

Descriptive Statistics

Arnaud Legrand and Jean-Marc Vincent

RICM4 Probabilities and Simulations
Grenoble, France, October 2015

① Descriptive statistics of an univariate sample

- Motivation

- Initial step

- Histograms of "Stable" samples

- Single mode: central tendency

- Dispersion: Variability around the central tendency

- Going further

- Summarizing a distribution

① Descriptive statistics of an univariate sample

- Motivation

- Initial step

- Histograms of "Stable" samples

- Single mode: central tendency

- Dispersion: Variability around the central tendency

- Going further

- Summarizing a distribution

Motivation

We have set up a world where we keep collecting data, **huge amount of data**...

Sweet, what knowledge can we extract from such data? How do we summarize a data set?

With a few numbers, some graphics? How? Why is this difficult?

There are three kinds of lies: lies, damned lies and statistics

– Mark Twain's Autobiography

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read or write

– Attributed to H. G. Wells

The only statistics you can trust are those you falsified yourself

– Winston Churchill

① Descriptive statistics of an univariate sample

- Motivation

- Initial step**

- Histograms of "Stable" samples

- Single mode: central tendency

- Dispersion: Variability around the central tendency

- Going further

- Summarizing a distribution

I just got new Tees!

- A series of **measurements** (one value per measurement)
- **Nature** of the measurements
 - Factors (**nominal data**)

```
1 [1] Red   Red   Black Green Blue  Black White Black Blue
2 [10] White Black White Red   Black Black Red   Red   Black
3 [19] Black Black
4 Levels: Black Blue Green Red White
```

- Ordered factors (**ordinal data**)

```
1 [1] XL M  S  XL M  M  M  XL M  L  M  L  M  M  M  L  M
2 [18] M  XL M
3 Levels: S < M < L < XL
```

- Numbers (e.g., price, duration, ...) (**numerical data**)

```
1 [1] 9.1 4.7 9.5 13.6 15.7 8.7 9.2 4.7 11.4 8.1
2 [11] 11.4 12.1 13.1 8.2 11.5 4.8 7.6 7.4 2.8 10.1
```

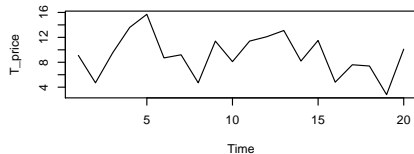
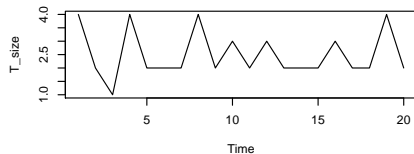
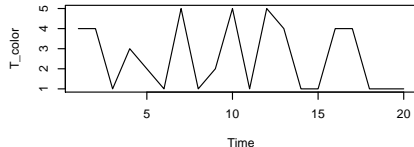
```
1 str(T_size); # May want to use the str function
```

```
1 Ord.factor w/ 4 levels "S"<"M"<"L"<"XL": 4 2 1 4 2 2 2 4 2 3 37.
```

Are these sample "structured"?

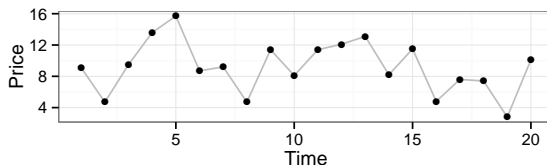
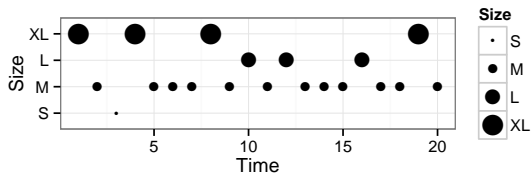
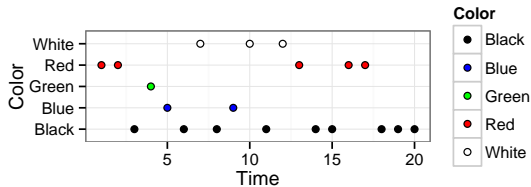
Use `plot.ts` (for **time series**)

```
1 par(mfrow=c(3,1));  
2 plot.ts(T_color,xy.lines=F);  
3 plot.ts(T_size,xy.lines=F);  
4 plot.ts(T_price,xy.lines=F);
```

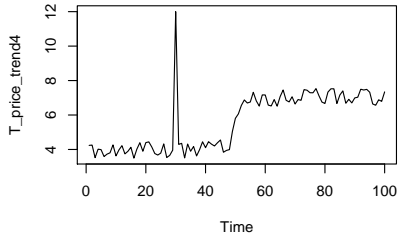
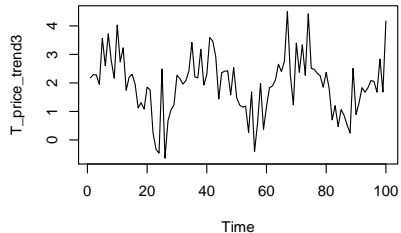
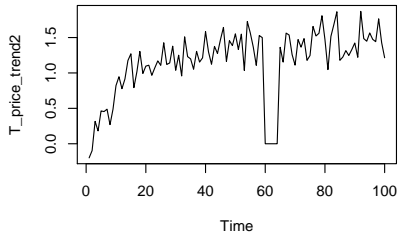
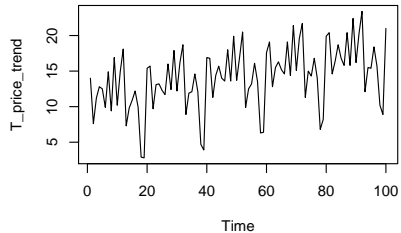


Are these sample "structured"?

Fancier output can be built using ggplot2



There could indeed be "trends"



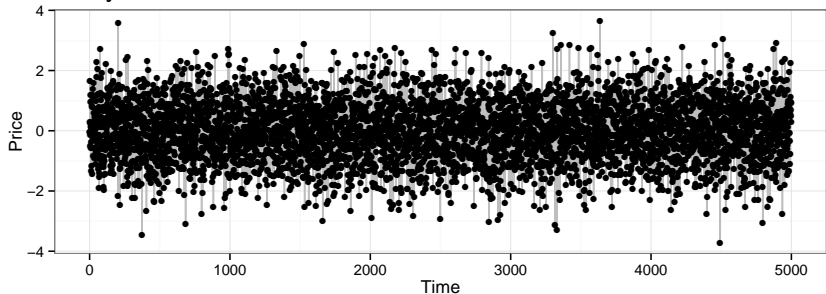
What should we look for?

- Structured/unstructured
- Trend, evolution
- Localization/order of magnitude
- Outliers, aberrant values

This preliminary study will:

- guide your analysis
- provide feedback on your experimental setup

This may be harder to do than it looks. . .



① Descriptive statistics of an univariate sample

- Motivation

- Initial step

- Histograms of "Stable" samples**

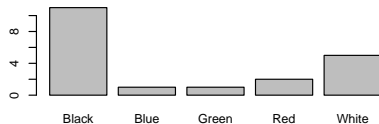
- Single mode: central tendency

- Dispersion: Variability around the central tendency

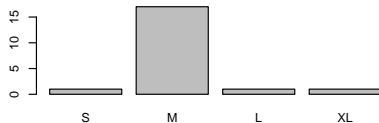
- Going further

- Summarizing a distribution

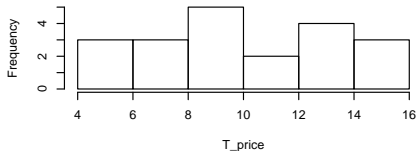
Bar charts vs. Histograms



```
1 par(mfrow=c(3,1));  
2 plot(T_color,xy.lines=F);  
3 plot(T_size,xy.lines=F);  
4 hist(T_price,xy.lines=F);
```

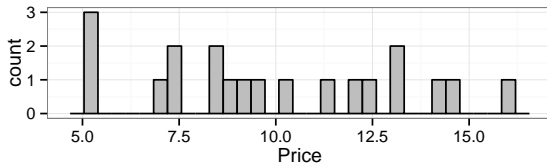
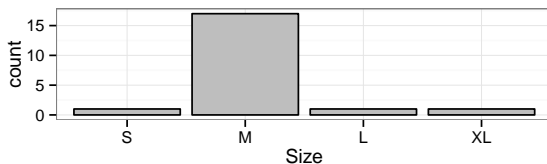
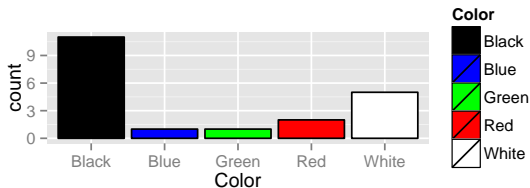


Histogram of T_price



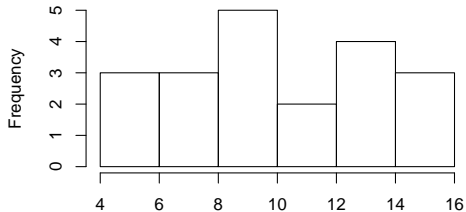
Bar charts vs. Histograms

Again, fancier output can be built using ggplot2

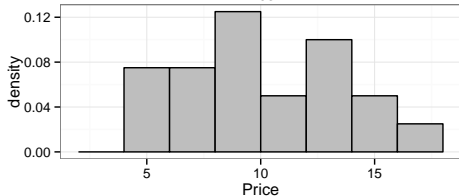
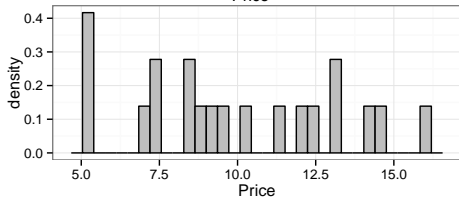
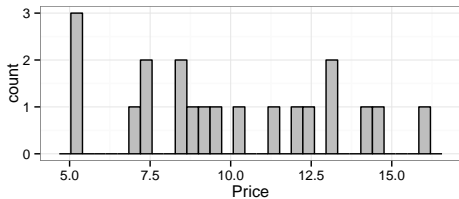
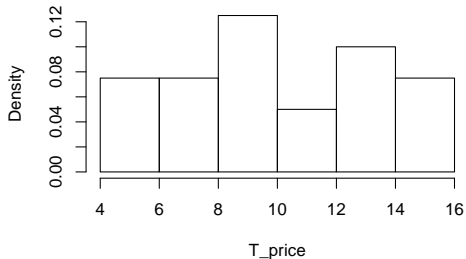


Wait, why are these histograms so different?

Histogram of T_price



Histogram of T_price



Rather indicate density than count

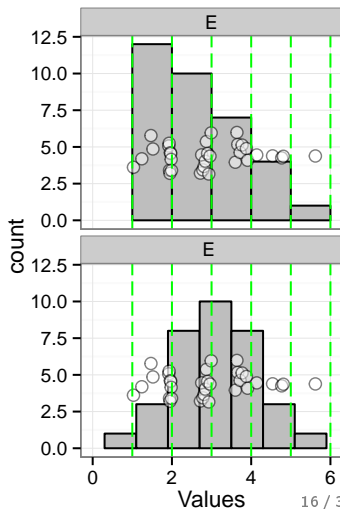
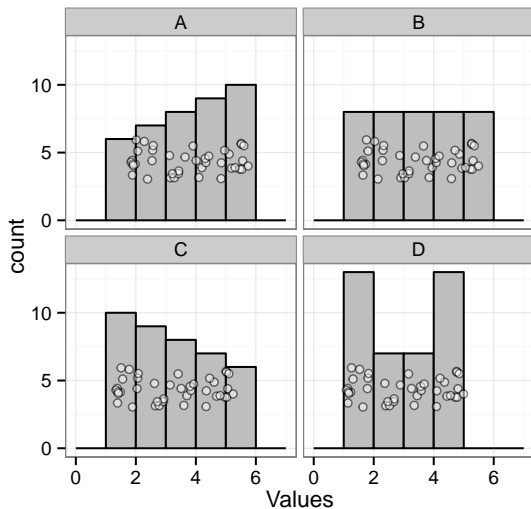
How many bins? Which binwidth?

- ggplot defaults to $k = 30$ bins of width $h = \text{range}/30$ 😞
- Square-root choice: $k = \sqrt{n}$ (Excel, 😞)
- Sturges: $k = \lceil \log_2 n + 1 \rceil$ (default for hist in R)
- Rice: $k = \lceil 2n^{1/3} \rceil$
- Scott: $k = \lceil \frac{\max x - \min x}{h} \rceil$, where: $h = \frac{3.5\hat{\sigma}}{n^{1/3}}$ (equivalent to Rice under some conditions)
- ...

Beware of Histograms

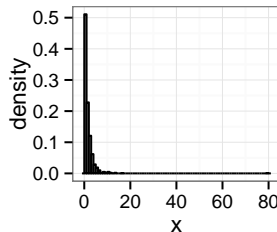
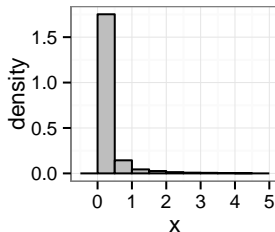
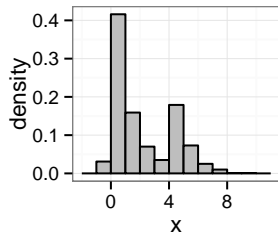
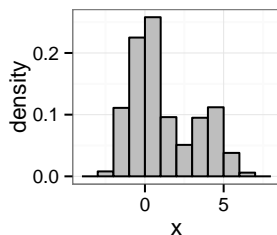
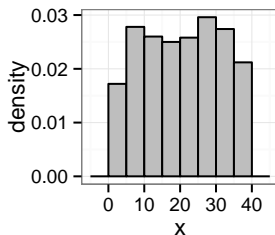
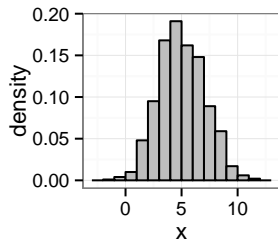
At which value should the bin start?

- In most cases, the binning is aligned on human readable values, which can create nasty artifacts (nice illustration from *stackexchange*)



What should we look for?

Shape: flat? symmetrical? multi-modal? Play with binwidth (and origin if you have few samples) to uncover the full story behind your data...



① Descriptive statistics of an univariate sample

- Motivation

- Initial step

- Histograms of "Stable" samples

- Single mode: central tendency

- Dispersion: Variability around the central tendency

- Going further

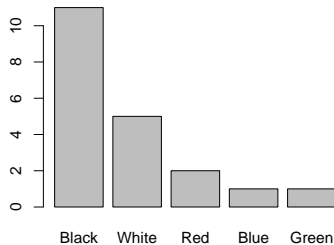
- Summarizing a distribution

Nominal Values

- What is the **mode** (most frequent value)?
- Sort values according to their frequency...

```
1 summary(T_color)
```

```
1 Black  Blue  Green   Red  White  
2     11     1     1     2     5
```



```
1 col_freq=table(T_color);  
2 T_color <- factor(T_color,  
3   levels = names(col_freq[order(col_freq, decreasing = TRUE)]));  
4 plot(T_color);
```

Ordinal Values

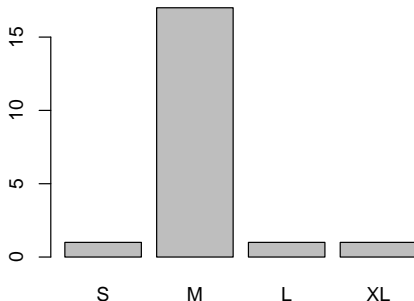
- What is the **mode** (most frequent value)?

```
1 summary(T_size)
```

```
1 S M L XL
```

```
2 1 17 1 1
```

- May still want to sort values according to their frequency...
- Median**: not implemented in standard R for ordinal values, as it's not well defined



```
1 median(T_size)
2 library(DescTools)
3 median(T_size) # :(
```

```
1 Error in median.default(T_size) : requires numerical data
2 [1] NA
```

Numerical Values

```
1 str(T_price);
```

```
1 num [1:20] 14.5 13.1 9.3 6.9 8.6 7.2 7.3 12.4 13.1 16 ...
```

```
1 summary(T_price);
```

```
1   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2 5.200   7.275   9.500   9.960  12.580  16.000
```

- min, max, median in R
- Median: 50% of values are smaller than 9.5
(a possible measure of **central tendency**)

Numerical Values

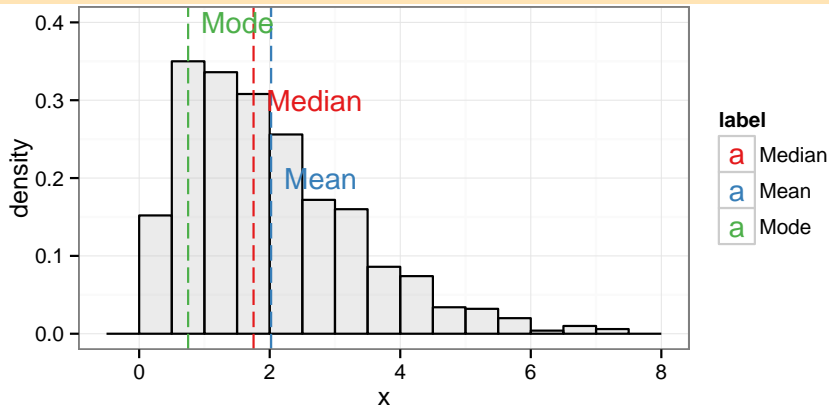
The **mode** and the **median** are measures of **central tendency** (typical value)

- **Note:** There may be several modes and it depends on binning...

There is also the (arithmetic) **mean**: $A = \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$

```
mean(T_price)
```

```
[1] 9.96
```



Things to know about the mean

- This measure is sensitive to "outliers".
 - One aberrant (say very large) value will drag the mean to the right while it would not change the median
- The key question is **what makes sense?**
 - Your favorite pair has been added a +20% mark-up in August but you have a -20% discount as a regular customer. Is the price the same?
 - No, you actually saved 4% of the original price ($1.2 \times .8 = .96$).
 - You drove half the way at 50mph and half of the way at 100mph. Did you drive on average at 75mph?
 - Obviously not...
 - Although you can compute the average of gains/loss, it is not at all what you would consider as the average gain.
 - May want to consider the geometric or the harmonic mean...

$$G = \sqrt[n]{\prod_{i=1}^N x_i} \text{ or } H = \frac{1}{\frac{1}{N} \sum_{i=1}^N \frac{1}{x_i}}$$

What should I look for?

- If the distribution is unimodal and symmetrical, then
 $\text{mean} = \text{mode} = \text{median}$
- Depending on the problem, one or the other may be more relevant
- Anyway, reporting such measure with no indication about variability is generally useless

① Descriptive statistics of an univariate sample

- Motivation

- Initial step

- Histograms of "Stable" samples

- Single mode: central tendency

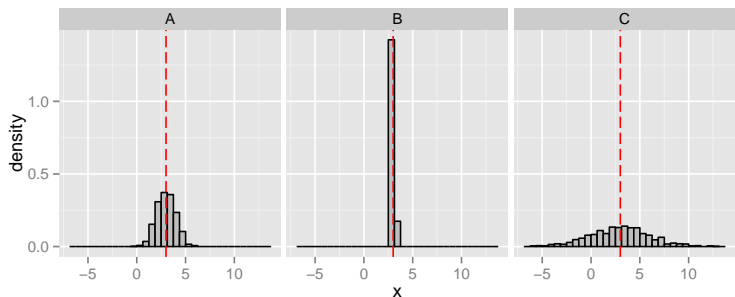
- Dispersion: Variability around the central tendency**

- Going further

- Summarizing a distribution

Variance

We expect most values to be "around" the mean



Departure from the mean:

- Mean absolute deviation: $\frac{1}{N} \sum_{i=1}^N |x_i - A|$
 - Rarely used
- **Variance:** $V = \frac{1}{N} \sum_{i=1}^N (x_i - A)^2$
 - only positive values and gives more importance to large deviations 😊
 - not homogeneous to the mean (units) 😞
- **Standard deviation:** $SD = \sqrt{V}$

Quantile

```
1 quantile(T_price,c(.05,.25,.5,.75,.95))
```

```
1      5%      25%      50%      75%      95%
2 4.605  7.550  9.150 11.425 13.705
```

Inter-Quantile Range:

- **Inter-quartile range:** $IQR = Q_{75} - Q_{25}$
- But other values are possible, e.g., $Q_{95} - Q_5$
- **Range:** $\max - \min$ (may grow unbounded)
 - \leadsto quite difficult to use

What about nominal or ordinal values?

There is for example the notion of **Entropy**: how many bits are required to encode the sample?

Say there is a fraction f_v of items with value v .

$$H = - \sum_{v \in V} f_v \log_2(f_v)$$

$-(x+y)\log_2(x+y) < -x\log_2(x) - y\log_2(y)$ so the smaller the entropy, the more condensed/predictable the sample distribution

- $H([0, 1, 0, 0]) = 0$
- $H([.25, .25, .25, .25]) = 2$
- $H([1/n, \dots, 1/n]) = \log_2(n)$ so you generally normalize H by $\log_2(n)$

This notion can be extended to numerical values (but the computation is complex as it depends on the binning...)

① Descriptive statistics of an univariate sample

- Motivation

- Initial step

- Histograms of "Stable" samples

- Single mode: central tendency

- Dispersion: Variability around the central tendency

- Going further**

- Summarizing a distribution

Skewness

Remember the **mean** and the **variance**:

- $A = \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$
- $V = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$

Could we measure the asymmetry of the samples around the mean?

- Proposal 1: $\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})$ (always 0... 😞)
- Proposal 2: $\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^3$ (not well normalized... 😞)

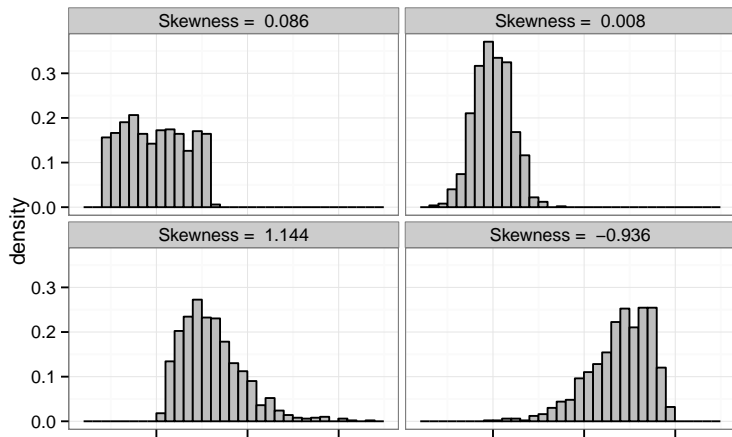
$$S = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left[\underbrace{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}_{\text{variance}} \right]^{3/2}}$$

Skewness

Could we illustrate this a bit?

```
1 library(moments)
2 skewness(runif(1000))
```

```
1 [1] 0.04626483
```



Kurtosis

- peakedness (width of peak), tail weight, lack of shoulders...
- measure infrequent extreme deviations, as opposed to frequent modestly sized deviations

$$K = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\underbrace{\left[\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^2}_{\text{variance}}} - 3$$

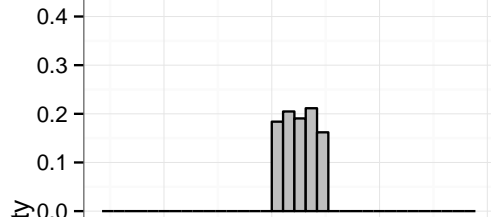
The **-3** is here so that normal distribution have a Kurtosis of 0

```
1 library(moments)
2 x = rnorm(1000) ; var(x);
3 kurtosis(x)-3
```

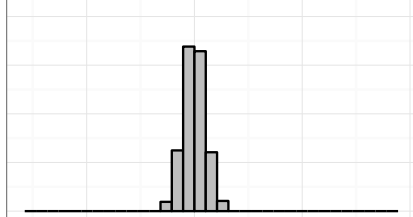
```
1 [1] 1.039743
2 [1] 0.01825114
```


Kurtosis

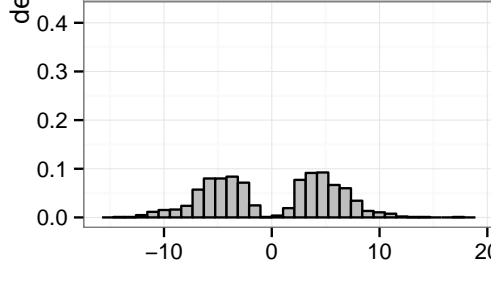
Kurtosis = -1.24



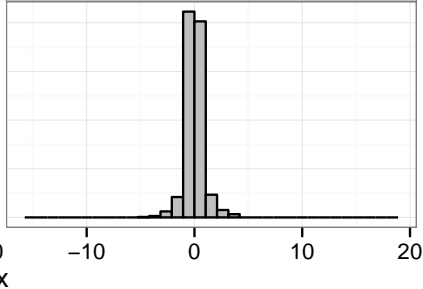
Kurtosis = -0.131



Kurtosis = -1.108



Kurtosis = 6.184



① Descriptive statistics of an univariate sample

- Motivation

- Initial step

- Histograms of "Stable" samples

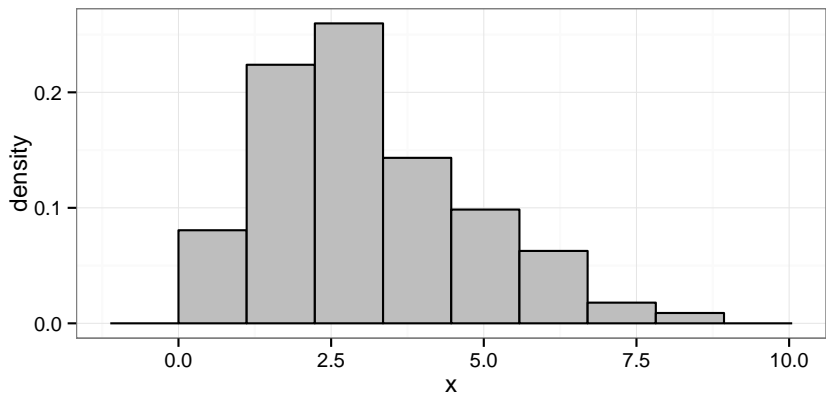
- Single mode: central tendency

- Dispersion: Variability around the central tendency

- Going further

- Summarizing a distribution

Classical information

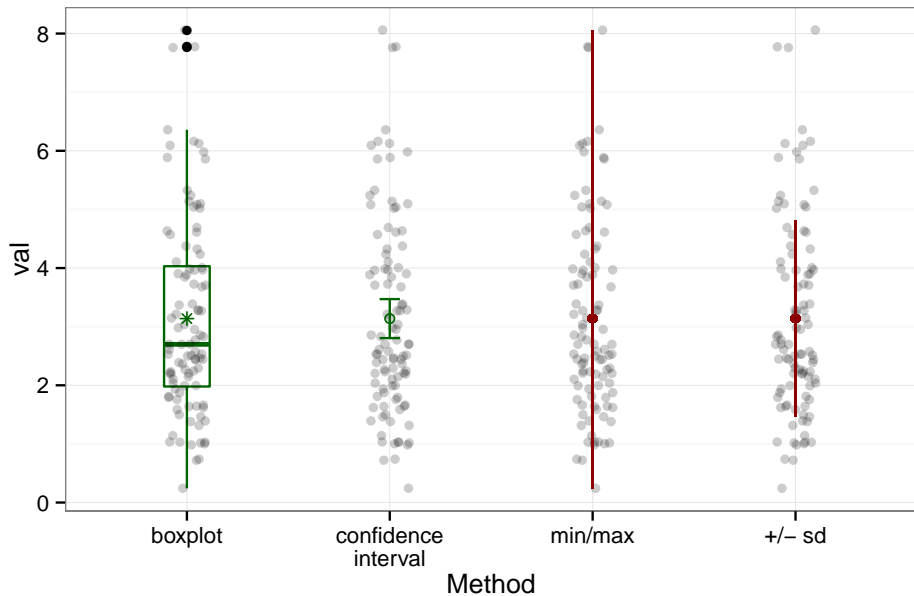


```
1 summary(x)
```

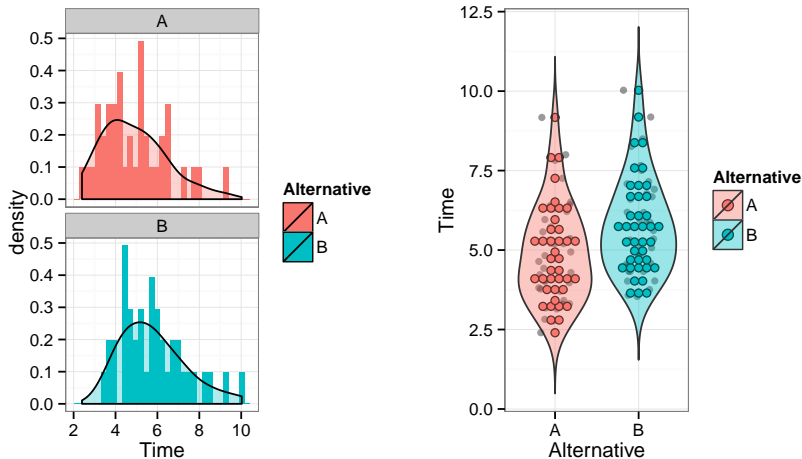
```
2 var(x)
```

```
1      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2  0.4065  1.8430   2.5020   2.8660  3.6310   7.0220
3 [1] 2.117541
```

Good and bad summaries



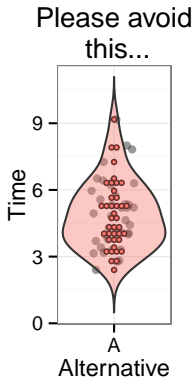
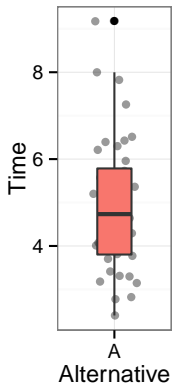
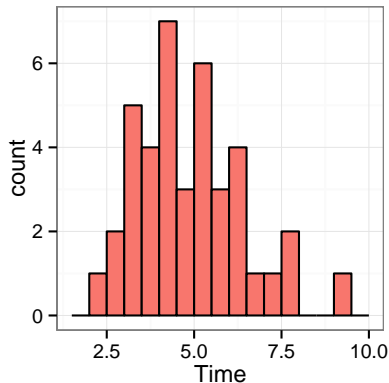
Be careful with fancy plots you do not fully understand!



The average human has one breast and one testicle

– Des McHale

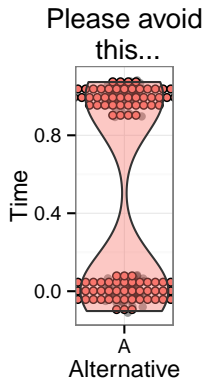
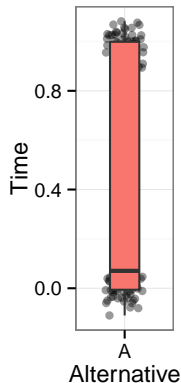
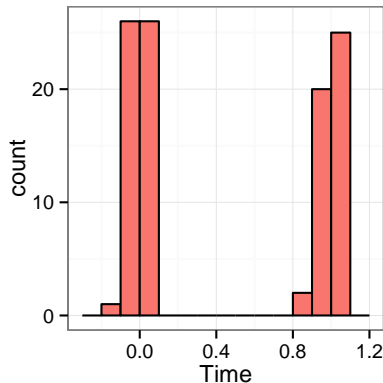
Be careful with fancy plots you do not fully understand!



The average human has one breast and one testicle

– Des McHale

Be careful with fancy plots you do not fully understand!



The average human has one breast and one testicle

– Des McHale