Descriptive statistics

Rapport package team @ https://github.com/aL3xa/rapport

2011-04-26 20:25 CET

## Description

This template will return descriptive statistics of a numerical, or a frequency table of a categorical variable.

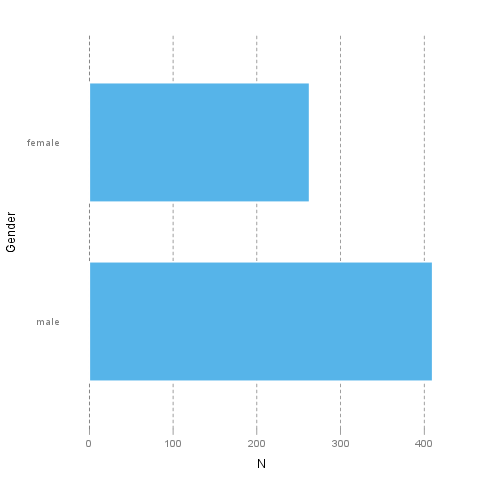
## *gender* ("Gender")

The dataset has *709* observations with *673* valid values (missing: *36*) in *gender* ("Gender"), which seems to be a qualitative variable.

### Base statistics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **gender** | **N** | **%** | **Cumul. N** | **Cumul. %** |
| male | 410 | 60.9212 | 410 | 60.9212 |
| female | 263 | 39.0788 | 673 | 100 |
| Total | 673 | 100 | 673 | 100 |

### Barplot

[](/tmp/RtmpI5pYwT/file542b6510-hires.png)

It seems that the highest value is *2* which is exactly 2 times higher than the smallest value (*1*).

The most frequent value is *male*.

## Description

This template will return descriptive statistics of a numerical, or a frequency table of a categorical variable.

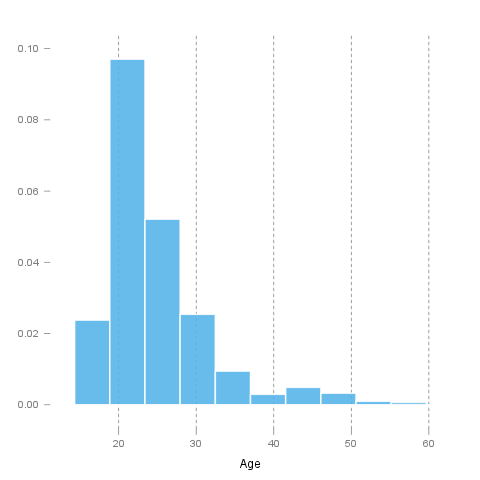
## *age* ("Age")

The dataset has *709* observations with *677* valid values (missing: *32*) in *age* ("Age"), which seems to be a quantitative variable.

### Base statistics

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **mean** | **sd** | **var** |
| Age | 24.5731 | 6.8491 | 46.9107 |

### Histogram

[](/tmp/RtmpI5pYwT/file62aaa0fa-hires.png)

It seems that the highest value is *58* which is exactly 3.625 times higher than the smallest value (*16*).

The standard deviation is *6.8491* (variance: *46.9107*). The expected value is around *24.5731*, somewhere between *24.0572* and *25.0891* with the standard error of *0.2632*.

If we suppose that *Age* is not near to a normal distribution (test: see below, skewness: *1.9254*, kurtosis: *4.463*), checking the median (*23*) might be a better option instead of the mean. The interquartile range (*6*) measures the statistics dispersion of the variable (similar to standard deviation) based on median.

### Normality tests

#### Introduction

In statistics, *normality* refers to an assumption that the distribution of a random variable follows *normal* (*Gaussian*) distribution. Because of its bell-like shape, it's also known as the *"bell curve"*. The formula for *normal distribution* is:

*Normal distribution* belongs to a *location-scale family* of distributions, as it's defined two parameters:

* *μ* - *mean* or *expectation* (location parameter)
* *σ2* - *variance* (scale parameter)

[](/tmp/RtmpI5pYwT/file5f342fe8-hires.png)

#### Normality Tests

##### Overview

Various hypothesis tests can be applied in order to test if the distribution of given random variable violates normality assumption. These procedures test the H0 that provided variable's distribution is *normal*. At this point only few such tests will be covered: the ones that are available in stats package (which comes bundled with default R installation) and nortest package that is [available](http://cran.r-project.org/web/packages/nortest/index.html) on CRAN.

* **Shapiro-Wilk test** is a powerful normality test appropriate for small samples. In R, it's implemented in shapiro.test function available in stats package.
* **Lilliefors test** is a modification of *Kolmogorov-Smirnov test* appropriate for testing normality when parameters or normal distribution (*μ*, *σ2*) are not known. lillie.test function is located in nortest package.
* **Anderson-Darling test** is one of the most powerful normality tests as it will detect the most of departures from normality. You can find ad.test function in nortest package.
* **Pearson Χ2 test** is another normality test which takes more "traditional" approach in normality testing. pearson.test is located in nortest package.

##### Results

Here you can see the results of applied normality tests (*p-values* less than 0.05 indicate significant discrepancies):

|  |  |  |
| --- | --- | --- |
|  | **Statistic** | **p-value** |
| Shapiro-Wilk normality test | 0.8216 | 0 |
| Lilliefors (Kolmogorov-Smirnov) normality test | 0.17 | 0 |
| Anderson-Darling normality test | 32.1645 | 0 |
| Pearson chi-square normality test | 625.8479 | 0 |

So, let's draw some conclusions based on applied normality test:

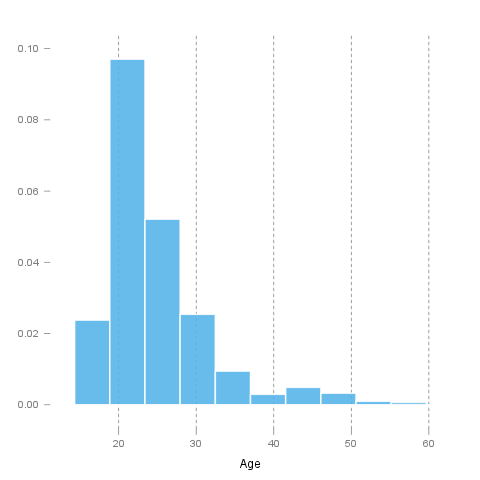
* according to *Shapiro-Wilk test*, the distribution of *Age* is not normal.
* based on *Lilliefors test*, distribution of *Age* is not normal
* *Anderson-Darling test* confirms violation of normality assumption
* *Pearson's Χ2 test* classifies the underlying distribution as non-normal

#### Diagnostic Plots

There are various plots that can help you decide about the normality of the distribution. Only a few most commonly used plots will be shown: *histogram*, *Q-Q plot* and *kernel density plot*.

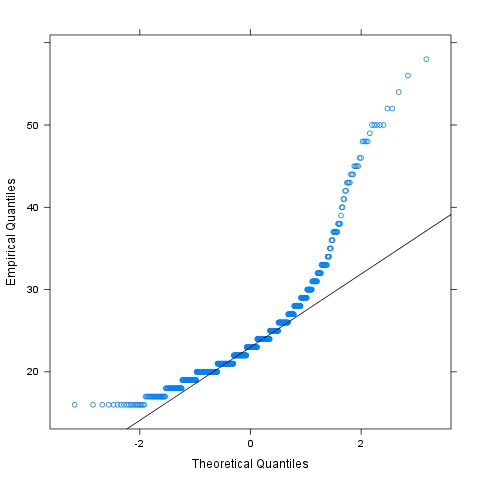
##### Histogram

*Histogram* was first introduced by *Karl Pearson* and it's probably the most popular plot for depicting the probability distribution of a random variable. However, the decision depends on number of bins, so it can sometimes be misleading. If the variable distribution is normal, bins should resemble the "bell-like" shape.

[](/tmp/RtmpI5pYwT/file2848d313-hires.png)

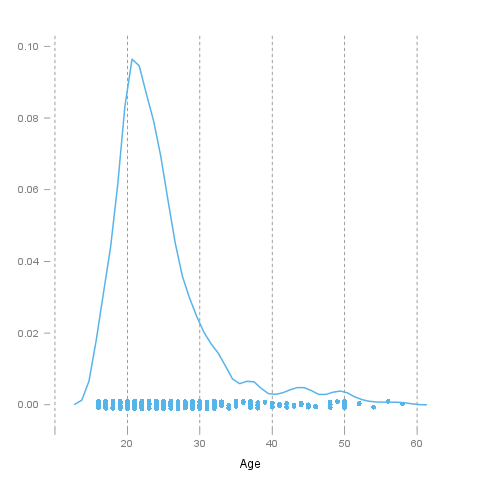
##### Q-Q Plot

"Q" in *Q-Q plot* stands for *quantile*, as this plot compares empirical and theoretical distribution (in this case, *normal* distribution) by plotting their quantiles against each other. For normal distribution, plotted dots should approximate a "straight", x = y line.

[](/tmp/RtmpI5pYwT/file6a81060f-hires.png)

##### Kernel Density Plot

*Kernel density plot* is a plot of smoothed *empirical distribution function*. As such, it provides good insight about the shape of the distribution. For normal distributions, it should resemble the well known "bell shape".

[](/tmp/RtmpI5pYwT/file80d8d57-hires.png)

This report was generated with [R](http://www.r-project.org/) (2.14.0) and [rapport](http://al3xa.github.com/rapport/) (0.2) in 1.816 sec on x86\_64-unknown-linux-gnu platform.

