Statistics One : Andrew Conway, Princeton

Stephane Genaud

October 1, 2012

Contents

1 Branches of Statistics: Descriptive and Inferential

1.1 Descriptive Statistics -> Experiment (L1S1)

1.1.1 Randomized

Define :

- a sample of the population
- the **dependent variables** :
  the characteristics to be measured, e.g does (yes or no) a patient has been injected the vaccine sex of the observed person (male, female)
- the **independent variable** :
  the measure, e.g what level of anticorps does a patient have e.g score on a test of spatial reasoning
- counfound
  avoid any bias in the experiment, e.g double-blind for polio

1.1.2 Causality

compare results in the **control group** and the administered

1.1.3 Ways of describing

- Histograms (L2S1)
  show an entire distribution example : Wine testing 30 experts rated overall quality of 4 different wine (scale 0-10) Example shows : rectangle, positive/negative skew (distribution +gros effectif d'abord/après)

- Summary Stastistics (L2S2)

  - Important concepts
    Central tendency (mean,median,mode) (mode = the score that happens most often Variability (std dev, variance) Skew Kurtosis Example in R: 'describe(ratings)', where rating the distribution Example for variability : Jeremy Lin, basket ball player arrray : points per game, X-M : deviation wrt mean, $(\text{X-M})^2$: square the deviation so that we can sum the deviation and divie by number of matches. M = mean = 22.7 = $\sum$ X / N SD = std dev = 9.6 SD^2 = variance = 92.21 = ($\sum$ $(\text{X-M})^2$)/ N (also known as MS : Mean Squares)

- Tools for inferential statistics (L2S3)

  - Important Concepts
    * The normal distribution
      bell shape, symetrical example: body temperature wand measurement : M=100.06, SD=.71
    * Z-scores
      A standardized unit of measurement : convert "raw" score to z-score: Z = (X-M)/SD
    * Percentile rank
      def: the percentage of scores <- a given score e.g body temperature: mean =100.06, std-0.71 mine=100.77, what the percentile rank ? Z=(100.77§100.06)/.71 = 1
      · Calculus of the area under the curve up to x-axis 100.77
      · or look at the Z-table : see 34,1% for std=1, so my percentile rank = 84.1%
    * Probability
      proba and normal distribution: if a choose a student at random, proba that his temperature >= 100.6 ? P(X>100.6) = .5 P(X>100.77)=.159 P(X>103) < .01

- Inferential Statistics
  Assume a normal distrib
    - assume certain values, such as the mean
    - conduct an experiment
    - do the assumptions hold ?

  Safe to assume a normal distrib ???
    - what are you trying to measure
    - what is the construct ?
    - how do you operationalize the construct (see lecture Measurement !)

1.2 Inferential Statistics -> Observational Study (L1S2)

1.2.1 Correlation

1.2.2 Quasi-independent variables

2 Correlation (L4)

2.1 Correlation Examples (L4S1)

def: a statistical procedure to measure and describe the relationship between 2 variables can range [-1;1]. -1 negative correlation, 1 perfect correlation. E.g working memory capacity (X) is strongly correlated with SAT score (Y) Graphically : scatterplot In R : plot(X˜Y) (X on the y-axis, Y on the x-axis)

Caution about correlation:

- accuracy of the prediction will depend on magnitude of the correlation => which depends on the reliability of X and Y, and sampling (random and representative ?)
- validity of the prediction : correlation is a **sample** statistics => does not apply to an individual

Example: Intelligence testing & WW1. Develop an aptitude test:

- multiple choice and short§answer questions (ASVAB today)
- R. Yerkes argued that “native intellectual ability” was unaffected by culture

Statistical analysis to support/refute claim ? Anwser: observe difference in predictibility = correlation. Take two groups: officers and soldiers, and observe if the test is predictive on the job.

Example: Baseball.

2.2 Correlation Calculations (L4S2)

2.2.1 Correlation coefficient r

(aka *Pearson product-moment correlation coef*)

- r = the degree to which X and Y vary together, relative to the degree X and Y vary independently
- $r = covariance(X, Y)/variance(X, Y)$

Fomulae for r: 2 different ways:

- Raw score formula
- Z-score formula

2.2.2 New concept : SP : Sum of Cross Products

- Review: Sum Squares: $SS = \sum_i (X_i - M)^2$
- SP:

– calculate deviation for X and Y
– for each subject, multiply the deviation scores of X and Y:
\$(X-M_X) × (Y - M_Y)\$
– then sum the cross-products: $SP = \sum_{i=1}^{n} (X - M_X) \times (Y - M_Y)$

2.2.3 Formula for r

- Using Raw score : $r = SP_{X,Y}/\sqrt{SS_X \times SS_Y}$
- Using Z-score : $r = \frac{\sum_{i-1}^{N}(Z_x Z_Y)}{N}$

2.2.4 Variance and Covariance

- Variance = SP /N
- Covariance= SP /N
- Correlation is standardized covariance (range -1 to 1)

2.3 Interpretation of Correlations (L4S3)

2.3.1 Validity of a correlation-based argumentation

Assumptions behind correlation analyses:

- normal distributions for X and Y. Detect violation by plotting, adn descriptive statistics.
- linear relationship between X and Y Detect violation by looking at the scatter plot, or more precise : residuals
- Homoskadesticity In a scatterplot the distance between a dot and the regression line reflects the amount of prediction error = **residual**. Homoskadesticity : def: the residuals are not a function of the values of X (residuals look like random values).

2.3.2 Reliability of a correlation

If i go to another sample, will i have the same correlation ?

- one approach is NHST : Null Hypothesis Significance Testing

Consider :

- H_0 = null hypothesis, e.g r=0
- H_A = alternative hypothese, e.g r>0

NHST Assume H_0 is true, then calculate the probability of observing data with these caracteristics, given H_0 is true

- Thus, $p = P(D|H_0)$
- if $p < \alpha$ then reject H_0 else retain H_0.

action	retain H_0	reject H_0
H_0 true	correct	false alarm
H_0 false	type II err	correct

retain H_0 reject H_0

————+————+————

H_0 true $p = 1 - \alpha$ \$p=α\$

————+————+—————

H_0 false	$p = \beta$ (Miss)	$p = 1 - \beta$

- $p = P(D|H_0)$
- Given that the null hypothesis is true, the probability of these, or more extreme data, is p. **NOT** : the probabilit of the null hypothesis being true is p. In other word :
  \$p = P(D|H_0) $\neq p = P(H_0|D)$

2.3.3 NHST application

NHST can be applied to:

- r : is the correlation significantly different from 0
- r1 vs. r2 : is one correlation significantly larger than another

2.4 Reliability and Validity of Correlation (L5S1)

2.4.1 Reliability

Classical test theory

- raw scores (X) are not perfect
- they are influenced by bias and chance error
- In a perfect world, we would obtain a "true" score X = true score + bias + error

A measure (X) is considered to be reliable as it approaches the true score
Methods to estimate reliablility

- test / re-test
  exemple measure temp body of everyone twice: X1 and X2
  However, if the bias is uniform, we wont't detect it
- parallel tests Measure temp body with the wand (X1) and oral thermometer (X2)
  The correlation would reveal a bias of the wand
- inter-item estimates Most commonly used in social sciences
  Example: suppose a 20-item survey is designed to measue extraversion

  - randomly select 10 items to get subset A (X1)
  - the other 10 items become subset B (X2)
  - the correlation between X1 and X2 is an estimate of the reliability

2.4.2 Validity

What is a construct?
An "object" that is not directly observable

- as opposed to "real" observable object
- example, "intelligence" is a construct

How do we operationalize a construct?
The process of defining the conostruct to make it observalbke and quantifiable

- Example: intelligence tests

  Construct Validity

- Example: construct: verbal ability in children
  one way to operationalize: vocabulary test
- content validity: does the test consists of words should know
- convergent validity Does the test correlate with other, established measures of verbal ability? For example, reading comprehension
- divergent validity Does the test correlates less with measures designed in a test of different type of ability? For example, spatial reasoning.
- nomological validity Are the scores on the test consistent with more general theories, for example, of child development and neuroscience For example, a child with neural disease should have smaller scores

2.5 Sampling (L5S2)

2.5.1 Sampling error

Example: Wine testing:

- suppose a population certified experts, N=300

- and suppose the ratings for RedTruck are normally distributed in te population

In that case, M=5.5 and SD=2.22 for N=300 Actually, observed was M=5.93 and SD=2.45 for N=30
Now, take a random sample of N=100 : M=5.47 and Sd=2.19
For a sample of N=10, we could have a large sampling error, M=6, and SD=1.7

The sampling error is the difference between the sample and the population.

- **Problem !**: we typically do not know the population parameters.
- So how do we estimate the sampling error ?

Clearly, depends

- on the size of the sample
- on the variance in the population

2.5.2 Standard error

Standard error is an estimate of amount of sampling error

- $SE = \frac{SD}{\sqrt{N}}$, where SD: std dev of the sample, N: size of the sample

3 R

3.1 Install packages

From console:

```
> install.package("pschy")
> library(psych)
> search() // list loaded pacakges
```

3.2 Script

Example: wine testing (file)

```
Ratings <- read.table("stats1_ex01.txt",header = T) # 1st line = row names
> class(ratings)
  [1] "data.frame"
> names(ratings)
[1] "RedTruck" "WoopWoop" "HobNob"   "FourPlay"
hist(ratings$RedTruck)
# --> plots histo
layout(matrix(c(1,2,3,4), 2, 2, byrow = TRUE))
hist(ratings$RedTruck, xlab = "Ratings", ylab="Number", main="RedTruck")
hist(ratings$HobNob, xlab = "Ratings", ylab="Number", main="HobNob")
hist(ratings$FourPlay, xlab = "Ratings", ylab="Number", main="FourPlay")
hist(ratings$WoopWoop, xlab = "Ratings", ylab="Number", main="WoopWoop")
describe(ratings)  # from the 'psych' package,
summary(ratings)
```