Statistics

Statistiek I

ATW, CIW, IK

John Nerbonne, J.Nerbonne@rug.nl,
spreekuur H1311.436, di. 11:15-12

Eleanora Rossi, E.Rossi@rug.nl
Mik van Es, M.van.Es@rug.nl

RuG

Statistics

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).

Why Statistics in Humanities Studies?

- Linguistics
  - Experiments *inter alia* in communications, information science, linguistics
  - Characterizing geographical, social, sexual Δ's
  - Processing uncertain input—speech, OCR, text(!)
- History, esp. social, economic
  - advantages of agriculture (over hunting)?
  - economic benefits of slavery (to slaveholders)
  - colonialism and development
- Literature
  - Characteristics of authors, genres, epochs diction; sentence structure, length
  - Authorship studies (e.g. *Federalist Papers*)
  - Stemmata in philology (RuG diss, J.Brefeld)

  Availability of online data increases opportunities for statistical analysis!

Statistics in Humanities

This Course

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

Formal Requirements

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

Grades

- Lectures (5%)
  Attendance required at all lectures. Check based on at least five (of seven) times.
- Quizzes (5%) `www.let.rug.nl/nerbonne/teach/Statistiek-I`
- SPSS Labs (15%); Complete/Incomplete (50% if late less one week)
- Exam (75%)

Role of 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 Statistics

Descriptive Statistics—describe data without trying to make further conclusions.

Example: describe average, high and low scores from a set of test scores.

Purpose: characterizing data more briefly, insightfully.

Inferential Statistics—describe data and its likely relation to a larger set.

Example: scores from **sample** of 100 students justify conclusions about all.

Purpose: learn about large **population** from study of smaller, selected **sample**, esp. where the larger population is inaccessible or impractical to study.

Note ‘sample’ vs. ‘population.’

Common Pitfalls

ignoratio elenchi: (missing the point) the most common error in arguments involving statistics is not mathematical or even technical.

Most common error: getting off track

- “L is a better cold medicine. It kills 10% more germs.”
- “Retail food is a rough business. Profit margins are as low as 2%!”
- “XXX is completely normal. 31.7% of the population reports that they have engaged in XXX.”

Of course, this is **not** limited to statistical argumentation!

Terminology

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**.

A More Formal View

Terminology: we speak of CASES, e.g., Joe, Sam, . . . and VARIABLES, e.g. height (h) and native language (l). 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 of a group 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 only score is numeric, & others nominal. Each row is a CASE.

Tables show *all* data, which is nice, but large tables are not insightful.

Coding

It is often necessary 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. This causes us to recode ‘area of origin’ and ‘sex’, since these were coded in letters.

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!

RuG

Classifying

It is also sometimes useful to group numeric values into classes. We'll group score 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

nonnumeric scales nominal, ordinal

numeric scales interval, ratio, etc.

Scale determines type of statistics possible.

We can average numeric data, but not non-numeric data. We speak of the average height of an individual (numeric), but not his average native language (nonnumeric).

Variable Subtypes—Non-numeric

nominal/categorical – categorized, but not ordered:

- male, female
- part of speech, POS in linguistics, e.g. noun, verb, . . .
- countries, languages, type of artefact, . . .

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

- rank listing of job candidates
- lots of test scores!
- marks of satisfaction, agreement, etc.

```
Circle the answer that most closely fits.
Taxes must decline.
        1          2          3          4          5
    "strongly                                   "strongly
      agree"                                     disagree"
```

Variable Subtypes—Numeric

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

- temperature (in Celsius of Fahrenheit)

ratio – like interval *plus* zero available

- frequency of occurence, e.g. 3 times per week
- height, weight, age
- elapsed time, reaction time

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

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

Measures of 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

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

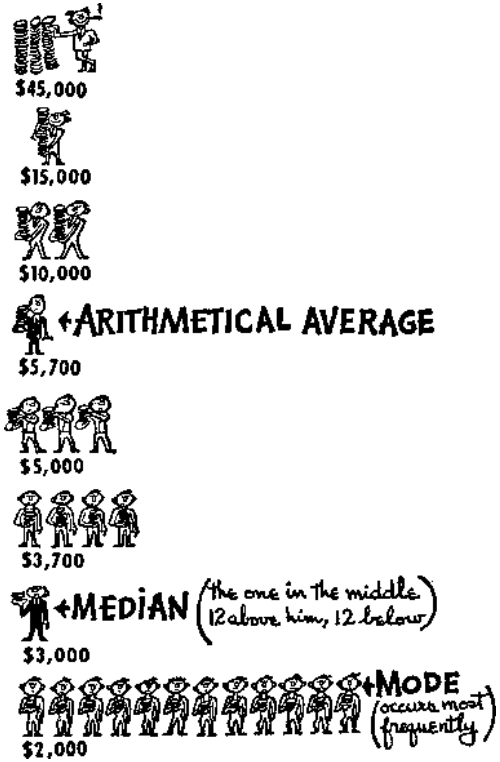
$$\frac{1}{n}\sum_{i=1}^{n} x_i$$

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

RuG

Measures of Central Tendency

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



RuG

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 $q_1 =$ 25th percentile, median $=$ 50th, ...

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

Measures of Variation

none for nonnumeric data!
  why?

minimum, maximum lowest, highest values

range difference between minimum and maximum

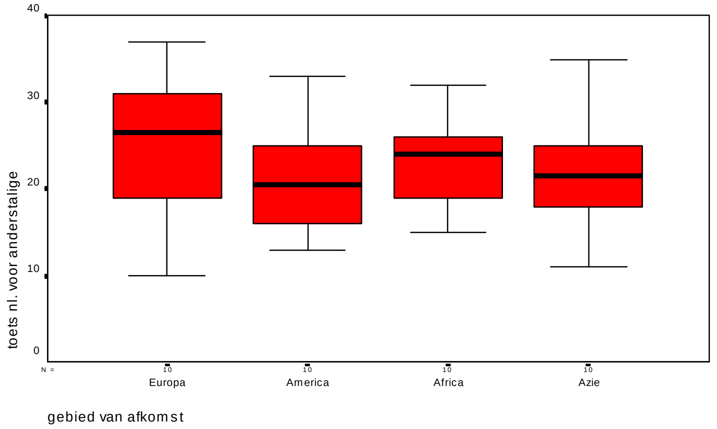
interquartile range $(q_3 - q_1)$ —center where half of all scores lie

semi-interquartile range $(q_3 - q_1)/2$

“box-n-whiskers” diagram showing q_2 & q_3, range
  sometimes median included

Visualizing Variation

“box-n-whiskers” diagram showing q_2 & q_3, range; sometimes median included



Test results “Dutch for Foreigners” for four groups of students.

“Boxes” show $q_3 - q_1$, line is median. “Whiskers” show first and last quartiles.

Measures of Variation

deviation is difference between observation and mean

variance average square of deviation

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

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 $e^2 = e_1^2 + e_2^2 + \cdots + e_n^2$

RuG

Other Statistical Measures

skew *"scheefheid"* measure of balance of distribution

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

kurtosis relative flatness/peakedness in distribution

$$= \begin{cases} - & \text{if relatively flat} \\ 0 & \text{if as expected} \\ + & \text{if peak is relatively sharp} \end{cases}$$

—seen in SPSS, not used further in this course

Other Measures

index numbers e.g., Consumer Price Index, Composite Index of Leading Indicators, Producer Price Index, ... — measures the value of a variable relative to its value at a base period

Example an apple cost Dfl 0.20 in 1990 but Dfl 0.22 in 1995 The apple price index in 1995 with 1990 as base is:

$$\frac{22}{20} \times 100 \ = \ 110$$

- always relative to some fixed base
- therefore *not* per annum percentage changes
  exception: one year after base
- real (composite) indices are weighted averages of simple indices
  weight reflecting relative share of costs, values

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**:

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

Standardized Scores

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

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

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

z shows distance from mean in number of standard deviations.

Standardized Scores

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: sampling, hypothesis testing

RuG

Toward 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			

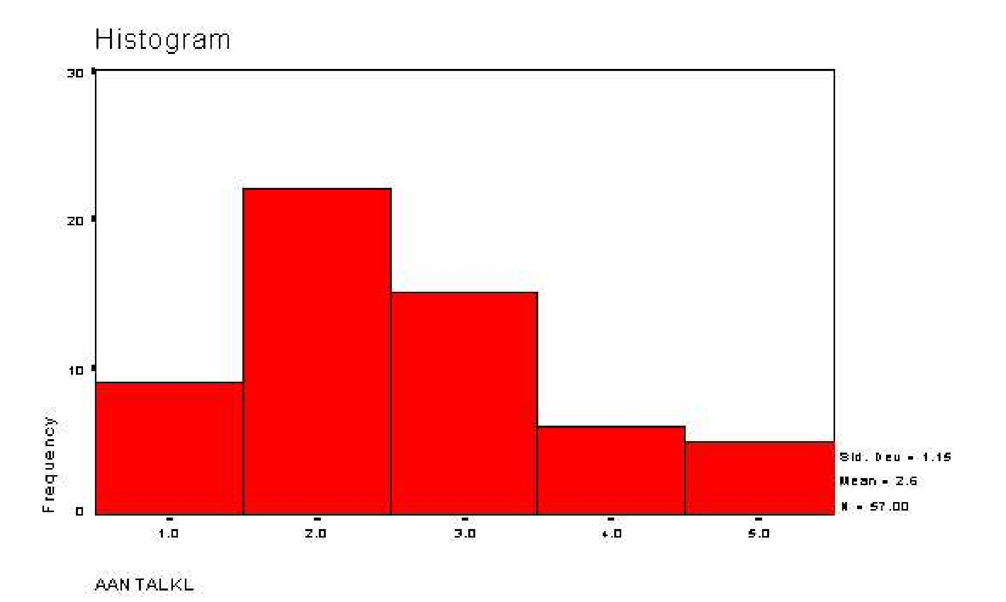
stem 'n leaf diagram sorts by most significant (leftmost) digit. As above, ignoring rightmost digit.

$$\begin{array}{lcl} 1 & | & 223344444556666677777888888889999 \\ 2 & | & 0000111122222344556777 \\ 3 & | & 00123 \end{array}$$

RuG

Displaying Distributions

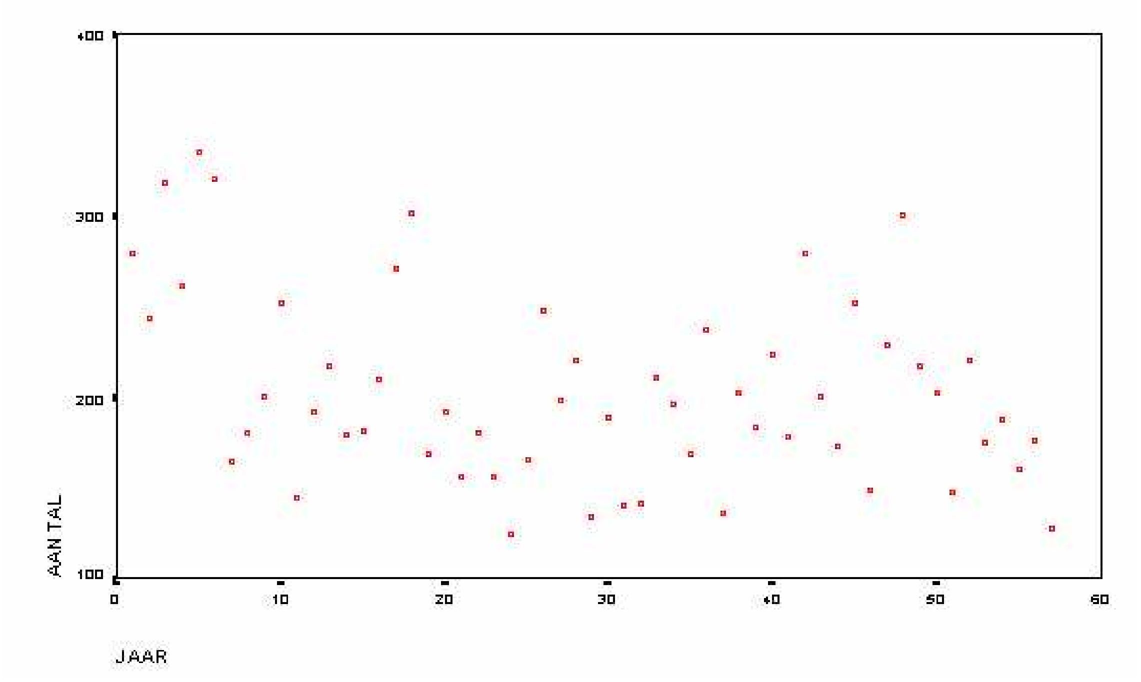
Histograms show how frequently all values appear, often require categorization into small number of ranges (≤ 10).



Look for general pattern, outliers, symmetry/skewness.

Time Series

Same variable at regular intervals e.g., indices, web site visits, ...



Change often focus of attention

RuG

Special—Moving Averages

Some measures fluctuate due to weather, business cycles, chance

moving average sums over overlapping intervals to eliminate some effects of fluctuation

Year	Export	5-yr Ave.	6-yr. Ave
1855	95.7		
1856	115.8		
1857	122.0	116.1	
1858	116.6	124.1	121.8
1859	130.4	126.0	125.0
1860	135.9	126.4	127.7
1861	125.1	132.4	133.4
1862	124.0	138.4	140.0
1863	146.5	144.4	
1864	160.4		
1865	165.8		

from J.T.Lindblad *Statistiek voor Historici*

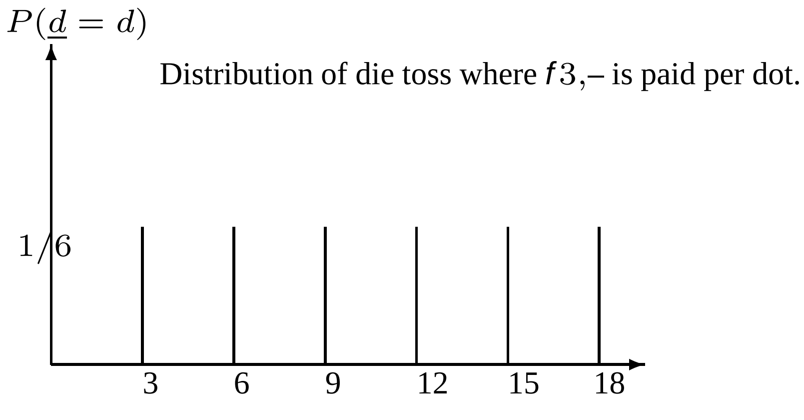
Distribution Functions

Frequency distributions "verdelingen" 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.

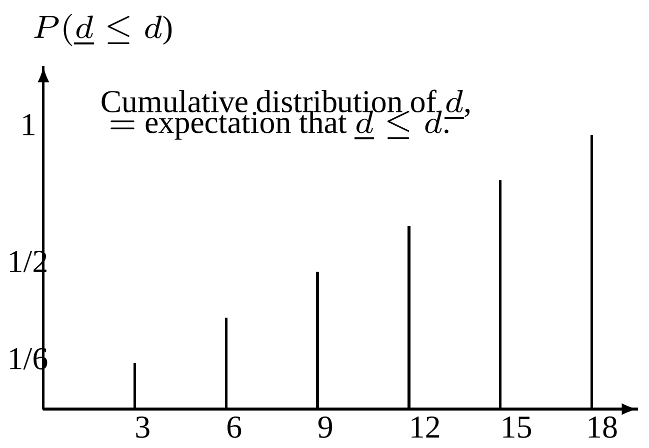


RuG

Distribution Functions

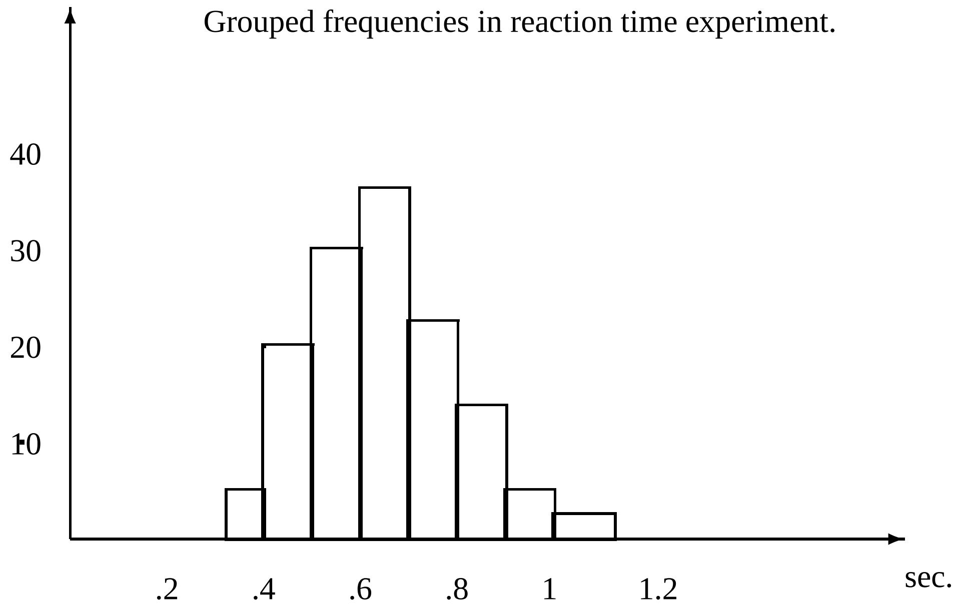
cumulative frequency how often values **at least as large as** a given value occur.

Example: cumulative relative frequency of an honest die.



Numeric Variables

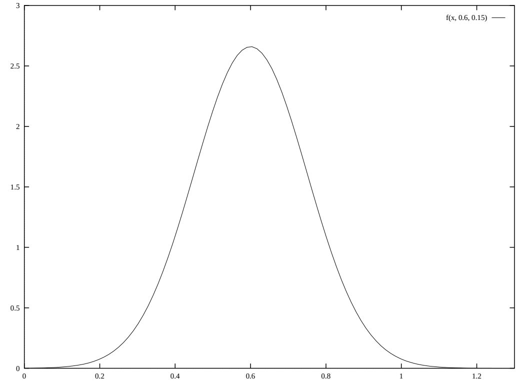
Most **numeric variables** take any number of values. (Ordinal) variables that take more than about 7 values are often analysed as numeric e.g., test scores. We display their frequency distributions by **grouping** values.



RuG

Density Displays

Example: reaction time results appear to fit on the curve

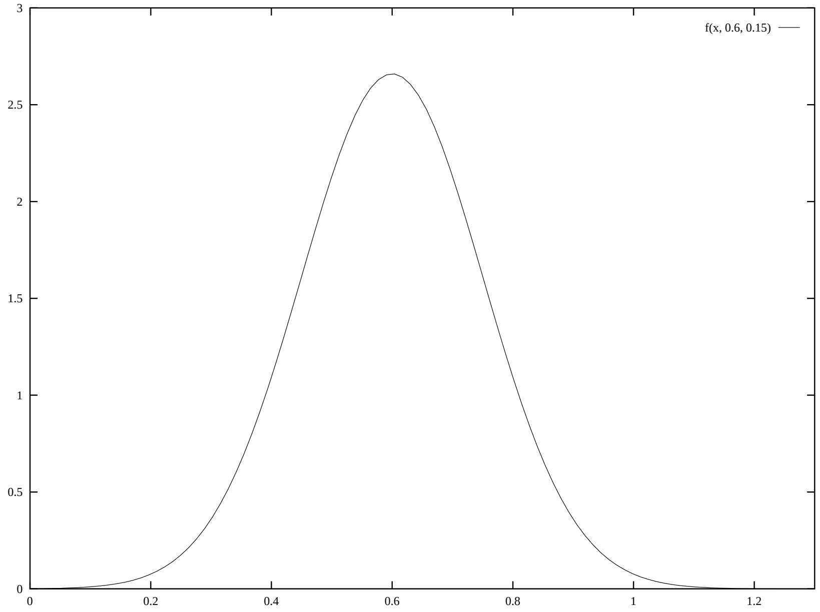


Most very close to 0.6 sec (600ms)

$\neg\diamond$ interpret as ‘p% of reaction times $= 600$ms.’
700ms reaction time ~ 25%

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

Density Displays

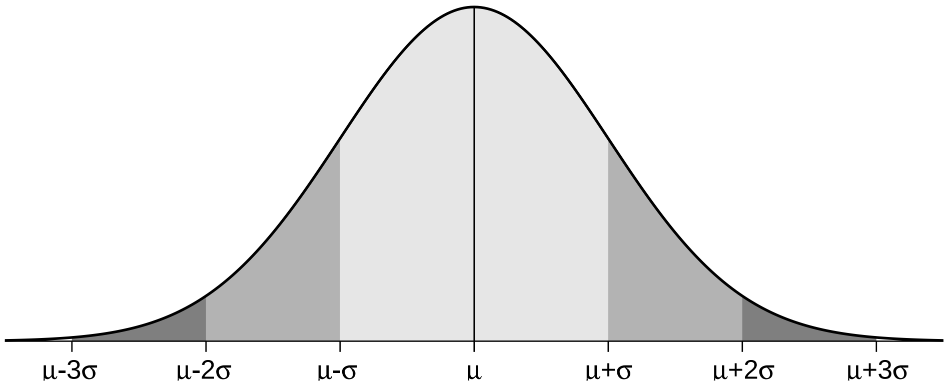


Interpretation: plot frequency DENSITY, so **area** under curve corresponds to percentage of values that fall within area.

Probability Density Functions

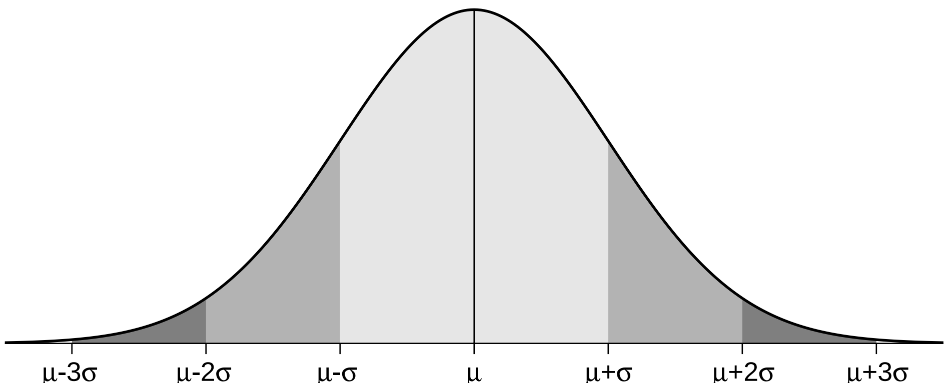
- assign (fractional) values to events, $0 \leq P(e) \leq 1$, where an event is a collection of (possible) occurences
- sum to one (all possible events) $\int_{-\infty}^{\infty} P(x)dx = 1$

lots of possibilities, most famously “normal” distributions—“bell-shaped” curve



Normal Curve

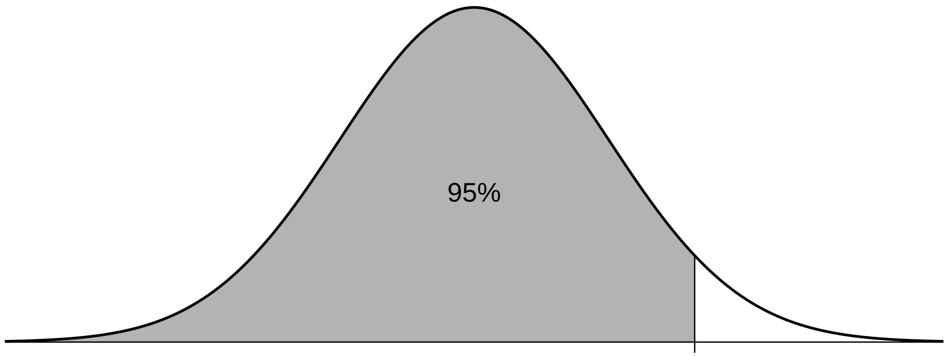
In normal distribution, the mean is always exactly at the center, and the standard deviations appear at fixed proportions. We refer to a particular normal curve using the mean and standard deviation, $N(\mu, \sigma)$, e.g., $N(100, 16)$ (the distribution of IQ's).



Very important in statistics because sample averages are **always** normally distributed.

Normal Curve

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



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

Normal Curve Tables

See M&M, Tabel A, pp.696-97

z	.00	.01	.02	.03	.04	.05	.06	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

where z is the standardized variable:

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

Interpreting z-Scores

If distribution is **normal**, then standardized scores correspond to percentiles

z	.00	.01	.02	.03	.04	.05	.06	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

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., $1 - ($Percentile $\div 100)$.

RuG

Interpreting 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"

Is the Distribution Normal?

Some statistical techniques can only be applied if the data is (roughly) normally distributed, e.g., t-tests, ANOVA.

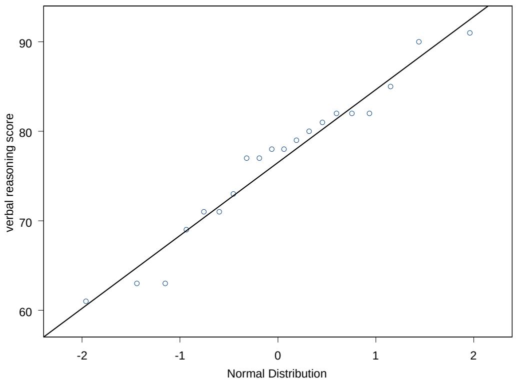
How can one check whether the data is normally distributed?

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^{th} percentile, $z = -1$ for 16^{th}, $z = 2$ for 97.5^{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.

M&M show normal quantile values on x-axis, SPSS on y — but check is always for straight line.

Next — Samples