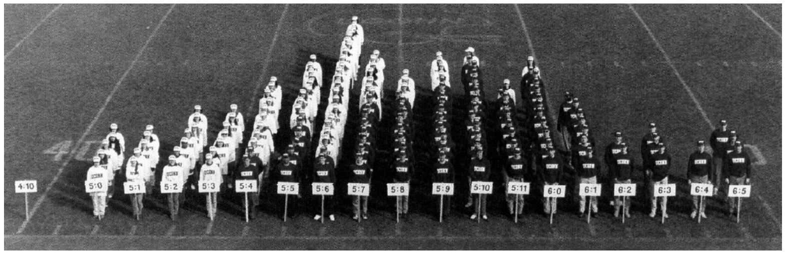
Introduction to Biostatistics



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

Introduction to Biostatistics
Lecture 1B and 2

Henrik Renlund

Uppsala Clinical Research Center

What shall we learn today?

- Data description
  - Graphs
  - Tables and summary measures
- Probability Models
  - Glimpse at theory (models/distributions)
  - The Normal distribution
  - Some properties of samples and the Central Limit Theorem.

Types of data

A data set contains one or more *variables* for each unit of study

ID	Sex	Age	Children	Albumin	Diabetes	Happiness
1	M	67	0	3.92	0	$\smile$
2	F	71	3	4.12	0	$\frown$
3	F	49	1	4.75	1	$-$
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Data categories:

- Categorical
  - nominal, e.g. `Sex`, `Diabetes`, *or*
  - ordinal, e.g. `Happiness`: $\frown, -, \smile$.
- Numerical
  - discrete; typically integer valued 0, 1, 2, . . . , like `Children`, *or*
  - continuous; i.e. any value in an interval, like `Albumin`.

The category determines what analyses are available.

Data management

Make sure you and your software agree on variable formats.

This is especially important if data has been transferred, e.g. between formats or operating systems.

Common problems:

- date- and categorical data stored as integers
- numerical values stored as text (due to ',' vs. '.')
- how are missing values represented? "Unknown"?

Descriptions Models Sampling Misc. Ref Polls

Visualization of (continuous) data

A sufficiently small data set might not need visualization.

The level (g/dL) 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

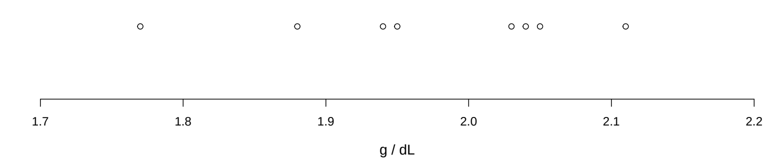
One simple way to get some handle on data is to order it:

1.77 1.88 1.94 1.95 2.03 2.04 2.05 2.11

Descriptions Models Sampling Misc. Ref Polls

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.

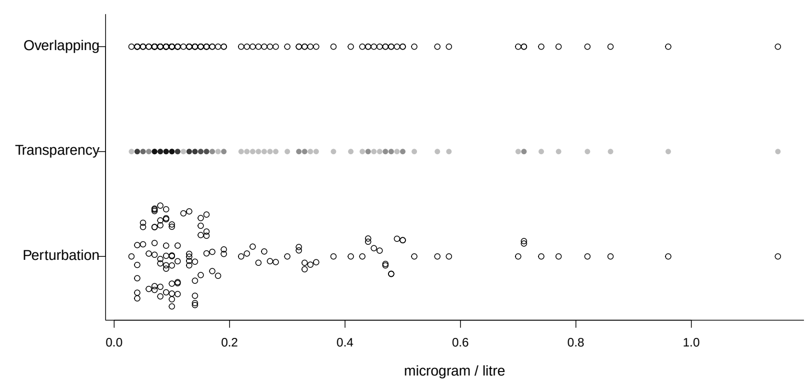
This can be alleviated by

- perturbation, or,
- (alpha) transparency.

Descriptions Models Sampling Misc. Ref Polls

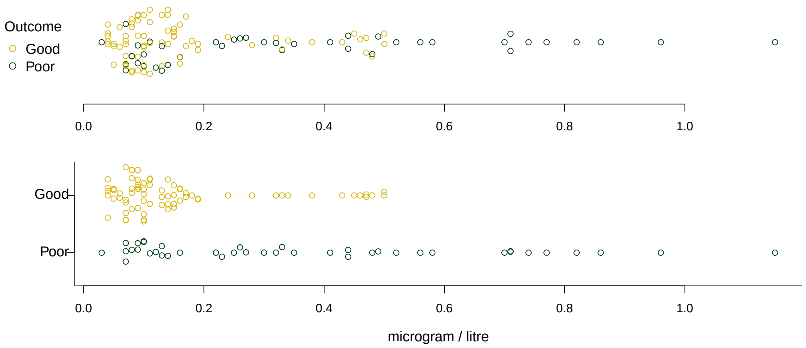
Subarachnoidal bleeding

A biomarker - the protein S100B - was measured for 113 individuals with aneurysmal subarachnoid hemorrhage.



Dotplot and groups

Dotplots can display groups.



"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	Diabetes: No		Diabetes: Yes	
value	*mean*	*sd*	*mean*	*sd*
Age	32.0	15.9	32.5	14.1
Albumin	4.20	0.37	3.80	0.50
	percent	*n*	*percent*	*n*
Sex				
M	64%	27	52%	22
F	36%	15	48%	20
Happiness				
⌢	61%	19	36%	15
—	23%	7	36%	15
⌣	16%	5	28%	12

Percentiles (Measure of location)

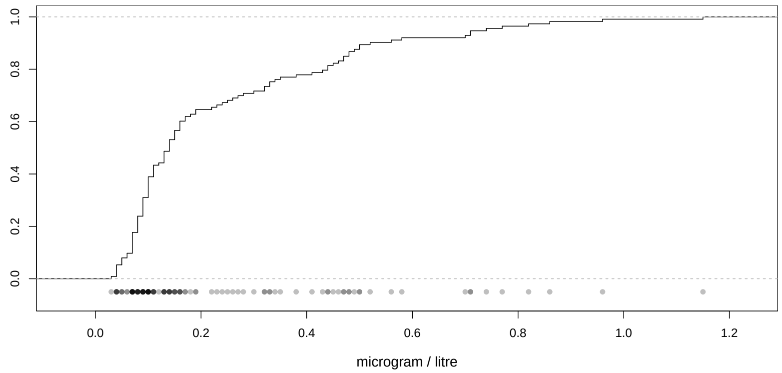
- The kth percentile is a value v such that k percent of your data lies below (or at) v. (Usually not uniquely defined.)
- The 50th percentile (the *median*) is the point which divides your ordered sample equally. (Only 'unique' if sample is odd, else use mean of the two midpoints.)
- *The Quartiles:* Q1 is the 25th percentile, Q2 is the 50th percentile and Q3 is the 75th percentile.
- We can describe all percentiles with the *cumulative frequency graph* (CF) also called the *empirical cumulative distribution function* (ECDF)

Creating a CF

A CF shows the cumulative frequency (or cumulative proportion) and thus starts at 0 for points smaller than the smallest point of the data set. Then it is a step-wise function with jumps according to:

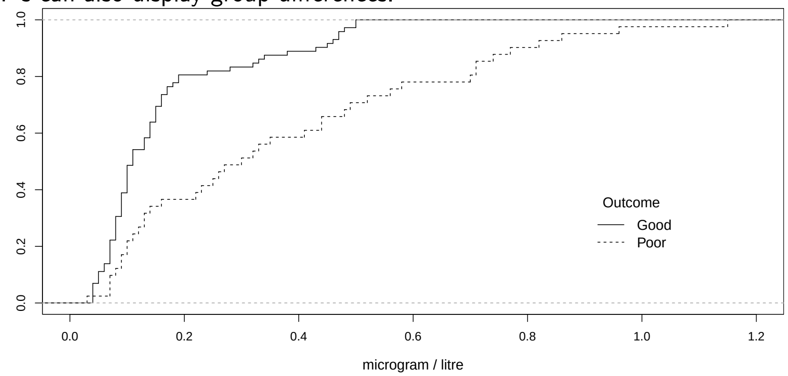
Unique points	Count	Cumulative count	Cumulative proportion
0.03	1	1	$\frac{1}{113} \approx 0.009$
0.04	5	6	$\frac{6}{113} \approx 0.053$
0.05	3	9	$\frac{9}{113} \approx 0.080$
⋮	⋮	⋮	⋮
1.15	1	113	1.000

Cumulative frequency for S100B



Cumulative frequency function

CF's can also display group differences.



Survival curves

A survival curve is a CF. Survival (time-to-event) data is typically *right censored* and the curve thus needs to be estimated (Kaplan-Meier) - more on that later in the course.

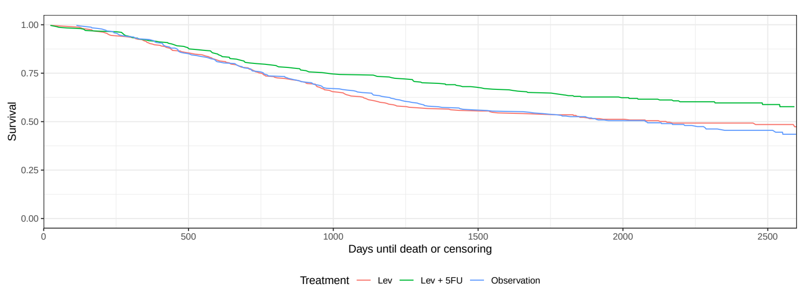
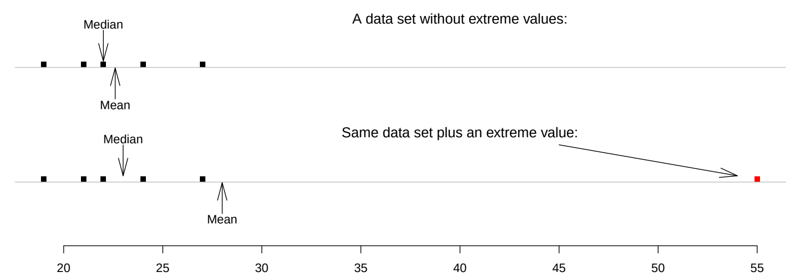


Figure: Survival curves for 3 different treatments of colon cancer; observation only, Levasimole, or Levasimole and 5-FU. (Moertel 1991)

Average value (Measure of location)

An average value should be representative of the entire data set.

- **The median:** is the midpoint of the ordered numerical sample when one iteratively cancelles the smallest and largest points.
- **The mean:** is the center of gravity of a data set.
  Note: unlike the median, it 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. (The numbers from the previous plot.)

Salaries	Excluding boss	Including boss
Median	22	23
Mean	22.6	28

Ex: The number of hospitalization days per individual in Uppsala is likely to be very skewed. The median might be of most interest on an individual level, wheras the mean (which is essentially the sum) is of more interest to whoever is in charge of the hospital budget (as well as other things).

In fact: the median is probably 0! One could e.g. describe this distribution by the median among those with non-zero hospitalization days.

Measures of spread

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

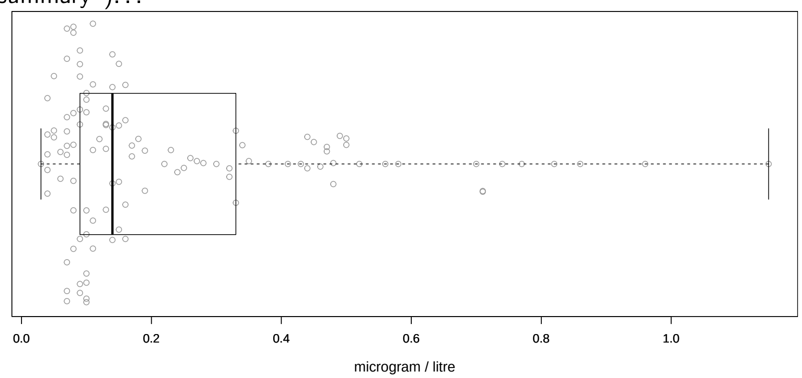
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

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 the typical distance between a value and the mean value.

Note: the sd in the previous salary example is 3.0 and 13.5 if the boss is excluded or included, respectively.

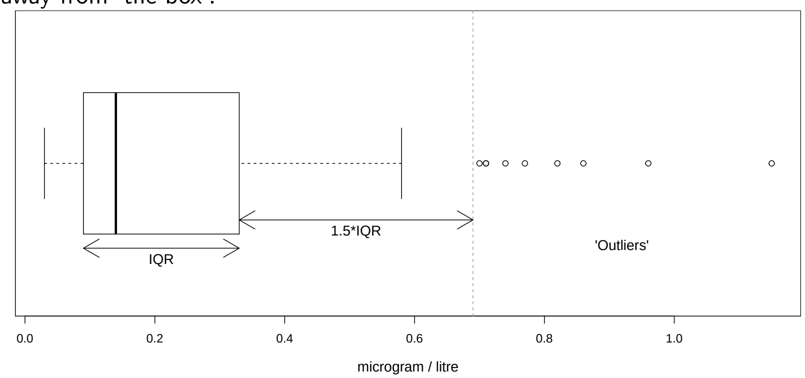
Boxplot of S100B

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



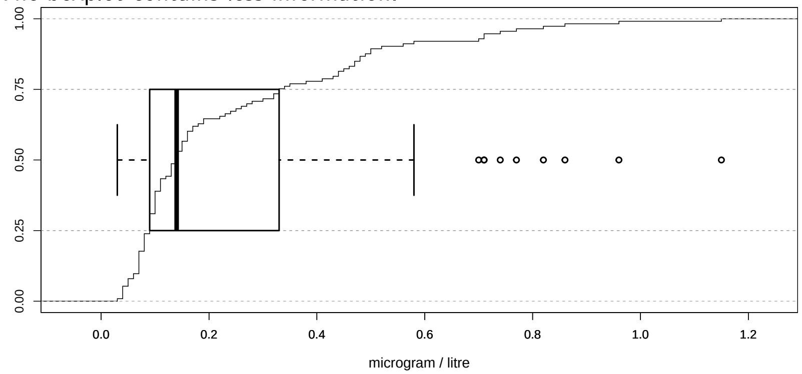
Boxplot

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



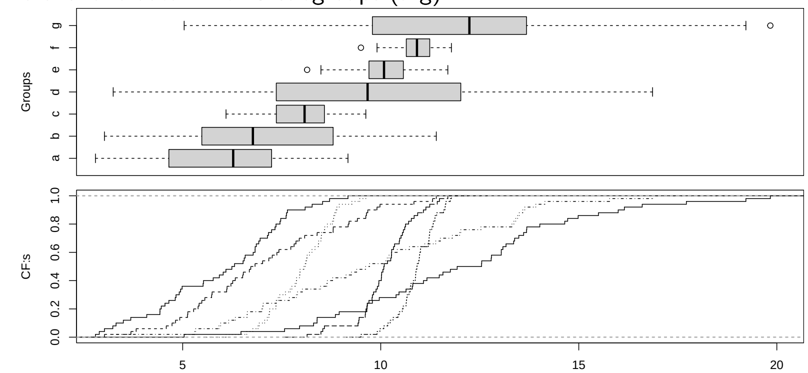
Connection between boxplot and cumulative frequency

The boxplot contains less information.



Pattern or detail?

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



Some rules of thumb for tables and graphs

Both:

- Table/graph + caption should be self-contained.

Tables:

- Captions *above* the table.
- Avoid excessive precision and use adequate measures of location and spread.

Graphs:

- Captions *below* the graph
- 'Economy' Do not make a graph which is more easily expressed in text or a small table, e.g. graph with a single boxplot.
- Avoid 2D graphs shown in 3D.
- Colors are tricky (colorblindness, black/white-printing, etc.) - a website like Colorbrewer (https://colorbrewer2.org) might guide color choice.

Probability theory and models

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

Probability theory: given model (model parameters, or other aspects) - describe how data behave. E.g.

- specific results: how likely are specific deviations
- 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).

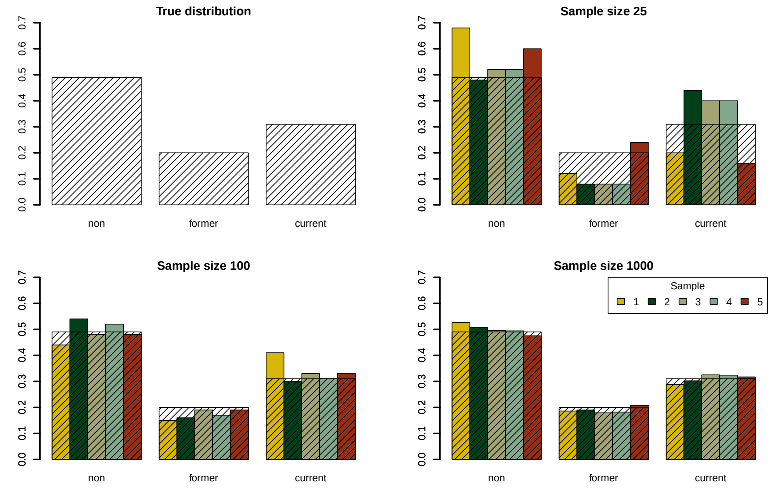
Let Z ='the outcome of the throw of a die'. Then $\text{Prob}(Z = k) = 1/6$ for all $k = 1, 2, \ldots, 6$, or, equivalently

Value k	1	2	3	4	5	6
$\text{Prob}(Z = k)$	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
$\text{Prob}(X = v)$	0.49	0.20	0.31

Five samples from three different sampling sizes from previous distribution



FEV data set

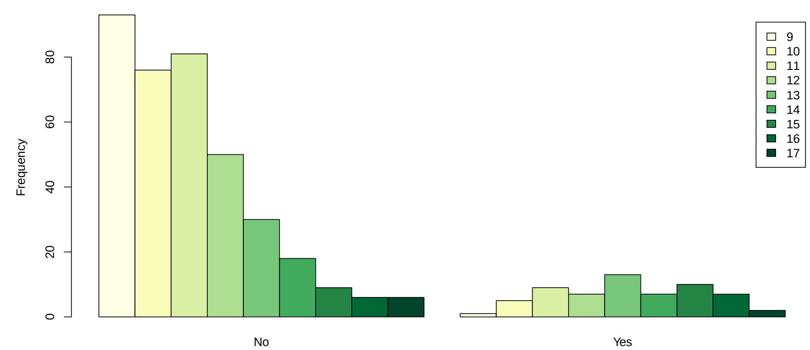
430 children (9-17 years of age) had their age, forced expiratory volume in 1 second (FEV) and smoking status recorded.

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'

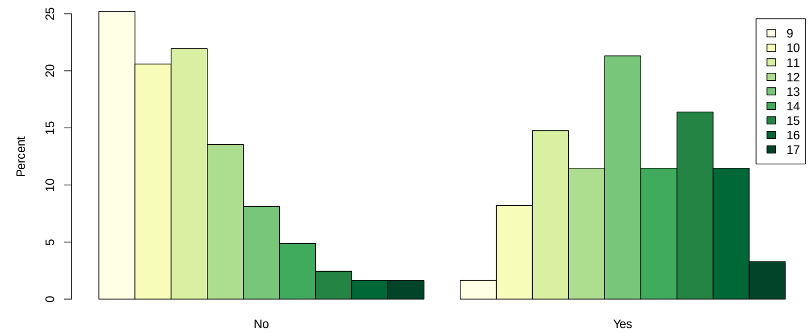


If the groups (smokers/non-smokers) aren't balanced it is difficult to compare the distributions.
Tabulate/plot the percentages within groups:

Age	**Smoking (proportion)**	
	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

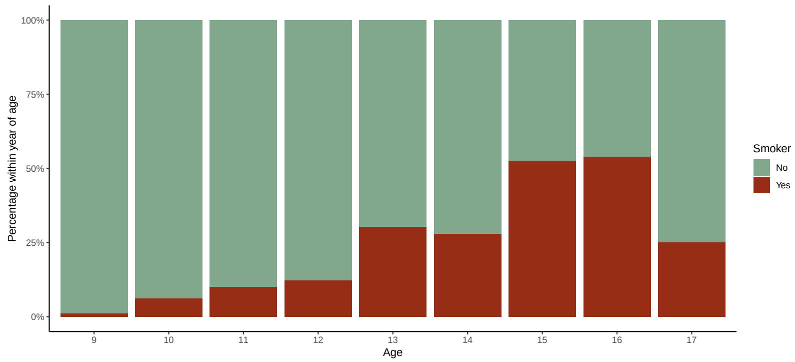
Percentages within smoking groups

The distributions are now easier to compare.



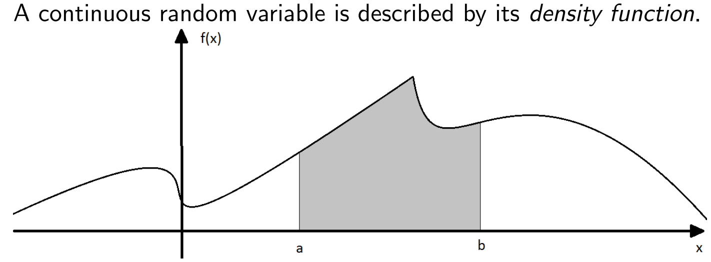
Flipping the axes

The "right" visualization might not be obvious. A binary categorical variable can be suitable to *stack* in a bar chart:



Probability model for continuous data

A continuous random variable is described by its *density function*.

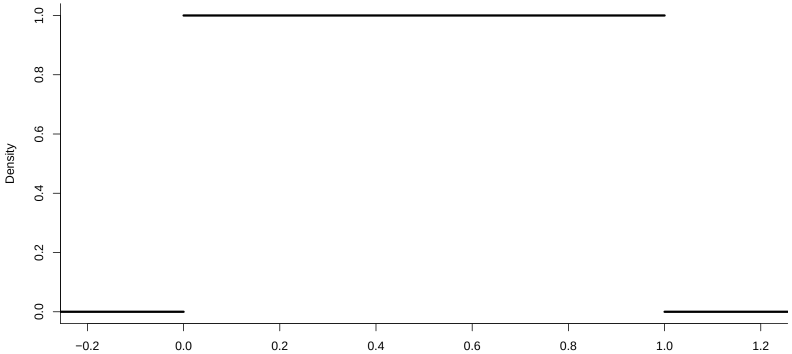


If X has density function f as above, then we compute probabilities as

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

Example: The Uniform distribution

A computer generated random number typically tries to mimic the *uniform distribution* (on the $[0, 1]$ interval).



Any number (or interval) within $[0, 1]$ is as likely as any other (of the same length).

Making a histogram

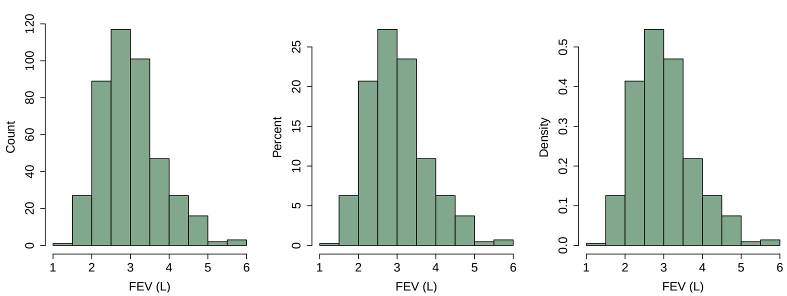
A histogram is a categorization of the x-axis into "bins", typically as intervals of the same range, and a statistic associated with each.

The FEV dataset has 430 (numercial) FEV-measurements between 1 and 6.

Data underlying a histogram:

Interval	Count	Proportion	Density
$1.0 - 1.5$	1	$\frac{1}{430} \approx 0.0023$	$\frac{1/430}{1.5-1.0} \approx 0.0047$
$1.5 - 2.0$	27	0.063	0.13
$2.0 - 2.5$	89	0.21	0.41
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$5.5 - 6.0$	3	0.0070	0.014

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

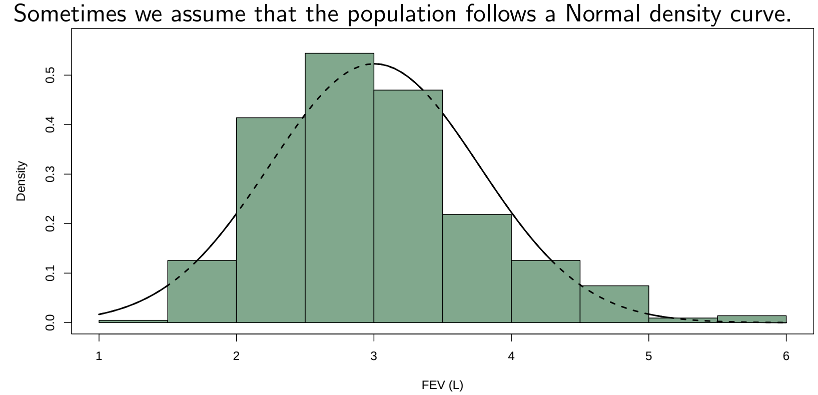


"Density" is more abstract but

- 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

The Normal Distribution

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



The Normal Distribution

The Normal, Gauss, or Bell, curve is centered at (the mean) μ with a standard devation of σ, according to the equation:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma}}.$$

The special case of $\mu = 0$ and $\sigma = 1$ is called a *standard normal distribution*.

Digression: Any sample can be "standardized" by subtracting the mean and dividing by the standard deviation.
If $x_1, x_2, \ldots, x_n$ is a sample (of size n) with mean $\bar{x} = \frac{1}{n}\sum_1^n x_i$ and $sd = \sqrt{\frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2}$, then the transformation

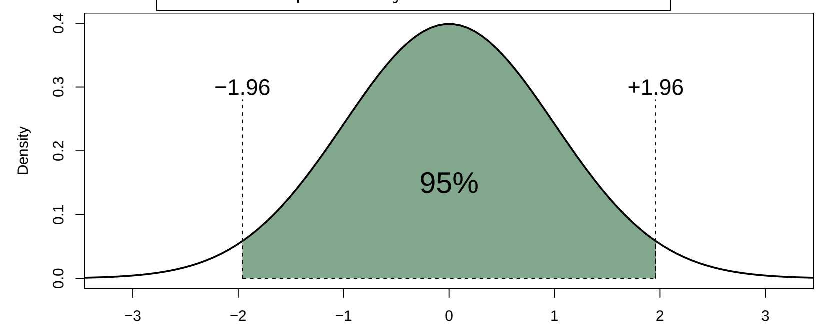
$$\frac{x_1 - \bar{x}}{sd}, \frac{x_2 - \bar{x}}{sd}, \ldots, \frac{x_n - \bar{x}}{sd},$$

is *standardized* (it has mean 0 and sd 1).

Properties of the standard Normal distribution

The standard Normal distribution is centered at 0 and has a standard deviation of 1.

95% of the 'probability mass' lies within ± 1.96.

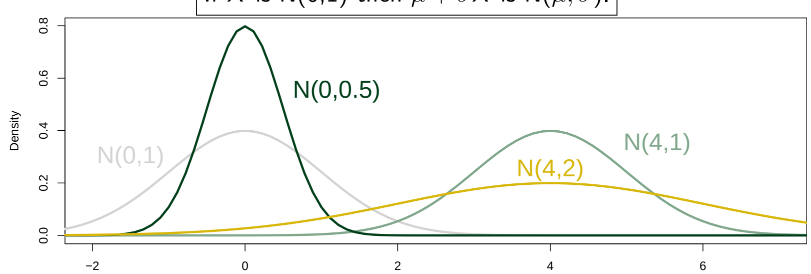


This will be useful when creating *confidence intervals*.

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).



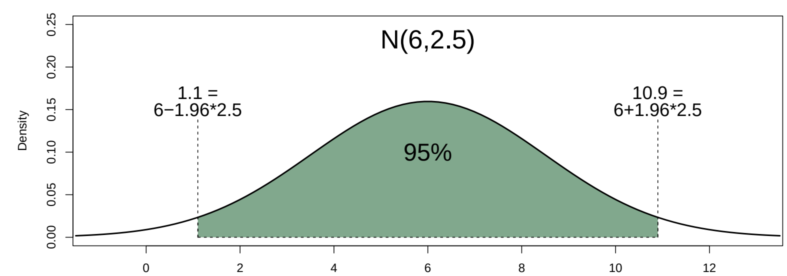
In fact, if X_1 and X_2 are Normal (and *independent*), then so is

$$a + bX_1 + cX_2.$$

Properties of the Normal distribution

If X is N(μ, σ), then 95% of observations will be between

$$\mu - 1.96\sigma \text{ and } \mu + 1.96\sigma.$$



The problem

Fact: to calculate e.g. a confidence interval for X, we need to know the distribution (at least approximately).

We want to estimate a characteristic of a population, we will use **the mean** as an example.

We may have no prior information about the population, but we get some information from a random sample. (Assume sample size > 1.)

We estimate the population mean by the sample mean.

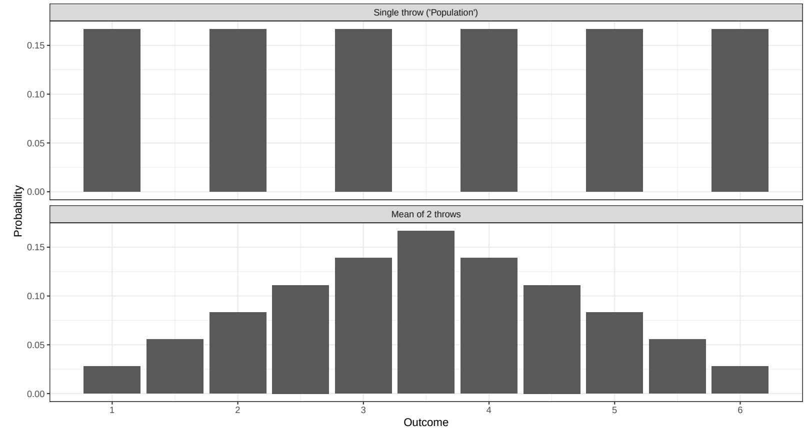
What is the distribution of the sample mean (considered as a random number)?

Hard to know - we only ever get 1 such number!

("Bootstrap" might be a work around.)

Mean value distribution $\neq$ population/sample distribution

Consider the simple case of the mean of 2 dice.



Understanding sampling using the Seeing Theory website

The Seeing Theory website
https://seeing-theory.brown.edu/probability-distributions
provides a way to visualize sampling (Section 'Central Limit Theorem').

A population distribution is given (with parameters α and β).

You simulate calculating a *sample mean* from a sample of size ≤ 15 up to 50 iterations at the time.

Note: at sample size 1, this means just recording the value sampled.

- At sample size 1, with many iterations the sample distribution should resemble the population.
- What happens to the distribution of the sample mean as the sample size increases?

Note: in a real study we only get the one sample, so we never observe the sampling distribution directly.

Important features of sampling

A population has a mean μ and sd σ.

A random (unbiased) sample will resemble the population (approximately same mean and sd).

But the distribution *of the sample mean* is more centered around this value μ, in particular

- it has a smaller standard devation, and
- it tends to be more symmetric.

But again: the sample mean distribution is unobserved, in general we must rely on theory to determine its properties.

Sampling from a normally distributed population?

$\boxed{\text{"Normal population"} \Rightarrow \text{"Normal sample mean"}}$

A random sample of size n from $N(\mu, \sigma)$ has the property that its mean is also normally around μ but with

$$\sigma_{\text{sample mean}} = \frac{\sigma}{\sqrt{n}}.$$

But, not everything is normally distributed to begin with!

The Central Limit Theorem (CLT)

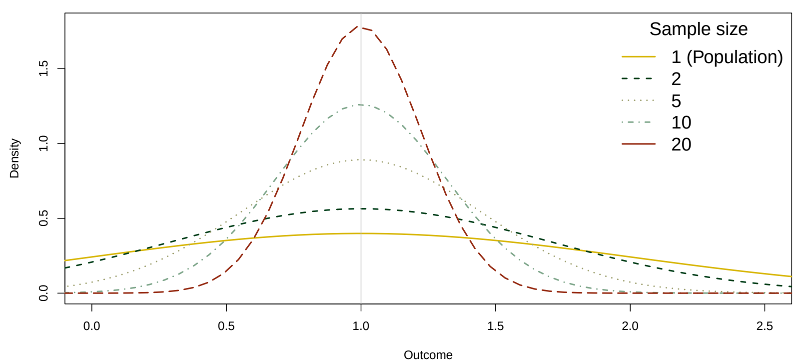
CLT: Regardless of the population density curve, the sample mean density can be made (with arbitrarily good approximation) Normal by choosing n large enough.

- 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).

The next 2 slides shows theoretical distributions for a population and sample means thereof, for varying sample sizes.

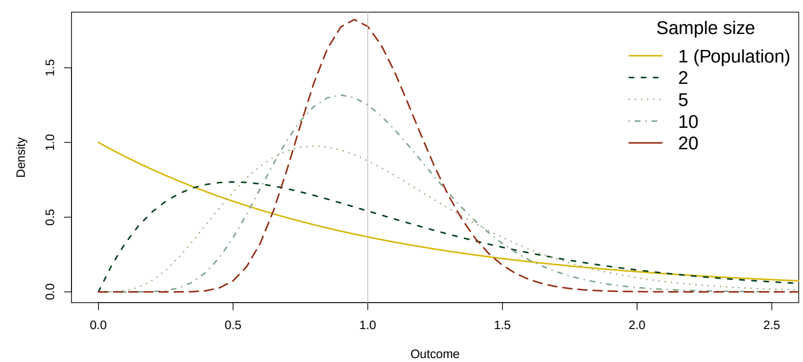
Visualization of sample mean from normal population

The distribution of sample means from a *normal* distribution. In this case the sample mean *is normally distributed* (but with a smaller sd).



Attempt to visualize CLT for the sample mean

The distribution of sample means from a skewed distribution. The sample mean is *not* normally distributed, but increasingly so.



Standard errors

The (estimated) standard deviation of an estimator is called its **standard error**.

The CLT allows us in many situations to ignore the actual distribution of the population and the estimator, but we typically need the standard error.

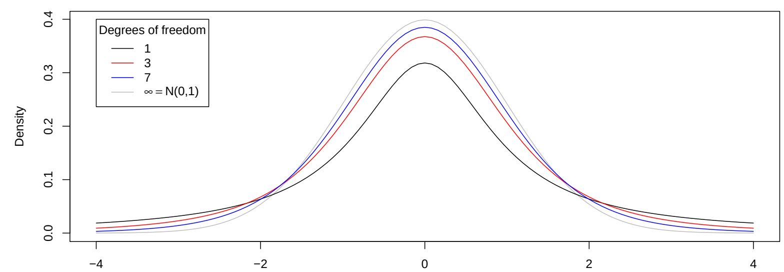
Exmple: suppose we want to estimate the height of individuals in a population. A sample of 50 people gave a mean 175 cm with a sample standard deviation of 6.5. Then $se = 6.5/\sqrt{50} \approx 0.92$.

As will be shown later the standard error help us to

- create confidence interval, and
- create *test statistics* for hypothesis testing.

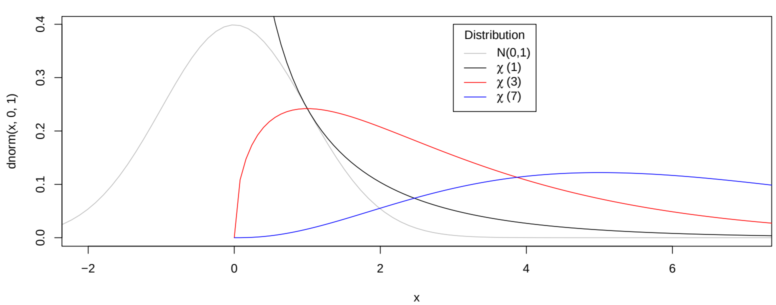
The t-distribution

Calculating standard test statistics from a normal distribution requires estimating the standard deviation. For **small samples** the change in distribution motivates making exact calculations. The t-distribution is "increasingly standard normal" as its parameter (degrees of freedom) increases.



The χ^2-distribution

If you add k standard normal random numbers, each squared, then the resulting distribution is $\chi^2(k)$. As we will see this is useful in the context of categorical variables.



Visual tests of normality

Perhaps surprisingly, quite often we rely on *visual* rather than *formal* tests of model assumptions.

A common formal test of normality is the Shapiro-Wilks test.

Many plots can provide a visual test of normality, but a common one is the Q-Q plot. 'Q' is for *quantile*.

Quantile? 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

13 patients had their peak expiratory flow (PEF, l/min) recorded 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 effect size.

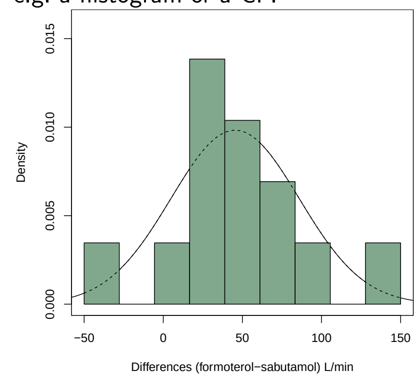
Data:

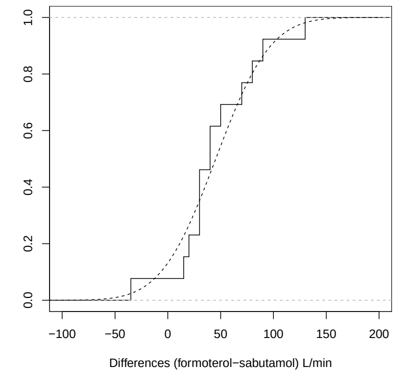
40, 50, 70, 20, 40, 30, -35, 15, 90, 30, 30, 80, 130

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

PEV

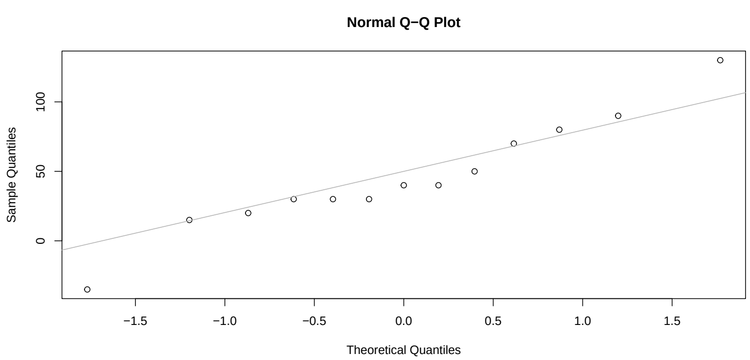
One could ascertain the plausibility of an underlying normal distribution via e.g. a histogram or a CF.



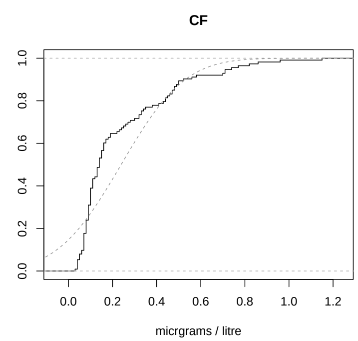


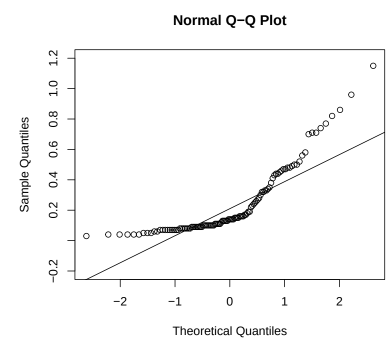
The Quantile-Quantile plot

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



The S100B measurements is certainly not normally distributed.





Caution

It is very rarely the actual data that is tested for normality!

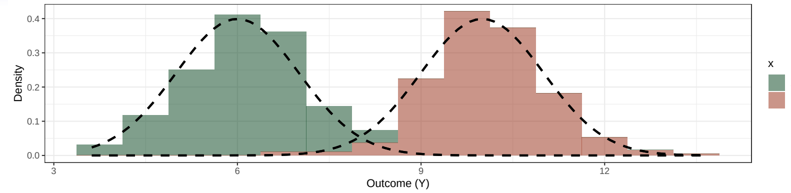
Most of the time the models that assume normality does so for the *error terms*, i.e. there is a model, depending on the covariates x, for the outcome Y such that

$$Y = \text{some deterministic function of x} + \text{random error}.$$

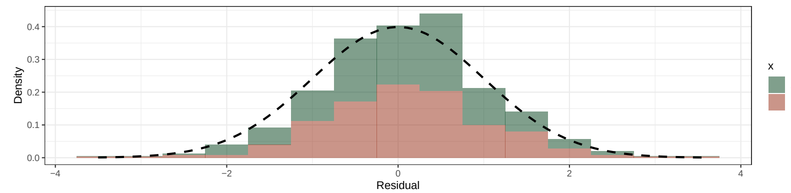
E.g. a 2-sample *t*-test assumes that an outcome is normally distributed around a group-specific mean. Data for such a test might look like this

outcome (Y)	5.1	6.2	7.9	9.2	4.7	...
group (x)	A	A	B	B	A	...

We cannot test the entire Y data for normality. This is evident if imagine the group effect to be very large...



It is the deviations (noise/error term) around each group-specific mean that is supposed to be normal, an estimation of which is called *residuals*. Subtract the (estimated) group effect from each datapoint to get:



What about missing data?

'Good' scenario: Suppose in an experiment a batch of samples are destroyed throught some random accident. Typically this only leads to a smaller sample size, but there is no problem running the analysis as planned.

'Bad' scenario: Suppose we study severity of myocardial infarctions with a model that includes sex, age, BMI (some missing) and smoking status (some missing). Worry: the reason for missing depends on the value.

The statistical software default is to include only those individuals with complete case data on all variables in the analysis.

This **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 (and/or the outcome).

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 (perhaps even "predicted" from other covariates).

> *"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.

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*, 2nd ed., Springer 2005.

Poll: Variables and driving

Q1: Tick the boxes you think are always true. (Multiple choice)

- ☐ An ordinal variable with values A, B, C, (etc) is such that the "increases" (although not numerical) B-A, C-B, (etc) can be considered to be of the same magnitude.
- ☐ Anything measured numerically must be analysed as such.
- ☐ Anything recorded as a categorization must be analysed as such.
- ☐ None of the above.

Q2: A poll showed that 80% of drivers thought their driving skills were above average. As a group, do these drivers overestimate their driving skills? (Single choice)

- ☐ No
- ☐ Yes
- ☐ Probably
- ☐ We can't say

Poll: Boxplots and means

Q1: A group of 24 individuals have been randomly assigned to 6 different exercise groups, and their blood pressure is measured after training. The researchers do a boxplot for the measurements for each group. Is this a good approach? (Single choice)

- ☐ Yes, a good way to visualize the overall pattern.
- ☐ No, the overall sample size is too small.
- ☐ No, the group sizes are too small.

Q1: Suppose the mean athletic skill is somewhat better in group A as compared to group B (of equal size). If a sport team is fairly admitting members by assessing athletic skill, we should expect more people from A being admitted. (Single choice)

- ☐ True
- ☐ False
- ☐ We can't say

Poll: LLN and CLT

Q1: Suppose you have flipped a (fair) coin 3 times, every time showing H ('Heads'). The Law of Large Numbers tells us that the relative frequency of H will tend to 50% in the long run. Does this imply that we now have an increased probability of T ('Tails')? (Single choice)

- ☐ Yes
- ☐ No

Q2: The Central Limit Theorem says that, althought the distribution of any numerical characteristic of individuals (income, blood pressure, etc) might be skewed in small populations, it will be (approx.) normally distributed in "large enough" populations. (Single choice)

- ☐ True
- ☐ False