Statistical Methods: Lecture 1

Lecture Overview

What is statistics?

- Statistics is the science of data, the main goal is to draw inference on the random mechanism that produced the data at hand.
- In particular, we use statistics (methods/tools/techniques) to gain information about a group of objects (population) and/or to make decisions and predictions.
- Census is collection of data from every member of population. Usually too large too collect.
- Typically, a sample, a selected subcollection from the population, is studied: Sample $\rightarrow$ Data $\rightarrow$ Analysis $\rightarrow$ Conclusion about population.

1.2 Statistical and critical thinking

- Statistical study:
  1. Prepare
     - Context
     - Source
     - Sampling method
  2. Analyse
     - Graph data
     - Explore data
     - Apply statistical methods
  3. Conclude
- Recall: sample is subcollection of population, different sample $\rightarrow$ different data.
- Hence, possibly different conclusions about population
- A sample should be representative (same characteristics as population) and unbiased (no systematic difference with population).

1.4 Collecting sample data

Different sampling methods:

- Voluntary response sample: subjects decide themselves to be included in sample.
- Random sample: each member of population has equal probability of being selected.
- Systematic sampling: after starting point, select every k-th member.
- Stratified sampling: divide population into subgroups such that subjects within groups have same characteristics, then draw a (simple) random sample from each group.
- Cluster sampling: Divide popluation into clusters, then randomly select some of these clusters.
- Convenience sampling: easily available results.

1.4 Collecting sample data

Important concepts:

- Variable: varying quantity

In cause and effect studies:

- Response (dependent) variable: representing the effect to study
- Explanatory (independent) variable: possibly causing that effect
- Confounding: mixing influence of certain (typically unobserved) variables on explanatory and response variables

Different types of study:

- Observational study: characteristics of subjects are observed; subjects are not modified.
  - Retrospective (case-control): data from past
  - Cross-sectional: data from one point in time
  - Prospective (longitudinal): data are to be collected
- Experiment: some subject treatment.
  - Sometimes control and treatment group; single-blind or double-blind,
  - To measure placebo effect or experimenter effect.

1.3 Types of data

Parameter: a population's characteristic. Notation: often Greek symbols, e.g., μ, σ.

Statistic: a data based measurement describing a characteristic of the sample. Notation: random variables, X, T, $\bar{X}$; realized (observed) values $x, t, \bar{x}$, etc.

- Quantitative (numerical): numbers representing measurements; discrete (countably many possible values), continuous (uncountably many).
- Qualitative (categorical): names or labels represent measurements.

The level of measurement of data determines which statistical methods are applicable.

- Qualitative data:
  - Nominal: names, labels, categories (no ordering). Examples: gender, eye colour.
  - Ordinal: categories with ordering, but no (meaningful) differences. Examples: U.S. grades (A-F), opinions (totally disagree / . . . / totally agree).
- Quantitative data:
  - Interval: ordering possible and meaningful differences, but no natural zero starting point. Examples: year of birth, temperature $^\circ$C/$^\circ$F.
  - Ratio: ordering possible and meaningful differences and natural starting point. Examples: body length, marathon times.

E.g., determine level of measurement: M&M colours, inauguration years of U.S. presidents, brain volumes (cm^3), level of lead in blood (low/medium/high).

Recap data

- Population vs. sample
- Different sample $\rightarrow$ possibly different conclusion about population
- Sample must be representative and unbiased
- Different data types.

Summarizing and graphing data

Until the slides about numerical summaries, the coming topics are not in the book.

Summarize the data. Consider dataset: amount of cotinine in blood.

```
> head(cotinine)
        Smoker Passive smoker Non-smoker
1            1            384          0
2            0              0          0
3          131             69          0
4          173             19          0
5          265              1          0
6          210              0          0
```

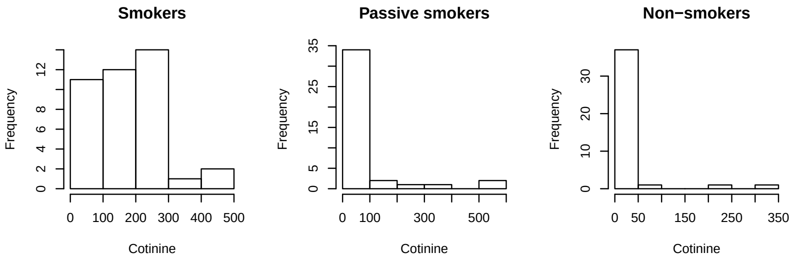
Example of numerical summary:

```
 > summary(cotinine)
     Smoker        Passive smoker     Non-smoker
 Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
 1st Qu.: 86.75   1st Qu.:  1.00   1st Qu.:  0.00
 Median :170.00   Median :  1.50   Median :  0.00
 Mean   :172.47   Mean   : 60.58   Mean   : 16.35
 3rd Qu.:250.75   3rd Qu.: 25.50   3rd Qu.:  0.00
 Max.   :491.00   Max.   :551.00   Max.   :309.00

> apply(cotinine,2,sd)  # standard deviations for 3 samples
        Smoker   Passive smoker     Non-smoker
      119.4983       138.0839        62.5335
```

Summarizing and graphing data

Example of graphical summary is histogram, consisting of bars whose heights are equal to the numbers of measurements in the corresponding intervals (cells).

```
par(mfrow=c(1,3))
hist(cotinine[,1],main="Smokers",xlab="Cotinine",ylab="Frequency")
hist(cotinine[,2],main="Passive smokers",xlab="Cotinine",ylab="Frequency")
hist(cotinine[,3],main="No smokers",xlab="Cotinine",ylab="Frequency")
```



Summarizing and graphing data

Choose summary most suitably for research question. Often interest in data distribution. Good summary shows:

- characteristics of data distribution: location, spread, range, extremes, accumulations, gaps, symmetry,...

Depending on context and goal:

- data sampled from a certain distribution?
- Different groups needed for further analysis?
- Influences of other variables, e.g. time?
- Dependence between variables?

Graphical summaries

Summarize → describe/find structure in data distribution:

- Graphical: tables, graphs, other figures
- Descriptive
  - Qualitative: describe shape, location and dispersion/variation
  - Quantitative: numerical summaries of location and variation

First step in every data analysis: if possible, make figures of data for own use → right choice of statistical methods.

Possible graphical summaries (not all applicable to all types of data):

- Frequency distribution (table)
- Bar chart
- Pareto bar chart
- Pie chart
- Histogram
- Time series

Summaries

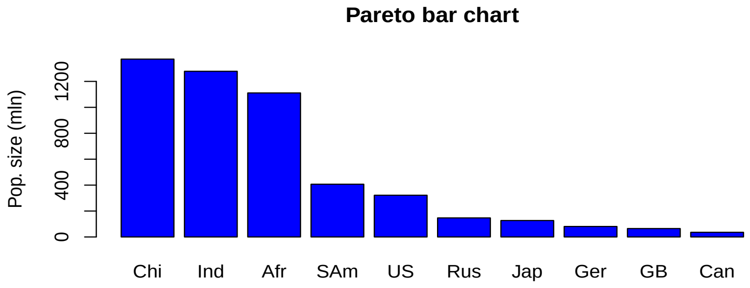
Data: exam grades

```
> grades=c(10,7,6,10,8,5,8,7,5,9,7); grades2=rbind(1:11,grades)
> rownames(grades2)=c("Student","Grade"); grades2

        [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
Student    1    2    3    4    5    6    7    8    9    10    11
Grade     10    7    6   10    8    5    8    7    5     9     7

> freq=table(grades2[2,]); freq   # frequencies of the grades

5  6  7  8  9 10
2  1  3  2  1  2

> freq2=cbind(freq,cumsum(freq),freq/length(grades),cumsum(freq)/length(grades))
> colnames(freq2)=c("Frequency","Cumulative","Rel.frequency","Cum.frequency")
> options(digits=2);freq2

      Frequency Cumulative Rel.frequency Cum.frequency
5           2          2         0.182          0.18
6           1          3         0.091          0.27
7           3          6         0.273          0.55
8           2          8         0.182          0.73
9           1          9         0.091          0.82
10          2         11         0.182          1.00
```

Graphical summaries

Data: countries population sizes (2015) in bar chart ordered w.r.t. frequency.

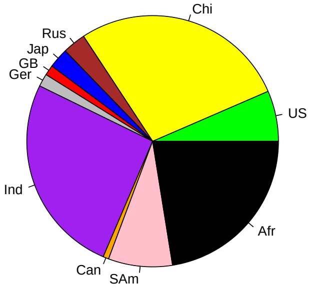
```
population=c(322,1372,147,127,65,81,1278,36,407,1111)
names(population)=c("US","Chi","Rus","Jap","GB","Ger","Ind","Can","SAm","Afr")
par(mfrow=c(1,1))
barplot(sort(population,decreasing=TRUE),main="Pareto bar chart",ylab="Pop. size (mln)",col="blue")
```



Graphical summaries

Pie chart of population sizes (2015)

```
> pie(population/sum(population),col=c("green","yellow","brown","blue","red","grey","purple","orange","pink","black")
```

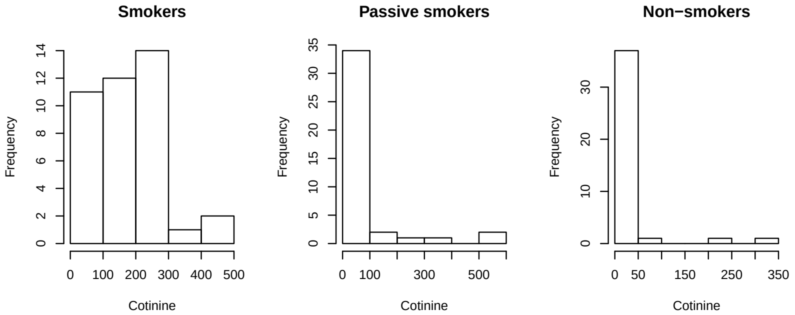


Pie piece size is proportional to the relative frequency of category (mainly: qualitative data).

Graphical summaries

Recall the data `cotinine` and its histograms.

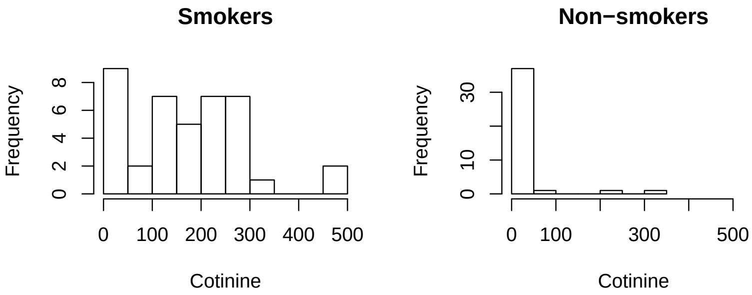
```
> par(mfrow=c(1,3)); hist(cotinine[,1],main="Smokers",xlab="Cotinine",ylab="Frequency")
> hist(cotinine[,2],main="Passive smokers",xlab="Cotinine",ylab="Frequency")
> hist(cotinine[,3],main="Non-smokers",xlab="Cotinine",ylab="Frequency")
```



Graphical summaries

Histograms depend on choices of number of cells (intervals) and bin locations.

```
> par(mfrow=c(1,2)); hist(cotinine[,1],main="Smokers",xlab="Cotinine",ylab="Frequency",breaks=8)
+ hist(cotinine[,3], main="Non-smokers",xlab="Cotinine",ylab="Frequency",xlim=c(0,max(cotinine)))
```

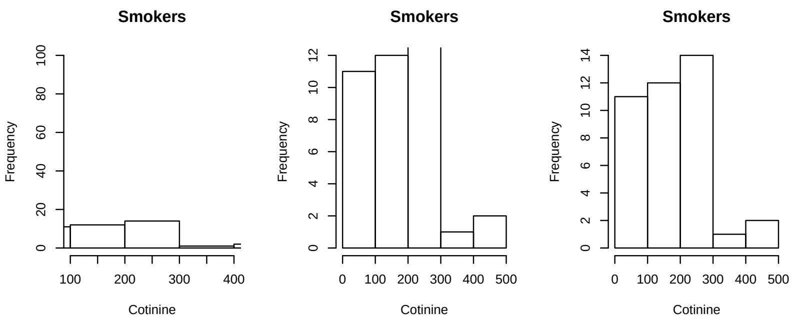


There exists the probabilistic representation of histogram (option `prob=TRUE` in the `hist`-command) constructed as follows: the areas of the bars equal to the frequencies of the measurements in the corresponding cells.

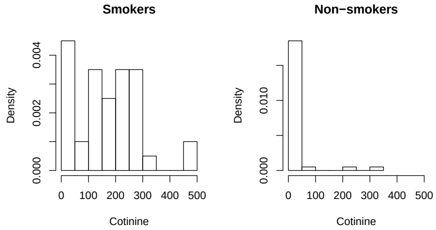
Graphical summaries

Presentation: reasonable dimensions (preferably square) and scale, appropriate title and axes labels.

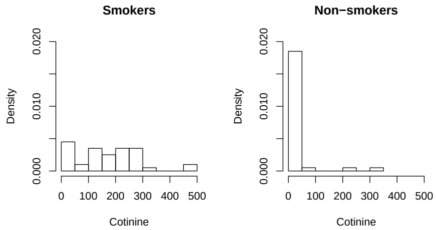
```
> par(mfrow=c(1,3))
> hist(cotinine[,1],main="Smokers",xlab="Cotinine",ylab="Frequency",xlim=c(100,400),ylim=c(0,100))
> hist(cotinine[,1],main="Smokers",xlab="Cotinine",ylab="Frequency",ylim=c(0,12))
> hist(cotinine[,1],main="Smokers",xlab="Cotinine",ylab="Frequency")
```



Graphical summaries



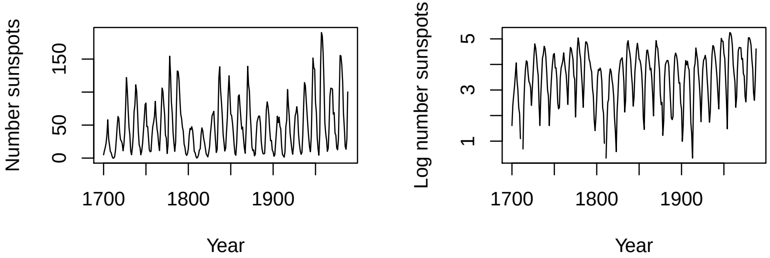
Or:



Graphical summaries

Time series plot: visualization of time-varying quantity; data: yearly number of sunspots:

```
> par(mfrow=c(1,2))
> plot(1700:1988,sunspot.year,xlab="Year",ylab="Number sunspots",type="l")
> plot(1700:1988,log(sunspot.year),xlab="Year",ylab="Log number sunspots",type="l")
```



Pay attention to scale.

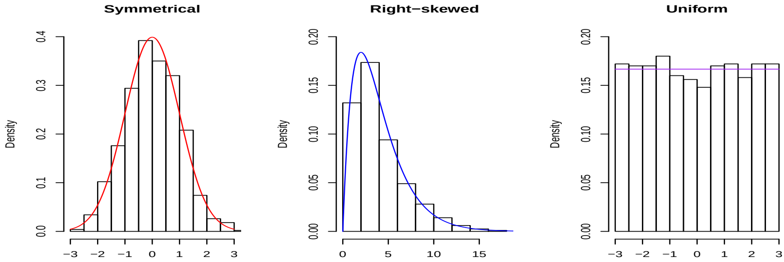
Graphical summaries

- choice of summary depends on type of data (level of measurement) and context
- appropriate figure dimensions and scale
- if comparing data sets, preferably the same scale

Describing data

Recall two ways to describe data: quantitatively (numerical summaries of location and variation) and qualitatively (shape, location, spread of data distribution).

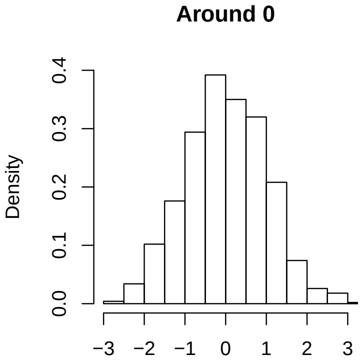
Qualitative description is for example shape: smooth approximation of histogram, relating the data distribution to familiar distributions.



Qualitative description

Location: position on x axis.

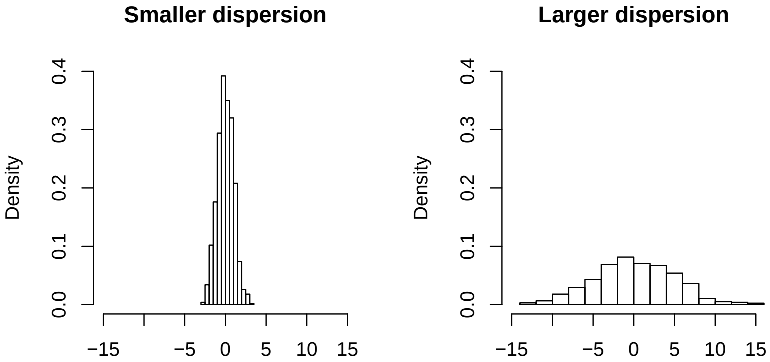
Same shape; different location:



Qualitative description

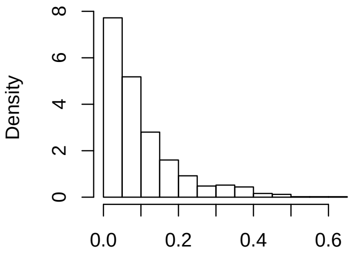
Spread (dispersion/variation): measure of variation within dataset.

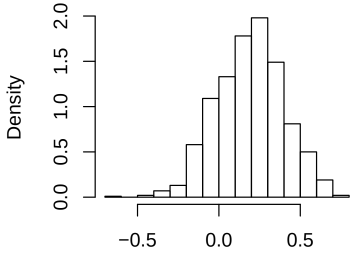
Same shape and location; different dispersion:



Qualitative description

Obviously different shape; but difference in location and/or dispersion?





Solution: numerical descriptions.

Numerical summaries

Describe data distribution with numerical values for

- location
- spread
- skewness
- ...

From now on we follow the book again in this lecture: Chapter 2.

2.2 Measures of center

Measure of center: value at center / middle of a dataset.
Different measures:

- Mean
- Median
- Mode

2.2 Measures of center

Let $(x_1, \ldots, x_n)$ be a dataset of size n.

The mean is the "average":

$$\text{mean} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + \ldots + x_n}{n}.$$

Every data value used.
Not "robust": strongly affected by extreme values.
In *R*: `mean()`

Sample mean denoted by $\bar{x} = (\sum_{i=1}^{n} x_i)/n$.

Population mean denoted by $\mu = (\sum_{i=1}^{N} x_i)/N$.

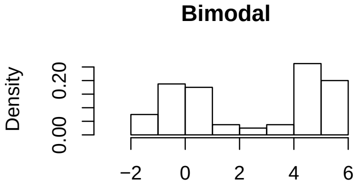
2.2 Measures of center

Median: "middle" value (after sorting).
Robust: not much affected by extreme values.
In *R*: `median()`

Mode: value with highest frequency.
Mainly for nominal data.

Dataset with unique mode: unimodal, else bimodal (2) / multimodal (>2 modes).

Graphs with different peaks are also called (multi-/)bimodal:



2.2 Measures of center

Recall the data set of exam grades.

```
>grades
[1] 10  7  6 10  8  5  8  7  5  9  7
>sort(grades) # ordered grades
 [1]  5  5  6  7  7  7  8  8  9 10 10
```

$$\text{Mean} = \frac{\sum_{i=1}^{11} x_i}{11} = 82/11 = 7.45454545\ldots \approx 7.5,$$
$$\text{Median} = 7 \quad \text{(middle in the ordered sample)},$$
$$\text{Mode} = 7 \quad \text{(most frequent)}.$$

Recall the data set `cotinine`:

```
> c(mean(cotinine[,1]),mean(cotinine[,2]),mean(cotinine[,3]))
[1] 172.47  60.58  16.35
> c(median(cotinine[,1]),median(cotinine[,2]),median(cotinine[,3]))
[1] 170.0   1.5   0.0
```

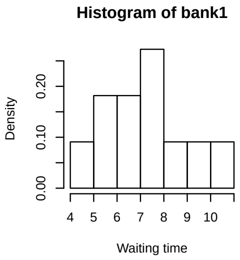
Remember that the distribution for non-smokers was skewed, hence difference in mean and median for non-smokers.

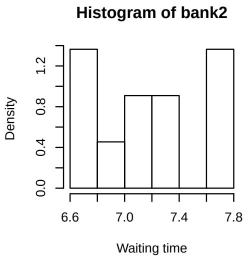
2.3 Measures of variation

Consider waiting times (min) at two banks:

```
> bank1
 [1]  4.1  5.2  5.6  6.2  6.7  7.2  7.7  7.7  8.5  9.3 11.0
> bank2
 [1] 6.6 6.7 6.7 6.9 7.1 7.2 7.3 7.4 7.7 7.8 7.8
```

For both banks: mean and median waiting time is 7.2 min.

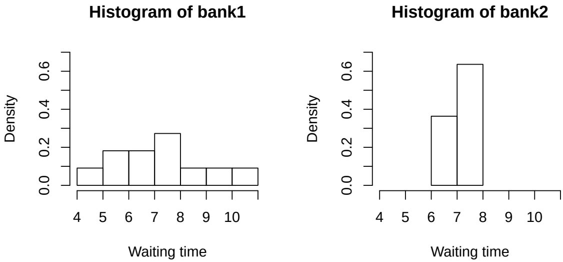




2.3 Measures of variation

Adjust scale for better comparisons:

```
> par(mfrow=c(1,2)); hist(bank1,xlab="Waiting time",prob=T,xlim=c(4,11),ylim=c(0,0.7))
> hist(bank2,xlab="Waiting time",prob=T,xlim=c(4,11),ylim=c(0,0.7),breaks=c(6,7,8))
```



Spread is smaller for bank2. How to quantify?

2.3 Measures of variation

The sample standard deviation s and the sample variance s^2 ("mean quadratic deviation from $\bar{x}$") are common measures of variation (or deviation from $\bar{x}$):

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

In R these are computed by the command `sd()` and `var()` respectively.

Do not confuse s and s^2 with population standard deviation σ and the population variance σ^2.

Let us compute s and s^2 for the both banks (customer preference?):

```
> c(sd(bank1),sd(bank2))
[1] 1.961 0.445
> c(var(bank1),var(bank2))
[1] 3.846 0.198
```

Another measure of variation: range = maximum - minimum.
Uses only two values $\Rightarrow$ very sensitive to extreme values / outliers.

2.4 Measures of relative standing and boxplots

Alternative measures of location **and** spread are percentiles.
Percentile P_i: $i\%$ of data values $< P_i$ **and** $(100 - i)\%$ of values $\geq P_i$.

Special percentiles: quartiles Q_1, Q_2 and Q_3.
Divide data set into four groups of $\approx 25\%$ of data values each.

- $Q_1 = P_{25}$
- $Q_2 = P_{50} =$ median
- $Q_3 = P_{75}$

In R: quartiles (and extrema) are found via `summary()` or `quantile()`:

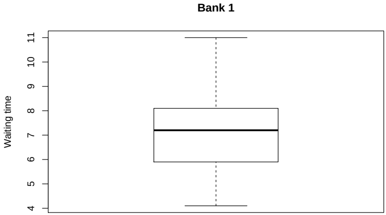
```
> summary(bank1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    4.1     5.9     7.2     7.2     8.1    11.0
> quantile(bank2)
  0%  25%  50%  75% 100%
6.60 6.80 7.20 7.55 7.80
```

In the output of summary-command we see 5 numbers:

1. Minimum
2. First quartile, Q_1
3. Median, Q_2
4. Third quartile, Q_3
5. Maximum

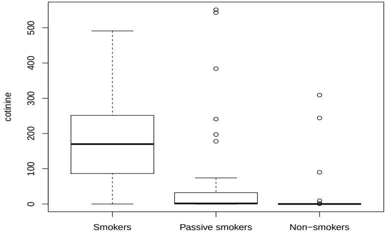
2.4 Measures of relative standing and boxplots

Useful graphical tool for presenting data: boxplot, in R we use `boxplot()`-command. Boxplots provide information about distribution: median = box's center? Are there outliers? Is distribution "asymmetrical"?



Output of `boxplot(bank1,main="Bank 1")`.

Top horizontal line: maximum,
Top of the box: Q_3,
Thick line: median,
Bottom of the box: Q_1,
Lowest horizontal line: minimum.



Output of `boxplot(cotinine)`.

Whiskers are the lines extending from the box. By default they end at values not exceeding $1.5 \times IQR$, where $IQR = Q_3 - Q_1$ is the interquartile range.
Outliers are all points not included between whiskers.

Summary

- If possible, first: make figures for own use (get impression of data)
- Summaries: graphical and/or numerical
- Choice of summary depends on data type and context
- Graphical summaries: choose right size and briefly comment: relevant aspects
- Numerical summaries: choose right measure of location and variation;
  briefly comment: what do the numbers reveal.