CHAPTER 2
DESCRIPTIVE STATISTICS

2.1 Data analysis

→ model specification

→ residuals after fitting a model

2.2 Univariate sample

$\underline{x} = (x_1, \ldots, x_n)$ generated by

$\underline{X} = (X_1, \ldots, X_n)$ i.i.d. F

Important summaries → LOCATION

→ DISPERSION

LOCATION → mean : $\bar{X} = \sum_{i=1}^{n} x_i / n$

→ median : a middle value in the sequence of sorted data

DISPERSION

- Variance or Standard deviation (for symmetric distrib.)
- Interquartile range (better for asymmetric distr.)

- Sample variance $s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

  Sample s.d. $= s_x$

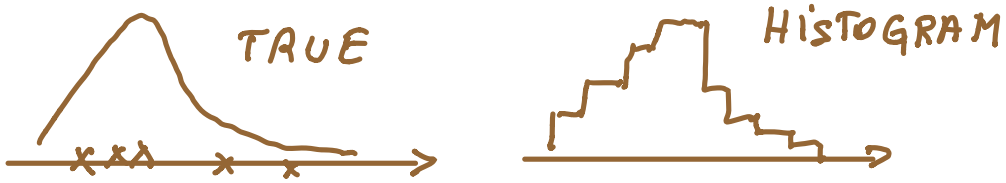
- IQR $= Q_{0.75} - Q_{0.25}$ (percentiles)

Graphical summaries

- histograms
- Box-plots
- QQ-plots

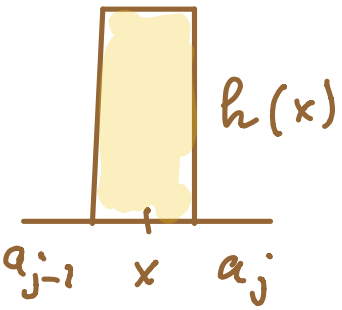
2.3 [Histograms]

To get an idea of the density



$a_0 \quad a_1 \quad a_2 \quad \dots \quad a_m$

$$x \in (a_{j-1}, a_j] \mapsto h(x) = \frac{\text{frequency in a class}}{(a_j - a_{j-1})n}$$



$h(x) =$ DENSITY

Relative frequency = AREA of rectangle

Total area = 1

- Histogram formally

Given $(x_1, \dots, x_n)$

$$h_n(x) = \frac{1}{a_j - a_{j-1}} \cdot \frac{1}{n} \sum_{i=1}^{n} 1_{a_{j-1} < x_i \leq a_j}$$

where 1 is an indicator function.

- A histogram is a good density estimator if
  - The partition $a_0, a_1, \dots, a_m$ is well chosen
  - Sample size is not too small.

Proof. $(X_1, \dots, X_n) \overset{i.i.d}{\sim} f(x)$ true density.

Assume that $a_{j-1} < x \leq a_j$ and $f(x) > 0$

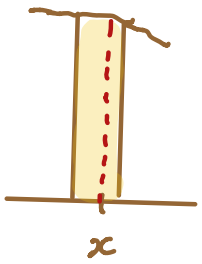
$$E\, h_n(x) = E \frac{1}{n(a_j - a_{j-1})} \sum_{i=1}^{n} 1_{a_{j-1} < X_i \leq a_j}$$

$$= \frac{1}{a_j - a_{j-1}} \frac{1}{n} \sum_{i=1}^{n} E\left(1_{a_{j-1} < X_i \le a_j}\right)$$

$$= \frac{1}{a_j - a_{j-1}} \frac{1}{n} \cdot n E\left(1_{a_{j-1} < X_1 \le a_j}\right)$$

$$= \frac{1}{a_j - a_{j-1}} P\left(a_{j-1} < X_1 \le a_j\right)$$

$$= \frac{\int_{a_{j-1}}^{a_j} f(s)\, ds}{a_j - a_{j-1}}$$

$E\, h_n(x) =$  $\approx f(x)$

By the Law of large numbers.

$$h_n(x) \xrightarrow{P} E\, h_n(x)$$

for $n \to \infty$.

See some examples in the notebook.

Box-plots

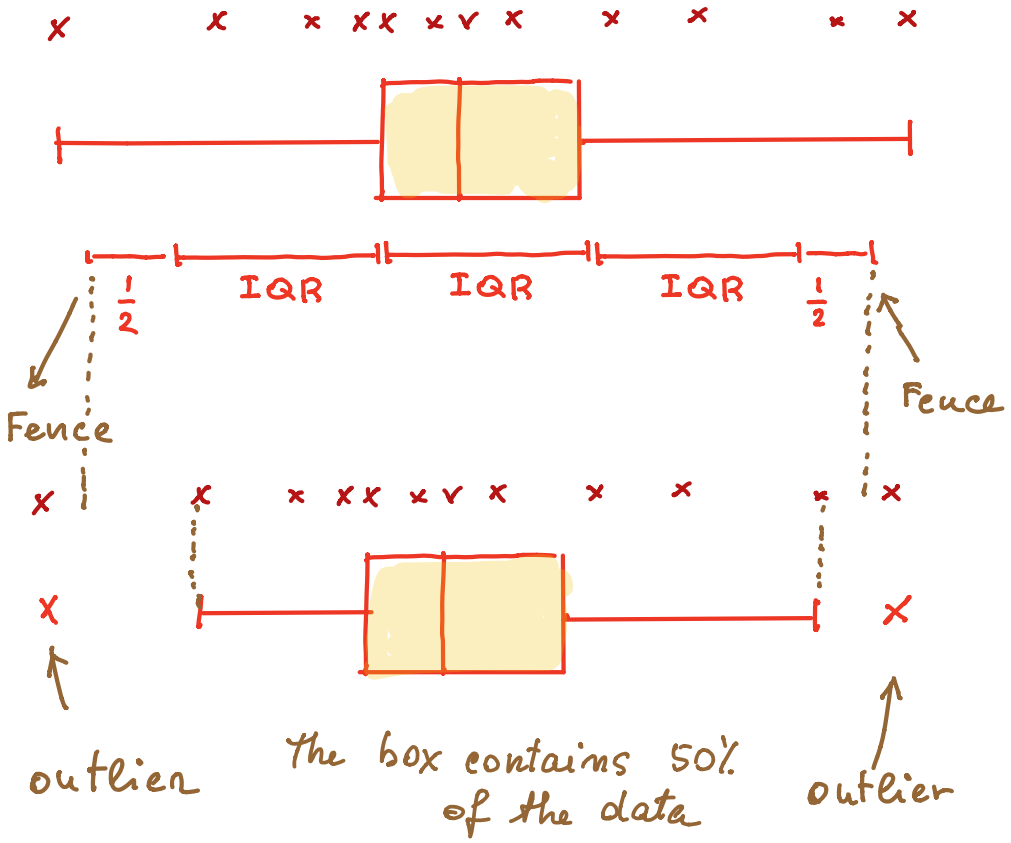
They can be vertical or horizontal

$$\text{Data} = (x_1, \ldots, x_n)$$

↓

5-number SUMMARY

min $Q_{0.25}$ Q_{50} $Q_{0.75}$ max



Example 2.5

Differences between normal and Cauchy.

Standard Cauchy density

$$f(x) = \frac{1}{\pi(1+x^2)} \qquad x \in \mathbb{R}$$

the expected value $E(X)$ does not exist since $E(|X|)$ is not finite.

LOCATION-SCALE FAMILY

- Preliminaries

If $X \sim F_X$ (distribution function)

Then $Y = a + bX$ has distribution

$$F_Y(y) = F_X\left(\frac{y-a}{b}\right) \qquad b > 0$$

Proof:

$$P(a+bX \le y) = P\left(X \le \frac{y-a}{b}\right)$$

$F_Y(y) \equiv F_{a,b}(y)$ is a family depending on $a \in \mathbb{R}$, $b > 0$.

called location-scale family generated by X.

- Density.

if F_X has a density f_X

then $F_{a,b}$ has a density $f_{a,b}$

$$f_{a,b}(y) = \frac{d}{dy} F_X\left(\frac{y-a}{b}\right) = \frac{1}{b} f_X\left(\frac{y-a}{b}\right)$$

Note

if X is standardized i.e.

$$EX = 0 \qquad \operatorname{var} X = 1$$

Then $a = E(Y)$

$b^2 = \operatorname{var}(Y)$

- Examples 1) $X \sim N(0,1)$

$$a + bX \sim N(a, b^2)$$

2) $X \sim \text{Cauchy}(0,1)$ $\quad f_X(x) = \dfrac{1}{\pi(1+x^2)}$

$$a + bX \to f_Y(y) = \frac{1}{b\pi\left[1 + \left(\frac{y-a}{b}\right)^2\right]}$$

• NOTE

$X \sim \chi^2_1$ chi-square with 1 degree of freedom

$Y \sim \chi^2_2$ 2 degrees of freedom.

They are not in the same location-scale family

but they have the same name

Verify. $X \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right)$

$Y \sim \text{Gamma}\left(1, \frac{1}{2}\right)$

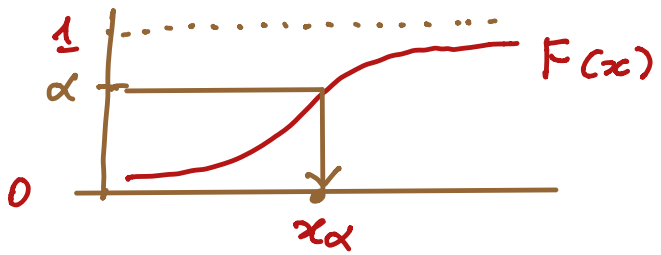
Different shape, same rate

$a = 0$; $bX \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{2b}\right)$

Then bX cannot change shape.

Quantile function

- Idea



Quantile $= F^{-1}(\alpha)$

x_α is an <u>α-quantile</u> of F

if exist a single value $\alpha \in (0,1)$ such that

$$F(x_\alpha) = \alpha$$

The α-quantile is denoted by $F^{-1}(\alpha)$

Note

$F^{-1}(0.25)$ = 25 percentile
$F^{-1}(0.75)$ = 75 percentile
$F^{-1}(0.5)$ = median

The function $\alpha \longmapsto F^{-1}(\alpha)$ is the <u>quantile function</u> provided that F^{-1} is well-defined.

If F is strictly increasing and continuous

$$F(F^{-1}(\alpha)) = \alpha \qquad \text{for all } \alpha \in (0,1)$$
$$F^{-1}(F(x)) = x \qquad \text{for all } x \in \mathbb{R}$$

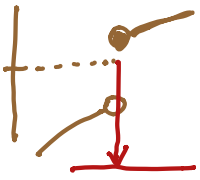
Example – Exponential (λ)

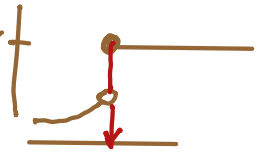
$$F(x) = 1 - e^{-\lambda x} \qquad x \geq 0$$

$$F^{-1}(x) = -\log(1-\alpha)/\lambda$$

In general the equation $F(x) = \alpha$ may have

- one solution 

- no solution 

- ∞ solutions 

To define well the quantile function in all cases let :

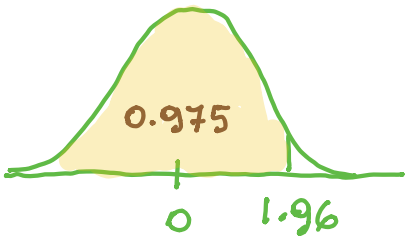
$$F^{-1}(\alpha) = \inf \{ x : F(x) \geq \alpha \}$$

In location-scale family.

$$F^{-1}_{a,b}(\alpha) = \alpha + \beta F^{-1}(\alpha)$$

Classical example

$X \sim N(0,1)$ $\qquad x_{0.975} = 1.96$



$Y \sim N(\mu, \sigma^2)$ $\qquad y_{0.975} = \mu + 1.96\sigma$

In general, the points with coordinates

$$\left(\bar{F}^{-1}(\alpha) \, , \; F_{a,b}^{-1}(\alpha) \right)$$

are on the line

$$y = a + b\,x$$

this is the base of the **QQ plot**

Order statistics

Given a sample $(X_1, \ldots, X_n)$
take the sorted sample

$$\underbrace{\left(X_{(1)}, X_{(2)}, \ldots, X_{(n)} \right)}$$

placed in increasing order

$X_{(1)}, \ldots X_{(n)}$ are called order statistics.

It can be proved that

$$E[F(X_{(i)})] = \frac{i}{n+1}$$

It is expected that the points

$$\left(\frac{i}{n+1}, F(X_{(i)})\right)$$

lie on the line $y = x$

The same must happen for

$$\left(F^{-1}\left(\frac{i}{i+n}\right), x_{(i)}\right)$$

QQ-plot.

Given a sample $(x_1 \dots x_n)$
the plot of the points

$$\left(F^{-1}\left(\frac{i}{i+n}\right), x_{(i)}\right)$$

is called a quantile-quantile plot with respect to the distribution F

Points $\left(F^{-1}\left(\frac{i}{n+1}\right), x_{(i)}\right)$

- $\simeq$ $y = x$ → sample comes from F
- $\simeq$ $y = a + bx$ → sample comes from $F_{a,b}$
- different curve → sample <u>not</u> from F.

2.3 CORRELATION

observations x_i can be on multiple variables, for same unit.

$$\underset{\sim}{x}_i = (x_{i1}, x_{i2})$$

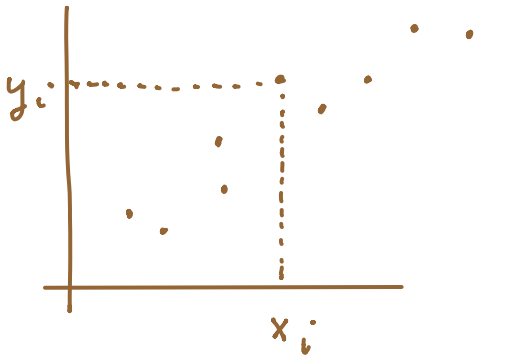
$$\downarrow$$

$$(x_i, y_i)$$

Ex: (height of mother , height of daughter)

SCATTER - a plot of (x_i, y_i)

$i = 1, \ldots, n$



If points lie close to a line

$$\downarrow$$

linear correlation between X, Y

POSITIVE NEGATIVE

- Association measure:
  sample correlation coefficient

$(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d

$$r_{xy} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)\sqrt{s_x^2}\sqrt{s_y^2}}$$

- Interpretation: strength of association

$$-1 \leq r_{xy} \leq +1$$

1) $r_{xy} = 1$: (x_i, y_i) on $y = \bar{y} + \frac{s_y}{s_x}(x - \bar{x})$

2) $r_{xy} = -1$: (x_i, y_i) on $y = \bar{y} - \frac{s_x}{s_y}(x - \bar{x})$

3) If $(X_1, Y_1), \dots, (X_n, Y_n)$ iid from (X, Y)
and X independent Y

then

$$r_{XY} \simeq 0$$

4) $|r_{XY}| \simeq 0 \not\Rightarrow X \perp\!\!\!\perp Y$

See book's example.

• AUTOCORRELATION

In time series $(x_1, .. x_h, \dots x_T)$ observations are not i.i.d (tipically)

this can be verified

1) With a scatter

SCATTER of (x_i, x_{i+1})
$i = 1, 2, \dots, n-1$

should have almost no structure.

2) with an index of autocorrelation

INDEX of autocorrelation of order
$h = 0, 1, 2, \ldots$

$$r_x(h) = \frac{\sum_{i=1}^{n-h} (x_{i+h} - \bar{x})(x_i - \bar{x})}{(n-h)\, s_x^2}$$

measures correlation
between X_i and X_{i-h}.