Statistiek I

Some Descriptive Basics

John Nerbonne and Martijn Wieling

CLCG, Rijks*universiteit* Groningen

www.let.rug.nl/nerbonne/teach/Statistiek-I/

Overview

Statistiek I, for ATW, CIW, IK

http://www.let.rug.nl/nerbonne/teach/Statistiek-I

John Nerbonne, J.Nerbonne@rug.nl, H1311.436, di. 11
Martijn Wieling, M.B.Wieling@rug.nl, H1311.434, vr. 10

Ingemarie Donker I.C.Donker@student.rug.nl

Saar Hommes SaarHommes@gmail.com

Jaap Nanninga, JaapNanninga@gmail.com

Annelot de Rechteren Anne.Recht@gmail.com

Lotte Verheijen L.J.J.Verheijen@student.rug.nl

Marieke Wijnbergen, M.Wijnbergen@student.rug.nl

Lab Sessions

- You must enroll for a lab session via Nestor
- Lab leaders listed in *studiehandleiding* w. email addresses.
- Turn in lab reports (all five) to lab leaders *not* to instructor.
- Lab reports due one week after lab. Late labs half credit; labs more than one week late count as zero.

Ingemarie Donker `I.C.Donker@student.rug.nl`
Saar Hommes `SaarHommes@gmail.com`
Jaap Nanninga, `JaapNanninga@gmail.com`
Annelot de Rechteren `Anne.Recht@gmail.com`
Lotte Verheijen `L.J.J.Verheijen@student.rug.nl`
Marieke Wijnbergen, `M.Wijnbergen@student.rug.nl`

- If you cannot enroll via Nestor, please contact the secretaries `seccnl@rug.nl`, *secretariaat cluster Nederlands (4e verdieping)*!

What? and Why?

Statistics—collecting, ordering, analyzing data
Why in general?

- Wherever studies are **empirical** (involving data collection), and where that data is **variable**.
- Most areas of applied science require statistical analysis.
- General education — e.g., political, economic discussion is statistical (see newspapers).

It's *not* mathematics, it's data analysis!

What? and Why?

Statistics—collecting, ordering, analyzing data
Why in general?

- Wherever studies are **empirical** (involving data collection), and where that data is **variable**.
- Most areas of applied science require statistical analysis.
- General education — e.g., political, economic discussion is statistical (see newspapers).

It's *not* mathematics, it's data analysis!

Statistics in Language Studies

- Experiments, large empirical naturalistic studies *inter alia* in communications, information science, linguistics
- Characterizing geographical, social, sexual Δ's
- Processing uncertain input—speech, OCR, text(!)
- Evaluating intuitions (grammatical theory)
  - Bresnan article on dative shift
- Literature
  - Characteristics of authors, genres, epochs diction; sentence structure, length
  - Authorship studies (e.g. *Federalist Papers*)
  - Stemmata in philology (RuG diss, J.Brefeld)
- Link to cultural history, other (digital) humanities

Availability of online data increases opportunities for statistical analysis!

This course

- Practical approach
  - Emphasis on statistical **reasoning**
  - Understand uses (e.g. in research reports)
  - Conduct basic statistical analysis
- Look at data before and during stat. analysis
- De-emphasis on mathematics — **no** prerequisite
- Use of software (SPSS)
  - Illustrates concepts, facilitates learning
  - Bridge to later use simpler
- Topics, examples from language and communications studies

- Weekly lecture (**attendance required**)
- Five exercises with SPSS (labs)
- Six weekly quizzes
- One exam (*in het Nederlands*)

Grades

- Lectures (5 points)
  Attendance required at all lectures. Check based on at least five (of six) times.
- Quizzes (5 points)
  `www.let.rug.nl/nerbonne/teach/Statistiek-I`
- SPSS Labs (10 points); Complete/Incomplete (50% if late less one week)
- Exam (100 points)

Role of SPSS Labs

- “Walk through” case studies
- Think through what statistical software is demonstrating
- Acquire facility with SPSS
- Practice statistical reporting

How to approach labs

- Chance to try out ideas from lecture, book
- Ask whether your labs jibe with theory

How to *waste time* with labs

- Copy results from others
- Go through the motions without thinking

Role of SPSS Labs

- “Walk through” case studies
- Think through what statistical software is demonstrating
- Acquire facility with SPSS
- Practice statistical reporting

How to approach labs

- Chance to try out ideas from lecture, book
- Ask whether your labs jibe with theory

How to *waste time* with labs

- Copy results from others
- Go through the motions without thinking

Role of SPSS Labs

- “Walk through” case studies
- Think through what statistical software is demonstrating
- Acquire facility with SPSS
- Practice statistical reporting

How to approach labs

- Chance to try out ideas from lecture, book
- Ask whether your labs jibe with theory

How to *waste time* with labs

- Copy results from others
- Go through the motions without thinking

Descriptive vs. Inferential Statistics

Descriptive Statistics—describe data without drawing conclusions.
Example: identify average, high & low scores from a set of tests.
Purpose: characterizing data more briefly, insightfully.

Inferential Statistics—describe data, likely relation to a larger set.
Example: reason from **sample** of scores to general conclusions.
Purpose: learn about large **population** from study of smaller **samples**, esp. where larger population is inaccessible or impractical to study.

Note 'sample' vs. 'population.'

Descriptive vs. Inferential Statistics

Descriptive Statistics—describe data without drawing conclusions.
Example: identify average, high & low scores from a set of tests.
Purpose: characterizing data more briefly, insightfully.

Inferential Statistics—describe data, likely relation to a larger set.
Example: reason from **sample** of scores to general conclusions.
Purpose: learn about large **population** from study of smaller **samples**, esp. where larger population is inaccessible or impractical to study.

Note ‘sample’ vs. ‘population.’

Variables and values

We refer to a property or a measurement as a **variable**, which can take on different **values**.

Variable	Typical Values
height	170 cm, 171 cm, 183 cm, 197 cm, ...
sex	male, female
reaction time	305 ms, 376.2 ms, 497 ms, 503.9 ms, ...
language	Dutch, English, Urdu, Khosa, ...
corpus frequency	0.00205, 0.00017, 0.00018, ...
age	19, 20, 25, ...

Variables tell us the the properties of **individuals** or **cases**.

Cases, variables, relations

Terminology: we speak of CASES, e.g., Joe, Sam, but also the FdL web site ... and VARIABLES, e.g. height (h), native language (l), visitors (to site). Then each variable has a VALUE for each case, h_j is Joe's height, and l_s is Sam's native language.

When we examine RELATIONS, we always examine the realization of two variables on each of a group of cases.

- height vs. weight on each of a group of Dutch adults
- effectiveness vs. a design feature of group of web sites, e.g. use of menus, use of frames, use of banners
- pronunciation correctness vs. syntactic category for set of words
- phonetic vs. geographic distance on a group of pairs of Dutch towns

Cases, variables, relations

Terminology: we speak of CASES, e.g., Joe, Sam, but also the FdL web site . . . and VARIABLES, e.g. height (h), native language (l), visitors (to site). Then each variable has a VALUE for each case, h_j is Joe's height, and l_s is Sam's native language.
When we examine RELATIONS, we always examine the realization of two variables on each of a group of cases.

- height vs. weight on each of a group of Dutch adults
- effectiveness vs. a design feature of group of web sites, e.g. use of menus, use of frames, use of banners
- pronunciation correctness vs. syntactic category for set of words
- phonetic vs. geographic distance on a group of pairs of Dutch towns

Tabular Presentation

Example: A test is given to students of Dutch from non-Dutch countries. Variables:

Variable	Values
area of origin	EUrope, AMerica, AFrica, ASia
test score	0-40
sex	Male, Female

Here is part of the results.

area	score	sex
EU	22	M
AM	21	F
⋮	⋮	⋮

Three variables, where score is numeric & each row CASE.
Tables show *all* data, but large tables are not insightful.

SPSS Coding

It is often advantageous to code information in a particular way for a particular software package.
In general, SPSS allows fewer manipulations and analyses for data coded in letters. Use numbers as a matter of course. We therefore recode ‘area of origin’ and ‘sex’ (into numbers).

area of origin	EUrope	AMerica	AFrica	ASia
	0	1	2	3
sex	Male	Female		
	1	2		

Notate bene: this is a weakness in SPSS. In general, it is good practice to use meaningful codings. But in SPSS, this will limit what you can do—use numbers!

Classifying

It is also sometimes useful to group numeric values into classes. We might group scores into 0-16 (beginner), 17-24 (advanced beginner), 25-32 (intermediate), and 33-40 (advanced).

area	score	sex	score class
0	22	1	1
1	21	2	1
2	15	2	0
3	26	1	2
⋮	⋮	⋮	⋮

Grouping numerical information into classes loses information. Care!

Reminder:

area of origin	EUrope	AMerica	AFrica	ASia
	0	1	2	3
sex	Male	Female		
	1	2		

Data/Measurement Scales

nl. *meetniveau*

categorical scales nominal, ordinal

quantitative scales interval, ratio, etc.

Scale determines type of statistics possible, e.g. mean (average).

We can average quantitative data, but not categorical data. We speak of the average (mean) height of an individual (quantitative), but not his average native language (categorical).

Non-Quantitative Measurement Scales

nominal/categorical – categorized, but not ordered:

- part of speech, POS in linguistics, e.g. noun, verb, . . .
- countries, languages, speech disorder, . . .

binary data is categorical, has only two values

- sex, pass/fail, survival/nonsurvival, ...

We COUNT instances of categories in analysis.

ordinal – ordered (ranked), but Δ's not comparable

- rank listing of job candidates
- satisfaction, agreement, etc. in LIKERT SCALES

Circle best answer. "Taxes must decline."

1	2	3	4	5
"strongly agree"				"strongly disagree"

Quantitative/Numerical Scales

interval – ordered, Δ's comparable, but no true zero (needed for multiplication)

- temperature (in Celsius of Fahrenheit)

ratio – like interval *plus* zero available

- height, weight, age
- elapsed time, reaction time

"logarithmic" – like ratio, but successive intervals multiply in size

- Richter scale in earthquakes
- loudness, pitch (auditory perception)
- improvement (in error) rates (often)

Log. scales often result from TRANSFORMATIONS to data.

Measurement (and other) Error

Measurements, but also other records of variable values

- All measurements (records) risk error
  - reaction times; number of site visits, “click throughs”, ..
  - judgments of “good” web sites, movies, products
  - ...
- All large data sets contain errors!

Two quality questions:

- VALIDITY: Do the var. values reflect what’s intended?
  - ’Readability’ often measured in sentence length, use of simple vocabulary
  - Free of errors? Grammatical simple? ...
- RELIABILTY: Do tests give similar (consistent) results? Important in questionnaires.
  - Necessary condition (for validity)!

Distribution

We are often interested not in a particular value of a variable for a particular individual, but rather all the values of the variable and how often they occur.

The DISTRIBUTION of a variable shows its values and how often they occur.

The CENTER and SPREAD refer to the variable's distribution.

Visualizing Distributions

DISTRIBUTION is the pattern of variation of a variable
Example: Number of health web-site visitors for 57 consecutive days.

279	244	318	262	335	321	165	180	201	252
145	192	217	179	182	210	271	302	169	192
156	181	156	125	166	248	198	220	134	189
141	142	211	196	169	237	136	203	184	224
178	279	201	173	252	149	229	300	217	203
148	220	175	188	160	176	128			

stem 'n leaf diagram sorts by most significant (leftmost) digit. As above, ignoring rightmost digit.

1 | 223344444556666677777888889999
2 | 00001111222234455677 7
3 | 00123

Visualizing Distributions

DISTRIBUTION is the pattern of variation of a variable
Example: Number of health web-site visitors for 57 consecutive days.

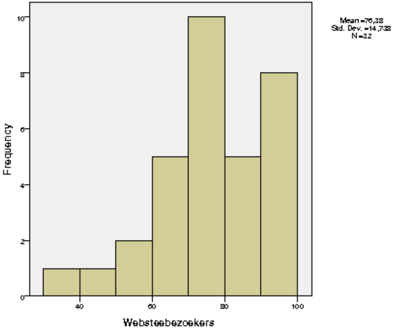
279	244	318	262	335	321	165	180	201	252
145	192	217	179	182	210	271	302	169	192
156	181	156	125	166	248	198	220	134	189
141	142	211	196	169	237	136	203	184	224
178	279	201	173	252	149	229	300	217	203
148	220	175	188	160	176	128			

stem 'n leaf diagram sorts by most significant (leftmost) digit. As above, ignoring rightmost digit.

1	|	223344444556666677777888888889999
2	|	0000111122222344556777
3	|	00123

Displaying Distributions via Histograms

Histograms show how frequently **all** values appear, often require categorization into small number of ranges (≤ 10).

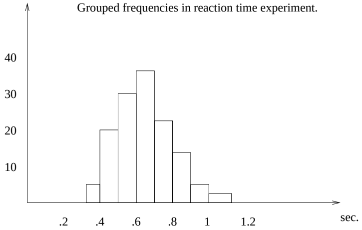


Look for general pattern, outliers, symmetry/skewness.

Distributions of Quantitative Variables

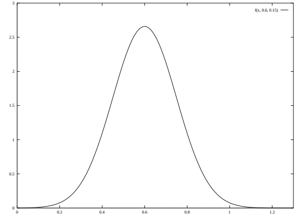
Most **quantitative variables** take any number of values. Variables that take more than about 7 values are often analysed as quantitative e.g., test scores. We often display their frequency distributions by **grouping** values.

Example: histogram of reaction times.



Density Curves

Smoothed curves also *plot area proportional to relative frequency*.
Same reaction time data:



Most very close to 0.6 sec (600ms)

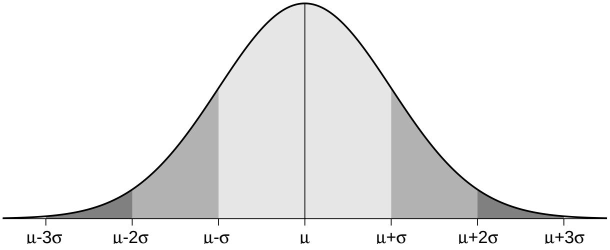
$\neg\Diamond$ (i.e., **don't!**) interpret as ‘*p*% of reaction times = 600ms.’
700ms reaction time $\sim$ 25%

—maybe **no** reaction time was exactly 600ms

Probability Density Curves

- assign fractional values to events, $0 \leq P(e) \leq 1$, (an event is a set of (possible) outcomes)
- area between x & y proportional to relative frequency of values between x & y
- sum to one (all possible events) $\int_{-\infty}^{\infty} P(x)dx = 1$

most famously “normal” distributions –“bell-shaped” curve



Frequency Distributions

Frequency distributions (*frequentieverdelingen*) show how **often** various values occur.

absolute frequency
: How many times values are seen, e.g., 16 *men*, 24 *women*

relative frequency
: What percentage or fraction of all occurrences, e.g., 40% ($= 16/40$) *men*, 60% ($= 24/40$) *women*

: Example: relative frequency of an honest die (flat graph,with each rel. frequency $\approx 1/6$)

Central Tendency

mode most frequent element
the **only** meaningful measure for nominal data

median half of cases are above, half below the median
available for ordinal data.

mean arithmetic average

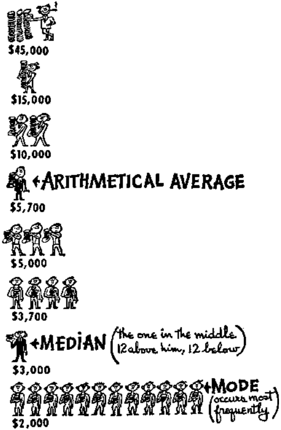
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i$$

μ for populations, m (and $\bar{x}$) for samples

Measures Of “Center”

... need not coincide—from *How to Lie with Statistics*



x-ile's

Quartiles, quintiles, percentiles–divide a set of scores into equal-sized groups

quartiles:

37	68	78	90
49	71	79	90
54	71	79	90
56	73	83	92
60	75	83	94
64	76	85	95
65	77	87	96
65	77	88	97

q_1 1st quartile—-dividing pt between 1st & 2nd groups; q_2—div. pt. 2nd & 3rd (= median!)

percentiles: divide into 100 groups—thus $q_1 = 25$th percentile, median $= 50$th, ...

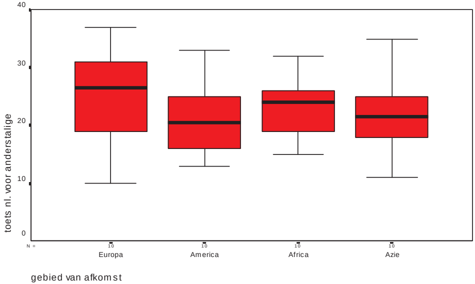
Score at nth percentile is better than n% of scores.

Measures of Spread/Variation

- **none** for nonnumeric data!
  why?
- **minimum, maximum** lowest, highest values
- **range** difference between minimum and maximum
- **interquartile range** $(q_3 - q_1)$ —center where half of all scores lie
- **semi-interquartile range** $(q_3 - q_1)/2$
- "**box-n-whiskers**" diagrams showing q_2 & q_3, range, median

Visualizing Variation—Box Plots

"box-n-whiskers" plot w. q_2, q_3, range, median



Results "Dutch for Foreigners" for four groups of students.
"Boxes" show $q_3 - q_1$, line is median. "Whiskers" show first and last quartiles.

Spread of Quantitative Variables

deviation is difference between observation and mean

variance average square of deviation

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

estimated in samples by dividing by $n - 1$:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

standard deviation square root of variance $\sigma = \sqrt{\sigma^2}$

σ^2 for population, s^2 for sample

—square allows orthogonal sources of deviation (error) to be analyzed $e^2 = e_1^2 + e_2^2 + \cdots + e_n^2$

Spread of Quantitative Variables

deviation is difference between observation and mean

variance average square of deviation

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

estimated in samples by dividing by $n - 1$:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

standard deviation square root of variance $\sigma = \sqrt{\sigma^2}$

σ^2 for population, s^2 for sample

—square allows orthogonal sources of deviation (error) to be analyzed $e^2 = e_1^2 + e_2^2 + \cdots + e_n^2$

Other Properties of Distributions

skew *"scheefheid"* measure of balance of distribution

$$= \begin{cases} - & \text{if more on left of mean} \\ 0 & \text{if balanced} \\ + & \text{if more on right} \end{cases}$$

—seen in SPSS, not used further in this course

Standardized Scores

"Tom got 112, and Sam only got 105"

—What do scores mean?

Knowing μ, σ one can **transform** raw scores into **standardized scores**, aka **z-scores**:

$$z = \frac{x - \mu}{\sigma} = \frac{\text{deviation}}{\text{standard deviation}}$$

Example

Suppose $\mu = 108$, $\sigma = 10$, then

$$z_{112} = \frac{112-108}{10} = 0.4$$

$$z_{105} = \frac{105-108}{10} = -0.3$$

z shows distance from mean in number of standard deviations.

Standardized Variables' Distributions

If we transform **all** raw scores into **z-scores** using:

$$z = \frac{x - \mu}{\sigma} = \frac{\text{deviation}}{\text{standard deviation}}$$

We obtain a **new** variable $\underline{z}$, whose

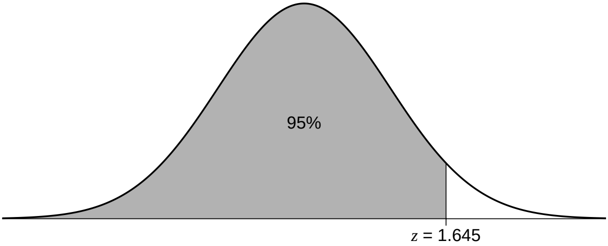
mean is 0
standard deviation is 1

z-score $=$ distance from μ in σ's

uses: interpretation, sampling, hypothesis testing

Interpretation via z-scores

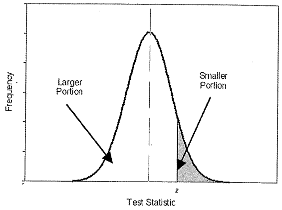
Interpretation of **normal curve** for standardized variables (z):



In every normal curve, 95% of the mass is under the curve below 1.645 standard deviations above the mean.

Normal Curve Tables (based on z-scores)

A.1. Table of the standard normal distribution



z	Larger Portion	Smaller Portion	y	z	Larger Portion	Smaller Portion	y
.00	.50000	.50000	.3989	.12	.54776	.45224	.3961
.01	.50399	.49601	.3989	.13	.55172	.44828	.3956
.02	.50798	.49202	.3989	.14	.55567	.44433	.3951

From Field, Appendix A, pp.797-802

Normal Curve Tables (based on *z*-scores)

z	larger portion	smaller portion	y
⋮	⋮	⋮	⋮
0	.5000	.5000	.3989
⋮	⋮	⋮	⋮
1.96	.9750	.0250	.0584

Note that larger and smaller portions always sum to 1.0!

Note, too, that -z and +z would have the same larger and smaller portions (the curve is symmetric)

where z is the standardized variable:

$$z = \frac{x - \mu}{\sigma} = \frac{\text{deviation}}{\text{standard deviation}}$$

Interpretation via z-scores

If distribution is normal, then standardized scores correspond to percentiles

z	larger portion	smaller portion	y
$\vdots$	$\vdots$	$\vdots$	$\vdots$
0	.5000	.5000	.3989
$\vdots$	$\vdots$	$\vdots$	$\vdots$
1.96	.9750	.0250	.0584

Table specifies the correspondence ($\div$100), containing the fraction of the frequency distribution less than the specified z value.
Tables in other books give, e.g., $1 - (\text{Percentile} \div 100)$.

Interpretation via z-scores

Typical questions, where tables can be applied

- $P(\underline{z} > 1.5) = ?$
  —What's the chance of a z value greater than 1.5?
- $P(\underline{z} \leq 1.5) = ?$
- $P(\underline{z} \leq -1.5) = ?$
- $P(-1 \leq \underline{z} \leq 1) = ?$

We assume normally distributed variables.
Exercises: "Interpretation of Normal Distribution"

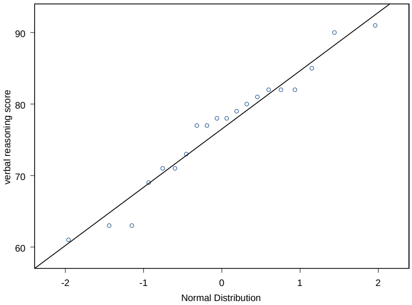
Checking Normality Assumption

Some statistical techniques can only be applied if the data is (roughly) normally distributed, e.g., t-tests, ANOVA.
How can one check whether the data is normally distributed?
Normal Quantile Plots show (roughly) straight lines if data is (roughly) normal.

- Sort data from smallest to largest—showing its organisation into **quantiles**
- Calculate the z-value that would be appropriate for the quantile value (normal-quantile value), e.g., $z = 0$ for 50$^{\text{th}}$ percentile, $z = -1$ for 16$^{\text{th}}$, $z = 2$ for 97.5$^{\text{th}}$, etc.
- Plot data values against normal-quantile values.

Normal Quantile Plots

Example: Verbal reasoning scores of 20 children



Plot expected normal distribution quantiles (x axis) against quantiles in samples. If distribution is normal, the line is roughly straight. Here: distribution roughly normal.

To do

This week:

- Reading (see schedule);
- "Pencil and paper" exercise 1;
- Quiz 1

Next Week: Samples, Sample Means