why statistics?

- allows us to derive general insights from a limited set of observations.
- ... which is the purpose of scientific research!

why statistics?

- derive general insights from a limited set of observations.
  - Statistics is the science of collecting, analyzing, and drawing conclusions from data.
    — Devore & Peck (2005: 8); also Wikipedia
  - [Statistics is the] science of making valid inferences about the characteristics of a group of persons or objects on the basis of [data] obtained from a ... sample of the group.
    — Encyclopædia Britannica 2002

some properties of science

- finite causality
  - events are caused by a finite number of factors, not everything is connected with everything else
- empiricity
  - insights are based on verifiable observations, combined with logic, not on (religious) authority or on common sense
- consistency, uniformity
  - similar observations under similar circumstances; reproducibility; this makes the world comprehensible; results are generalizable to larger population
- objectivity
  - focus on object of observation, not on observer
- parsimony, simplicity, elegance
  - seek the most simple explanation (Occam’s Razor)

variables

- science attempts to comprehend and explain our world.
- permanent, constant properties tend to be less interesting for this purpose.
- variable properties tend to be more interesting.
  - links with evolution: dangers and opportunities?
    - ... body length, weight, foot size, average F0, zip code, income, sexual preference, age, number of traffic violations, number of speech errors, ...

overview

- why statistics
- levels of measurement
- scientific method
- validity
- large numbers
- Central Limit Theorem
- hypothesis testing
- significance and power
- regression and correlation
- generalizing

variables

- science attempts to comprehend and explain our world.
- permanent, constant properties tend to be less interesting for this purpose.
- variable properties tend to be more interesting.
  - links with evolution: dangers and opportunities?
    - ... body length, weight, foot size, average F0, zip code, income, sexual preference, age, number of traffic violations, number of speech errors, ...

why statistics?

- allows us to derive general insights from a limited set of observations.
- ... which is the purpose of scientific research!

why statistics?

- derive general insights from a limited set of observations.
  - Statistics is the science of collecting, analyzing, and drawing conclusions from data.
    — Devore & Peck (2005: 8); also Wikipedia
  - [Statistics is the] science of making valid inferences about the characteristics of a group of persons or objects on the basis of [data] obtained from a ... sample of the group.
    — Encyclopædia Britannica 2002

overview

- why statistics
- levels of measurement
- scientific method
- validity
- large numbers
- Central Limit Theorem
- hypothesis testing
- significance and power
- regression and correlation
- generalizing
**good variables**

Scientific research attempts to find relations between variables under study...

- **independent (IV):** manipulated
  - factor, condition, **treatment**
  - predictor

- **dependent (DV):** measured
  - perceived phoneme
  - MLU, word duration

**bad variables**

Relations between variables-under-study may be obscured by...

- **extraneous variables:** a nuisance
  - not of interest, but affect DV
  - may be controlled (blocked, fixed)

- **confounded variables**
  - extraneous, and relevant
  - and not controlled

**level of measurement**

- variable can be observed at 4 (incremental) levels of measurement.
- higher levels can be degraded
- lower levels can **not** be promoted!

**level of measurement**

1. **Nominal**
   - categorical, discrete, unordered
     - sex, eye color, fuel type
     - only **names** of categories are available, there is no meaningful ordering of male-female, blue-brown-grey, wood-coal-oil-gas-electricity, etc.

2. ** Ordinal**
   - categorical, discrete, **ordered**
     - salary rank, military rank, clothes sizes (S, M, L, XL)
     - **Order** of categories is meaningful: general outranks major (not vice versa)

**bad variables**

<table>
<thead>
<tr>
<th>group</th>
<th>N</th>
<th>X</th>
<th>s</th>
</tr>
</thead>
<tbody>
<tr>
<td>test</td>
<td>1</td>
<td>40</td>
<td>220</td>
</tr>
<tr>
<td>control</td>
<td>2</td>
<td>40</td>
<td>200</td>
</tr>
</tbody>
</table>

Results show significant effect of treatment
unmatched t=1.83, df=39, p=.038
but treatment **confounded** with sex
effects impossible to distinguish
level of measurement

3 interval
numerical, continuous, no zero
- temperature in Celsius (arbitrary zero)
- observations can be ordered, plus intervals can be calculated and compared meaningfully:
  ✓ the interval from 20° to 60° is twice as large as the interval from 40° to 60°.
  but ratios between observations are not meaningful:
  » a temperature of 60° is twice as large as one of 30°

level of measurement

4 ratio
numerical, continuous, with zero
- temperature in Kelvin (0°K = -273°C)
- intervals can be compared, and ratios are meaningful:
  ✓ the number of cylinders in the engine of car A is twice that of car B.
  ✓ the age of person A is one-third of the age of person B.

why statistics?

- two types of statements:
  - estimation (descriptive statistics)
    People in the Netherlands are having sex for the first time at an average age of 16.6.
    — Global Sex Survey 2005 (www.durex.com)
  - inference (inferential statistics)
    The number of boys who regularly ate vegetables remained more or less constant, while the number of girls doing so showed a significant decrease.
    — Gezondheidsraad, Report 2002/12 (www.gr.nl)

empirical reasoning

- scientific research can be regarded as a cyclical process: the empirical cycle

... is very important!
- it determines which statistical operations and tests are allowed...
...and hence, which conclusions you can draw from your data!
empirical cycle

From a general idea, the hypothesis (H1) is induced as a testable statement. Positive evidence for this H1 is not convincing, because of the inference problem (David Hume, 1711-1776). Hence, the scientist attempts to prove the null hypothesis (H0), the logical opposite of H1. Testable predictions are deduced from H0. Data are obtained, and the predictions are tested by means of statistics (more later). If H0 is very implausible, given the data obtained, then H0 is rejected. Rejection of H0 implies positive evidence (support) for H1, and hence for the general idea. The conclusion, combined with other observations, may generate new ideas to investigate.

falsificationism

Karl Popper (1902-1994): knowledge only grows by falsification of hypotheses – because of induction problem – "... science advances by unjustified, exaggerated guesses followed by unstinting criticism. Only hypotheses capable of clashing with observation reports are allowed to count as scientific. "Gold is soluble in hydrochloric acid" is scientific (though false) ... Falsifiable theories ... enhance our control over error while expanding the richness of what we can say about the world."

--- The Karl Popper Web (http://www.eeng.dcu.ie/~tkpw/)
validity

- property of a statement, claim, conclusion (not of the underlying data)
- validity = quality
- every step in a study determines the validity of its conclusions!
- the whole chain is only as strong as its weakest link

2. concluding validity

- how credible are the conclusions of the study?
- can the study show any relation between IV and DV?
  - reliability of observations,
  - effect size,
  - statistical power,
  - appropriateness of statistical test used,
  - etc etc

1. construct validity

- “construct validity involves generalizing from your program or measures to the concept of your program or measures.” — W. Trochim
- is the dependent variable a good measure of the theoretical concept (construct) in which you are interested?

3. internal validity

- can the study show a causal relation between independent and dependent variables?
- threatened by extraneous factors, “plausible alternatives”, inappropriate design!
- and by maturation (pos, neg), memory, fatigue, bias, drop-out, instrumentation, etc. etc.

example 1: confounding

- AGE is a between-subject factor.
- effect of AGE cannot be separated from differences between GROUPs; AGE and GROUP are confounded.
- Are “age-related” differences really due to children’s age, or to sample fluctuations between the age groups?
problem

• this confounding may lead to incorrect decisions and conclusions!

solution

• follow age groups over a longer time (repeated measurement); repeat experiment on same children across time; within-subjects design
• to separate group differences from age effects.

problem

• confounding is a threat to internal validity
  – can sample fluctuations be separated from interesting differences?
  – are separate sources of sample variance identified and kept separate?

internal validity

example 2: bias

data are observed for a sample out of a larger population;
sample should be representative, for domain of generalization.
• selection bias
  – CHILDES: non-random, opportunistic sampling; possible bias in selected children (e.g. parents tend to be highly educated & verbally talented)
example 2: bias

- non-response bias
  - non-random drop-out of sampling units (e.g. more utterances untranscribed in group X than in group Y)
- response bias
  - self-reported values for prestigious variables (income, sexual activity, etc) are often biased.
  - participants want to "look good" and cooperate, more than non-participants


4. external validity

- "external validity is the degree to which the conclusions in your study would hold for other persons in other places and at other times."
  — W. Trochim

- this depends on
  - sampling method
  - statistical analysis

---

example 2: bias

- experimenter bias
  - investigator should not double as experimenter, coder, transcriber, observator
  - investigator’s own observations might be biased towards his/her expectations, wishes, interests
  - use "double-blind" procedure where possible

4. external validity

- increases with repetitions of your study, with other children, other languages, other times, and preferably with other methods.
- decreases with unnatural, strictly controlled experiment (having good internal validity).
- more on generalizing later!

---

protecting validity

- randomization: randomly assign persons/objects to treatment conditions!
  - uncontrolled extraneous variables will be evenly distributed
- blocking: distribute confounding variables evenly over persons/objects and conditions
- control over nuisance variables, e.g. by keeping them fixed throughout study.
- repetition: repeat your observations and measurements, where possible
large numbers
- dice is unbiased cube, with 6 sides numbered 1 to 6, uniform distribution
- expected mean outcome is
  \[ \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = \frac{1}{6} \cdot (1 + 2 + 3 + 4 + 5 + 6) = \frac{21}{6} = 3.5 = \mu \]

large numbers
- \( n = 3 \) draws from Scrabble set blanks removed; 100 letters in set
- DV is number of vowels in sample of \( n = 3 \)
- not symmetrical

large numbers
- \( n \) sample mean error
  \begin{align*}
  n & \quad \text{sample mean} & \text{error} \\
  10 & \quad 2.7 & -0.8 \\
  100 & \quad 3.59 & +0.09 \\
  1000 & \quad 3.542 & +0.042 \\
  10000 & \quad 3.4845 & -0.0155 \\
  1 M & \quad 3.50069 & -0.00039 \\
  \end{align*}

• larger sample yields better estimate of true population mean: smaller error
• error is always present, even for very large samples!

aside: probabilities
\begin{align*}
  p & = P(V) = 0.4 \\
  q & = P(C) = 0.6 \\
  P(DV=0) & = q^3 = 0.238 \\
  P(DV=1) & = pqq + pqq + qpp = 3pq^2 = 0.438 \\
  P(DV=2) & = ppq + pqp + qpp = 3p^2q = 0.269 \\
  P(DV=3) & = ppp = p^3 = 0.055 \\
\end{align*}
large numbers

- larger sample provides better estimate of true (population) mean
  - for a larger sample, the sample mean is less likely to deviate much from the true mean
  - that’s why opinion polls can have ‘only’ N=1000 participants
  - if units are selected at random from population into sample!

Central Limit Theorem

- take a random sample of females (at 50th birthday, n=10 persons)
- DV is number of children per person
  \{0 3 2 1 0 2 1 2 1 0\}
- not symmetrical

Central Limit Theorem

- the means \( \bar{x} \) of infinitely repeated samples are always normally distributed,
- irrespective of the distribution within each sample (normal, binomial, uniform, poisson, ...)
- mathematical proof, without simulations

• compute sample mean \( \bar{x} = 1.2 \)
• now repeat this study, many times
• histogram of many sample means
standard normal distribution
- important
- total area under curve = 1
- mean \( z = 0 \);
- width \( s = 1 \)
- symmetrical
- continuous curve

Central Limit Theorem

distribution of sample means
- its mean is the best unbiased estimate of the population mean
- its width is called standard error of the mean; this SE expresses uncertainty of estimated mean; SE depends on \( n \)

Central Limit Theorem, v.v.
- CLT also works in reverse:
  - if the sample size is large, then one may assume that the sample mean is drawn at random from a standard normal distribution!
  - only for large samples, with \( n > 30 \) or so

Central Limit Theorem
- standard error of the mean \( s_e = \frac{s}{\sqrt{n}} \)
- decreases as sample size \( n \) increases;
- and decreases as standard deviation \( s \) decreases
- larger sample yields a better estimate of population mean \( \mu \) and of population variance \( s^2 \)
statistical testing
• either reject H0 convincing evidence against H0
• or fail-to-reject H0 no convincing evidence against H0
• decision based on incomplete sample risk of error

principle
• if H0 is true, then a particular sample statistic (eg. \( \bar{x} \) or \( t \)) has a known distribution
• if the probability of finding this value for the sample statistic is very low, under H0, then we should reject H0: convincing evidence against H0.

two possible errors
• reject H0 incorrectly, convincing evidence against H0 although H0 is true in reality Type I error
• keep (fail-to-reject) H0 incorrectly, no convincing evidence against H0 although H0 is false in reality Type II error

asymmetric hypotheses
• remember: we’re working to falsify H0
• trying to find convincing evidence against H0
• positive evidence in favor of H0 does not strengthen H0 (Hume’s problem of induction)
• “proving your H0” is very in-efficient!
  — e.g. proving that groups are similar
• how similar do groups have to be (given inevitable sampling error) to reject H0?

chances of errors
• \( P(\text{Type I error}) = a = \text{level of significance} \)
• \( P(\text{Type II error}) = \beta \)
  • \( a \) smaller \( \Rightarrow \) \( \beta \) larger
  • \( a \) larger \( \Rightarrow \) \( \beta \) smaller
• “don’t make \( a \) smaller than it needs to be.” — Devore & Peck (2005: 413)

significance and power
• \( P(\text{Type I error}) = a = \text{level of significance} \)
  - \( P(\text{reject } H0 \mid H0 \text{ true}) \)
• \( P(\text{Type II error}) = \beta \)
  - \( P(\text{not reject } H0 \mid H0 \text{ false}) \)
• power = \( P(\text{reject } H0 \mid H0 \text{ false}) \)
• if H0 false, then power equals\( P(\text{reject } H0 \mid H0 \text{ false}) = 1 - P(\text{not reject } H0 \mid H0 \text{ false}) = 1 - \beta \)
hypothesis testing
H1: dice is biased: $\mu \neq 3.5$
H0: dice is unbiased: $\mu = 3.5$

use $t$ statistic, with $n=6$: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

$t = \frac{2.67 - 3.50}{1.21/0.49} = -0.83/1.68$

$n=60$, df=59
$t = \frac{3.23 - 3.50}{1.23/0.16} = -0.27/1.70$

t=1.70, $p=.047 < .05$, reject H0

OK or not OK?
• in reality, dice was biased so that 6-side was never thrown ($\mu=3.0$).
• test with small sample failed to detect the difference between expected and observed mean outcomes.
• test with large sample successfully detected that the observed mean outcome differed from its expected value.
• large samples are better for detecting a difference: more powerful statistical testing
• "large" means $N>30$

comparing means
• this example showed one-sample hypothesis testing
• the same line of reasoning also applies if sample means from two or more samples are compared: pairwise t-test, ANOVA, etc.
power
depends on...
• d effect size
  (observed–expected value, normalized)
• a level of significance
  (not discussed today)
• n sample size
  (degrees of freedom)

testing difference \( d \)

\[
t = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}}
\]

difference is zero, according to H0
denominator is standard error of difference

• test statistic \( t \) follows \( t \) distribution, with \( df = n-1 \)

• simple case:
  two conditions \( (k=2) \), within-subject (paired), just one score per child per condition, \( N=100 \) children, \( a=.05 \)

• \( m_1=100, s_1=10, m_2=105, s_2=10 \)
  \( m_D = 5, s_D = 10, se_D = 1 \)

smaller sample

• \( m_1=100, s_1=10, m_2=105, s_2=10 \)
  \( m_D = -5, s_D = 10, N=10 \)
• same effect size \( d \approx 0.5 \)

\[
t = \frac{|5 - 0|}{10/\sqrt{10}} = 1.58
\]

• \( t=1.58 \), now \( df=9, p=.074 \), not significant
### Large Sample, Small $d$

- $m_1 = 100, s_1 = 10, m_2 = 101, s_2 = 10$
- $m_D = -1, s_D = 10, N = 500$
- Now effect size $d^* \approx 0.1$
  
  $t = \frac{|m_1 - m_2|}{s_D} = 2.24$  
  
  $t = 2.24$, now $df = 499, p = .013$, significant

- But... meaningful?? interesting??

### Small Sample, Less Variation

- $m_1 = 100, s_1 = 5, m_2 = 105, s_2 = 5$
- $m_D = -5, s_D = 5, N = 10$
- Now effect size $d^* \approx 1.0$

- $t = \frac{|m_1 - m_2|}{s_D} = 3.16$  
  
  $t = 3.16$, now $df = 9, p = .006$, highly significant
power: summary

• if $N$ is very large, then even a very small effect $d$ is significant, but not necessarily interesting!
• smaller effect $d$ yields increment of $\beta$, less power
• reduced variation $s$ yields larger effect size $d$, hence more power.
• larger sample yields decrement of $\beta$, more power

what $N$ is enough?

• children are difficult to obtain
  children that have been most intensively studied
  were bred by investigators themselves
• estimate $D$ and $s$ from pilot studies,
  eg. difference $D = 5$ and $s_D = 10$
• so, effect $d = 5/10 = 0.5$
• choose desired power,
  eg. $1-\beta = 0.8$ so $\beta = 0.2$
• two-sided testing, $a=.05$
• read df from graph ...

... or do it the lazy way ...

on the web: google for power sample size calculator
e.g. http://www.stat.uiowa.edu/~rlenth/Power/
**relation between variables**

Research aims at insight in relations between variables
- between IV and DV
- between two DVs

First: inspect relations in **scattergram**

2D plot of each unit, in dimensions \( x \) and \( y \)

---

**scattergram**

- tarief per KLM naar Europese bestemmingen
- online boeken 2003.12.14, vliegen 2004.01 heen 15 terug 22
- Brussel, Las Palmas, Londen, Sofia

---

**correlation**

- How accurate can \( y \) be predicted from \( x \)? Which part of total variance is predicted, which is residual?
- \( r^2 = \frac{SS_{pred}}{SS_{tot}} \) coefficient of determination
- \( r = \sqrt{r^2} \) with sign of slope \( b_1 \)
- \( r \) indicates strength of relation
  - \(<.5 \) weak, \(.5<r<.8 \) moderate, \( r>.8 \) strong

---

**regression**

- Use \( x \) to predict \( y \), via linear relation
- \( y \) : observed values
  - \( \hat{y} \) : predicted values
    - \( \hat{y} = b_0 + b_1 x \)
- Residual: deviance between (predicted) line and (observed) point
scattergram

tarief per KM naar Europese bestemmingen
afstand (km) prijs (EUR)

scattergram

tarief per KM naar Europese bestemmingen
afstand (km) prijs (EUR)

causality

$r = 0.70$

causality

$r = 0.70$

warning #1

• causality
• a statistical correlation does not imply a causal relation!
• let’s investigate the relation between density of TV sets (people per television) and life expectancy (with countries as units).

causality

$r = -0.86$

causality

$r = -0.86$

warning #2

• do not extrapolate outside domain of x – there is no reason to assume that the linear relationship holds for other values of x (distance <200km or distance >2500km)

causality

$r = -0.61$

causality

$r = -0.61$

warning #1

• causality
• a statistical correlation does not imply a causal relation!
• let’s investigate the relation between density of TV sets (people per television) and life expectancy (with countries as units).

causality

$r = -0.86$

causality

$r = -0.86$

warning #2

• do not extrapolate outside domain of x – there is no reason to assume that the linear relationship holds for other values of x (distance <200km or distance >2500km)
warning #3

• regression is **sensitive to outliers**
  outliers produce a lot of variance, hence
  they exert a strong “pull” on the regression
  line
  e.g. Las Palmas
• always plot and inspect residuals

interlude

insert interlude about here

generalization

• Inductive generalization (from instance
to rule) is an essential learning
mechanism. Humans learn and
generalize easily. This trait has been
favored during evolution of the human
species. In general, we tend to descend
from ancestors who were good at
learning and generalizing.

generalization

• “All generalizations are dangerous,
even this one”
  — Alexandre Dumas Père

• Men are more apt to be mistaken in
  their generalizations than in their
  particular observations.”
  — Niccolo Macchiavelli

generalization

if sampling units
  (intended votes, cheese, 1m speech fragment)
are drawn at random from population
(voting population, batch, speech corpus),
then results may be generalized
to population of interest.

“units” constitute random factor, iff
  – sampled at random, and
  – sample is small fraction (<5%) of population
    because of law of large numbers
two random factors

In studies with two random factors, we would like to **generalize over both random factors** simultaneously... ... but such generalizations are not always warranted.

questions to ask

- What is my range of generalization?
- Is my range of generalization identical to the range (population) of interest?
- In what ways has my experiment affected my possible generalizations?
- Is my empirical base sufficient for valid generalization?

---

two random factors

- **nested** random factors: repeated measurements
  - subject, and measurement within subject, are two random (nested) factors (e.g. phrase within interview; tempo is DV)
- **crossed** random factors
  - subject and item are two random (crossed) factors (e.g. test items and subjects; RT is DV)

hasty generalization

- “jumping to conclusions”
- A tendency to over-generalize from an empirical base which is too small to warrant these conclusions.
- Seductive (and maybe hard-wired)
  - “every child is born from two married parents”
  - “all birds can fly”

QUESTIONS

- What is my range of generalization?
- Is my range of generalization identical to the range (population) of interest?
- In what ways has my experiment affected my possible generalizations?
- Is my empirical base sufficient for valid generalization?
some further reading


some further browsing

- http://www.uvm.edu/ Web Interface for Statistics Education, Claremont Graduate University, Claremont, CA. Contains many tutorials, applets, and other goodies.
- http://www.cuyamaca.net/bruce.thompson/Fallacies/intro_fallacies.asp Fallacy Page (taxonomy of bad reasoning), by Bruce Thompson, Cuyamaca College, El Cajon, CA.