Statistical inference (MVE155/MSG200)

Summarising data

Summary

- Estimating distribution function
- Quantiles and quantile-quantile plots
- Estimating density function
- Skewness and kurtosis
- Measures of dispersion
- Boxplot

Empirical distribution function

Distribution function F of a random variable X is defined as

$$F(x) = \mathrm{P}(X \leq x) = \begin{cases} \int\limits_{-\infty}^{x} f(y)\, dy \\ \sum\limits_{y \leq x} \mathrm{P}(X = y) \end{cases}$$

The empirical distribution function based on the sample $(x_1, ..., x_n)$ is

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_i \leq x),$$

where $\mathbb{I}$ is an indicator function, is an unbiased and consistent estimator for $F(x)$.

Empirical distribution function

As a function of x, $\hat{F}(x)$ is a distribution function for a uniform random variable Y with

$$\mathrm{P}(Y = x_i) = \frac{1}{n}, \quad i = 1, \dots, n$$

if $x_i \neq x_j$ for all $i \neq j$. We have that (even when some of the x_i's coinside)

$$\mathbb{E}(Y) = \sum_{i=1}^{n} x_i \, \mathrm{P}(Y = x_i) = \sum_{i=1}^{n} \frac{x_i}{n} = \bar{x}$$

and

$$\begin{aligned} \mathrm{Var}(Y) &= \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2 \\ &= \sum_{i=1}^{n} x_i^2 \, \mathrm{P}(Y = x_i) - \left(\sum_{i=1}^{n} x_i \, \mathrm{P}(Y = x_i) \right)^2 \\ &= \overline{x^2} - (\bar{x})^2 = \frac{n-1}{n} s^2 \end{aligned}$$

and $\frac{n-1}{n} s^2$ is called empirical variance.

Survival function

When studying life length L, survival function

$$S(x) = \mathrm{P}(L > x) = 1 - \mathrm{P}(L \leq x) = 1 - F(x), \quad x \geq 0$$

and its empirical counterpart

$$\hat{S}(x) = 1 - \hat{F}(x)$$

are often used since it gives you the probability of living longer than a certain age x.

Hazard function

Mortality rate at age x can be defined as $f(x)/S(x)$ since

$$\frac{\mathrm{P}(x < L \leq x + \delta | L \geq x)}{\delta} \to \frac{f(x)}{S(x)}$$

as $\delta \to 0$. This is called a hazard rate, i.e.

$$h(x) = \frac{f(x)}{S(x)} = \frac{f(x)}{1 - F(x)}.$$

For the exponential distribution, the hazard rate is constant.

Hazard rate is also the negative slope of the log survival function since

$$-\frac{d}{dx} \ln(S(x)) = -\frac{S'(x)}{S(x)} = \frac{f(x)}{S(x)}.$$

Quantiles

The inverse of the distribution function F, F^{-1}, is called a quantile function

$$Q(p) = F^{-1}(p), \quad 0 < p < 1$$

and the p quantile is defined as

$$x_p = Q(p).$$

For an ordered sample $(x_{(1)}, ..., x_{(n)})$, where $x_{(1)} = \min\{x_1, ..., x_n\}$ and $x_{(n)} = \max\{x_1, ..., x_n\}$,

$$\hat{F}(x_{(k)}) = \frac{k}{n} \quad \text{and} \quad \hat{F}(x_{(k)} - \epsilon) = \frac{k-1}{n}$$

(where $\epsilon > 0$ small). Therefore, $x_{(k)}$ is called the empirical $\frac{k-0.5}{n}$-quantile.

Most commonly used quantiles

- median $m = x_{0.5} = Q(0.5)$
- lower quartile $x_{0.25} = Q(0.25)$
- upper quartile $x_{0.75} = Q(0.75)$

QQ-plot

We have two independent equal sized samples, $(x_1, \ldots, x_n)$ with the distribution function F_1 and quantile function Q_1 and $(y_1, \ldots, y_n)$ with the distribution function F_2 and quantile function Q_2. We want to test whether these two samples come from the same distribution, i.e. we test

$$H_0 : F_1 \equiv F_2$$

or equivalently,

$$H_0 : Q_1 \equiv Q_2.$$

Graphically, this can be done by a QQ plot which is a scatter plot with coordinates $(x_{(k)}, y_{(k)})$, $k = 1, \ldots, n$.

QQ-plot

If the QQ plot is approximately a 45° line, the distributions are approximatively the same (and H_0 is not rejected).

If the QQ plot is a straight line with some other angle, X and Y have a linear relationship

$$Y = a + bX$$

meaning that

$$F_1(x) = F_2(a + bx)$$

and

$$Q_2(p) = a + bQ_1(p).$$

Normal QQ-plot

To test whether the data can be considered being normally distributed, we define

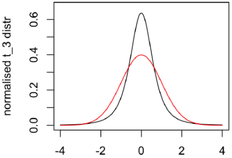
$$y_k = \Phi^{-1}\left(\frac{k - 0.5}{n}\right), \quad k = 1, ..., n,$$

where Φ^{-1} is the quantile function for the standard normal distribution $N(0, 1)$. Then, plot

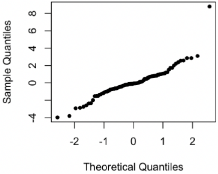
$$(x_{(1)}, y_{(1)}), ..., (x_{(n)}, y_{(n)})$$

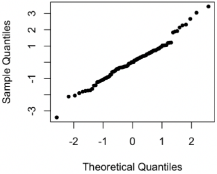
and check whether the QQ-plot is a straight line.

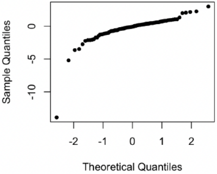
Normal QQ-plot: heavy tails



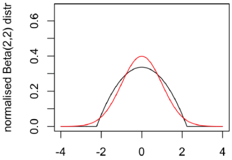
Red curve: $N(0,1)$
Black curve: distribution with heavy tails



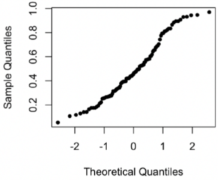


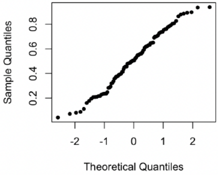


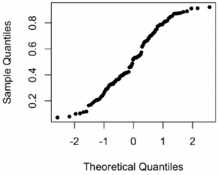
Normal QQ-plot: light tails



Red curve: $N(0,1)$
Black curve: distribution with light tails







Median

Median is the 0.5 quantile, $m = x_{0.5} = Q(0.5)$, and

$$\mathrm{P}(X < m) = \mathrm{P}(X > m) = 0.5.$$

It can be estimated from the ordered sample $x_{(1)}, \ldots, x_{(n)}$ by

$$\hat{m} = \begin{cases} x_{(k)} \quad \textit{if} \quad n = 2k-1 \quad \text{(odd)} \\ \\ \frac{1}{2}(x_{(k)} + x_{(k+1)}) \qquad \textit{if} \quad n = 2k \quad \text{(even)} \end{cases}$$

Confidence interval for median

Let us have a sample $(x_1, ..., x_n)$ (without ties) from a continuous population distribution and let

$$Y = \sum_{i=1}^{n} \mathbb{I}(x_i \leq m).$$

Then, $Y \sim \text{Bin}(n, 0.5)$ and a $100(1-2p_k)\%$ confidence interval for m becomes

$$I_m = (x_{(k)}, x_{(n-k+1)}),$$

where

$$p_k = \mathrm{P}(Y < k).$$

Y is also used as a test statistic in the sign test, which can be used instead of the one sample t-test when the sample size is small and the data not normal. Reject the $H_0 : m = m_0$ in the favour of $H_1 : m \neq m_0$ if the value of Y is not within the confidence interval.

Density estimation

A density function can be estimated by using the histograms of the observed counts

$$c_j = \sum_{i=1}^{n} \mathbb{I}(x_i \in \text{cell}_j),$$

where the observation interval is divided into adjacent cells of width h. The density function can be estimated by the scaled histogram

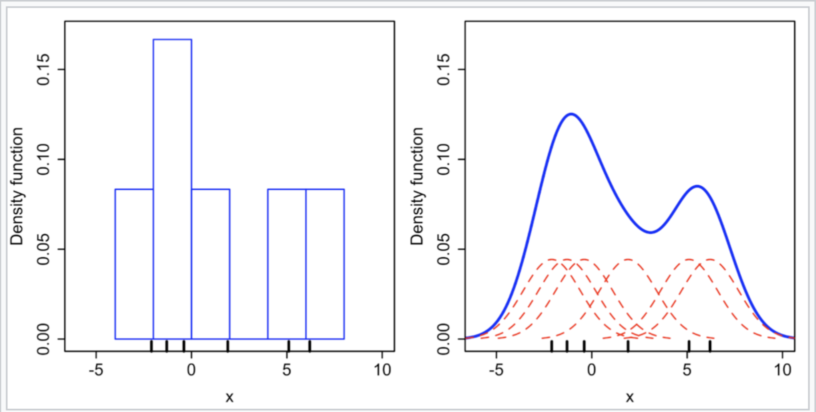
$$f_h(x) = \frac{c_j}{nh} \quad \text{for} \quad x \in \text{cell}_j.$$

Often, the scaled histogram is smoothened resulting in a kernel estimate with bandwidth h:

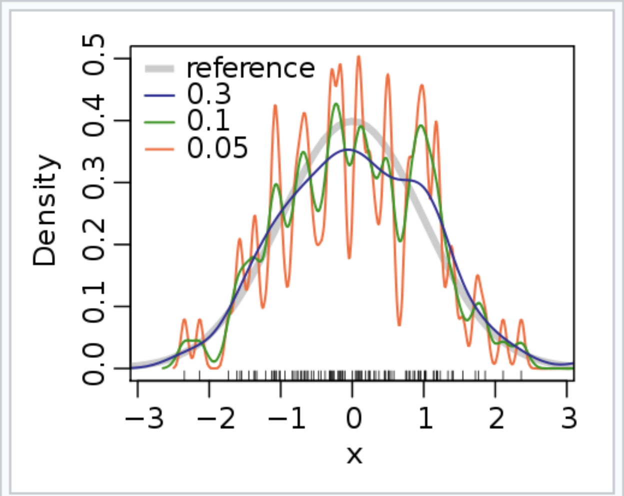
$$f_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \phi\left(\frac{x - x_i}{h}\right),$$

where (for example) $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)$.

Kernel estimation



Kernel estimation: selection of bandwidth



Skewness and kurtosis

Let the random variable X come from a distribution with mean μ and variance σ^2. For the standardized version of X,

$$Z = \frac{X - \mu}{\sigma},$$

the first moment (mean) $\mu_1 = \mathbb{E}(Z) = 0$ and the second moment $\mu_2 = \mathbb{E}(Z^2) = 1$.

Population coefficient for skewness and population kurtosis are defined as

$$\mu_3 = \mathbb{E}(Z^3) \quad \text{and} \quad \mu_4 = \mathbb{E}(Z^4),$$

which, given $\bar{x}$ and s, can be estimated by

$$m_3 = \frac{1}{ns^3}\sum_{i=1}^{n}(x_i - \bar{x})^3 \quad \text{and} \quad m_4 = \frac{1}{ns^4}\sum_{i=1}^{n}(x_i - \bar{x})^4.$$

Skewness and kurtosis

For normal distribution, $\mu_3 = 0$ (symmetric) and $\mu_4 = 3$ (reference, not heavy or light tails).

Skewness

- $\mu_3 = 0$, symmetric
- $\mu_3 > 0$, skewed to the right
- $\mu_3 < 0$, skewed to the left

Kurtosis

- $\mu_4 > 3$, heavy tails
- $\mu_4 < 3$, light tails

Measures of dispersion

- Sample variance and standard deviation: S^2 and S
- Sample range: $x_{(n)} - x_{(1)}$
- Interquartile range, IQR: $x_{0.75} - x_{0.25}$
- Median of the absolute values of deviations, MAD: sample median of $\{|x_i - \hat{m}|,\ i = 1, \ldots, n\}$

Boxplots

- Box: from the lower quartile to the upper quartile with the median indicated with the thicker line.
- Whiskers: Based on the 1.5 IQR value.
- Outliers: Observations outside the whiskers.