MS-A0503 First course in probability and statistics

3B Statistical datasets

Jukka Kohonen

Deparment of mathematics and systems analysis
Aalto SCI

Academic year 2020–2021
Period III

Contents

What is statistics (as a science)?

- Applying and developing methods for studying random or uncertain phenomena in the *real world*.
  - The methods are based on the mathematical laws of probability.
  - Sources of uncertainty are many: physical randomness; unknown properties of the real world phenomenon; random sampling on purpose; measurement errors; missing data ...
  - ...generally, the same math applies.

- Roughly:
  - Probability theory tells: How a certain process produces data.
  - Statistics tells: What was the process that produced the data.

- Statistics is applicable whenever you have data; especially if there is any kind of uncertainty or randomness.
  - Most fields of engineering and business have data, so they can (and do) use statistics.

Two basic approaches of statistics

Descriptive statistics

Present and describe the data "as it is", either fully, or in a summary way.

- Tables ("the raw data")
- Graphs (visualization)
- Numerical summaries or "statistics" (e.g. average, minimum, maximum)

Statistical inference

Infer facts about the real phenomenon that lies "behind" the data.
Generalize, e.g. sample $\rightarrow$ population; or measurements $\rightarrow$ universal physical law.

- Stochastic models
- Parameter estimation
- Hypothesis testing

Contents

Statistical data

Typically (not always), statistical data is in a table, the data frame, where

- rows correspond to units, e.g. people
- columns are the variables observed for each unit, e.g. height

Caveat: Different fields of science/engineering use different words. E.g. “units” may be “objects”, “items”, “data points” (geometrically thinking), “records” (in databases)

Depending on number of variables, we may call our data *univariate*, *bivariate*, *multivariate*.

Levels of measurement = What kind of values

- nominal scale ≈ categorical: just distinct classes
  gender: {male, female}
  piece of DNA: {**A**denine, **C**ytosine, **G**uanine, **T**hymine}
- ordinal scale: classes have meaningful order
  cloth size: { XS < S < M < L < XL }
  Likert scale: { str. disagree < disagree < neutral < agree < str. agree }
- numerical: values have arithmetic meaning
  - interval scale: differences $x - y$ are meaningful
    calendar dates, Celsius temperature
  - ratio scale: also quotients x/y are meaningful
    length, weight, distance, Kelvin temperature

Notes:

- all can be *represented* as numbers, e.g. adenine=1, cytosine=2, guanine=3, but arithmetic might not make sense.
- nominal sometimes called "qualitative", but other meanings
- this is *not* the discrete/continuous distinction. Numerical data can be well discrete; e.g. counts (frequencies)

Data set (Data frame)

- data set = sequence of elements (units) of the same type, e.g. numbers, identifiers, or lists of values (one for each variable)
- Often arranged in a table; (R terminology) "data frame"
- Order of units often not meaningful, so we could treat it as a set (or *multiset*, if many identical observations possible)

E.g. course feedback: ((12345A, 5, 1, 5), (98759K, 1, 5, 2), (33312K, 4, 4, 3), (23453B, 4, 4, 3), (21453U, 3, 3, 3))

One string variable (student id), three numerical variables (general satisfaction, workload, usefulness)

Student ID	General	Workload	Usefulness
12345A	5	1	5
98759K	1	5	2
33312K	4	4	3
23453B	4	4	3
21453U	3	3	5

5 units, 4 variables

Average and standard deviation

If we have univariate numerical data: $\vec{x} = (x_1, \ldots, x_n)$

Average (sample mean) $m(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$

Variance $\mathrm{var}(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - m(\vec{x}))^2$

Standard deviation $\mathrm{sd}(\vec{x}) = \sqrt{\mathrm{var}(\vec{x})}$

Eg. $\vec{y} = (0, 0, 1, 1, 2, 2)$

$$m(\vec{y}) \;=\; \frac{1}{6}\Big(0 + 0 + 1 + 1 + 2 + 2\Big) \;=\; 1$$

$$\mathrm{var}(\vec{y}) \;=\; \frac{1}{6}\Big((0-1)^2+(0-1)^2+(1-1)^2+(1-1)^2+(2-1)^2+(2-1)^2\Big) \;=\; \frac{2}{3}$$

$$\mathrm{sd}(\vec{y}) \;=\; \sqrt{\frac{2}{3}} \;\approx\; 0.8165$$

Caveat: Sometimes $n-1$ used as divisor for variance and sd, for technical reasons; more about this later (in parameter estimation).

Example

Calculate sample mean and standard deviation for the following data sets

$$\begin{aligned}
\vec{x} &= (1, 1, 1, 1, 1), \\
\vec{y} &= (0, 0, 1, 1, 2, 2), \\
\vec{z} &= (0, 2, 0, 2, 0, 2, 0, 2, 0, 2), \\
\vec{w} &= (\underbrace{0, 0, 0, 0, \ldots, 0, 0, 0, 0}_{\text{666666 times}}, 1000000, \underbrace{0, 0, \ldots, 0, 0}_{\text{333333 times}}).
\end{aligned}$$

Dataset	Mean	SD
$\vec{x}$	1	0.0000
$\vec{y}$	1	0.8165
$\vec{z}$	1	1.0000
$\vec{w}$	1	999.9995

Average and standard deviation are *summaries*, they do not tell everything about the data. (Just like in probability distributions.)

Computing the summary statistics *from data*

Notation	Name	R	Matlab	Excel
$m(\vec{x})$	Average	`mean()`	`mean()`	`AVERAGE()`
$\mathrm{sd}(\vec{x})$	Standard deviation	`sqrt(1-1/n)*sd()`	`std(,1)`	`STDEV.P()`
$\mathrm{sd}_s(\vec{x})$	(Corrected) std.dev.	`sd()`	`std()`	`STDEV.S()`
$\mathrm{var}(\vec{x})$	Variance	`(1-1/n)*var()`	`var(,1)`	`VAR.P()`
$\mathrm{var}_s(\vec{x})$	(Corrected) variance	`var()`	`var()`	`VAR.S()`
$\mathrm{cov}(\vec{x}, \vec{y})$	Covariance	`(1-1/n)*cov()`	`cov(,1)`	`COVARIANCE.P()`
$\mathrm{cov}_s(\vec{x}, \vec{y})$	(Corrected) covariance	`cov()`	`cov(1)`	`COVARIANCE.S()`
$\mathrm{cor}(\vec{x}, \vec{y})$	Correlation	`cor()`	`corrcoef()`	`CORREL()`
$q_{0.5}(\vec{x})$	Median	`median()`	`median()`	`MEDIAN()`
$q_{0.25}(\vec{x})$	Lower quartile	`quantile( ,.25)`	`quantile( ,.25)`	`PERCENTILE.INC( ,.25)`
$q_{0.75}(\vec{x})$	Upper quartile	`quantile( ,.75)`	`quantile( ,.75)`	`PERCENTILE.INC( ,.75)`

Caveat. Terminology and notation varies across sources. For technical reasons, many computer programs offer the so-called "unbiased" or "Bessel-corrected" *sample variance* and *sample standard deviation*, where the divisor is $n - 1$ instead of n. (Don't worry now – we will talk about this in parameter estimation.)

Contents

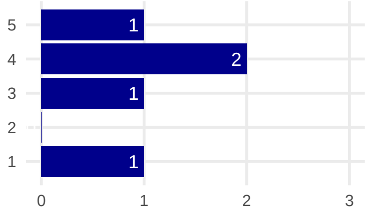
Counts = Frequencies of values

The count, or (absolute) frequency of a value x, in the univariate dataset $\vec{x}$, is

$$n_{\vec{x}}(x) \;=\; \#\{i : x_i = x\}$$

Course feedback, pick one variable "General" $\to$ univariate data $(5, 1, 4, 4, 3)$. Frequency as a table and a (horizontal) bar chart:

x	1	2	3	4	5
$n_{\vec{x}}(x)$	1	0	1	2	1



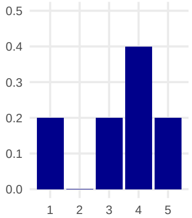
Relative frequencies

The proportion, or relative frequency of value x in dataset $\vec{x}$ is

$$f_{\vec{x}}(x) = \frac{n_{\vec{x}}(x)}{n} = \frac{\#\{i : x_i = x\}}{n}$$

Course feedback, pick "General", dataset $(5, 1, 4, 4, 3)$, relative frequencies as a table and (vertical) bar chart

x	1	2	3	4	5
$f_{\vec{x}}(x)$	$\frac{1}{5}$	0	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{1}{5}$



Observation: $\sum_x f_{\vec{x}}(x) = 1$, thus $f_{\vec{x}}(x)$ is a probability distribution!
It is the empirical distribution of the dataset $\vec{x}$.

Empirical distribution

Proposition

If an element X is chosen uniformly at random, from the dataset $\vec{x} = (x_1, \ldots, x_n)$, then X is a discrete random variable, whose density corresponds to the empirical distribution: $f_X(x) = f_{\vec{x}}(x)$. Furthermore,

$$\mathbb{E}(X) = m(\vec{x}), \quad (1)$$

$$\mathrm{SD}(X) = \mathrm{sd}(\vec{x}), \quad (2)$$

$$\mathrm{Var}(X) = \mathrm{var}(\vec{x}). \quad (3)$$

Also, for any function g, we have

$$\mathbb{E}[g(X)] = \frac{1}{n}\sum_{i=1}^{n} g(x_i). \quad (4)$$

Example

For the dataset $\vec{y} = (0, 0, 1, 1, 2, 2)$, determine the empirical distribution, and its mean and standard deviation.

The relative frequencies are

y	0	1	2
$f_{\vec{y}}(y)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

If random variable Y has density $f_{\vec{y}}(y)$, then

$$\mathbb{E}(Y) \;=\; \sum_{y=0}^{2} y f_{\vec{y}}(y) \;=\; 0 \times \frac{1}{3} + 1 \times \frac{1}{3} + 2 \times \frac{1}{3} \;=\; 1,$$

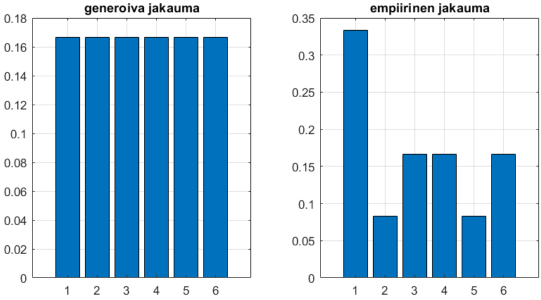
$$\text{Var}(Y) \;=\; \sum_{y=0}^{2} (y-1)^2 f_{\vec{y}}(y) \;=\; (0-1)^2 \times \frac{1}{3} + (1-1)^2 \times \frac{1}{3} + (2-1)^2 \times \frac{1}{3} \;=\; \frac{2}{3}$$

$$\implies m(\vec{y}) = \mathbb{E}(Y) = 1$$

$$\implies \text{sd}(\vec{y}) = \sqrt{\text{var}(\vec{y})} = \sqrt{\text{Var}(Y)} = \sqrt{\tfrac{2}{3}} \approx 0.8165$$

Generating vs. empirical distribution

Rolling a fair die. Generating distribution (stochastic model) versus empirical distribution from 12 rolls that were (5,1,6,4,3,1,1,6,2,4,1,3).



- The empirical distribution *is* a probability distribution and you can use all tools of probability distributions.
- But *different distribution* from the generating one.
- For large *n* they are reasonably close (by LLN).

Contents

Binning (Grouping)

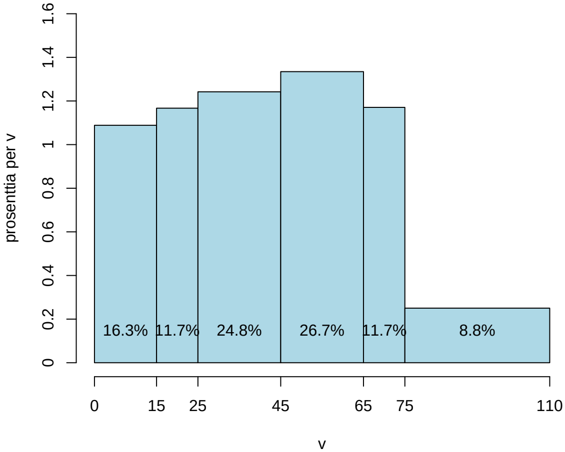
Eg. ages of all Finns 31.12.2015.
$n = 5\ 487\ 308$ units (data points)
Not probably good idea to draw as individual points (especially if ages are expressed in 1-day precision)
Let us group the data into bins at predefined boundaries.

Age (yr)	Frequency
0–14	896 023
15–24	640 387
25–44	1 363 155
45–64	1 464 640
65–74	642 428
75–	480 675

In fact we have transformed our real-valued variable (age) into a discrete one (group or bin index).

Example: Histogram with unequal widths

Finnish age distribution 31.12.2015 [Source: Tilastokeskus]



Age (yr)	Frequency
0–14	896 023
15–24	640 387
25–44	1 363 155
45–64	1 464 640
65–74	642 428
75–	480 675

The bars are an *approximation* of the true density function.

Could you find the proportion of Finns in the 1-year interval $[13, 14)$?

What about the the interval $[109, 110)$ years? Would it be accurate?

How to draw a histogram (allowing unequal bin widths)

- One bar for each bin (interval of possible values)
- Bar width = width of the interval
- Bar height = relative frequency *divided* by width

Example: Age distribution, first bar:

- Represents Finns with ages in interval $[0, 15)$
  Note: values *strictly* smaller than 15; age in whole years 0–14
- Bar width = 15 years
- Frequency = 896023, relative frequency $896023/5487308 \approx 16.3\%$
- Bar height = $16.3/15 \approx 1.09$ (unit: % per year).
- Then bar *area* is the relative frequency.

(Typically, we use equal-width intervals, but not always.)

Contents

Bivariate data

Bivariate data = sequence (or multiset) of *pairs*

$$\vec{xy} \;=\; ((x_1, y_1), \ldots, (x_n, y_n)).$$

Alternatively, a pair $(\vec{x}, \vec{y})$, where $\vec{x} = (x_1, \ldots, x_n)$ and $\vec{y} = (y_1, \ldots, y_n)$ are univariate data (note: as ordered sequences, so we know which x_i and y_i belong together).

Course feedback: Two variables "General" and "Usefulness" composed as a bivariate dataset ((5,5), (1,2), (4,3), (4,3), (3,3))

Univariate statistics $m(\vec{x}), m(\vec{y}), \mathrm{sd}(\vec{x}), \mathrm{sd}(\vec{y})$ are surely useful, but they tell nothing about the dependence between variables. Covariance and correlation tell (some aspects of) dependence.

$$\mathrm{cov}(\vec{x}, \vec{y}) \;=\; \frac{1}{n} \sum_{i=1}^{n} (x_i - m(\vec{x}))(y_i - m(\vec{y}))$$

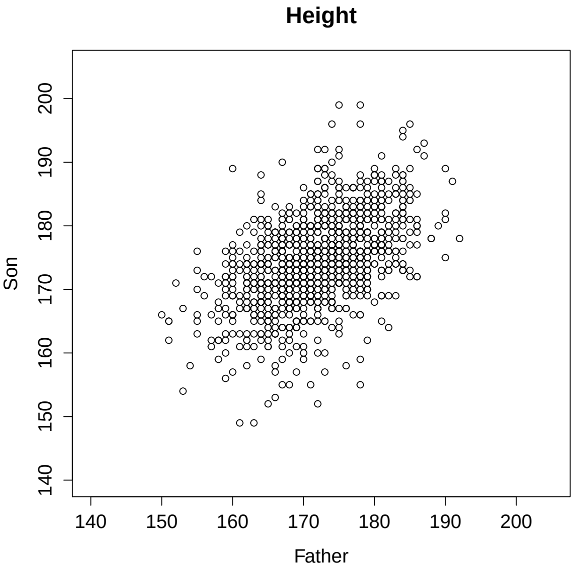
$$\mathrm{cor}(\vec{x}, \vec{y}) \;=\; \frac{\mathrm{cov}(\vec{x}, \vec{y})}{\mathrm{sd}(\vec{x})\,\mathrm{sd}(\vec{y})}$$

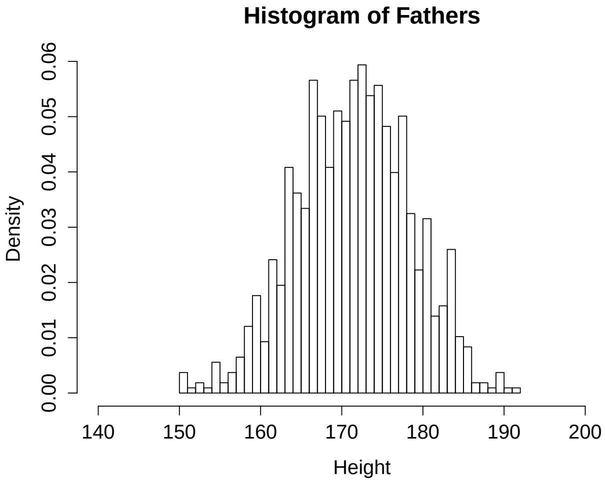
Example: Heights of father-son pairs

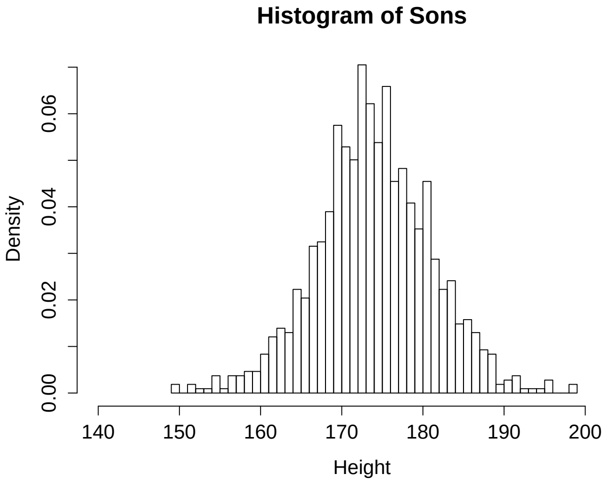
I	P	I	P	I	P	I	P	I	P	I	P	I	P	I	P	I	P	I	P	I	P	I	P	I	P	I	P	I	P	I	P	I	P	I	P	I	P	I	P
165	152	165	171	179	174	178	178	184	187	174	168	171	174	165	178	178	184	173	165	178	170	172	176	178	181	170	161	168	170	182	173	179	179	175	199	180	168	173	173
161	161	166	171	181	172	177	178	173	192	173	168	169	172	169	178	181	183	175	165	177	170	175	175	178	182	158	162	167	170	185	173	179	179	178	155	157	172	173	172
165	161	167	170	180	174	179	177	172	192	174	167	170	172	168	177	183	182	158	168	156	172	175	175	177	180	161	163	167	171	161	176	184	178	184	184	160	171	173	173
167	159	167	171	182	173	180	178	185	196	181	169	169	173	167	177	184	182	160	169	160	174	175	175	177	182	166	164	167	170	163	176	186	179	151	165	163	170	173	174
155	163	167	172	186	172	181	178	159	156	159	170	173	173	169	178	170	186	164	169	163	172	175	176	182	181	165	163	171	170	165	175	163	181	186	185	162	171	173	173
160	163	167	171	160	175	183	179	173	189	160	170	172	172	169	179	172	187	162	168	162	172	175	176	183	181	169	164	169	171	166	176	164	180	173	182	166	171	174	173
166	163	167	170	162	177	183	178	165	180	163	171	173	172	174	177	175	186	164	167	165	173	176	175	184	180	169	164	171	171	167	177	167	182	160	166	164	171	174	173
164	163	170	172	166	177	192	178	187	191	162	172	173	174	173	178	177	186	164	168	164	173	177	176	184	182	151	165	171	170	168	175	170	180	168	155	164	171	176	173
168	164	169	171	164	176	164	181	160	157	165	171	175	173	173	178	177	186	168	169	167	174	177	177	164	184	160	166	171	171	167	176	170	181	181	183	171	173	176	173
170	163	171	170	169	177	169	180	159	163	166	171	176	172	173	180	179	185	168	168	168	172	178	176	170	184	163	167	173	170	167	177	174	182	173	180	170	159	176	174
150	166	171	171	167	176	167	181	178	159	165	171	175	174	175	179	184	185	167	169	167	173	178	175	172	182	163	167	173	170	170	176	174	181	166	157	165	169	176	174
160	166	169	171	167	176	170	181	181	187	174	175	175	174	175	179	185	186	170	168	168	172	181	177	173	183	164	165	173	170	170	175	173	182	179	171	176	167	178	173
162	167	173	172	168	177	170	182	181	169	166	158	176	173	175	179	173	188	171	169	168	173	181	175	176	183	166	167	174	170	171	176	174	182	174	164	162	174	177	172
163	166	172	171	169	176	173	181	176	158	163	163	177	173	176	178	174	190	171	168	171	173	182	176	175	184	168	167	175	171	173	176	174	181	157	161	164	172	179	173
164	166	174	172	169	177	171	180	178	166	168	168	179	172	177	177	181	188	174	169	170	172	163	179	178	183	168	167	177	171	172	176	175	181	165	161	168	171	181	173
165	165	173	171	169	175	172	181	166	163	175	167	178	174	177	178	183	189	172	169	170	173	166	178	178	183	170	165	178	172	174	176	175	182	168	162	168	171	183	174
169	165	175	172	172	176	175	180	158	159	164	170	180	174	179	179	190	189	172	169	169	173	168	179	180	184	173	165	177	171	174	175	178	180	151	162	169	171	184	173
167	166	175	171	174	176	176	181	162	161	168	170	179	173	178	179	181	191	174	168	171	174	167	178	181	184	174	167	178	172	174	176	177	182	159	162	168	170	160	177
171	165	177	170	172	176	175	180	167	161	167	170	184	173	180	177	174	196	178	169	170	173	166	177	183	182	156	169	155	173	175	176	178	180	162	162	167	172	162	174
170	166	179	169	172	177	177	181	173	160	169	171	184	174	181	179	153	154	155	170	173	174	169	179	185	184	161	168	159	174	175	175	178	181	164	163	170	170	164	175
172	165	178	170	173	176	179	182	157	162	167	170	159	176	181	178	162	158	159	172	173	174	171	178	173	186	162	167	163	174	176	175	181	181	167	164	170	169	166	175
177	166	185	172	176	175	178	181	162	162	169	171	162	175	182	177	160	169	163	171	172	173	171	178	175	186	163	168	163	173	175	176	179	181	168	164	169	170	167	175
159	169	160	173	174	176	178	180	165	164	171	172	164	176	186	177	163	149	163	170	173	173	173	177	175	186	165	168	165	174	175	175	184	181	179	162	170	171	168	175
162	169	160	174	174	176	180	181	167	162	170	171	165	175	162	180	188	178	162	171	174	172	172	180	177	186	164	168	165	174	177	175	186	180	158	165	170	170	167	175
164	168	162	173	174	176	181	182	169	164	169	172	166	176	168	181	183	180	164	170	174	174	173	178	181	187	168	168	168	172	177	176	164	185	164	166	172	171	169	175
165	168	164	174	174	175	184	181	175	163	171	170	167	176	168	181	155	165	164	170	175	173	176	177	183	185	169	168	168	173	179	176	170	183	163	166	172	170	171	176
164	168	164	174	179	176	186	180	157	166	171	170	168	175	171	180	166	179	173	173	175	173	175	179	185	185	169	167	169	173	179	175	172	184	165	166	171	170	171	175
167	169	168	174	178	175	190	182	163	165	173	171	167	177	169	180	183	176	164	162	176	173	176	178	167	190	171	169	168	174	180	176	174	184	166	166	173	170	170	175
166	167	167	173	178	176	171	184	162	167	172	171	170	176	173	181	181	165	168	160	174	173	174	177	174	188	171	169	167	174	181	176	176	184	166	166	174	170	174	177
167	168	168	174	178	176	171	183	165	165	173	170	171	177	172	180	166	175	166	167	178	174	176	178	180	189	171	168	170	174	183	175	174	183	168	167	174	171	172	176
170	167	166	173	181	177	172	184	166	165	173	170	171	175	173	180	175	177	177	171	177	174	178	178	184	188	172	169	171	174	161	179	177	182	168	164	176	172	173	176
169	167	167	174	180	176	176	183	165	166	175	171	172	175	174	180	164	159	166	171	178	172	177	179	184	188	174	168	170	174	165	179	178	184	171	165	177	172	172	175
169	169	170	173	183	176	175	184	167	167	176	170	172	175	176	181	163	161	166	172	182	174	178	179	175	192	172	167	169	174	167	178	179	184	172	167	178	171	172	175
172	167	169	174	190	175	178	183	168	167	176	170	174	176	176	181	169	161	168	171	184	174	180	177	187	193	176	169	171	172	168	177	180	183	178	166	179	172	176	176
173	169	170	173	165	177	177	184	170	165	179	170	172	177	177	181	170	160	168	170	183	174	180	177	178	199	177	169	170	173	169	178	182	184	161	169	160	172	175	177
172	168	170	173	167	178	179	184	172	166	178	172	174	175	178	181	158	162	168	170	155	176	181	179	173	157	152	171	172	173	168	179	185	184	162	167	161	173	175	176
174	168	170	174	167	178	181	184	175	164	179	170	174	176	177	182	162	163	168	170	160	176	184	178	164	177	158	171	174	174	170	177	173	186	163	168	164	174	175	177
176	169	170	172	167	178	184	183	158	167	159	172	175	176	178	180	166	163	168	170	164	177	185	179	154	158	162	171	172	173	170	178	172	187	164	168	163	174	175	176
182	169	172	173	167	179	184	183	159	167	160	172	174	175	180	181	167	164	170	170	166	176	164	181	190	181	162	170	173	173	172	179	176	186	165	169	164	172	175	175
159	171	173	173	171	179	172	185	162	169	161	174	176	176	179	181	168	163	170	170	165	177	165	181	179	182	164	170	172	174	173	178	179	186	167	169	164	174	178	175
164	171	172	173	171	177	173	186	166	169	162	174	174	176	184	181	172	162	171	172	166	176	167	180	155	166	164	170	175	172	174	177	181	187	167	169	168	173	179	176
164	170	174	172	172	179	175	185	164	169	164	174	174	175	182	180	153	167	171	172	167	176	171	180	166	183	166	170	176	174	173	178	184	187	168	168	167	173	178	175
162	170	172	174	174	178	178	186	167	168	165	174	177	176	185	181	160	165	171	172	168	175	171	180	181	181	165	173	176	174	176	179	183	185	171	169	167	173	177	175
166	170	175	173	174	177	182	187	168	168	167	174	179	175	168	183	161	167	173	171	171	176	172	182	182	164	176	171	176	174	176	178	164	188	169	168	167	173	181	176
165	170	176	173	173	180	180	185	168	169	168	174	178	175	171	183	165	164	174	170	169	175	173	182	159	166	172	160	175	173	174	179	172	189	169	168	169	173	181	177
167	176	175	173	176	178	183	186	167	169	169	173	178	176	172	183	165	165	172	170	171	175	174	180	181	177	165	171	177	174	175	179	178	188	170	169	170	174	185	177
174	175	176	174	175	178	160	189	171	168	166	173	181	176	175	182	165	165	173	170	171	176	172	180	169	165	176	175	177	173	177	179	180	188	173	168	170	174	166	179
169	157	176	173	176	178	173	189	170	169	170	173	181	177	176	183	168	167	175	171	173	175	176	182	171	155	165	170	178	172	177	177	185	189	172	168	171	174	168	178
165	167	174	174	176	179	178	187	171	169	170	173	184	176	177	183	168	166	175	171	172	175	176	181	159	160	165	171	179	173	179	179	175	191	174	167	170	174	166	179
177	175	177	173	177	177	180	188	172	167	170	174	164	178	178	184	169	165	179	172	173	175	176	182	165	162	167	171	184	173	181	179	184	194	176	167	170	172	167	179

Table: 1000 pairs of heights (father, son) from Pearson's data. Note that this not just two data sets, but a set of *pairs* that belong together.

Scatterplot (scatter diagram)



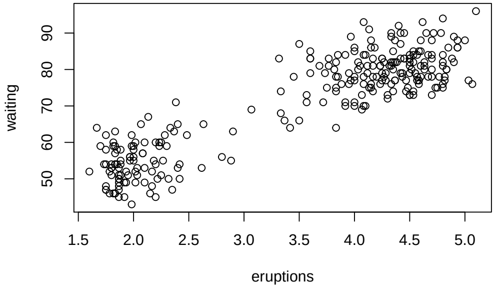

Histogram of Fathers
Density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
140
150
160
170
180
190
200
Height



Histogram of Sons
Density
0.00
0.02
0.04
0.06
140
150
160
170
180
190
200
Height

Example: Eruptions of Old Faithful geysir

Scatterplot of 272 eruptions of *Old Faithful* (Yellowstone).
Two variables: eruption length and waiting time to next eruption.



Already a visual inspection ("eyeballing") reveals interesting patterns (that are not captured by mean and standard deviation).

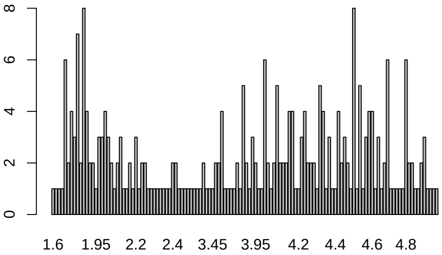
You can find the data in R, try `faithful` and `help("faithful")`

Old Faithful: bar chart of one variable. . .

We could try listing *all different values* of eruption length, and collect their frequencies (within $n = 272$)

x	1.6	1.667	1.7	1.733	1.75	...	5.1
$n_{\vec{x}}(x)$	1	1	1	1	6	...	1

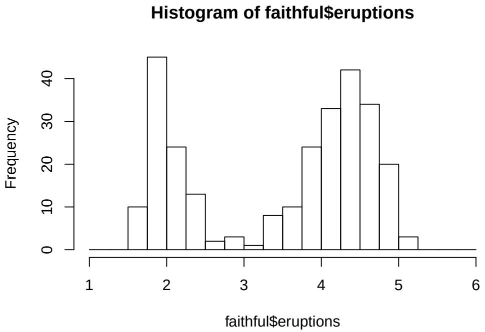
and draw a bar chart



Neither the table or the bar chart seems very informative.

Old Faithful: histogram

But if we group the data into 0.25-minute intervals such as $[2.00, 2.25)$, and plot the counts, we have a better picture of the distribution.



Try this on your own! What happens if you use more (finer) intervals? What if you use less (coarser) intervals?

Cross-tabulation (Contingency table)

The frequency of the pair (x, y), in the data set, is

$$n_{\vec{x}\vec{y}}(x, y) \; = \; \#\{i \; : \; x_i = x \;\text{ and }\; y_i = y\}$$

Course feedback: "General" and "Usefulness" as bivariate dataset ((5,5), (1,2), (4,3), (4,3), (3,3)) has this contingency table:

	y					
x	1	2	3	4	5	Sum
1	0	1	0	0	0	1
2	0	0	0	0	0	0
3	0	0	1	0	0	1
4	0	0	2	0	0	2
5	0	0	0	0	1	1
Sum	0	1	3	0	1	

Cross-tabulation of relative frequencies

The relative frequency of the pair (x, y) is

$$f_{\vec{x}\vec{y}}(x, y) \;=\; \frac{\#\{i \;:\; x_i = x \;\text{ and }\; y_i = y\}}{n}$$

	y					
x	1	2	3	4	5	Sum
1	0	$\frac{1}{5}$	0	0	0	$\frac{1}{5}$
2	0	0	0	0	0	0
3	0	0	$\frac{1}{5}$	0	0	$\frac{1}{5}$
4	0	0	$\frac{2}{5}$	0	0	$\frac{2}{5}$
5	0	0	0	0	$\frac{1}{5}$	$\frac{1}{5}$
Sum	0	$\frac{1}{5}$	$\frac{3}{5}$	0	$\frac{1}{5}$	

$\sum_{x,y} f_{\vec{x}\vec{y}}(x, y) = 1$, thus $f_{\vec{x}\vec{y}}(x, y)$ is a probability distribution.
$f_{\vec{x}\vec{y}}(x, y)$ is the empirical joint distribution of the dataset $\vec{x}\vec{y}$.

Empirical joint distribution

Proposition

If a pair (X, Y) is chosen uniformly at random, from the dataset $\vec{x}\vec{y} = ((x_1, y_1), \ldots, (x_n, y_n))$, it is a discrete random variable whose (joint) density is the empirical distribution $f_{X,Y}(x, y) = f_{\vec{x}\vec{y}}(x, y)$ and also

$$\begin{aligned} \mathbb{E}(X) &= m(\vec{x}), & \mathbb{E}(Y) &= m(\vec{y}), \\ \mathrm{SD}(X) &= \mathrm{sd}(\vec{x}), & \mathrm{SD}(Y) &= \mathrm{sd}(\vec{y}), \\ \mathrm{Var}(X) &= \mathrm{var}(\vec{x}), & \mathrm{Var}(Y) &= \mathrm{var}(\vec{y}), \end{aligned} \tag{5}$$

and

$$\mathrm{Cor}(X, Y) = \mathrm{cor}(\vec{x}, \vec{y}), \tag{6}$$

$$\mathrm{Cov}(X, Y) = \mathrm{cov}(\vec{x}, \vec{y}). \tag{7}$$

Also, for any two-argument function g, we have

$$\mathbb{E}[g(X, Y)] = \frac{1}{n} \sum_{i=1}^{n} g(x_i, y_i). \tag{8}$$

Contents

Quantiles/Percentiles of data

Suppose data can be ordered from smallest to largest.
(OK for numerical or ordinal data; not for nominal data.)

If $0 < p < 1$, then the p-quantile (or $100p$-percentile) $Q(p)$ is roughly the point x such that *proportion* p of the data is smaller than x, and $1 - p$ is greater.

- $Q(0.25)$ is lower (first) quartile; 25% of data is below
- $Q(0.5)$ is median or second quartile; 50% of data is below
- $Q(0.75)$ is upper (third) quartile; 75% of data is below

Note that half of data is between lower and upper quartiles.

R: `quantile(x,p), summary(x), median(x)`

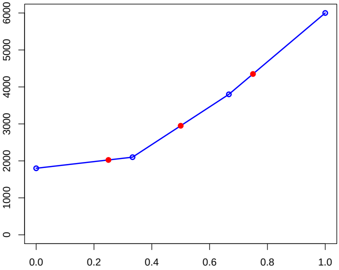
The "roughly" is because in finite data, you may not find exact quarters. There are some (varying) conventions for this.

Quantile function

One way to define the quantile function of dataset $(x_1, \ldots, x_n)$:

- Order the data as $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$
- Divide the horizontal unit interval $[0, 1]$ into equal parts, at points $p_k = (k-1)/(n-1)$, $k = 1, \ldots, n$
- Plot the points $(p_k, x_{(k)})$ and connect with lines

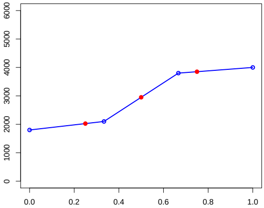
Example. Four salaries (eur/month): $3800, 1800, 2100, 6000$



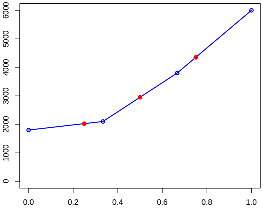
Quartiles = Evaluate the quantile function at 0.25, 0.50, 0.75

Example: Three small datasets

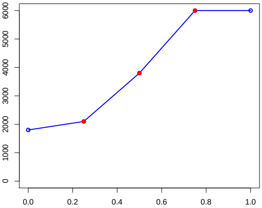
$$\vec{x} = (1800, 2100, 3800, 4000)$$
$$\vec{y} = (1800, 2100, 3800, 6000)$$
$$\vec{z} = (1800, 2100, 3800, 6000, 6000) \qquad (n \text{ not divisible by four})$$



$Q_x(0.50) = 2950$,
$m(x) = 2925$



$Q_y(0.50) = 2950$,
$m(x) = 3425$



$Q_z(0.50) = 3800$,
$m(x) = 3940$

Contents

Sample, population and “population”

The finite dataset you have, the “sample”, is often thought to “represent” qualities of a larger “population”.

sample	population
Pearson’s 1000 fathers and sons	All father-son pairs in . . . ?
1000 poll responses	Opinions of 5 million Finns now
272 eruptions of Old Faithful	All its eruptions (past? future? potential?)
Drug effect on 30 patients	Drug effect on future patients
100 rolls of a loaded die	Potential infinite sequence of rolls

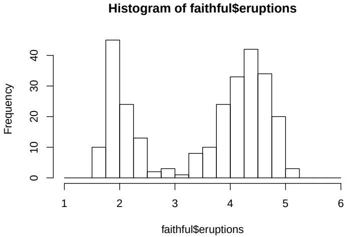
Population is statistical jargon for

- where your data came from (data-generating mechanism; data source)
- what you are trying to understand by looking at the data

Hence, terms such as *sample mean* and *population mean*.
The “population” may be quite concrete, or a figure of speech.

Old Faithful, once again

We have a *sample* of 272 eruption lengths. The real physical mechanism may be complicated, but perhaps we can think the lengths are as if they come from one particular distribution f. But what distribution?



One eruption length is a random variable X from some generating distribution or underlying distribution or true distribution. ("Population" if you want.)

Empirical distribution (shown above) approximates the generating distribution. Why? Think of the event $\{2.0 \leq X < 2.25\}$. We know by LLN that its *relative frequency* (long-run average) $\approx$ its *probability* (in generating distribution).

Next lecture is about inference.

From the data, we will estimate some parameters concerning the reality “behind” the data.