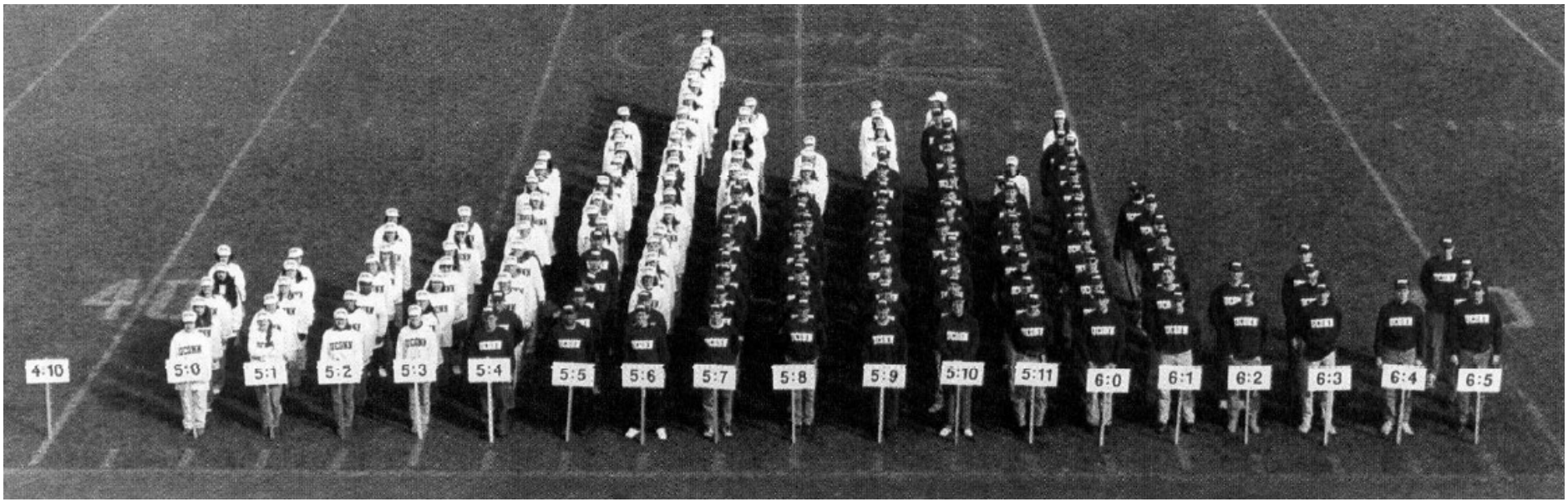
Introduction to Biostatistics



Arranged by Linda Strausbaugh (Genetics 147:5, 1997)

Introduction to Biostatistics
Lecture 1B and 2

Henrik Renlund

October 10-11, 2016

Contents of Lecture 1-2

Data and descriptive statistics

Probability theory and models

Sampling, SE(M) and CLT.

Visual tests (of normality)

What shall we learn today (and tomorrow)?

- Data description
  - Graphs
  - Tables and summary measures
- Probability Models
  - Glimpse at theory
  - Normal distribution and some properties
  - Some properties of samples and the Central Limit Theorem.

Types of data

A data set contains one or more *variables*, typically arranged as:

Individual	Gender	Age	Albumin	Diabetes	Happiness
1	M	29	3.92	0	⌣
2	F	37	4.12	0	⌢
3	F	25	4.75	1	—
⋮	⋮	⋮	⋮	⋮	

Theoretical data categories:

- Numerical
  - discrete; typically integer valued 1, 2, . . . , like `Age`, *or*
  - continuous; i.e. any value in an interval, like `Albumin`.
- Categorical
  - nominal, e.g. `Gender`, *or*
  - ordinal, e.g. `Happiness`: ⌢, −, ⌣.

How statisticians spend their time

In practice data might be stored in the wrong format.

Check this prior to analysis! This is especially important if data has been transferred (between formats, different OS, etc).

Common problems:

- date- and categorical data suddenly stored as integers
- numerical values stored as text (due to ',' vs. '.')

Usually one only needs to distinguish

- measurements (`Age`, `Albumin` are numerical), and
- categories (Gender, `Diabetes`, `Happiness` are categorical).

"Table 1"

It is useful to provide a summary table of the variables you are working with. Choice of descriptive measures may be context dependent.

variable / *value*	Diabetes No *mean*	Diabetes No *sd*	Diabetes Yes *mean*	Diabetes Yes *sd*
Age	32.0	15.9	32.5	14.1
Albumin	4.20	0.37	3.80	0.50
	percent	*n*	*percent*	*n*
Gender				
M	64%	27	52%	22
F	36%	15	48%	20
Happiness				
⌒	61%	19	36%	15
—	23%	7	36%	15
⌣	16%	5	36%	12

Some points on tables (for publication)

- Table and caption should be self contained.
- Every table should be referred to in text.*
- Put captions *above* the table.
- Avoid excessive precision. (Mean age 65.63?)
- Use adequate measures of location and spread.

*The table should illustrate some part of the data relating to the hypothesis of the paper. Referring to a table should fit the narrative.

Generally bad: "Table 3 displays values of biomarkers in group A and B".

Better: "The mean level of measured biomarkers are higher in group A compared to B (Table 3)."

What about missing data?

Missing data is usually a problem for observational studies.

Typically data is missing in explanatory variables.

Suppose we investigate death within one year of myocardial infarction with a model that includes gender, age, BMI (some missing) and smoking status (some missing).

The statistical software default is to include only those individuals with complete case data on all variables that you include.

Complete case analysis will only give an unbiased result if the reason that a variable is missing has nothing to do with the actual value.

If the missing mechanism is known this might be utilized in the analysis.

Solutions. . . ish

There is no trick that guarantees a non-biased analysis.

Single imputation e.g. replace missing value with a "typical" value for that variable (e.g. mean or median). This method underestimates the variance in the variable and will give overly optimistic results.

Multiple imputation create multiple imputed data sets where the missing values are replaced differently in each iteration (e.g. drawn at random from the non-missing values).

> *"It is not that multiple imputation is so good; it is really that other methods for adressing missing data are so bad."*
>
> *(Donald Rubin)*

N.B. we typically do not impute our outcome data!

Visualization of (continuous) data

A sufficiently small data set might not need visualization.

The level of the protein albumin was recorded in a sample (of size 8) of mice (56 days old):

1.88 2.03 2.11 1.77 2.04 2.05 1.94 1.95

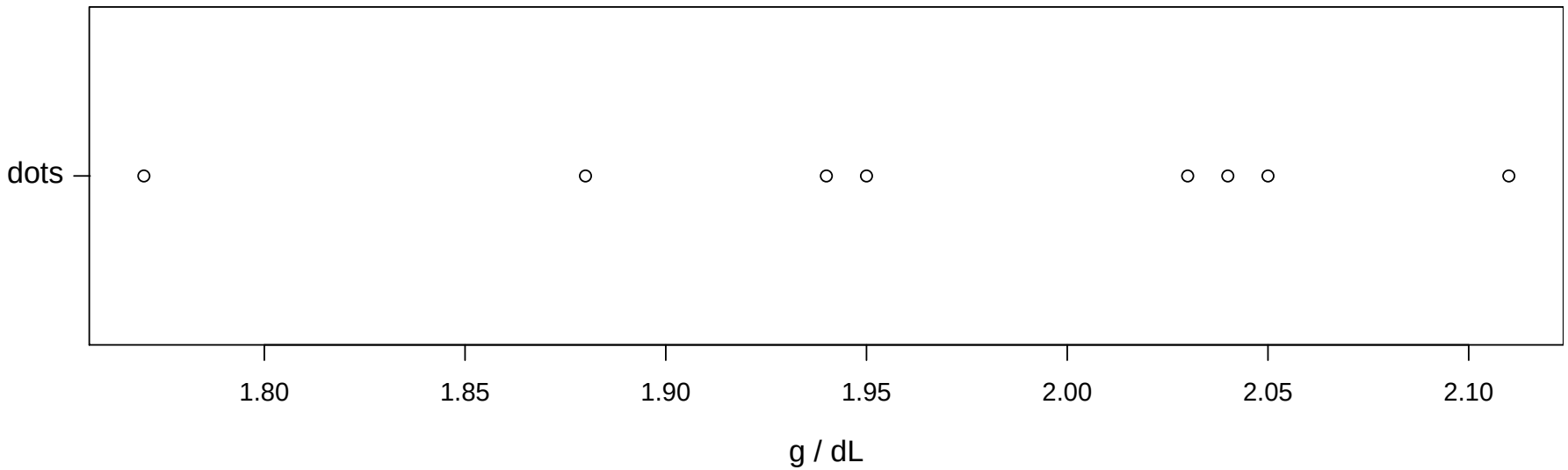
(measured in g/dL - this data set will return in lecture 3).

One simple way to get some handle on data is to order it:

1.77 1.88 1.94 1.95 2.03 2.04 2.05 2.11

Dotplot of albumin data

A dotplot is a one dimensional plot of the data.



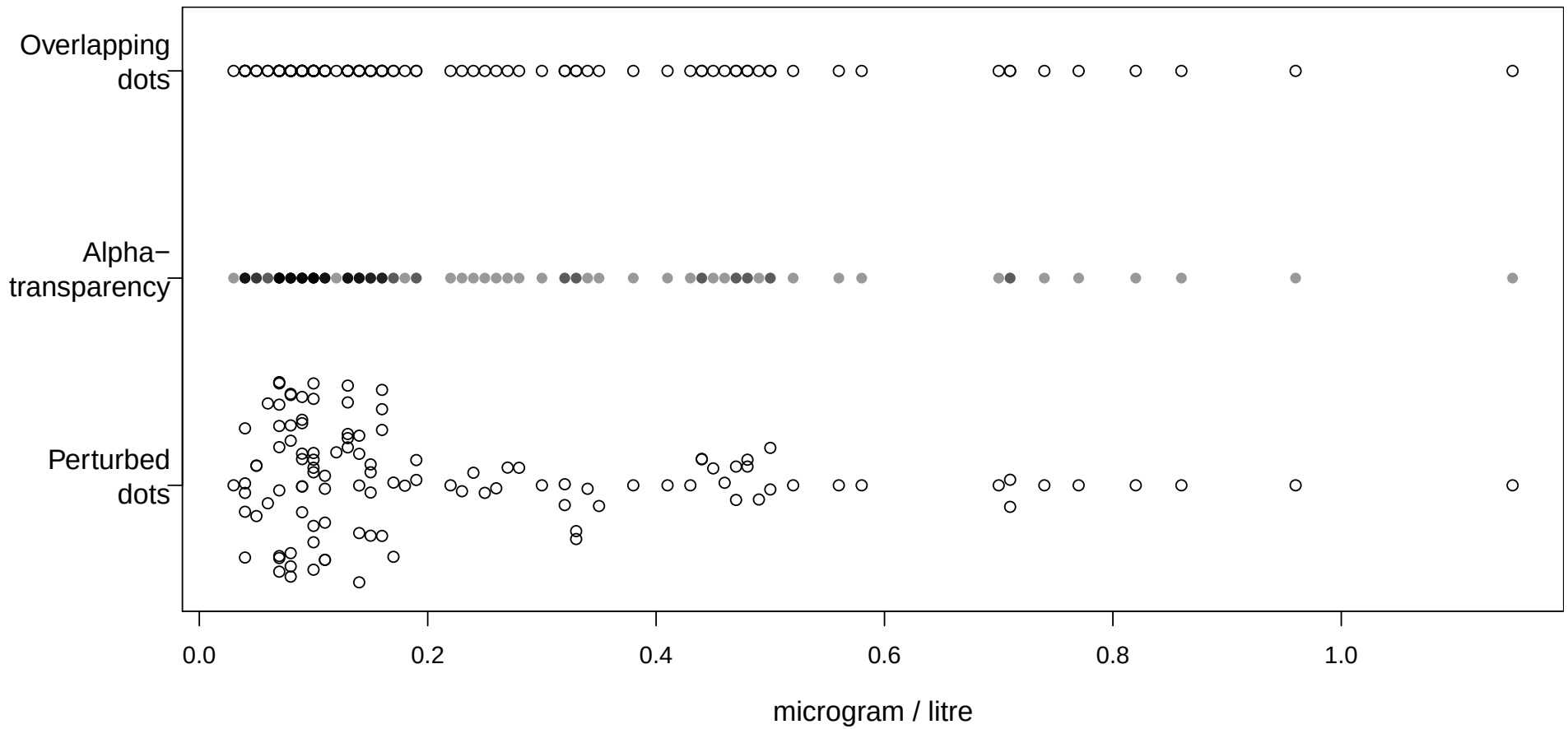
If there are non-unique (or close) points, the data set may appear smaller than it really is.

This can be alleviated by

- perturbation, or,
- alpha transparency.

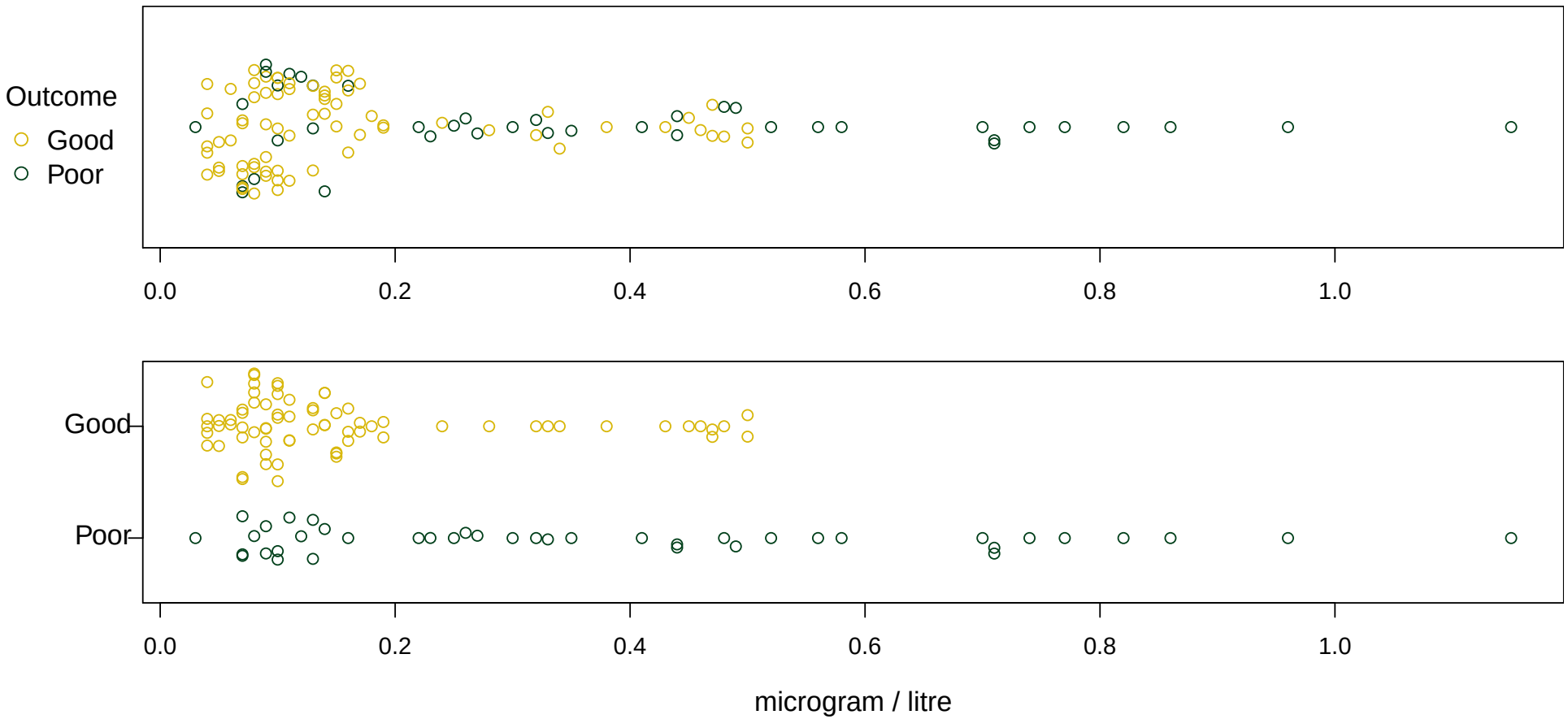
Subarachnoidal bleeding

A biomarker - the protein S100β - was measured for 113 individuals with aneurysmal subarachnoid hemorrhage. (To return in lecture 10.)



Dotplot and groups

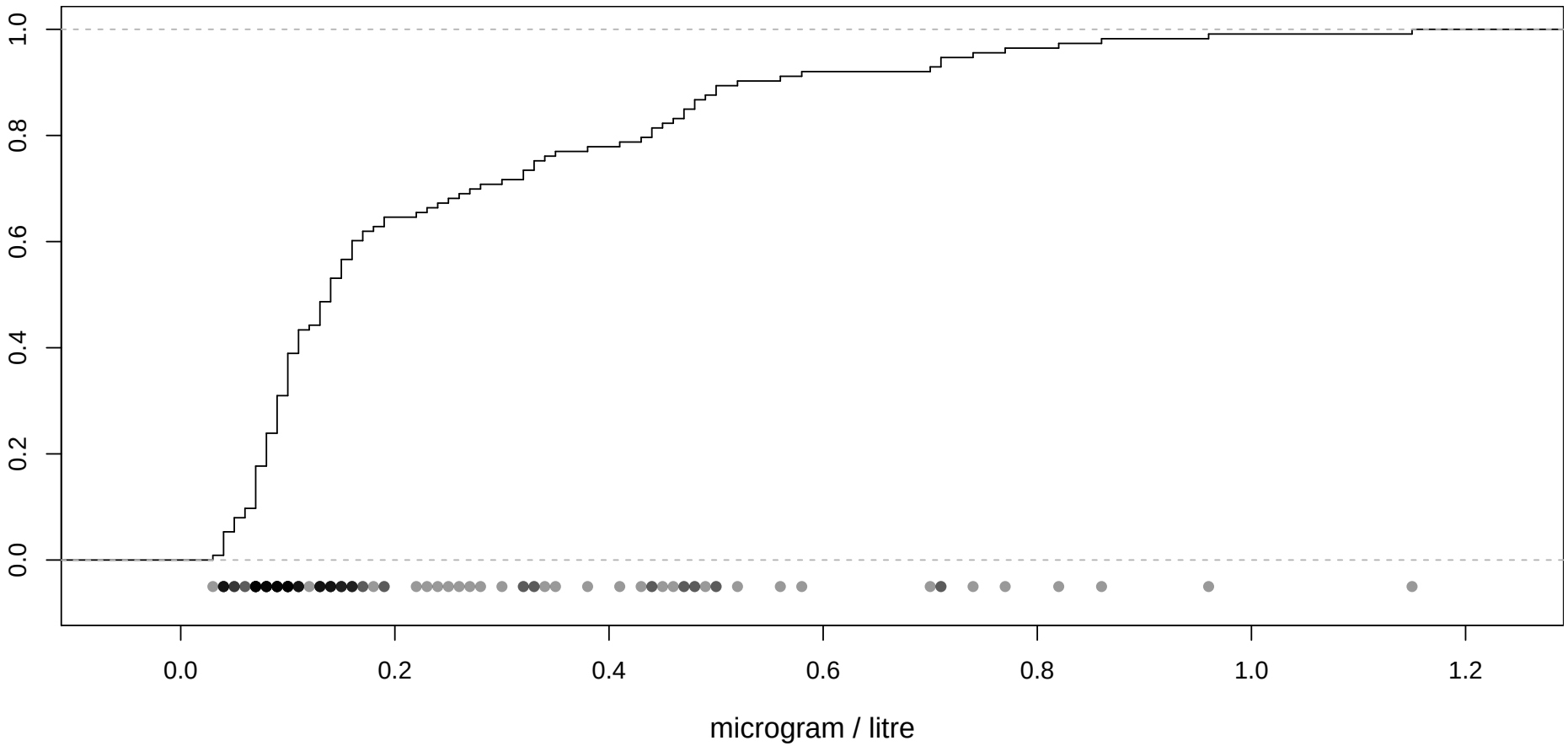
Dotplots can display groups.



Percentiles

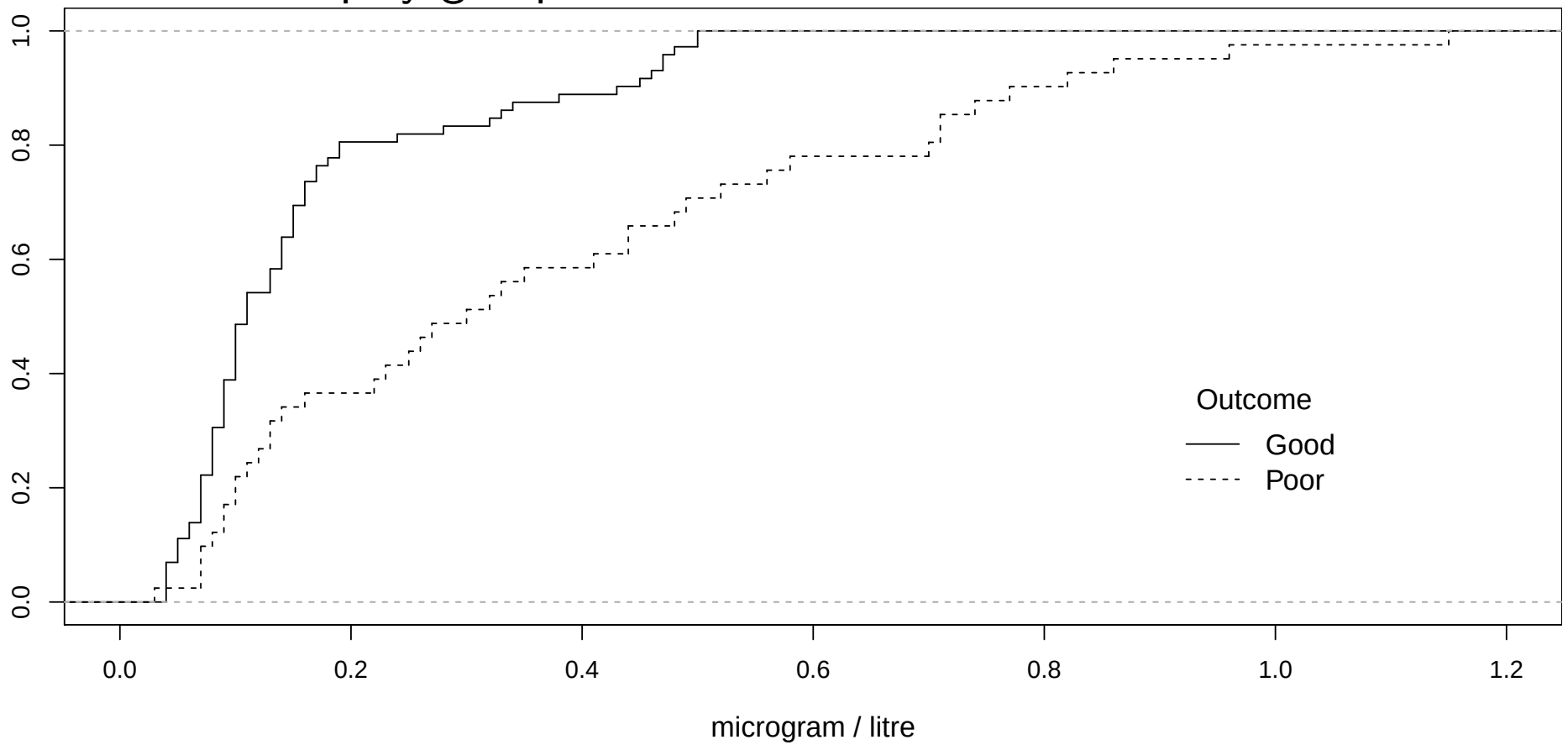
- The kth percentile is a value v such that k percent of your data lies below (or at) v. (Usually not uniquely defined.)
- The 50th percentile (the *median*) is the point which divides your ordered sample equally. (Only 'unique' if sample is odd, else use mean of the two midpoints.)
- *The Quartiles:* Q1 is the 25th percentile, Q2 is the 50th percentile and Q3 is the 75th percentile.
- We can describe all percentiles with the *empirical cumulative distribution function* (ecdf).
  When the sample size is 113 the jumps in the ecdf will be multiples of $1/113 \approx 0.01$.

Empirical cumulative distribution function for S100β



Empirical cumulative distribution function

Ecdf's can also display group differences.



Average value / Measure of location

An average value is a single value meant to be representative of the entire data set.

- **The mode:** is the single most common data point.
  (Mostly meaningful for categorical data.)
- **The median:** is the midpoint of the ordered numerical sample.
- **The mean:** is the "center of gravity" of a data set.
  Note: unlike the median, the mean value is sensitive to "extreme" values.

Mean or median?

Ex: A small company has 5 employees, who earns 19, 21, 22, 24, 27 (K SEK) and a boss who earns 55.

Salaries	Employees	Entire company
Median	22	23
Mean	22.6	28

Some points:

- Small data sets might not need summary measures.
- Symmetric data has mean $\approx$ median.
- (Easy enough to calculate both.)

Measuring the spread of a data set

- **Range** The difference between the maximum and the minimum value.
- **Interquartile range (IQR)**: Q3-Q1.
- **Standard deviation (sd)** is given by the formula,

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

Where $x_1, x_2, \ldots, x_n$ is the sample and

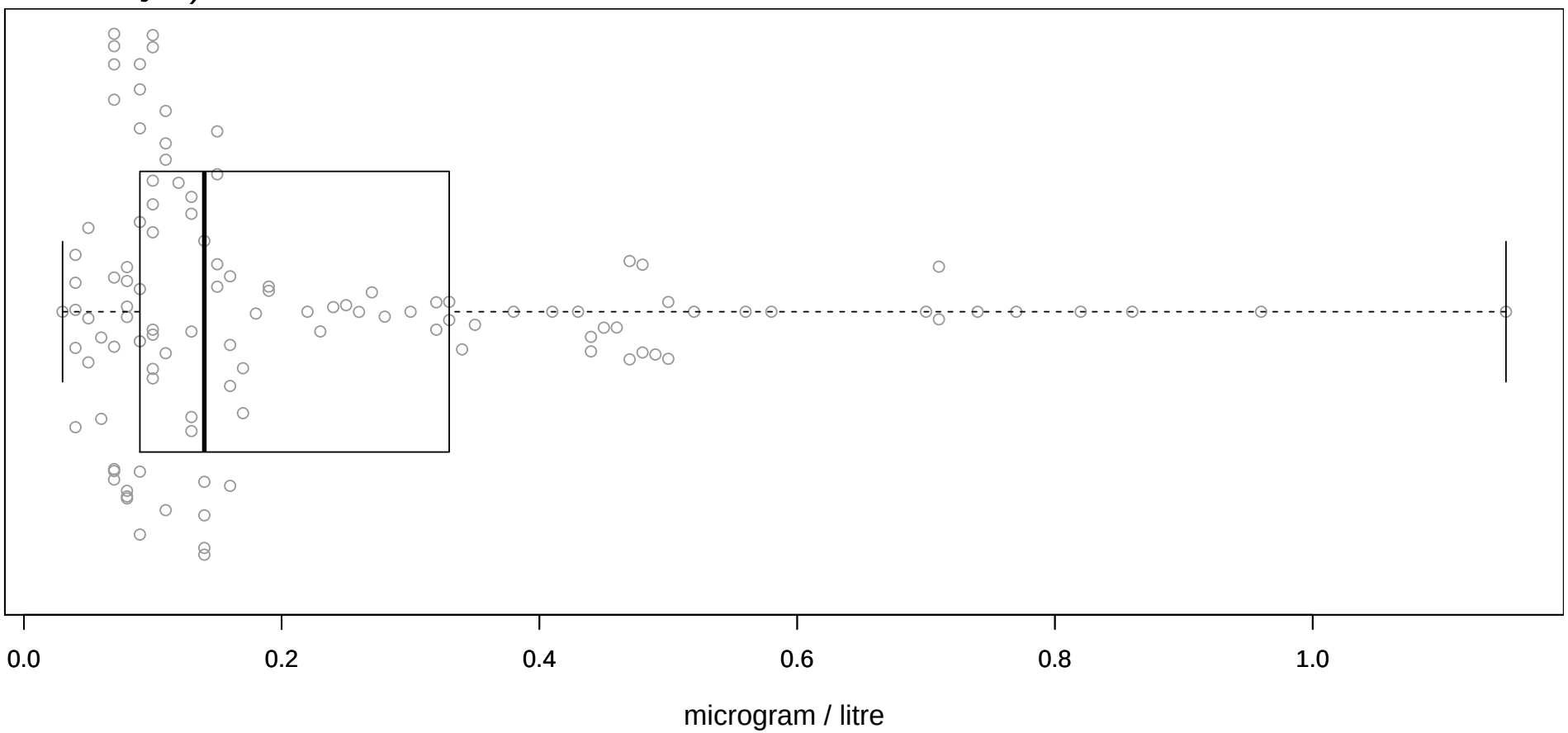
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

is the (sample) mean.

It is (approximately) the mean distance to the mean value.

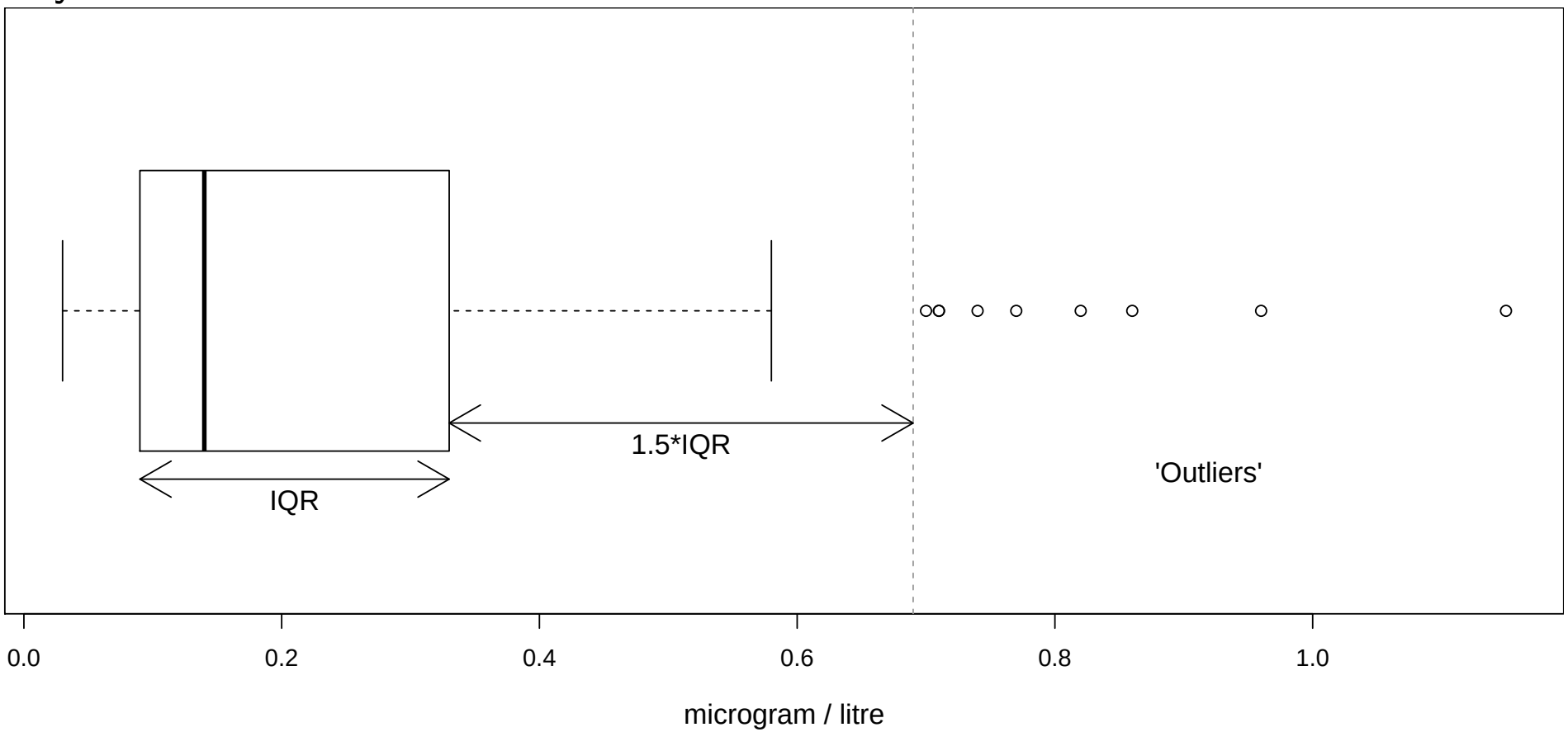
Boxplot of S100β

The boxplot usually show min, Q1, med, Q3 and max (the "5-point summary")...



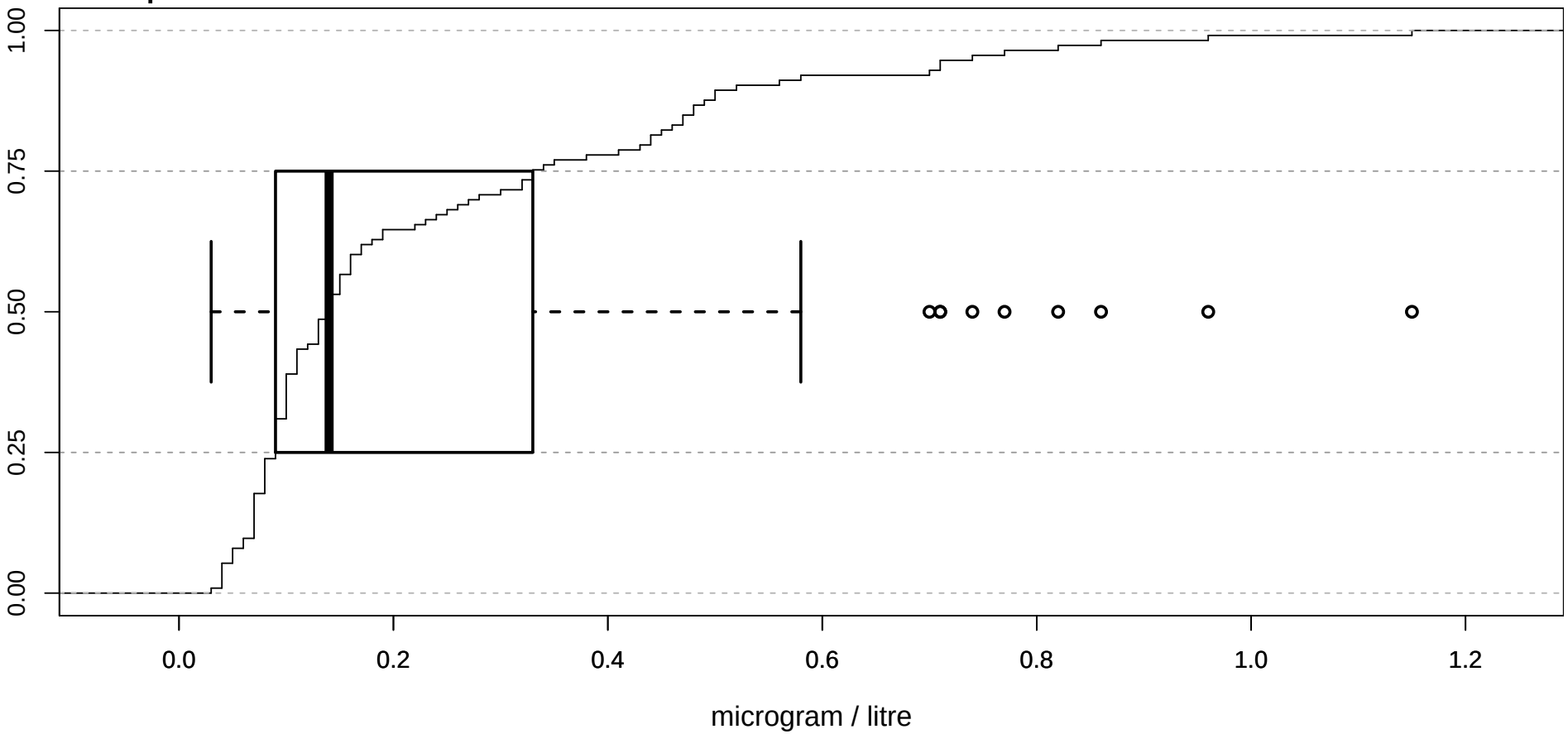
Boxplot

. . . but most software mark points that are more than 1.5 times the IQR away from 'the box'.



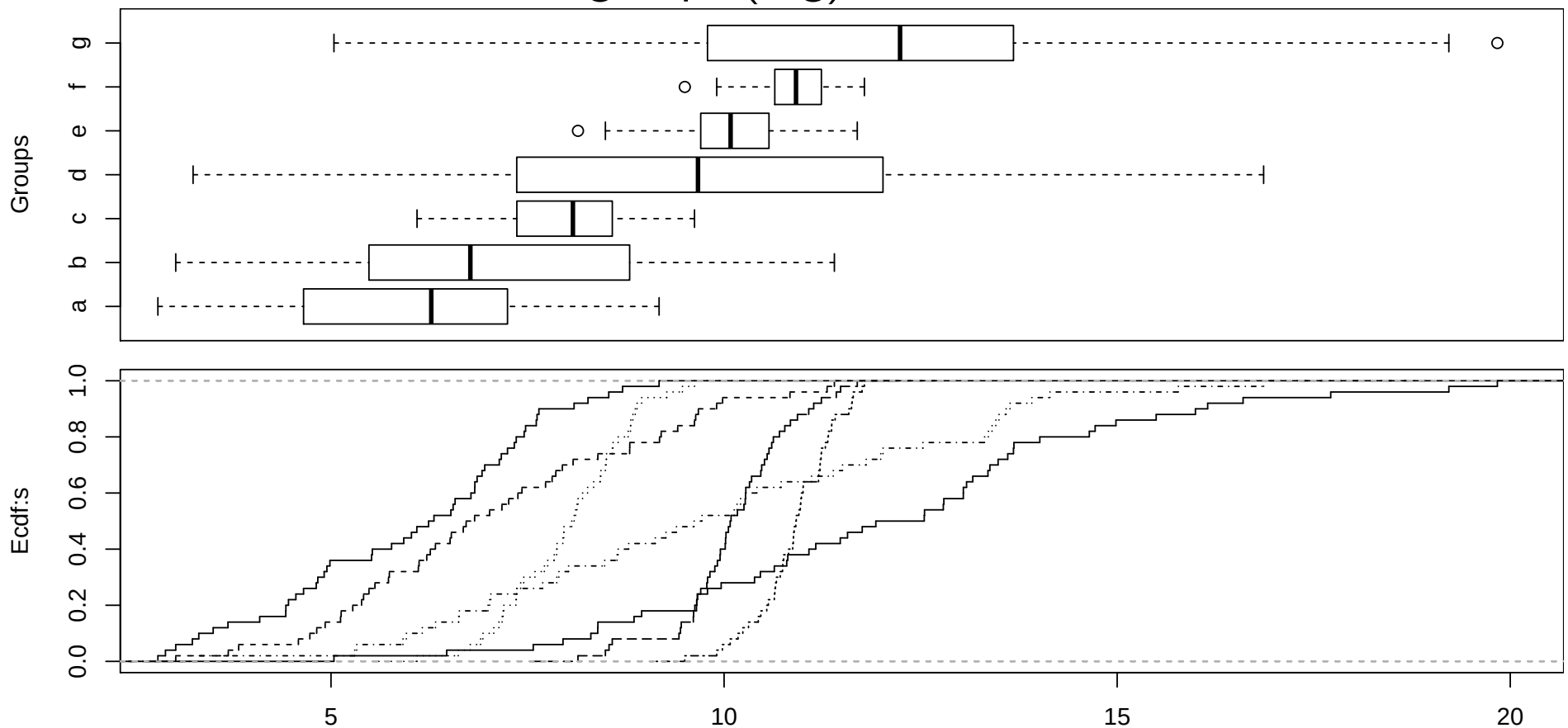
Connection between boxplot and ecdf

The boxplot contains less information.



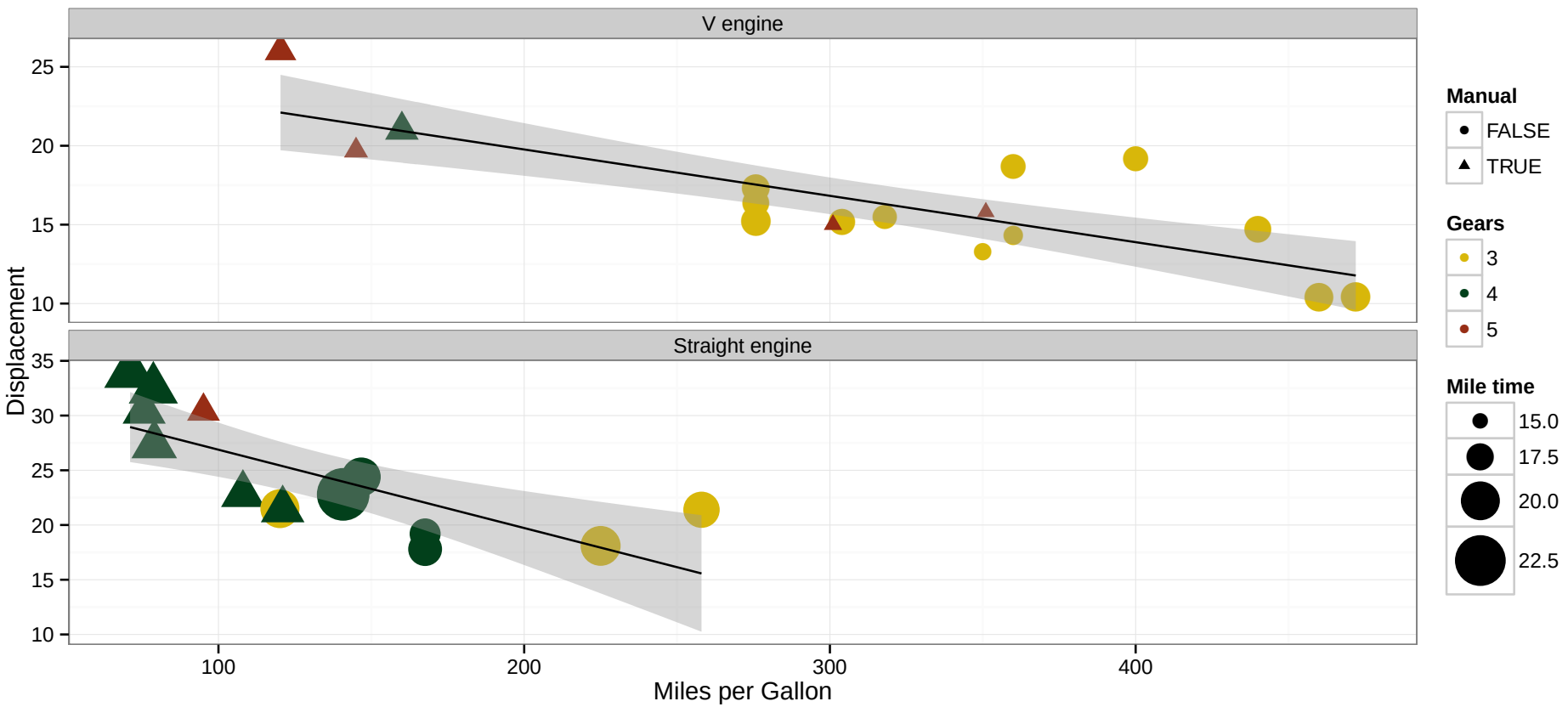
Pattern or detail?

Here is fake data whith 7 subgroups (a-g).



Advanced graphics...

Anyone seriously interested in graphics might want to look at Wilkinson's *The Grammar of Graphics*, which provides a framework for thinking about data visualization. These ideas are implemented in the R package **ggplot2** - which provides tools for mapping variables to aestic properties, adding layers of statistical computations, faceting, etc.

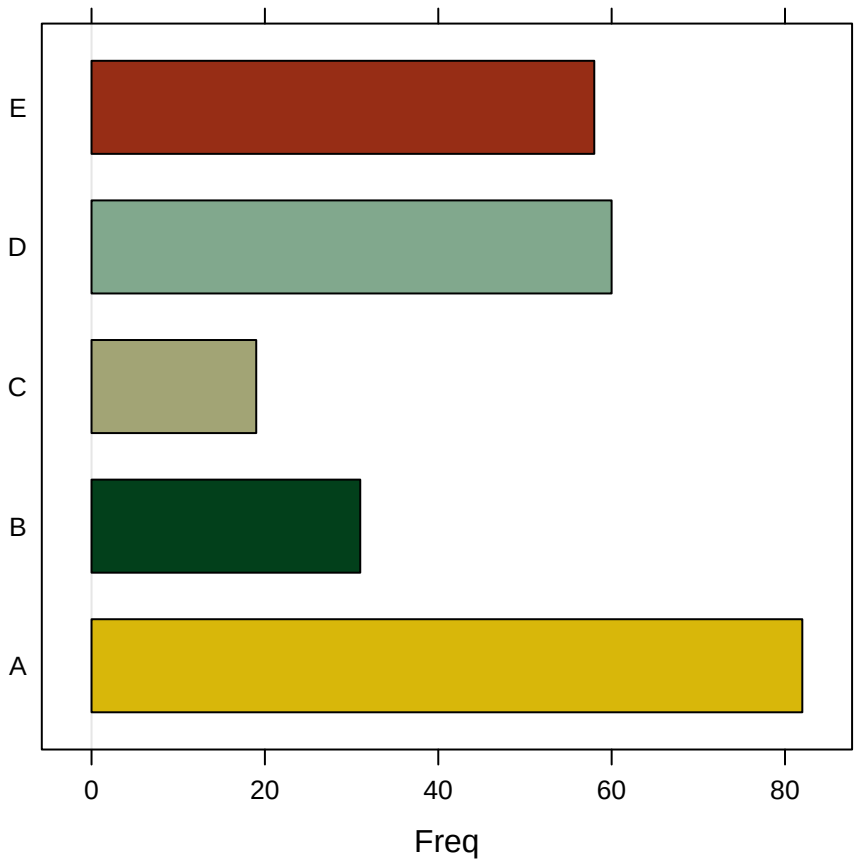


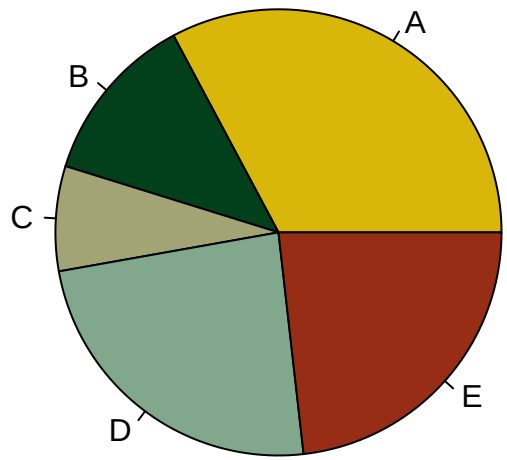
Some points on graphs (for publication)

- Graph and caption should be self-contained.
- Every graph should be referred to in the text.
- 'Economy' Do not make a graph which is more easily expressed in text or a small table.
  E.g. graph with a single boxplot.
- Pattern or detail?
  E.g.
  - a ecdf gives a lot of detail of a data set.
  - graph with multiple boxplots can reveal pattern among subgroups.
- Avoid 2D graphs shown in 3D.
- Avoid pie charts? It depends...

In defense of pie charts?

Is it of interest to be able to compare summation of levels?
(Is A + B + C smaller than D + E?)





Probability theory studies models of random data. A **model** is a way of specifying the range of possible values and the probability with which these occur.

- **Probability functions** describe discrete numeric/categorical data
- **Density functions** describe continuous (numeric) data

The **cumulative distribution function** is a universal way of describing all random (numeric) data. (Will not be discussed.)

- **Probability theory:** given model (model parameters, or other aspects) - describe how data behave. E.g.
  - specific results: how likely are deviations from what we expect
  - general results: Law of Large Numbers, Central Limit Theorem, etc.
- **Inference theory:** given data what is a likely model/parameters or other aspects of the underlying distribution (without specifying model = non-parametric statistics).

Probability models for categorical or integer-valued data

A yet undetermined random value is called a *random variable* (RV).

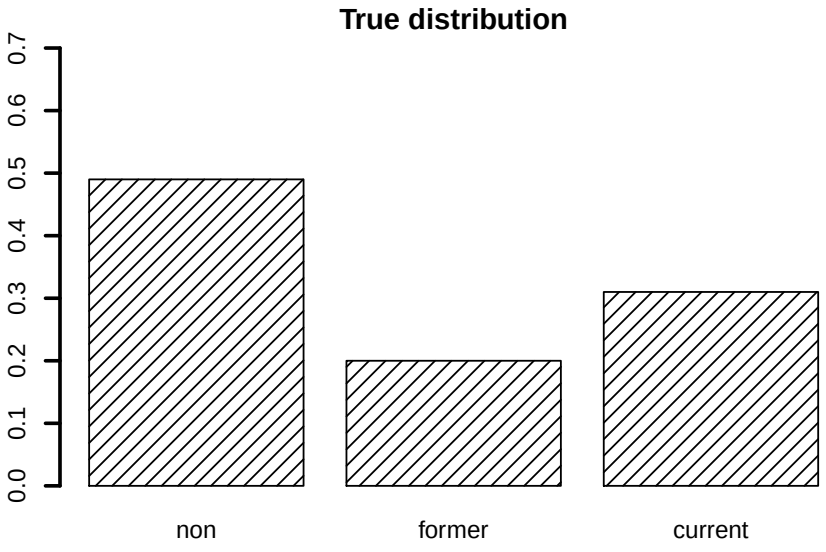
Let $Z =$'the outcome of the throw of a die'. Then $\mathbf{Prob}(Z = k) = 1/6$ for all $k = 1, 2, \ldots, 6$, or, equivalently

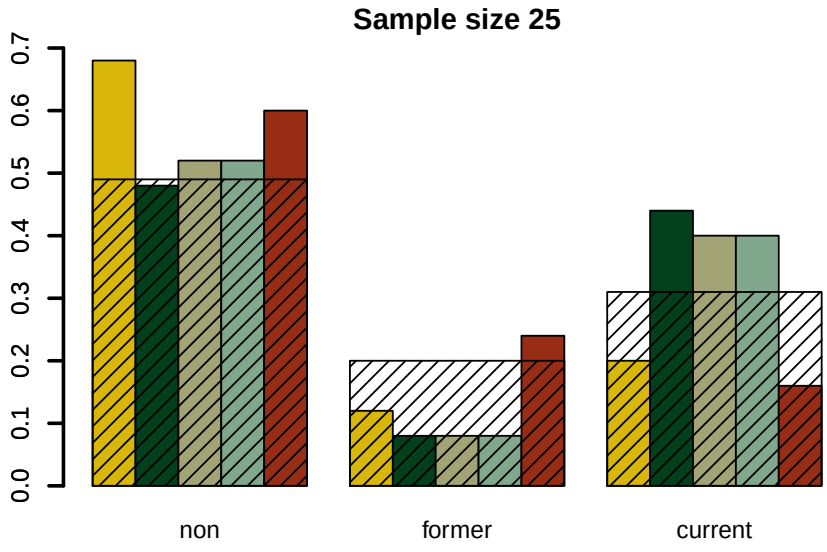
Value k	1	2	3	4	5	6
$\mathbf{Prob}(Z = k)$	1/6	1/6	1/6	1/6	1/6	1/6

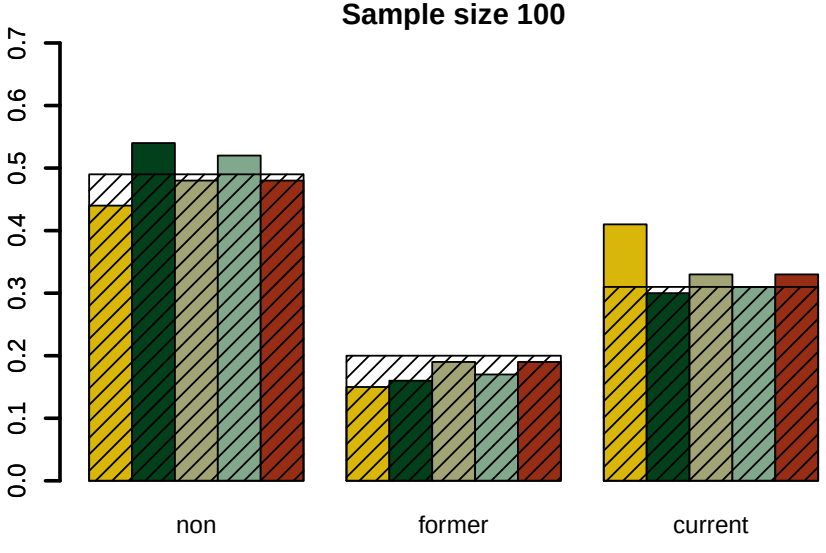
Suppose that in the population there are 49 % non-smokers, 20 % former smokers and 31% current smokers. Then the smoking status X of a person selected at random is a RV with a probability function

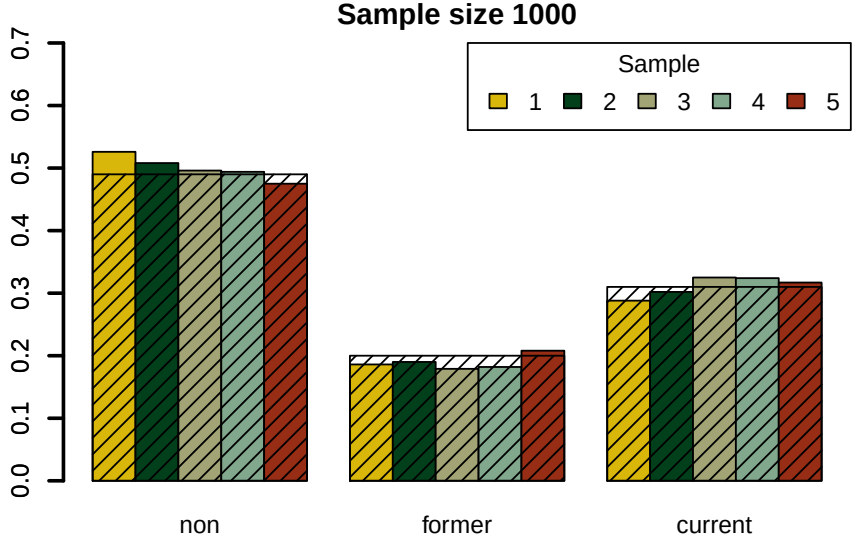
Value v	non	former	current
$\mathbf{Prob}(X = v)$	0.49	0.20	0.31

Five samples from three different sampling sizes from previous distribution









FEV data set

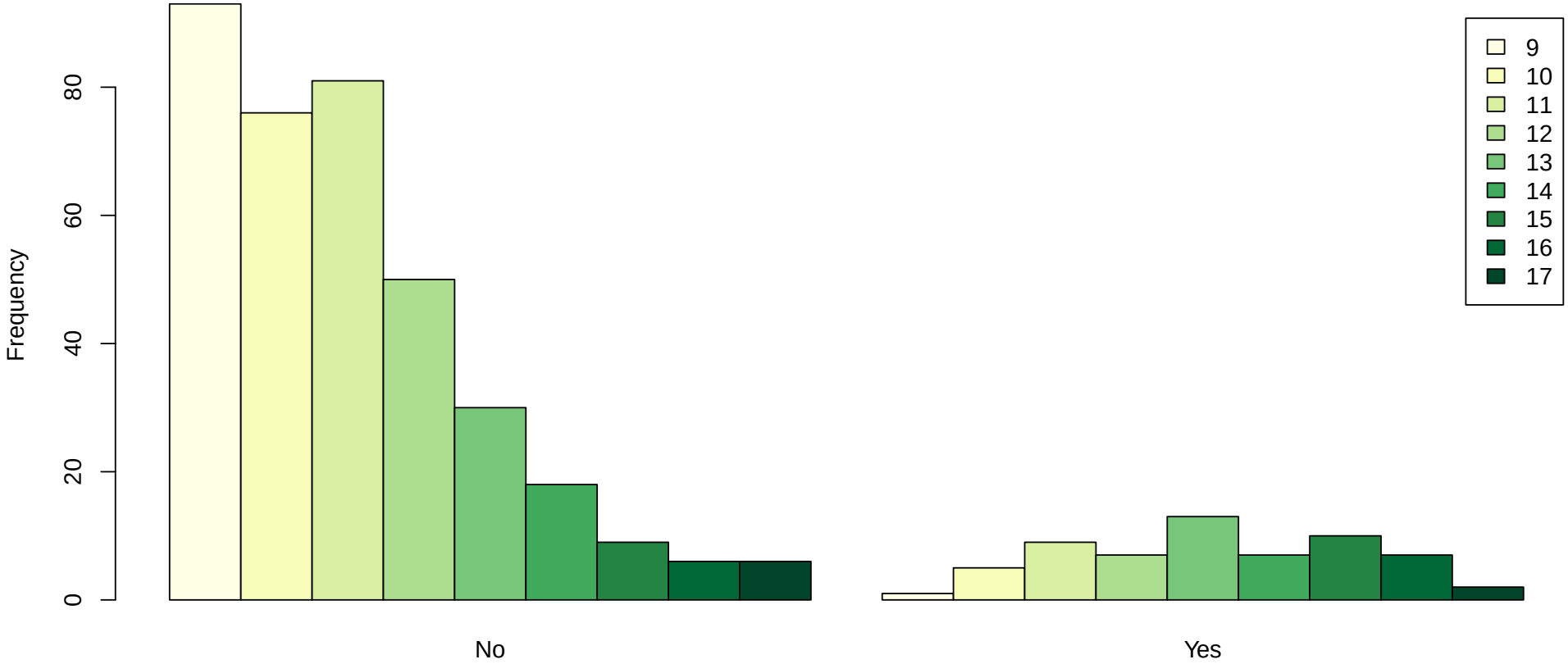
430 children (9-17 years of age) had their age, forced expiratory volume in 1 second (FEV) and smoking status (!) recorded. (To be seen again in lecture 8.)

A barchart is a way to visualize a variable with a small number of unique values (often categorical). They are visual analogous of tables.

Ex: how do the ages distribute over smoking status?

Smoking	**Age**								
	9	10	11	12	13	14	15	16	17
No	93	76	81	50	30	18	9	6	6
Yes	1	5	9	7	13	7	10	7	2

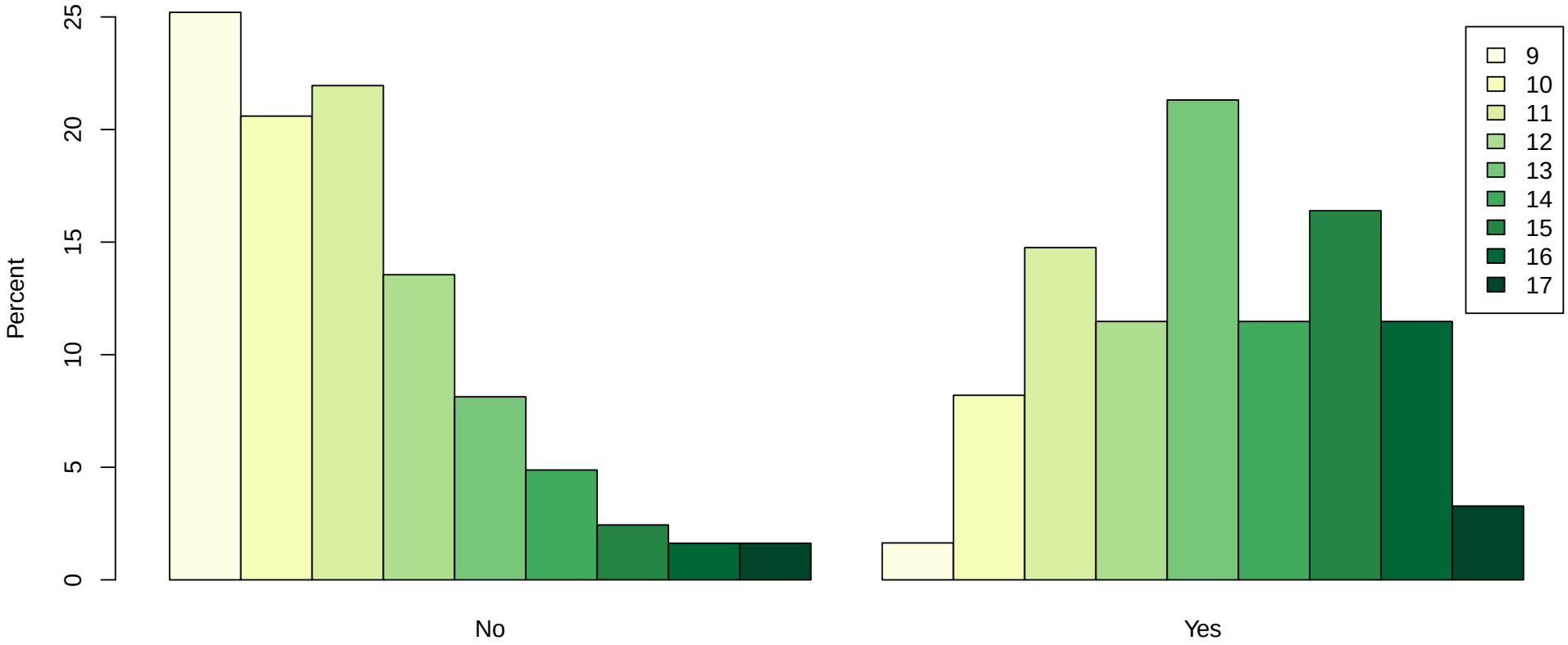
Visualizing 'age' versus 'smoking'



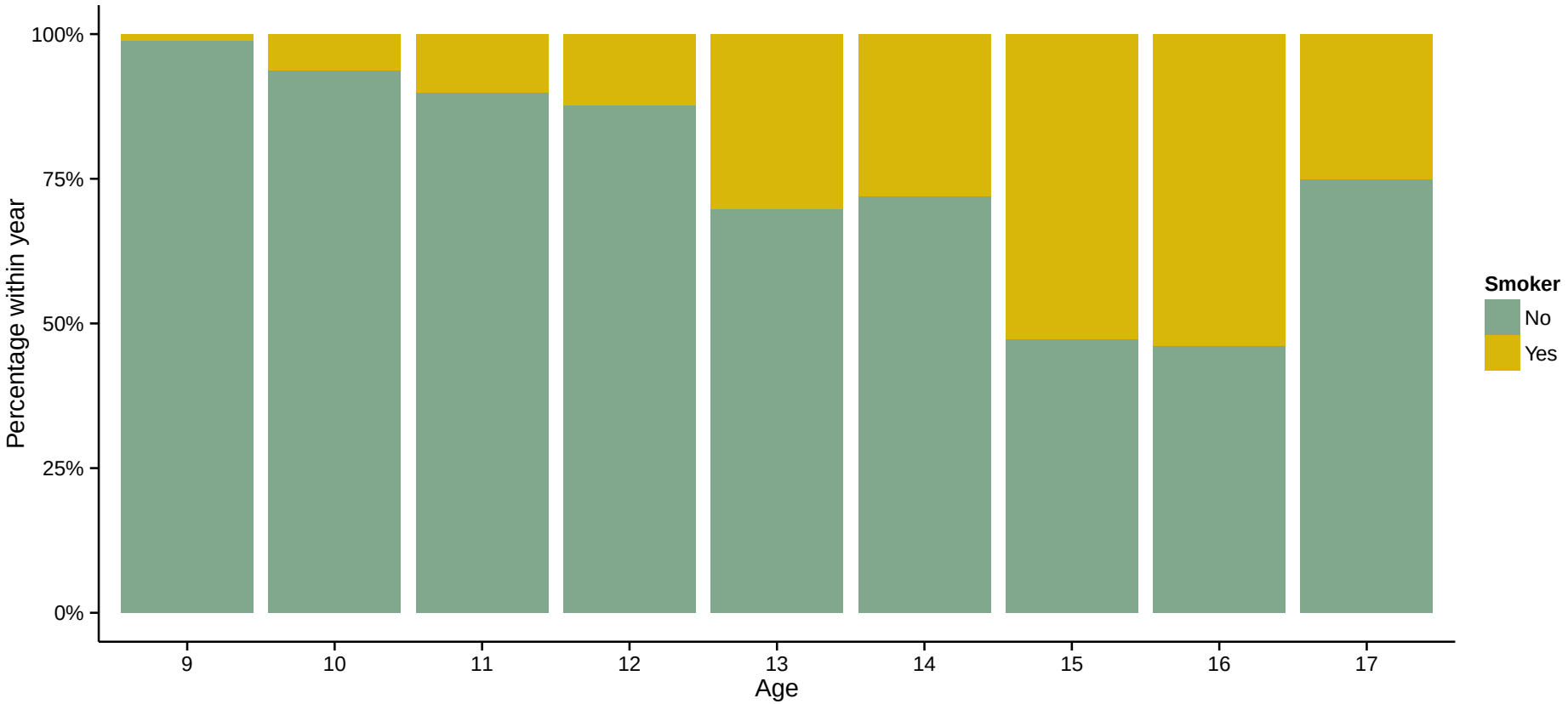
If the groups (smokers/non-smokers) aren't balanced it is difficult to compare the distributions.
Tabulate/plot the percentages within groups:

Age	**Smoking (%)**	
	No	Yes
9	0.252	0.016
10	0.206	0.082
11	0.220	0.148
12	0.136	0.115
13	0.081	0.213
14	0.049	0.115
15	0.024	0.164
16	0.016	0.115
17	0.016	0.033
Sum	1.000	1.000

Visualizing 'age' versus 'smoking' as percentages within smoking groups



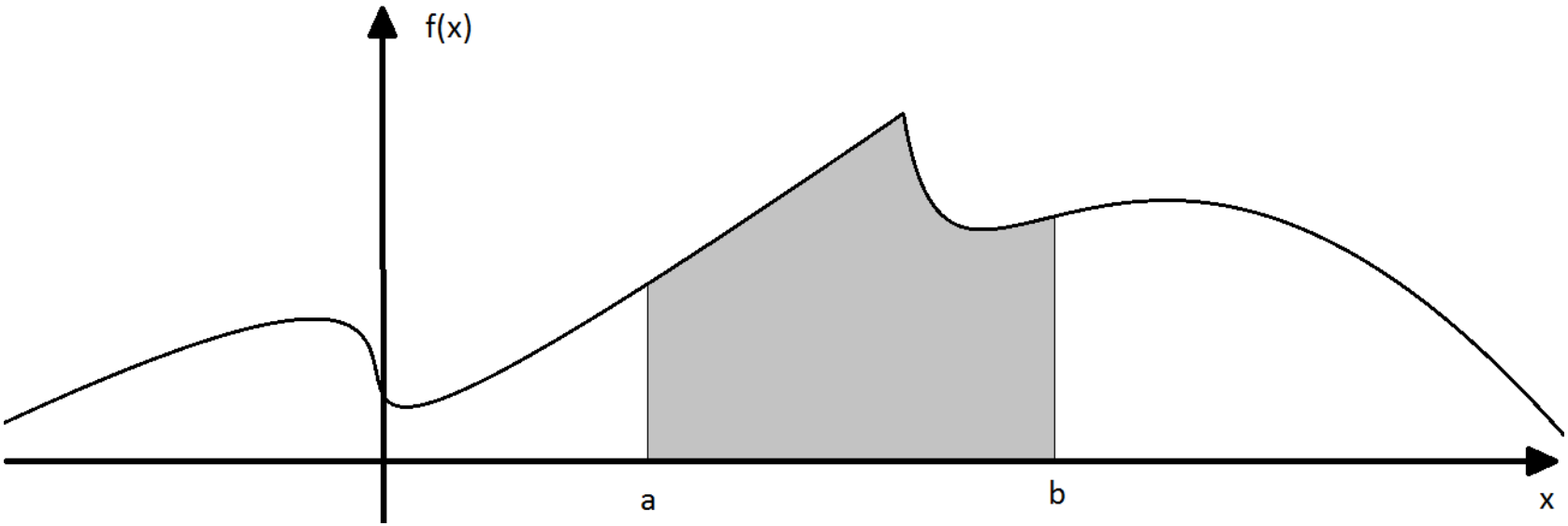
Visualizing 'smoking' versus 'age' as percentages within age groups, stacked



Probability model for continuous data

Recall that a RV is a yet undetermined value among several possible numbers.

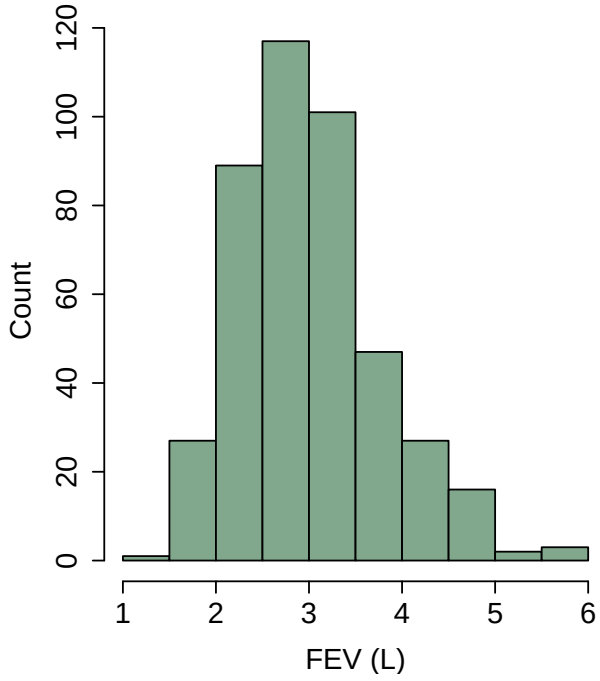
A continuous RV is described by its *density function*.

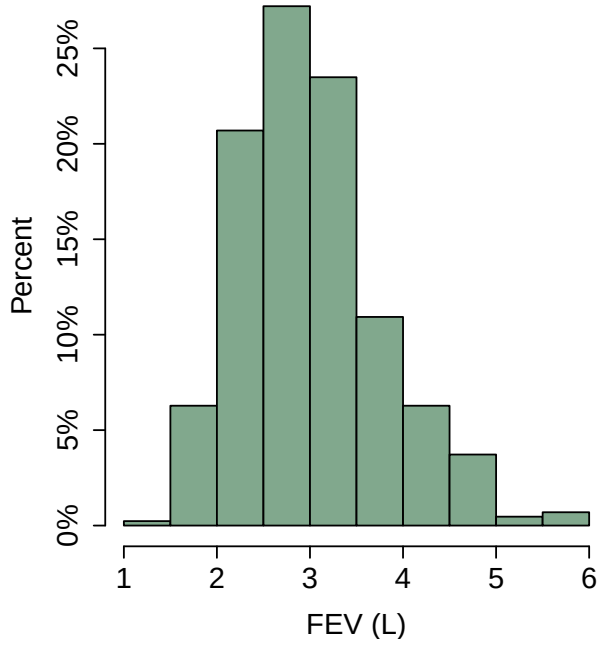


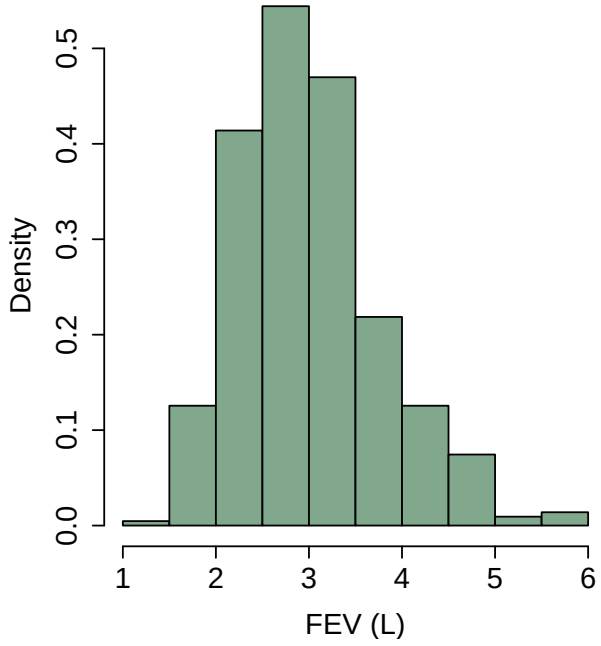
If X has density function f as above, then we compute probabilities as

$$\textbf{Prob}(a \leq X \leq b) = \text{ Area(a,b).}$$

The shape of a histogram estimates the shape of the density function.





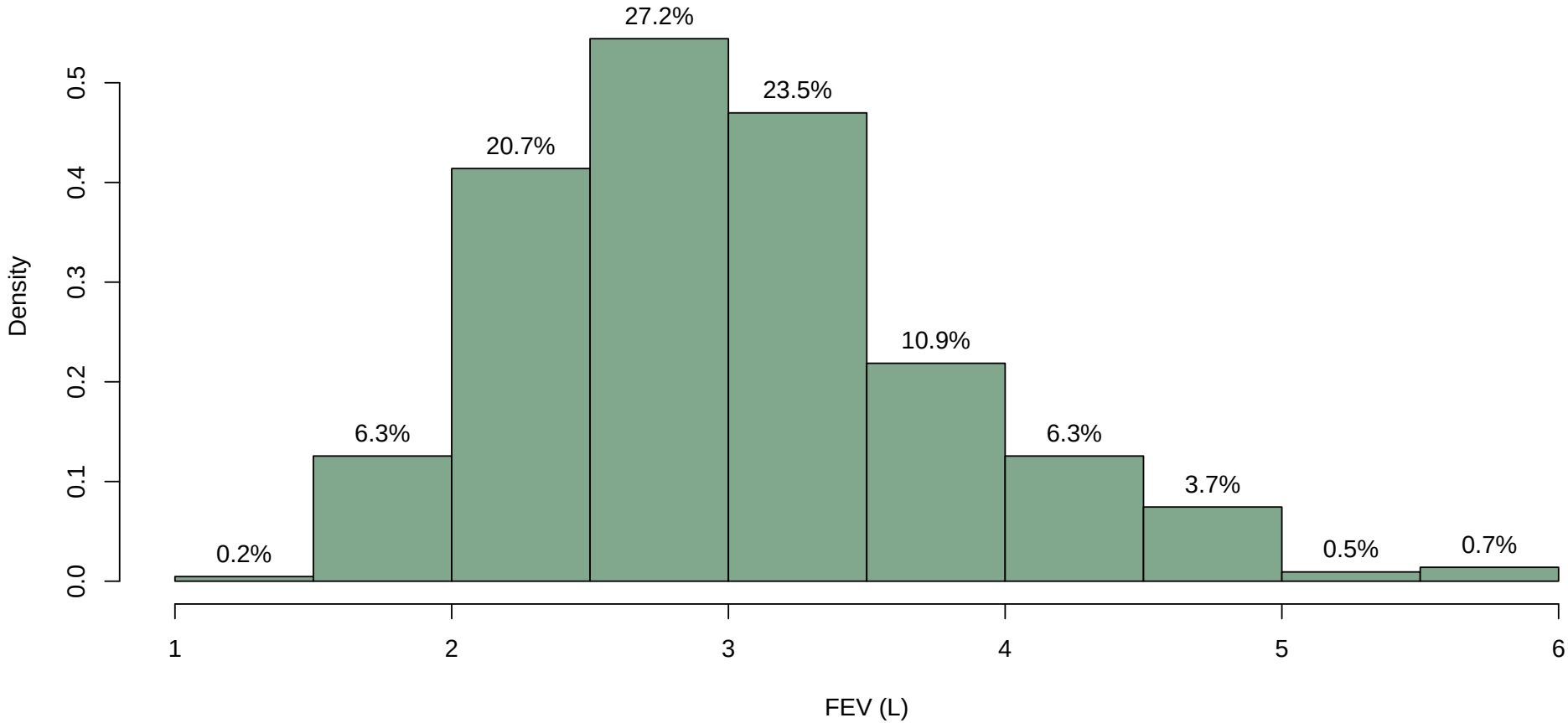


"Density" is more abstract but

- gives right scale for density function estimate (easy to correctly plot candidate model on top of histogram)
- allows for varying "bins"
- allows for comparison between very different sample sizes

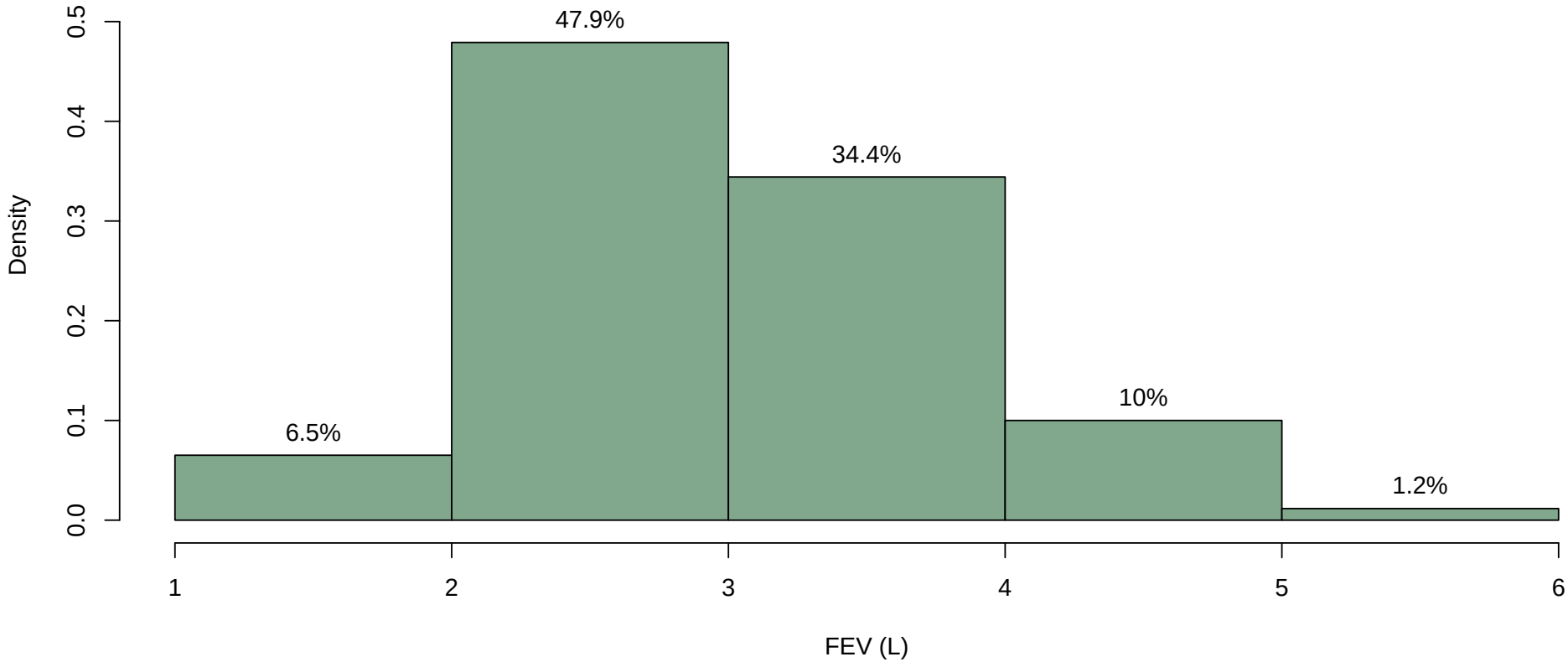
Histogram (with density) of FEV

The histogram (with density on the y-axis) estimates the density function.



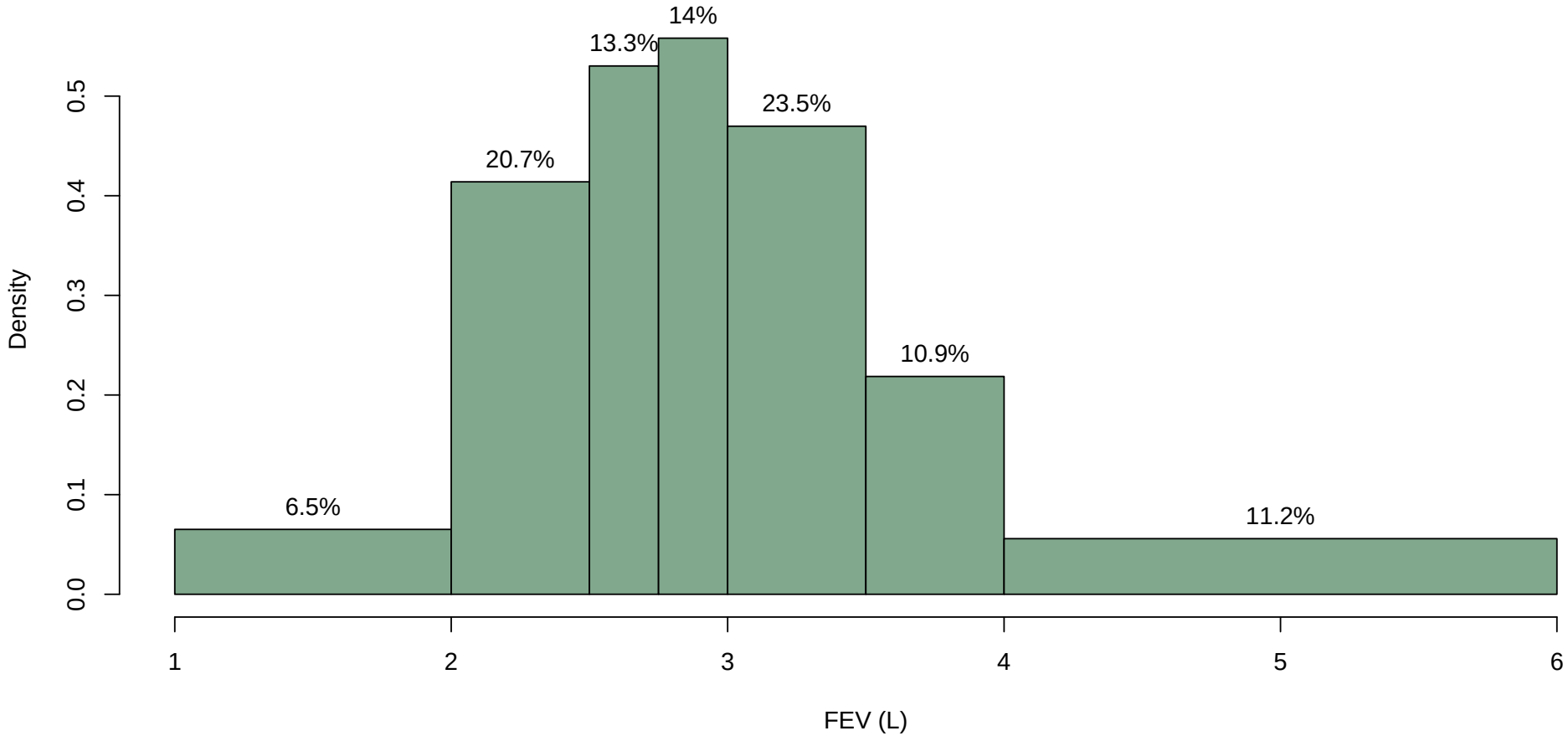
Histogram of FEV

The interval width is arbitrary. . .



Histogram of FEV

. . . and could be varying.

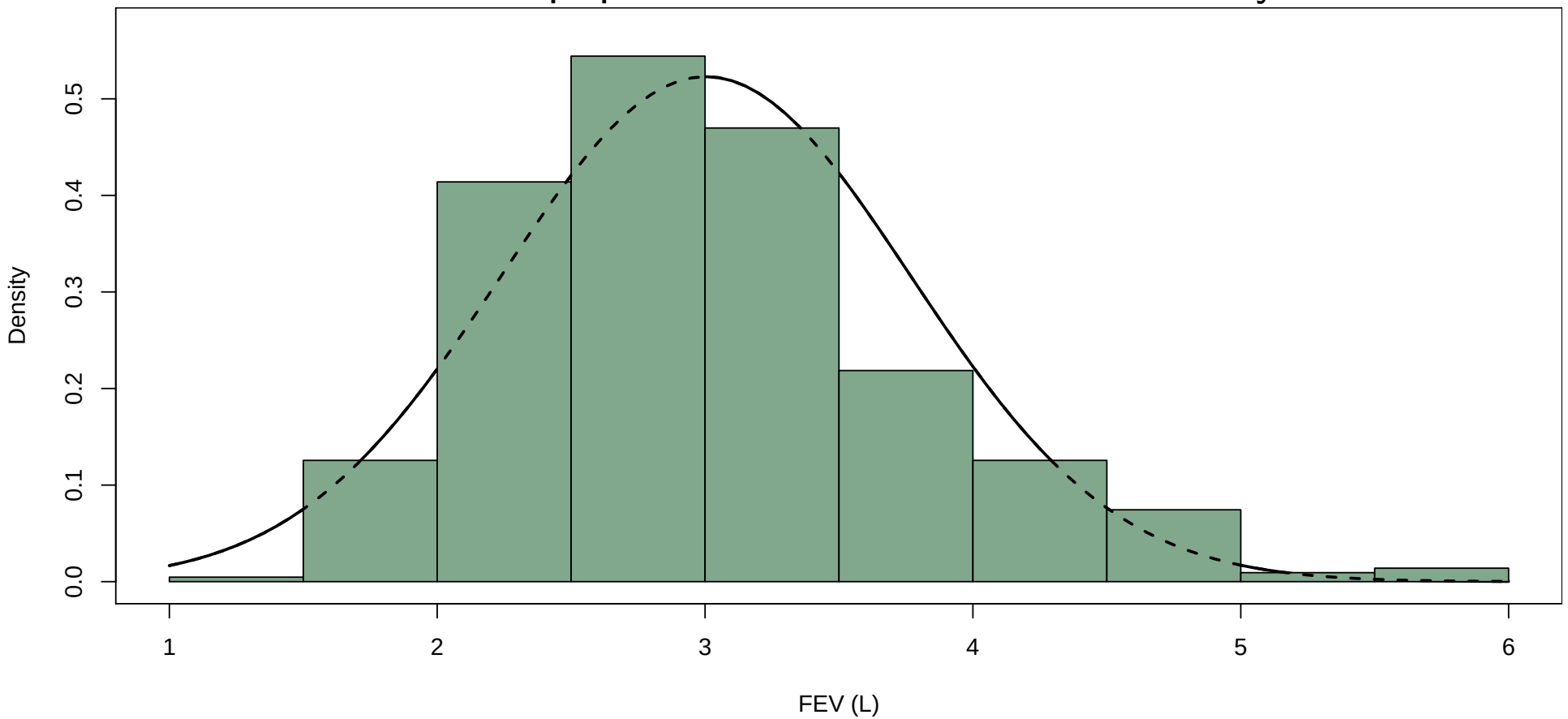


Summary

Graph	Summary Measure	Theory
Ecdf	percentiles	(Cdf)
Boxplot	min, Quartiles, max	
Bar charts		Probability functions (discrete RV)
Histograms		Density functions (continuous RV)
	median, IQR	any distribution
	mean, s.d.	symmetrical distribution ($\approx$ Normal distribution)

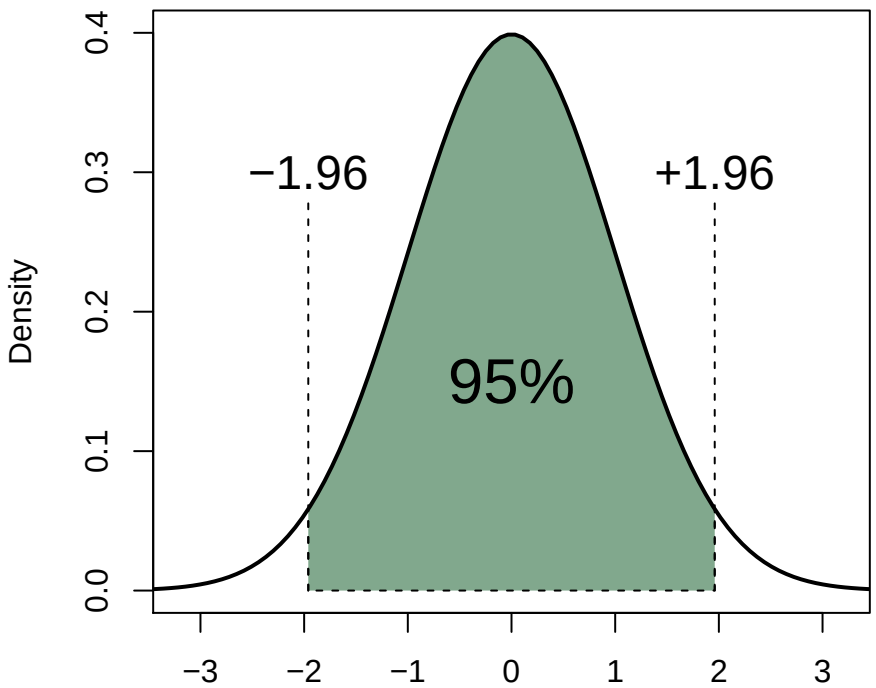
Assuming Normality

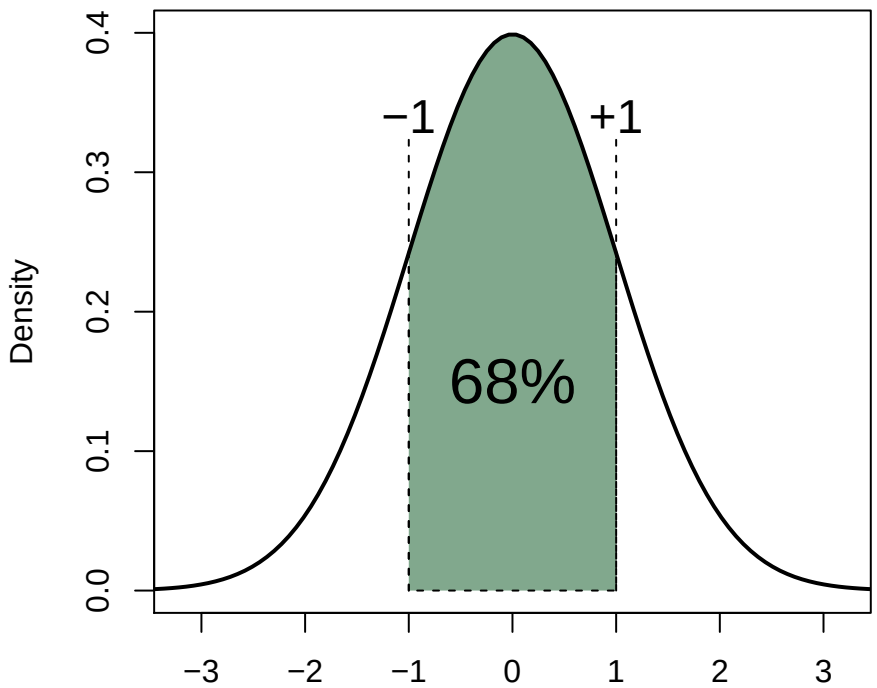
Often we assume that the population follows a Normal density curve.



Properties of the standard Normal distribution (and why you need to know the number 1.96)

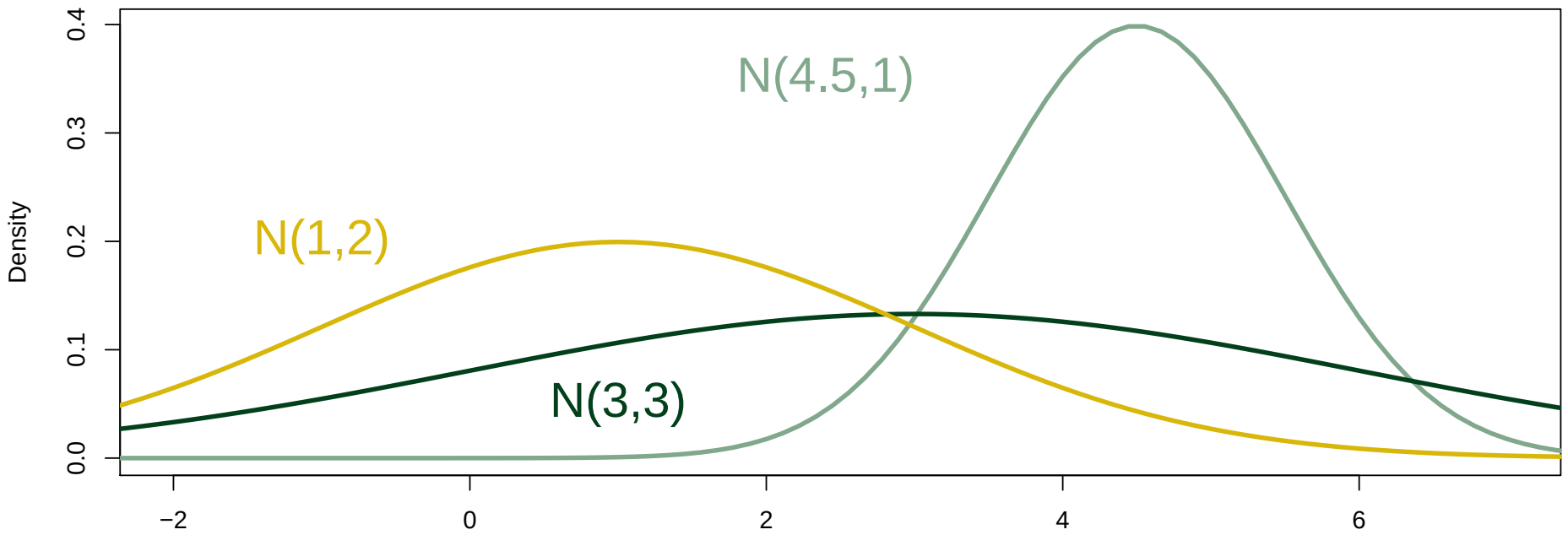
The standard Normal distribution has a standard deviation of 1.
68% of the 'probability mass' lies within ± 1.
95% of the 'probability mass' lies within ± 1.96.





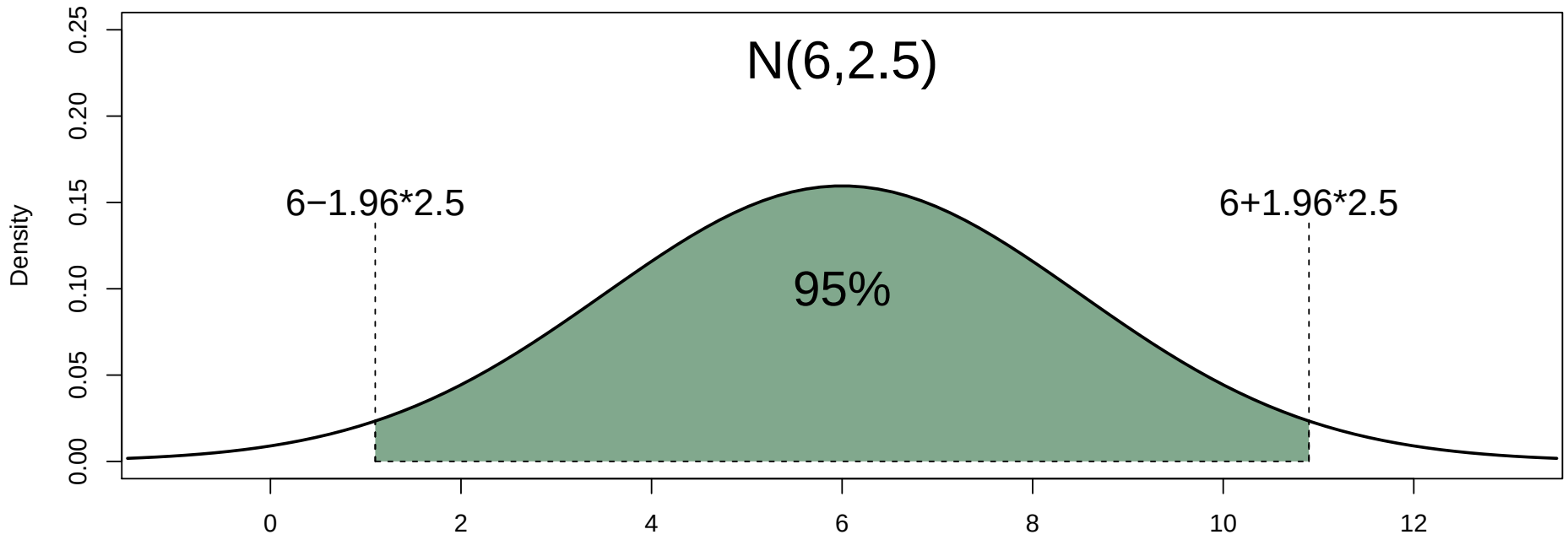
The Normal distribution $N(\mu, \sigma)$

is determined by its mean (μ) and standard deviation (σ).
If X is N(0,1) then $\mu + \sigma X$ is N(μ, σ).



Properties of the Normal distribution

If X is $\mathrm{N}(\mu, \sigma)$, then 95% of observations will be between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$.



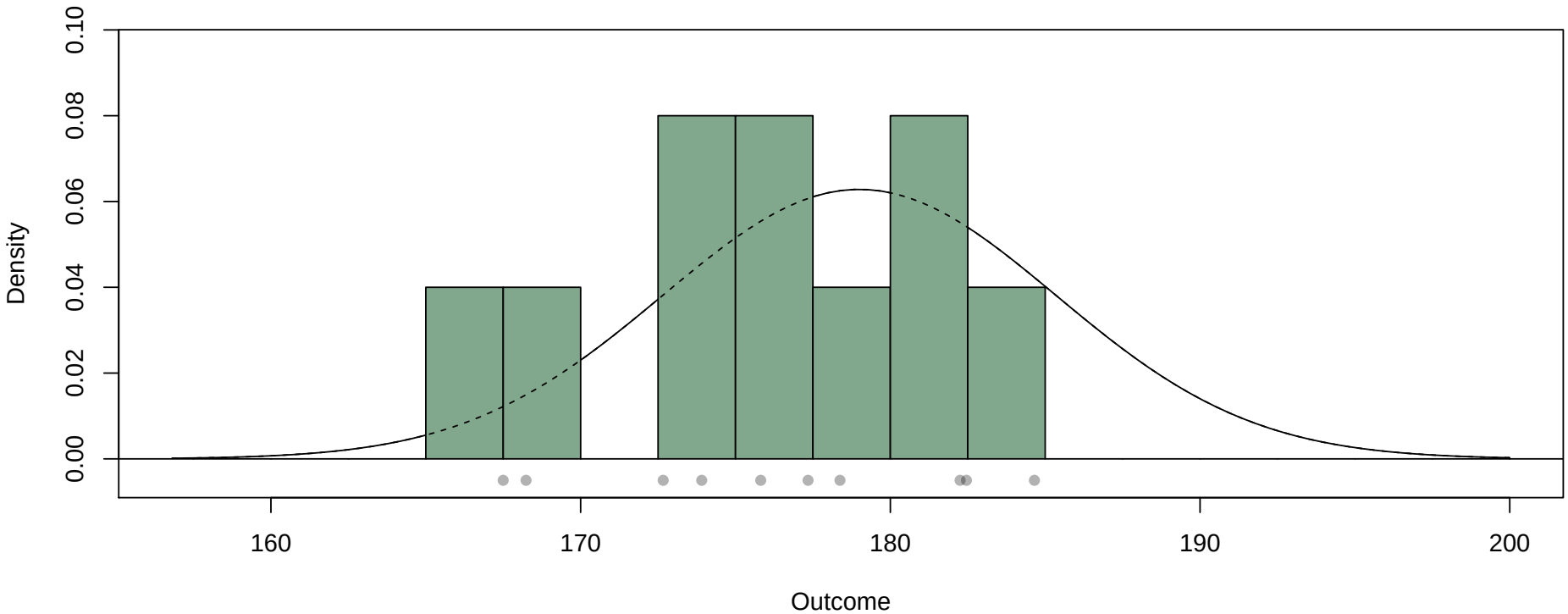
Attempt to visualize sampling from a given model

Assume that height of individuals in some population

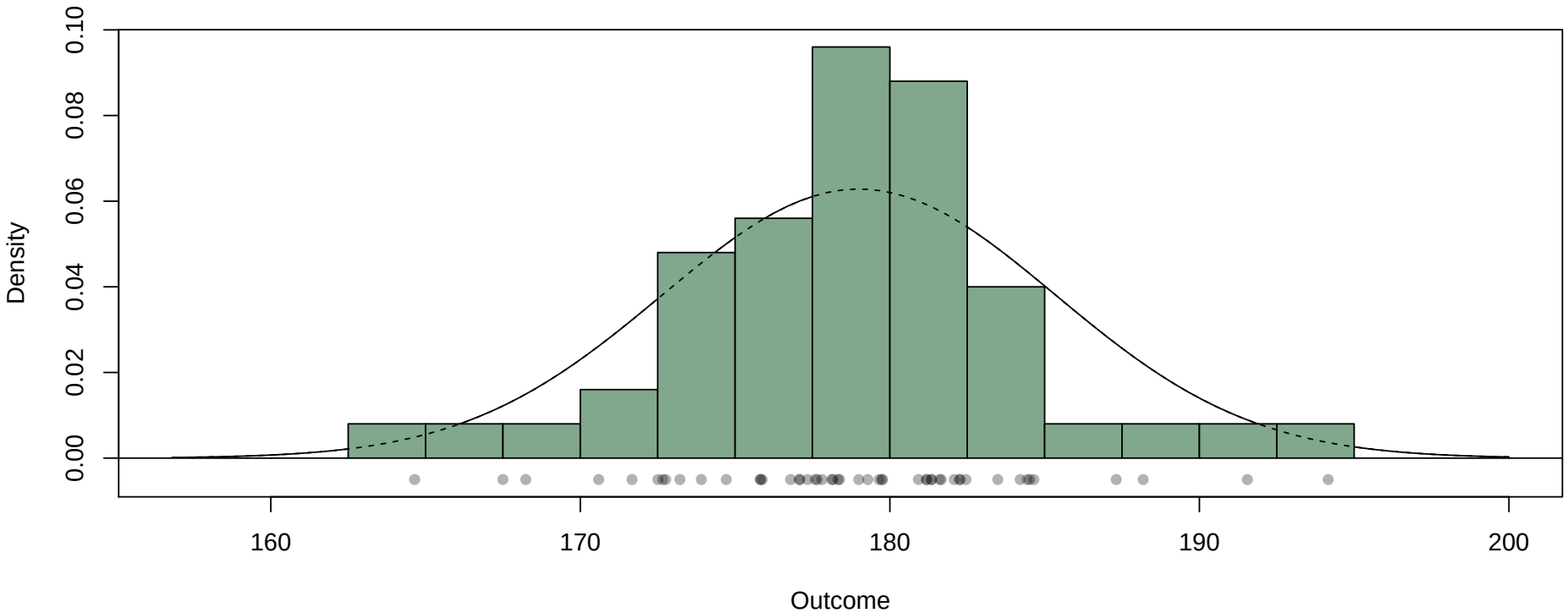
- is Normally distributed,
- has mean (μ) 179 (cm), and,
- has s.d. (σ) 6.35 (cm).

The following three slides show histogram of samples of sizes 10, 50 and 500, respectively.

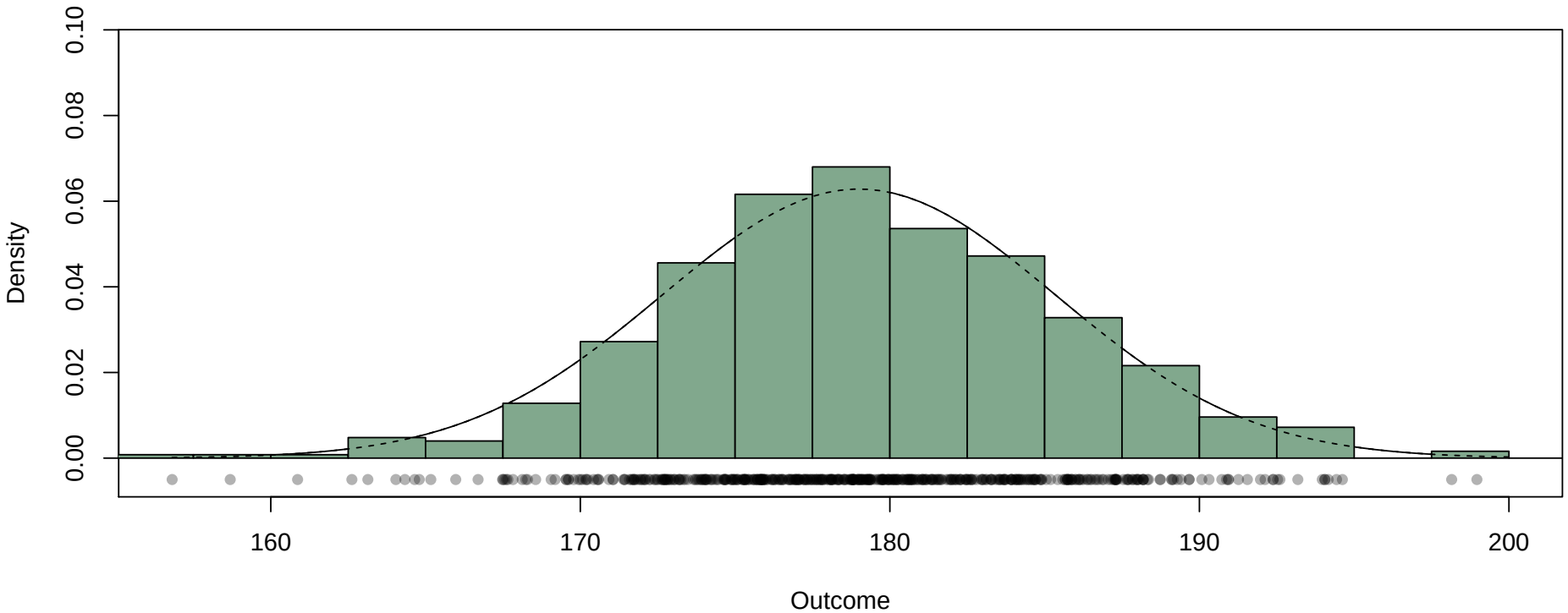
$n = 10$



$n = 50$



$n = 500$



Properties of samples and sample means

Suppose we have a sample of size 10 from the population.

We want to know the *population mean* (μ).

We can estimate with the *sample mean*(μ^*).

But how good a guess is μ^*?
Well. . . if the sample is drawn at random[1] then μ^* is *unbiased* (on average correct).

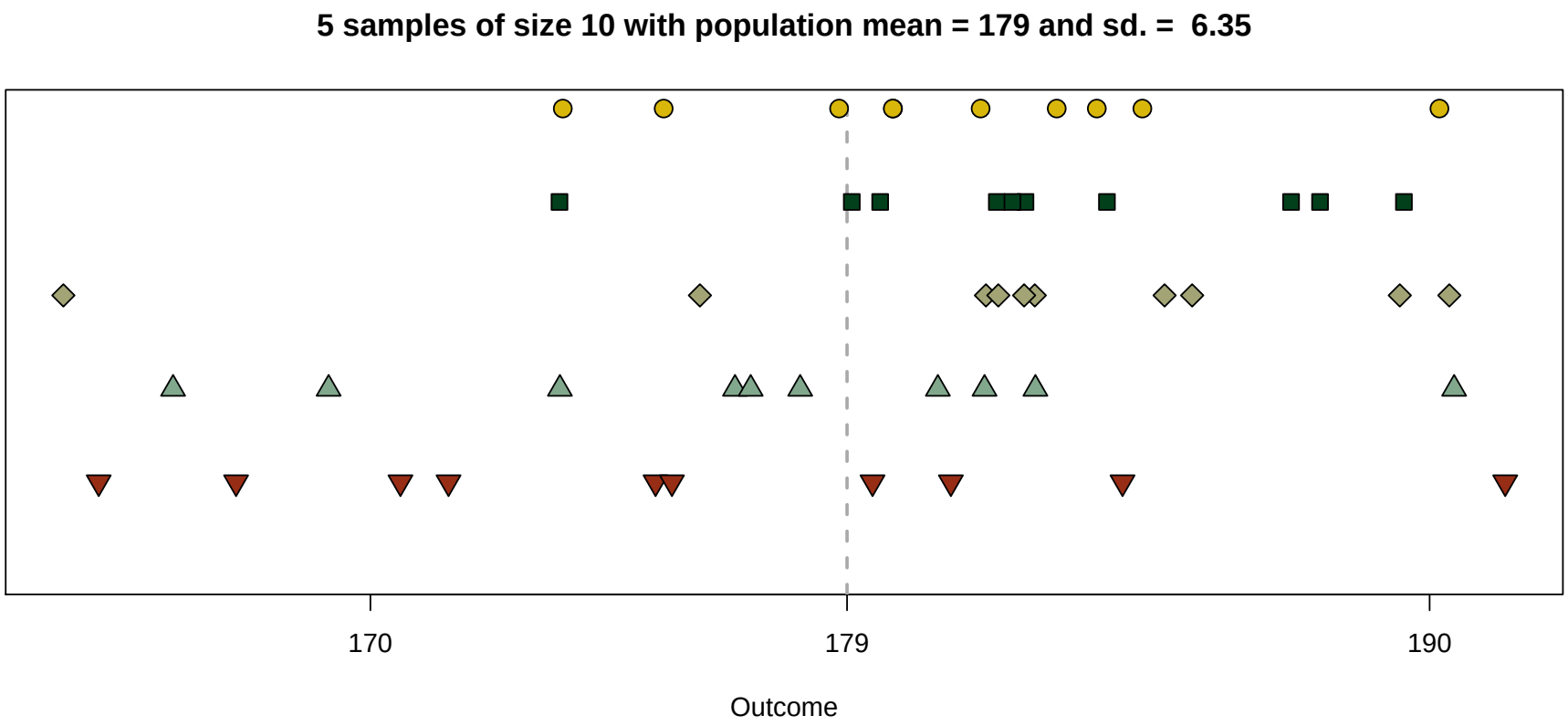
If we knew that the sd of $\mu^* = 180$ was, say 3.0, then there is an (approximate!) interpretation.

The estimated value 180 (cm) is on average off by 3.0 (cm).

But what *is* the standard deviation of a sample mean?

Consider several iterations of the procedure of drawing a sample of size 10.

[1] All possible samples are equiprobable


5 samples of size 10 with population mean = 179 and sd. = 6.35
170
179
190
Outcome

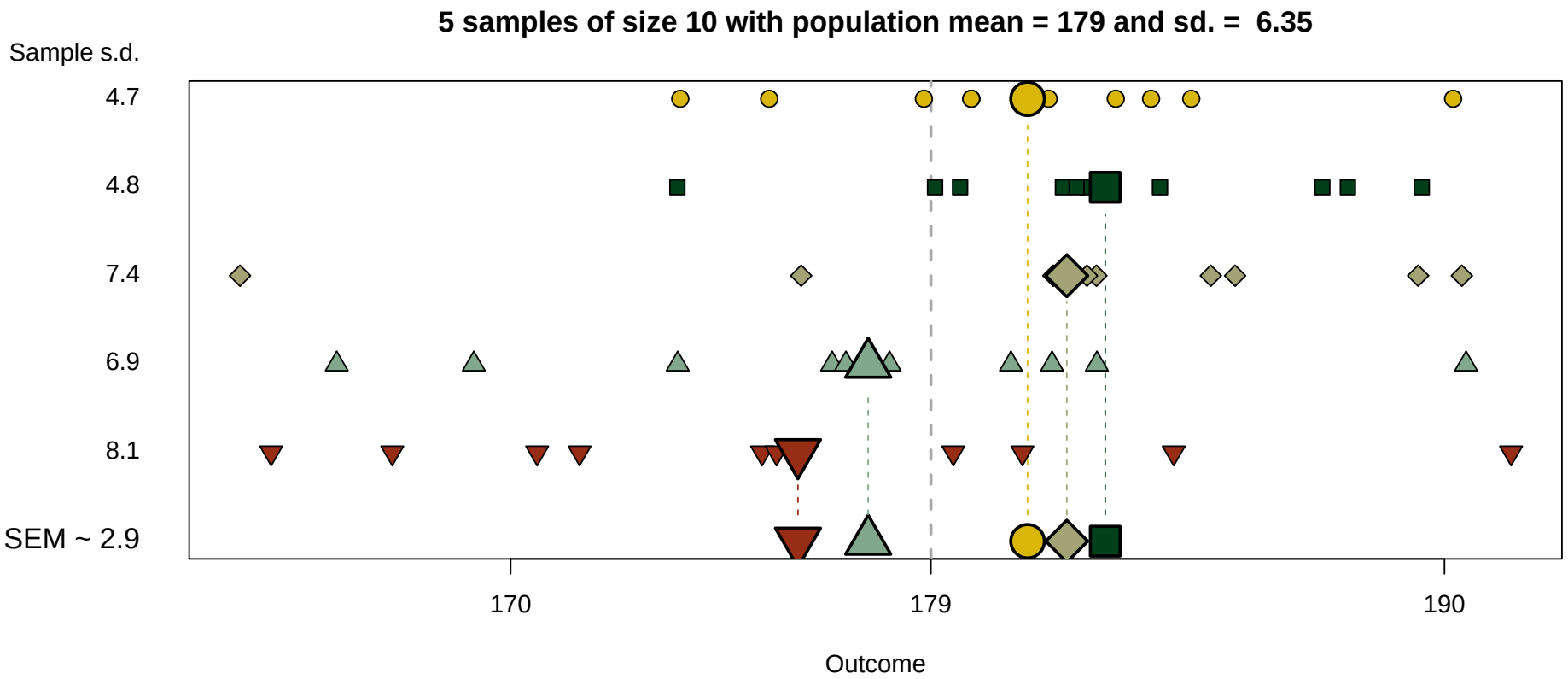
Properties of sample means

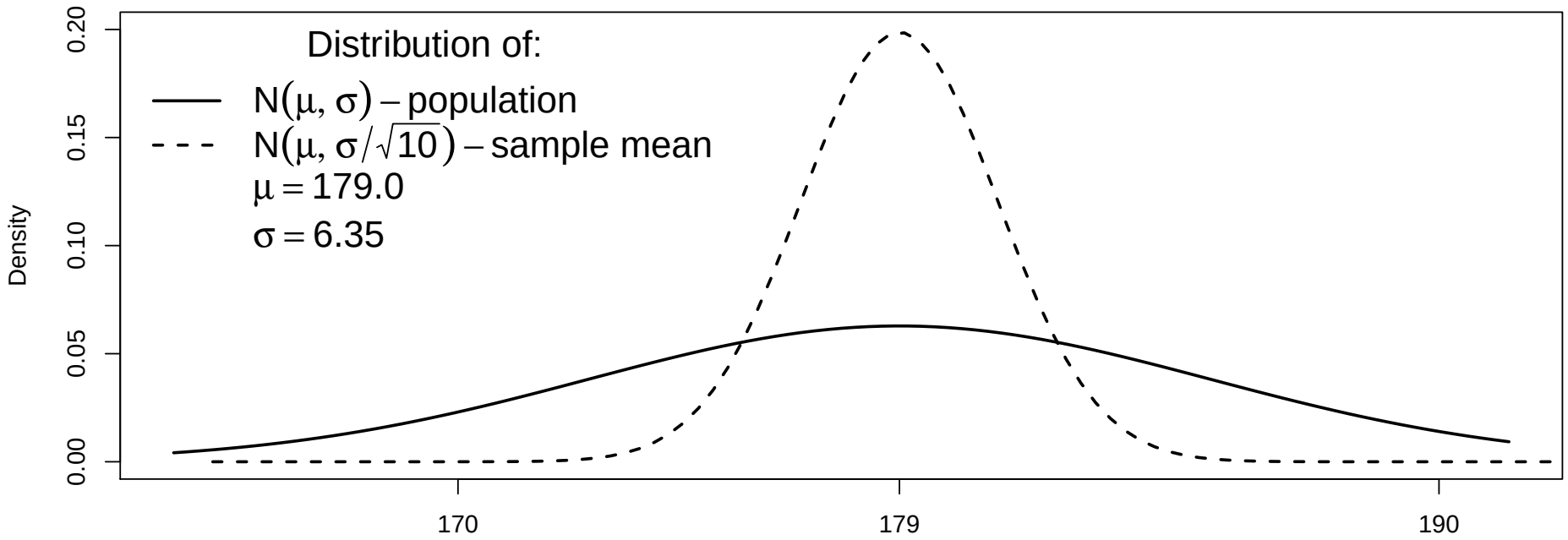
We can see that samples and their means and standard deviations vary (of course).

Standard Error of the Mean (SEM)
The standard deviation of the sample mean is referred to as SEM and a measure of (approx.) the average distance the population mean. Thus a measure of 'how good' the sample mean is as an estimator of the population mean.

In a Normal model the SEM is the population s.d. dived by $\sqrt{n}$.

The next slide tries to visualize this.


5 samples of size 10 with population mean = 179 and sd. = 6.35
Sample s.d.
4.7
4.8
7.4
6.9
8.1
SEM ~ 2.9
170
179
190
Outcome


Distribution of:
N(μ, σ) – population
N(μ, σ/√10) – sample mean
μ = 179.0
σ = 6.35
Density
0.00
0.05
0.10
0.15
0.20
170
179
190

Normal population ⇒ Normal mean

On the previous slide you saw

- a Normal population curve, and
- the Normal curve for the sample mean (for sample size 10).

In this case this is the true (derived from theory) density curve for the sample mean.

Properties of the Normal distribution will aid further analysis using the mean of the sample.

What happens if the population follows some non-Normal curve?

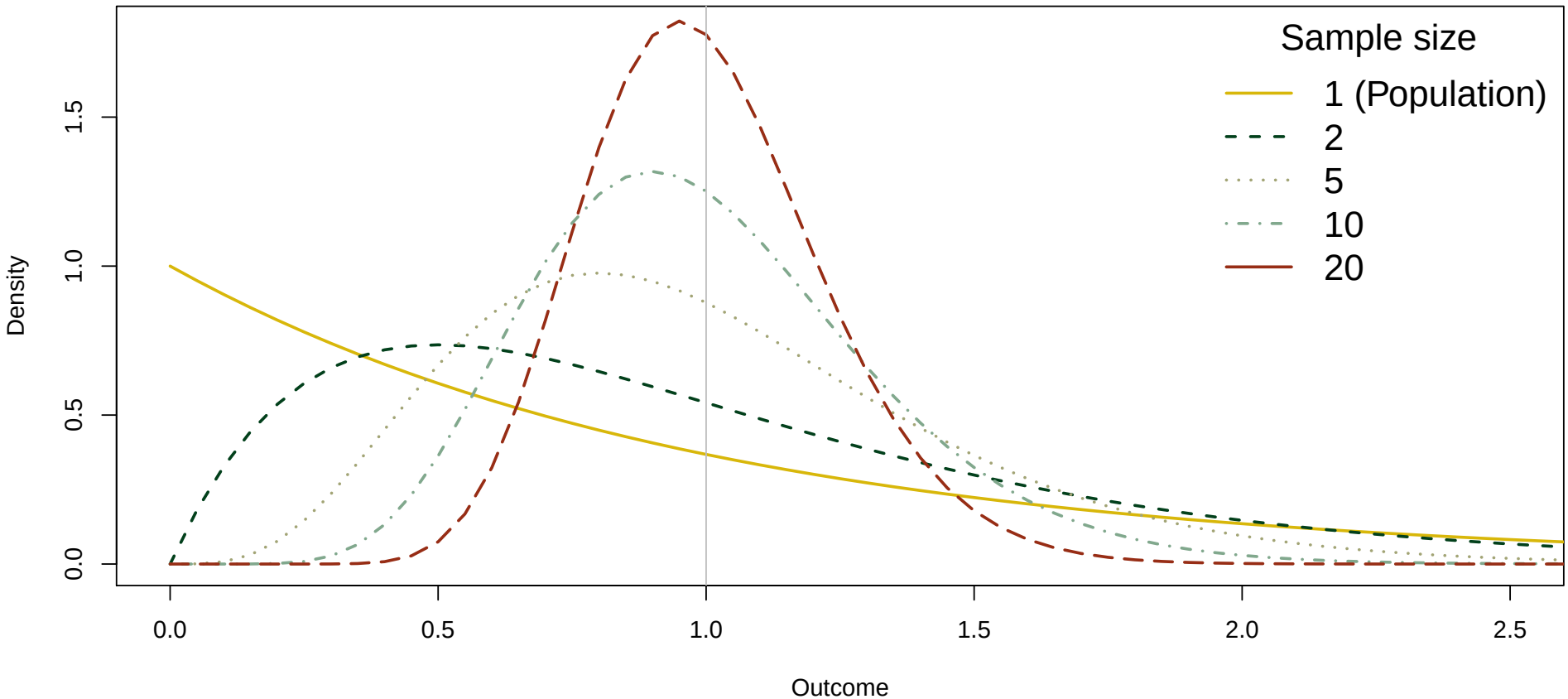
The Central Limit Theorem (CLT)

CLT: Regardless of the true population density curve, the sample mean density can be made (with arbitrarily good approximation) Normal by choosing n large enough.

- How large does n have to be?
  Depends on how skew the population density is.
  - In general $n = 20$ will suffice.
  - If population is Normal then $n = 1$ is enough.
- CLT applies to many 'statistics' (= functions of samples).

Attempt to visualize CLT for the sample mean

The distribution of sample means from a skewed distribution.



Quantiles?

Quantiles divides your data into (roughly) equal piles.

- the median is the 2-quantile
- the tertiles are the 3-quantiles (the $33\frac{1}{3}$ percentile and the $66\frac{2}{3}$ percentile)
- the quartiles (Q1, Q2 and Q3) are the 4-quantiles.
- . . . and so on.

Cross-over data (from lecture 4)

13 patients had their peak expiratory flow (PEF, l/min) recorded once after inhaling each of two different asthma drugs (the order of which were random).

In *paired* data one usually look at the 13 differences as a measurement of differences in individual effect size.

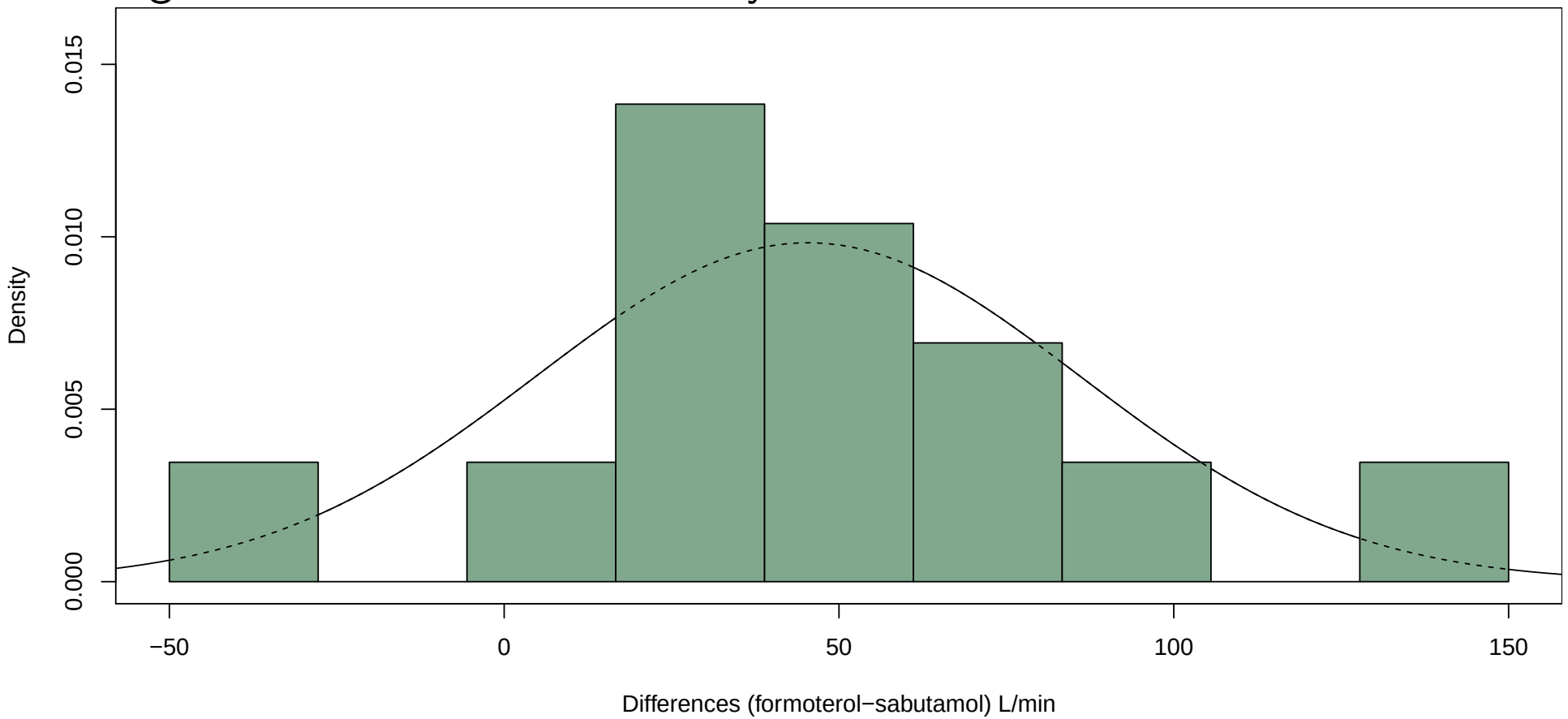
Data:

40, 50, 70, 20, 40, 30, -35, 15, 90, 30, 30, 80, 130

Is the normal distribution a good model for these 13 numbers?

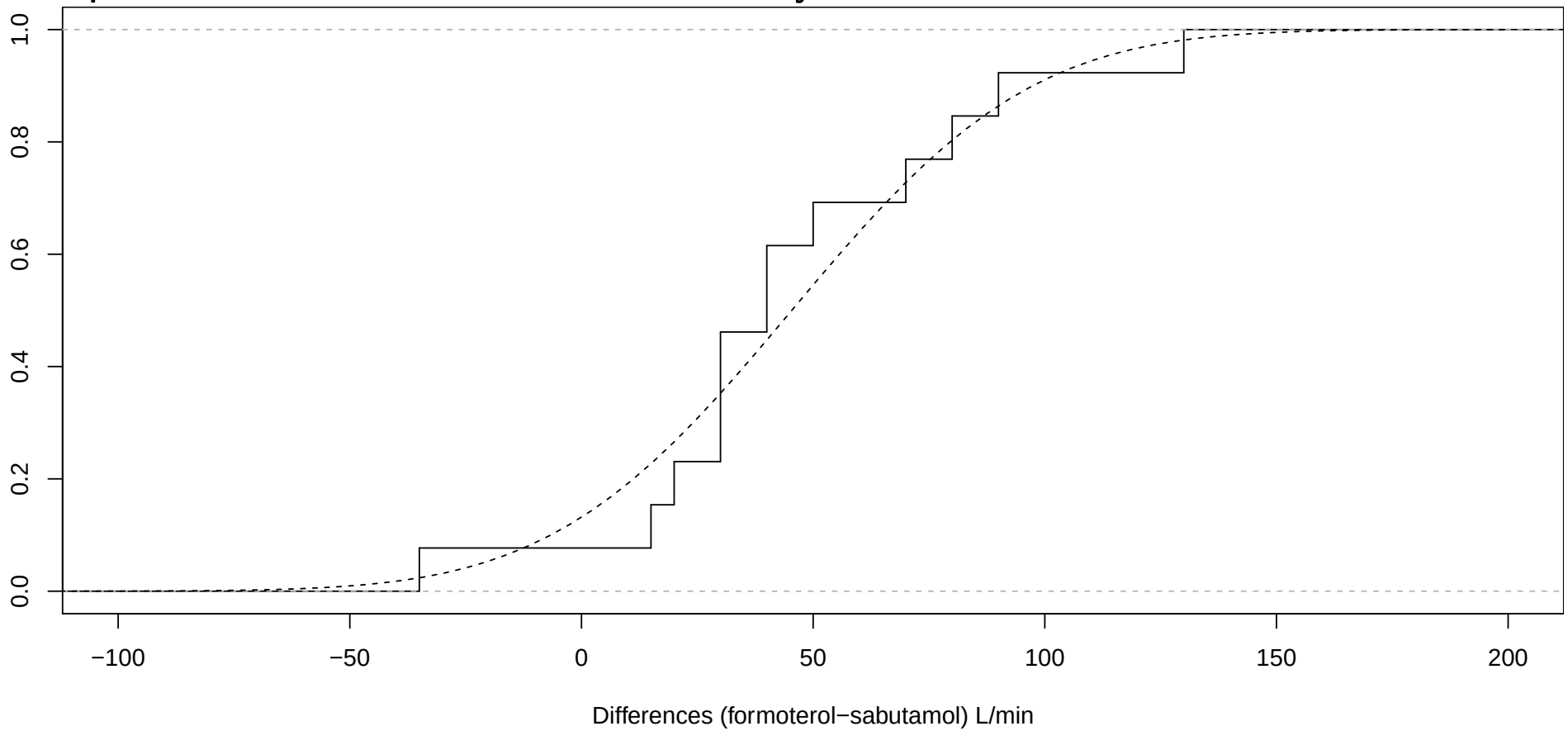
PEV

A histogram with best normal density fit.



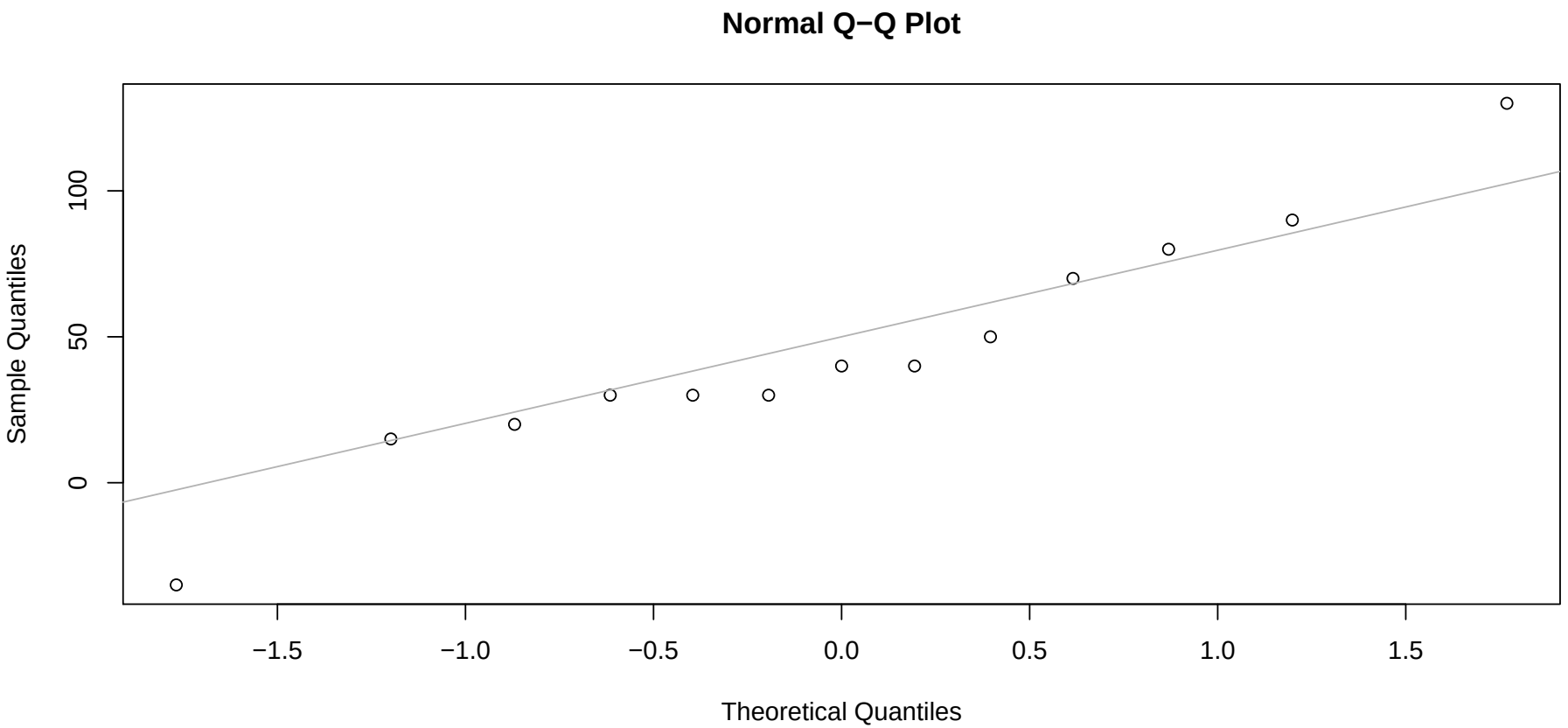
PEV

A empirical cdf with best normal density cdf.

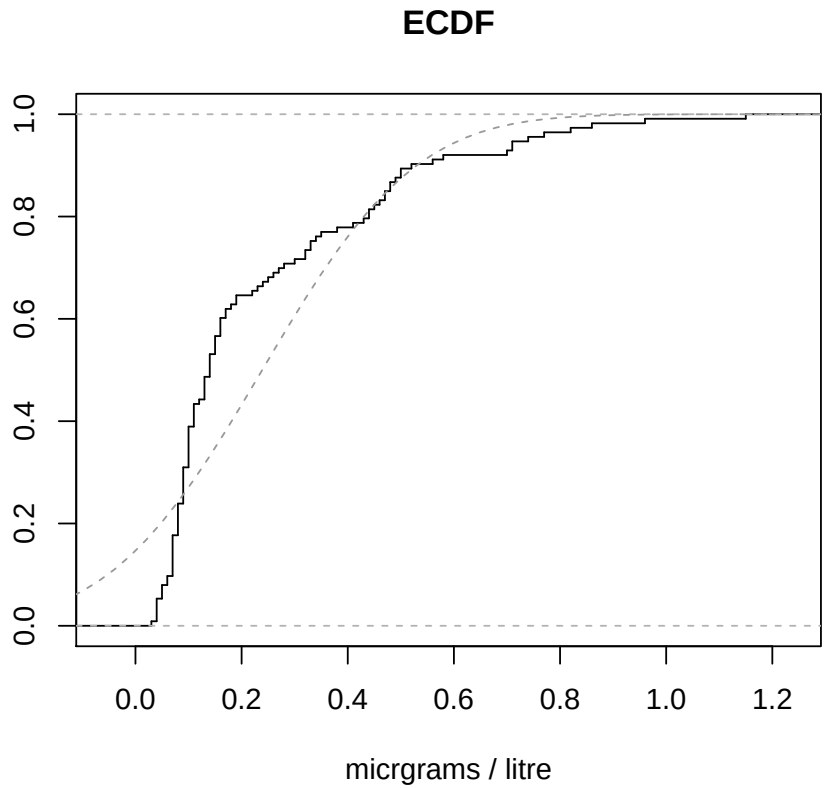


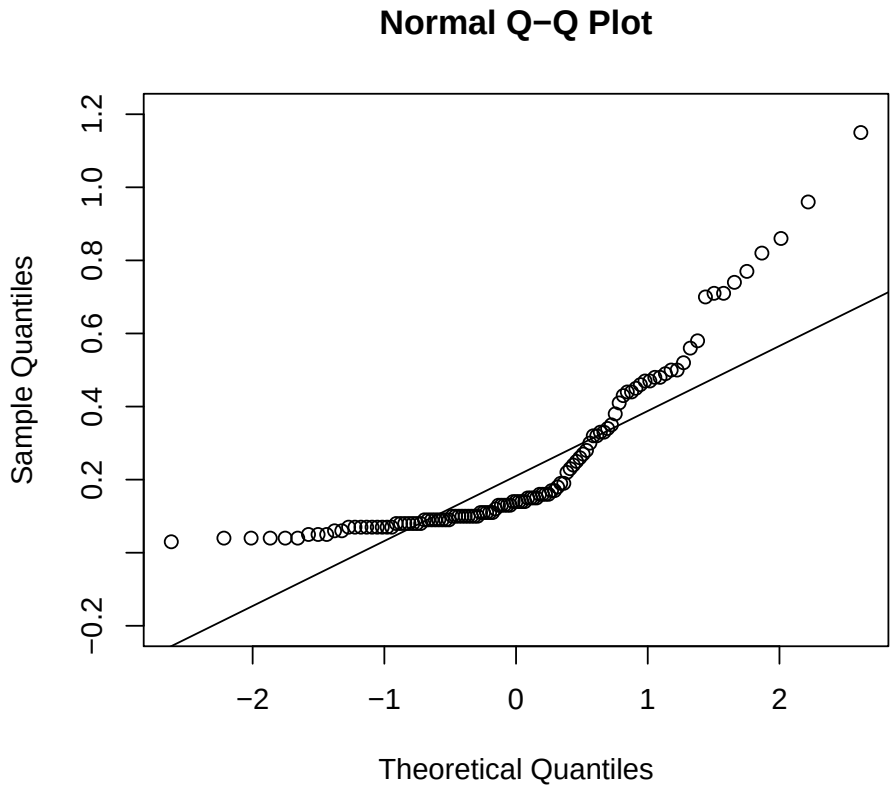
The Quantile-Quantile plot

If the differences in effect size is sampled from a Normally distributed population its QQ-plot should be a straight line (approximately).
A QQ-plot plots the sample of size n against the (slightly shifted) n-quantiles of the (standard) Normal distribution.



These visual test require some training.
There are also formal tests of normality (e.g. Shapiro-Wilks)
The S100β measurements from the subarachnoidal bleeding example is certainly not normally distributed.





Second Summary

95% of observations from a Normal population lies within 1.96 multiples of the (population) s.d. from the (population) mean.

Means and in particular s.d. must be distinguished on three levels

- population,
- sample, and
- estimate.

The s.d. of the latter is called the Standard Error
(Standard Error of the Mean if the mean is used as estimate.)

(The sample s.d. is an estimate of the population s.d.)

The CLT explains why many estimates ('statistics') are (approx.) Normally distributed even though the population may not be.

References

- Chapters 1-8, 10: Petrie & Sabin. *Medical Statistics at a Glance*, Wiley-Blackwell (2009).
- Puhan et al. *More medical journals should inform their contributors about three key principles of graph construction*, Journal of Clinical Epidemiology, **59** (2006) 1017-1022.
- Franzblau & Chung. *Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter*, American Society for Surgery of the Hand, **37A** (2012) 591-596.
- Kelleher & Wagener. *Ten guidelines for effective data visualization in scientific publications*, Environmental Modelling & Software **26** (2011) 822-827.
- L. Wilkinson, *The Grammar of Graphics*, 2$^{\text{nd}}$ ed., Springer 2005.