with 10 as an answer for the same questions
    - What can you say about the two people in relation to each other?

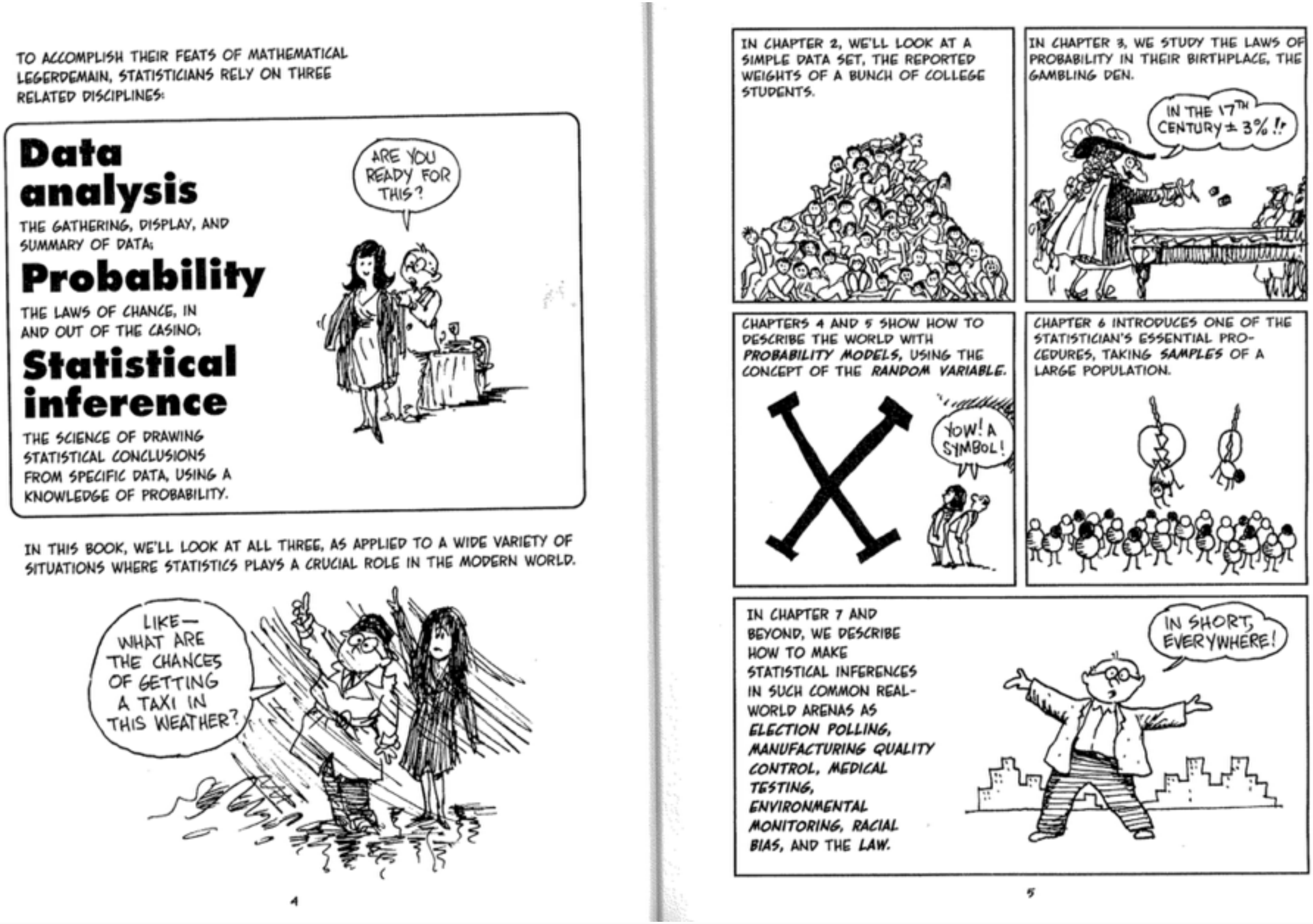
Spatz, Chris. Basic statistics: Tales of distributions. Cengage Learning, 2007.

Scales of measurement

- 4 scales:
  - Nominal scale, Ordinal scale, Interval scale, Ratio scale
- Knowing the distinctions among the four scales of measurement will help you in two tasks:
  - The kind of descriptive statistics you can compute from numbers depends, in part, on the scale of measurement the numbers represent.
  - Understanding scales of measurement is sometimes important in choosing the kind of inferential statistic that is appropriate for a set of data. If the dependent variable is a nominal variable, then a chi square analysis is appropriate. If the dependent variable is a set of ranks (ordinal data), then a nonparametric statistic is required

Spatz, Chris. Basic statistics: Tales of distributions. Cengage Learning, 2007.

To undertake statistical analysis (or mathematical trickery)



TO ACCOMPLISH THEIR FEATS OF MATHEMATICAL LEGERDEMAIN, STATISTICIANS RELY ON THREE RELATED DISCIPLINES:

Data analysis

THE GATHERING, DISPLAY, AND SUMMARY OF DATA;

Probability

THE LAWS OF CHANCE, IN AND OUT OF THE CASINO;

Statistical inference

THE SCIENCE OF DRAWING STATISTICAL CONCLUSIONS FROM SPECIFIC DATA, USING A KNOWLEDGE OF PROBABILITY.

IN THIS BOOK, WE'LL LOOK AT ALL THREE, AS APPLIED TO A WIDE VARIETY OF SITUATIONS WHERE STATISTICS PLAYS A CRUCIAL ROLE IN THE MODERN WORLD.

IN CHAPTER 2, WE'LL LOOK AT A SIMPLE DATA SET, THE REPORTED WEIGHTS OF A BUNCH OF COLLEGE STUDENTS.

IN CHAPTER 3, WE STUDY THE LAWS OF PROBABILITY IN THEIR BIRTHPLACE, THE GAMBLING DEN.

CHAPTERS 4 AND 5 SHOW HOW TO DESCRIBE THE WORLD WITH ***PROBABILITY MODELS***, USING THE CONCEPT OF THE ***RANDOM VARIABLE***.

CHAPTER 6 INTRODUCES ONE OF THE STATISTICIAN'S ESSENTIAL PROCEDURES, TAKING ***SAMPLES*** OF A LARGE POPULATION.

IN CHAPTER 7 AND BEYOND, WE DESCRIBE HOW TO MAKE STATISTICAL INFERENCES IN SUCH COMMON REAL-WORLD ARENAS AS ***ELECTION POLLING, MANUFACTURING QUALITY CONTROL, MEDICAL TESTING, ENVIRONMENTAL MONITORING, RACIAL BIAS,*** AND THE ***LAW.***

ARE YOU READY FOR THIS?
LIKE— WHAT ARE THE CHANCES OF GETTING A TAXI IN THIS WEATHER?
4
IN THE 17TH CENTURY ± 3% !?
X
YOW! A SYMBOL!
IN SHORT, EVERYWHERE!
5

Gonick and Smith 1993

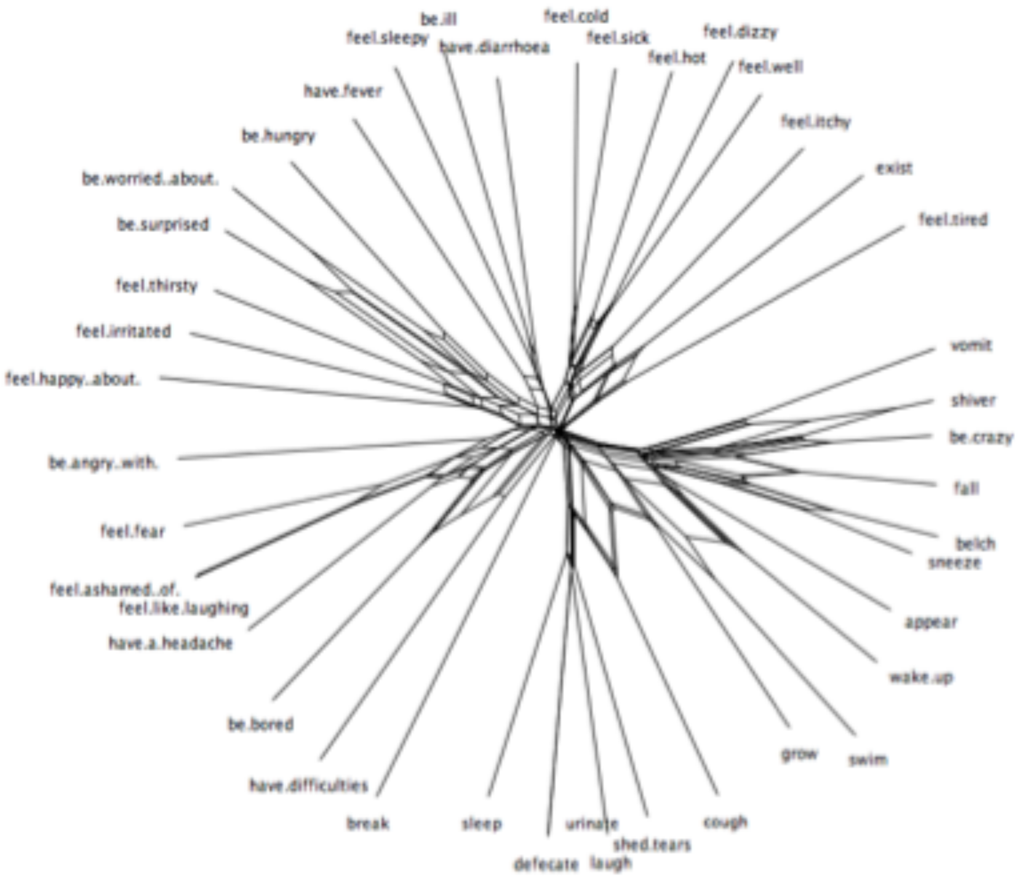
Relevance of quantitative methods

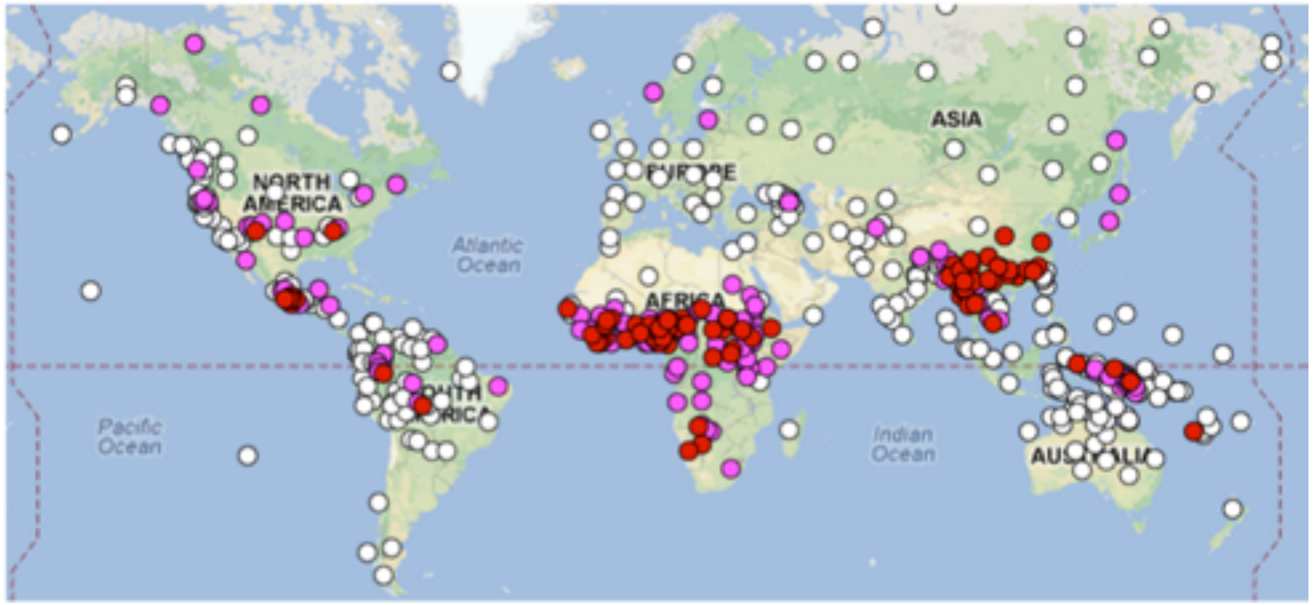
- First goal: description of your data
- Second goal: explanation of your data (usually on the basis of hypotheses about what kind(s) of relations you expected to find in the data
    - Often this is sufficient
- Third goal: prediction, i.e. what is going to happen in the future if you look at other data
- Problems:
    - Adequacy of the sample - particularly with linguistic data (the ability to generalize findings depend on have a representative sample)
    - Observations are complex with many potentially important variables

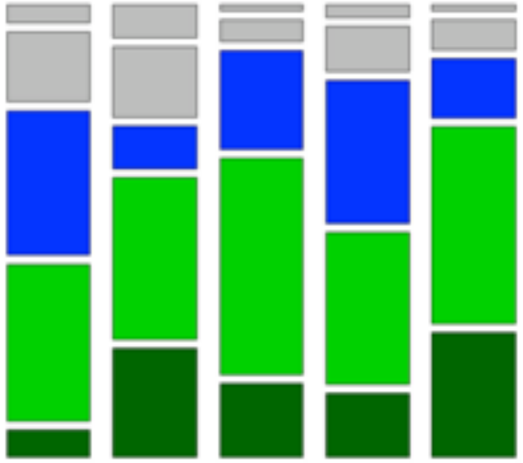
Th. Gries 2013, Johnson 2008

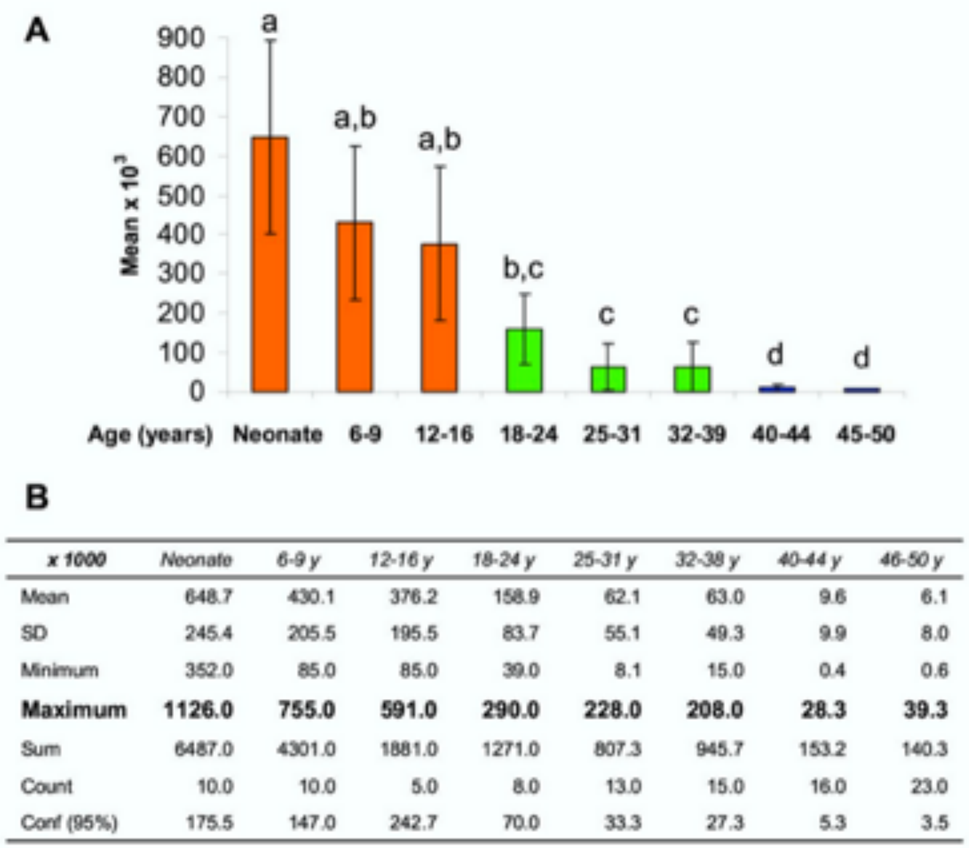
Descriptive statistics

- Descriptive statistics includes methods for organizing, summarizing and visualizing data









x 1000	*Neonate*	*6-9 y*	*12-16 y*	*18-24 y*	*25-31 y*	*32-38 y*	*40-44 y*	*46-50 y*
Mean	648.7	430.1	376.2	158.9	62.1	63.0	9.6	6.1
SD	245.4	205.5	195.5	83.7	55.1	49.3	9.9	8.0
Minimum	352.0	85.0	85.0	39.0	8.1	15.0	0.4	0.6
Maximum	**1126.0**	**755.0**	**591.0**	**290.0**	**228.0**	**208.0**	**28.3**	**39.3**
Sum	6487.0	4301.0	1881.0	1271.0	807.3	945.7	153.2	140.3
Count	10.0	10.0	5.0	8.0	13.0	15.0	16.0	23.0
Conf (95%)	175.5	147.0	242.7	70.0	33.3	27.3	5.3	3.5

Descriptive statistics

- Univariate statistics
  - statistics that summarize the distribution of one variable, of one vector, of one factor
- Bivariate statistics
  - statistics that characterize the relation of two variables, two vectors, two factors to each other
- Multivariate
  - statistics that encompass the simultaneous observation and analysis of more than one dependent variable

Spatz, Chris. Basic statistics: Tales of distributions. Cengage Learning, 2007.

Measures of central tendency - central points, typical values

- Three measures of the the central tendency, or mid-point, of a skewed distribution of data.

- The *mode* of the distribution is the most frequently occurring value in the distribution - the tip of the frequency distribution.

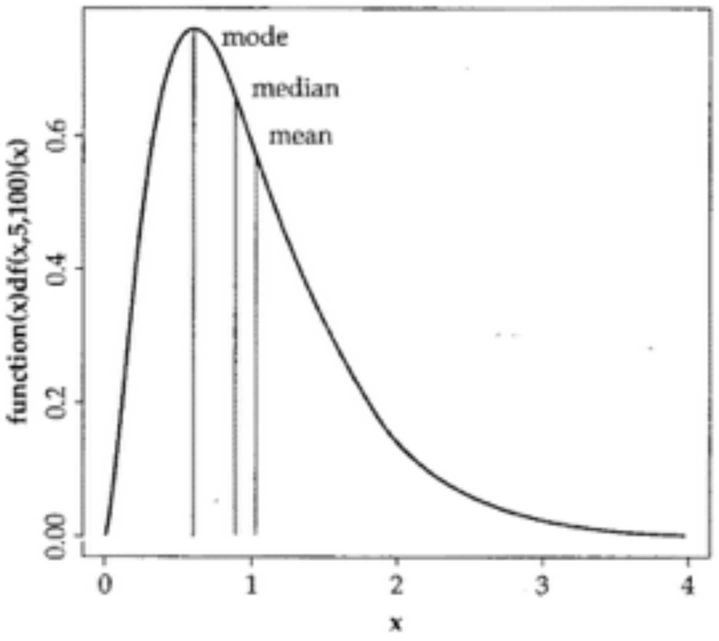


Figure 1.14 The mode, median, and mean of a skewed distribution.

- The *median* of the distribution is the value in the middle of the ordered list (“center of gravity”).

- The *mean* value (arithmetic average) is the *least squares estimate* of central tendency. We take the difference between the mean and each value in our data set, square these differences and add them up, we will have a smaller value than if we were to do the same thing with the median or any other estimate of the "mid-point" of the data set.

- Mean and median useful for comparison of datasets

An example with height

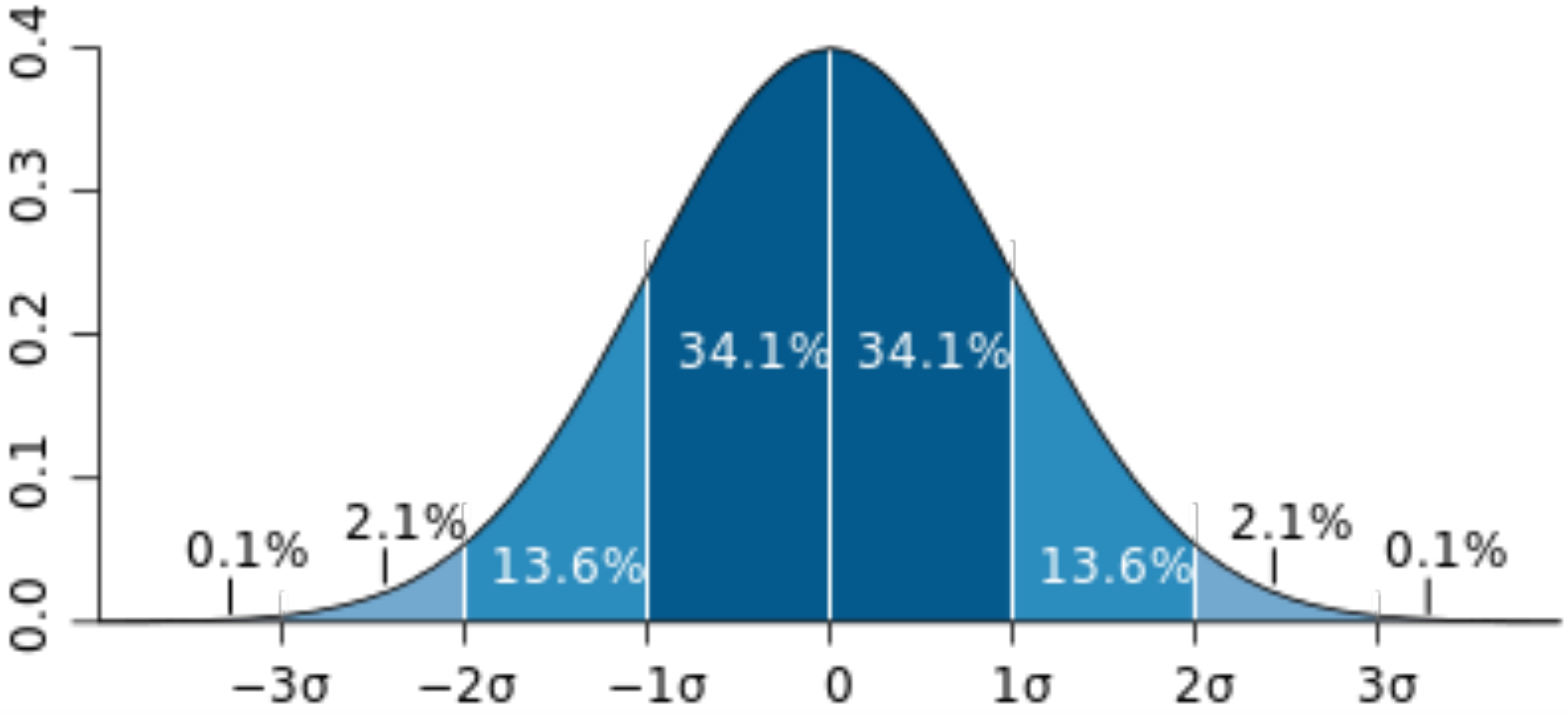
- Gather data (e.g. 5 heights)
- Summarize data, e.g. find the mean, variance and standard deviation
  - get the mean (average)
  - standard deviation (square root of the variance — what is variance?)
  - variance (average of the squared differences of the mean)
    - work out mean (simple average of the numbers)
    - get squared difference (subtract the mean and square the result)
    - work out average of squared differences

An example with height

- Sample vs population data (small difference in calculation)
    - Population (the 5 heights we were interested in)
    - Sample (a selection taken from a bigger population)
        - What does this equate to when we look at cross-linguistic phenomena?
- If you have N data values
    - Population: divide by N when calculating Variance (like we did)
    - Sample: divide by N-1 when calculating Variance
- Think of it as a “correction” when your data is only a sample

Normal distribution (aka normal curve; bell curve; Gaussian distribution)

- Derived from just two numbers: the mean value and a measure of how variable the data are
- The sum of area under the curve fx is 1
- Instead of terms of a "frequency" distribution, the normal curve gives us a way to calculate the probability of any set of observations by find the area under the proportion of the curve



Normal distribution

- is a useful way to describe data
- embodies reasonable assumptions about how we end up with variability in our data sets
- gives us mathematical tools to use in two important goals of statistical analysis:
    - data reduction - we can describe the whole frequency of distribution with just two numbers - the mean and the standard deviation
    - normal distribution provides a basis for drawing inferences about the accuracy of our statistical estimates
- it is essential to know whether or not the frequency distribution of your data is shaped like the normal distribution

Types of distributions

- Uniform - if every outcome is equally likely
    - e.g. six sides of a dice, equally likely
- Normal - measurements tend to congregate around a typical value and values become less and less likely as they deviate further from the central value
    - Two parameters:
        - (μ) central tendency (“mu”)
        - (σ) how quickly probability goes down as you move away from the center of the distribution (“sigma”)

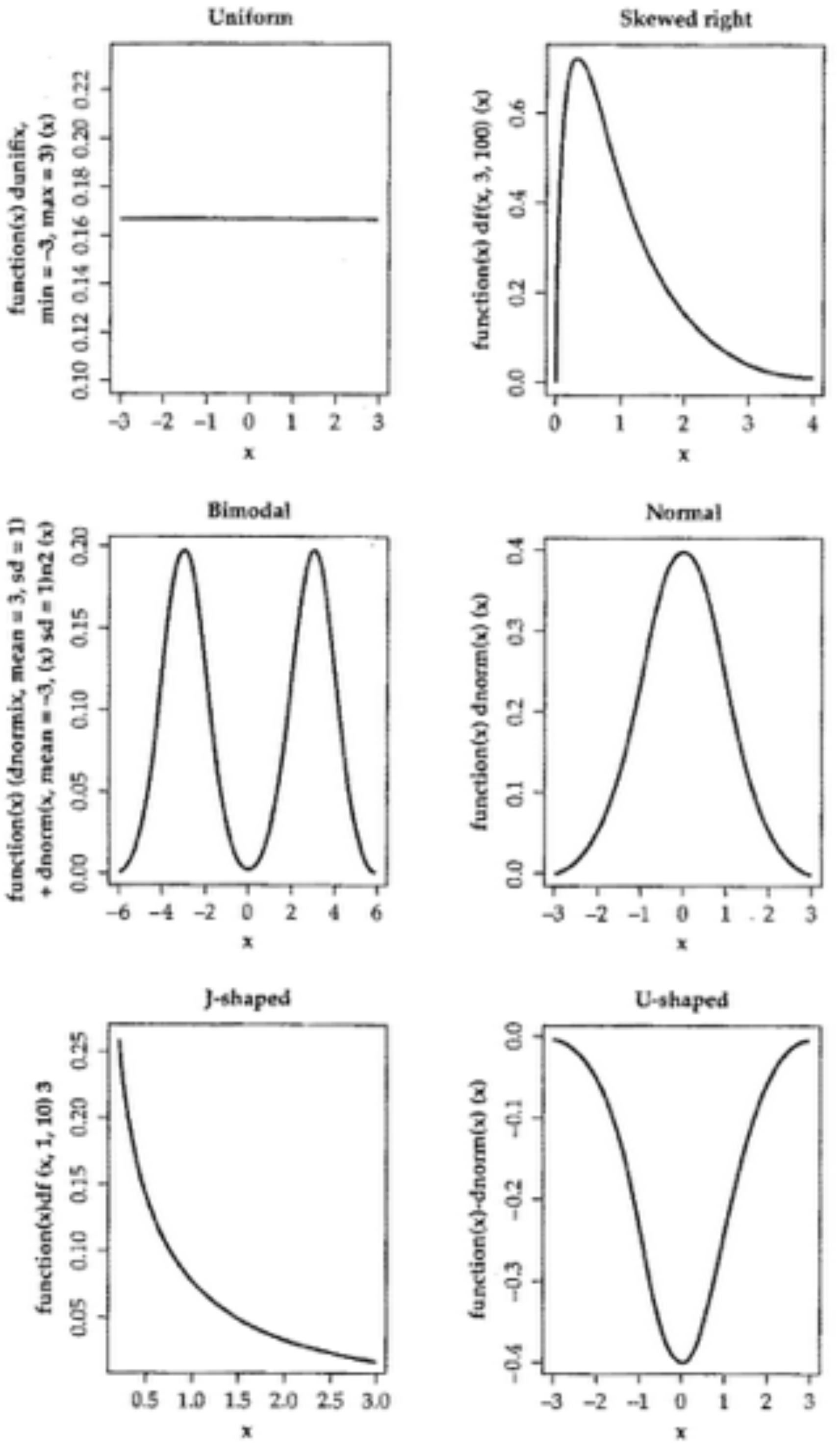


Figure 1.6 Types of probability distributions.

Hypothesis testing

- hypotheses = statements about the potential and/or suggested relationship between at least two variables
  - the older a learner, the less swear words they use
  - age and gender influence language use
- hypotheses must be proven right or wrong
  - must be well defined, i.e. must be falsifiable and not tautological (e.g. "age can either influence a person's language use or not" - it will always be true!)
- hypotheses should relate to something that can be measured; should be specific; should be clear

Spatz 2008.

Independent vs dependent variables

- variables can stand in relationship to each other
- independent vs dependent variables
  - latter can be influenced by the former, but not vice versa
- example in linguistics research: age as independent variable
  - L2 acquisition and influence from age
  - L2 acquisition success (dependent variable) depends on age (independent variable)
  - L2 proficiency influences age?

Hypothesis testing

- Null hypothesis significance testing (NHST)
- A hypothesis is a statement that makes a prediction about the distribution of one variable (or the relation between two or more variables) in a population and that has the implicit structure of a conditional sentence (if..., then... or the more/less..., the more, less)
- If (IV) part, then (DV) part
  - Independent variable (IV) - the variable in the “if” part of the hypothesis - often the cause of the changes/effects on the “then” part
  - Dependent variable (DV) - the variable in the “then” part of the the hypothesis and whose values, variation, or distribution is to to be explained

Experimental design variables

- A simple experiment has two major variables, the independent variable and the dependent variable.
    - independent variable - variable controlled by the researcher; changes in this variable may produce changes in the dependent variable.
    - dependent variable - the observed variable that is expected to change as a result of changes in the independent variable in an experiment.
- The basic idea is that the researcher finds or creates two groups of participants that are similar except for the independent variable.
- These individuals are measured on the dependent variable.
- The question is whether the data will allow the experimenter to claim that the values on the dependent variable depend on the level of the independent variable.

Spatz 2008.

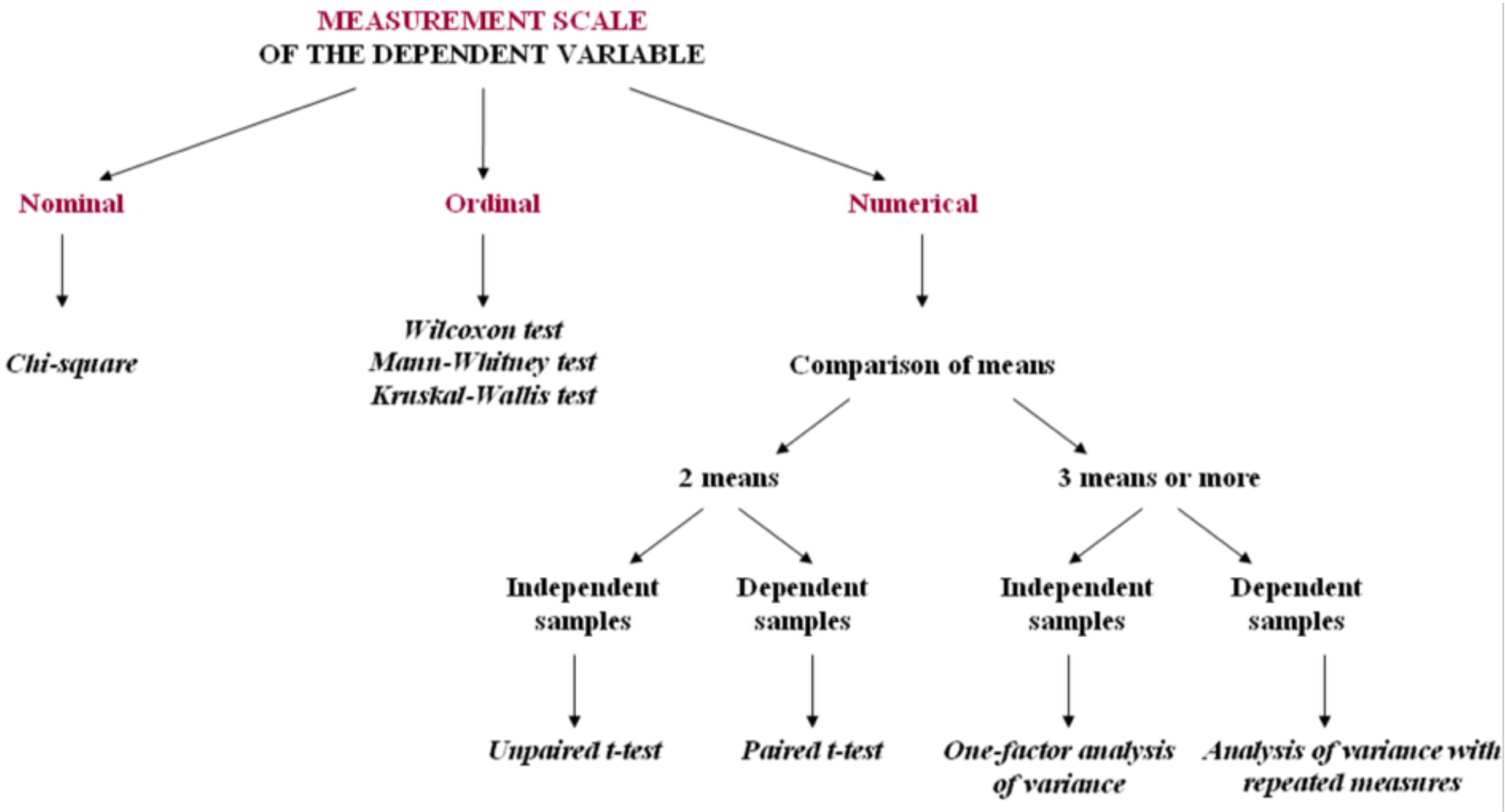
Choosing (a) significance test(s)

- The decision for a particular statistical test is typically made on the basis of a set of questions that cover various aspects of the study you are conducting, the number and types of variables that are involved, and the size and distribution of the dataset(s) involved.
  - What kind of study is being conducted?
  - How many and what kinds of variables are involved?
  - Are data points in your data related such that you can associate data points to each other meaningfully and in a principled way?
  - What is the statistic of the dependent variable in the statistical hypotheses?
  - What does the distribution of the data look like? Normally or another way that can be described by a probability density function or some other way?
  - How big are the samples to be collected? $n<30$ or $n\geq30$?
- Once all the above questions have been answered and all other requirements have been checked, they usually point to one or two tests that address your question exactly.

When to use a particular statistical test?

- http://www.csun.edu/~amarenco/Fcs%20682/When%20to%20use%20what%20test.pdf

Choosing a test



MEASUREMENT SCALE
OF THE DEPENDENT VARIABLE

Nominal

Chi-square

Ordinal

Wilcoxon test
Mann-Whitney test
Kruskal-Wallis test

Numerical

Comparison of means

2 means

Independent samples

Unpaired t-test

Dependent samples

Paired t-test

3 means or more

Independent samples

One-factor analysis of variance

Dependent samples

Analysis of variance with repeated measures

Chi-square

- You compare the observed frequencies with the theoretical (expected) frequencies
  - https://www.youtube.com/watch?v=SvKv375sacA
- Is variation beyond what we would expect due to chance alone?
  - http://www.r-tutor.com/elementary-statistics/goodness-fit/chi-squared-test-independence

Chi-square

- Compare the observed frequency of occurrence of an event with its theoretically expected frequency of occurrence
  - number of men and women in this class should be equal because there are about as many men as there are women in the world
  - in a class of 20, we expect 10 females and 10 males
- If the difference between observed and expected frequency is much greater than chance you might begin to wonder what is going on. Perhaps an explanation is called for.

Chi-square

- The "goodness-of-fit test" is a way of determining whether a set of categorical data came from a claimed discrete distribution or not. The null hypothesis is that they did and the alternate hypothesis is that they didn't. It answers the question: are the frequencies I observe for my categorical variable consistent with my theory?
    - The goodness-of-fit test is used if you have two or more categories.
- The "test of independence" is a way of determining whether two categorical variables are associated with one another in the population, like race and smoking, or education level and political affiliation.

Chi-square: goodness of fit

- Variables
  - dependent: nominal
  - 1 sample
- Compare the observed frequencies with theoretical (expected) frequencies
- Example: in a corpus, number of cases of different grammatical categories

http://phdmonkey.com/2011/07/23/chi-square-goodness-of-fit-vs-test-of-independence/

Chi-square: test of independence

- Variables
  - dependent: nominal
  - 2 samples
- Determine whether the difference between two frequency distributions is due to chance or is statistically significant
- Example: number of successes and failures in group A and in group B
- Use contingency table

http://phdmonkey.com/2011/07/23/chi-square-goodness-of-fit-vs-test-of-independence/

- Example of goodness-of-fit:
  - We might compare the proportion of M&M's of each color in a given bag of M&M's to the proportion of M&M's of each color that Mars (the manufacturer) claims to produce. In this example there is only one variable, M&M's. M&M's can be divided into many many categories like Red, Yellow, Green, Blue, and Brown, however there is still only one variable... M&M's. Hungry?
- Example of Test of Independence:
  - To continue with the M&M's example, we might investigate whether purchasers of a bag of M&M's eat certain colors of M&M's first. Here there are two variables: (1) M&M's (2) The order based on color that an M&M bag holder/purchaser eats the candies.

http://phdmonkey.com/2011/07/23/chi-square-goodness-of-fit-vs-test-of-independence/

Hypothesis testing - an example

- We often want to ask questions about mean values
  - Is this average voice onset time (VOT) different from that one?
  - Do these two constructions receive different average ratings?
  - Does one variant occur more often than another?
- These all boil down to the question is $\overline{x}$ (the mean of the data x_i) different from $\overline{y}$ (the mean of the data y_i)?
- Is the population parameter estimated by $\overline{x}$ and $\overline{y}$?
  - We know the sample mean values for $\overline{x}$ and $\overline{y}$, can we say with some *degree of confidence* that μ_x and μ_y (population parameters) are different?
  - If we can measure the error of the sample mean, then we could put a confidence value on how well it estimates μ

Summing up...

Four levels of statistical sophistication

- Category 1: those who understand statistical presentations
- Category 2: those who understand, select, and apply statistical procedures
- Category 3: applied statisticians who help others use statistics
- Category 4: mathematical statisticians who develop new statistical techniques and discover new characteristics of old techniques

Kirk 1999

Research methods... summing up

- (1) our ability to generalize our findings depends on having a representative sample of data - good statistical analysis can't overcome sampling inadequacy - and

- (2) the observations that we are exploring in linguistics are complex with many potentially important variables. The balancing act that we attempt in research is to stay aware of the complexity, but not let it keep us from seeing the big picture.

Conclusion: studying languages is one big puzzle...