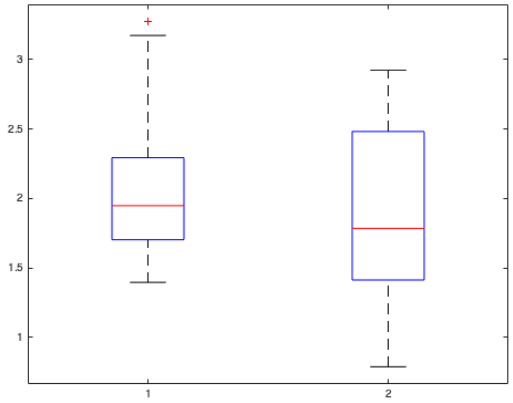
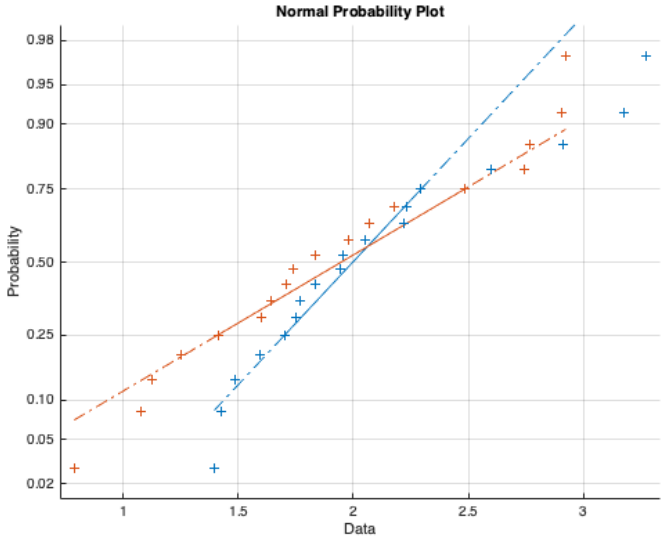
Slides 9: Empirical distribution

- Empirical distribution function
- Empirical variance
- Hazard function
- Empirical quantiles
- QQ-plots
- Normal probability plot
- Skewness and kurtosis
- Boxplots





Empirical distribution function

Population distribution $F(x) = \mathrm{P}(X \le x)$ and its density $f(x) = F'(x)$.

For a given random sample $(x_1, \ldots, x_n)$, define

$$\boxed{\text{Empirical distribution function } \hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{x_i \le x\}}}$$

For a fixed x,

$$\hat{F}(x) = \hat{p}$$

is the sample proportion estimating the population proportion $p = F(x)$.

On the other hand, for variable x, the function $\hat{F}(x) = \mathrm{P}(Y \le x)$ describes the discrete distribution

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

assuming that all sample values x_i are pairwise different. Clearly,

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

is the sample mean. This holds even if some of x_i coincide.

Empirical variance

Since

$$\mathrm{E}(Y^2) = \sum_{i=1}^{n} \frac{x_i^2}{n} = \overline{x^2},$$

we get what we call the empirical variance

$$\mathrm{Var}(Y) = \overline{x^2} - (\bar{x})^2 = \frac{n-1}{n} s^2 = \hat{\sigma}^2.$$

Question. I the empirical an unbiased estimate of σ^2? Is it a consistent estimate?

Hazard function

If a life length T has distribution function $F(t) = \mathrm{P}(T \leq t)$, then its survival function is

$$S(t) = \mathrm{P}(T > t) = 1 - F(t).$$

The hazard function $h(t) = \frac{f(t)}{S(t)}$ gives the mortality rate at age t:

$$\tfrac{1}{\delta}\mathrm{P}(t < T \leq t + \delta | T \geq t) = \tfrac{\mathrm{P}(t<T\leq t+\delta)}{\delta \mathrm{P}(T \geq t)} = \tfrac{F(t+\delta)-F(t)}{\delta S(t)} \to \tfrac{f(t)}{S(t)}, \quad \delta \to 0.$$

A constant hazard rate $h(t) = \lambda$ corresponds to the exponential distribution Exp(λ). The lack of memory property. Indeed, with $f(t) = \lambda e^{-\lambda t}$ and $S(t) = e^{-\lambda t}$, we get

$$h(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

The hazard function can be viewed as the negative of the slope of the log survival function:

$$h(t) = -\frac{d}{dt} \log S(t) = -\frac{d}{dt} \log(1 - F(t)).$$

In terms of the empirical survival function

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

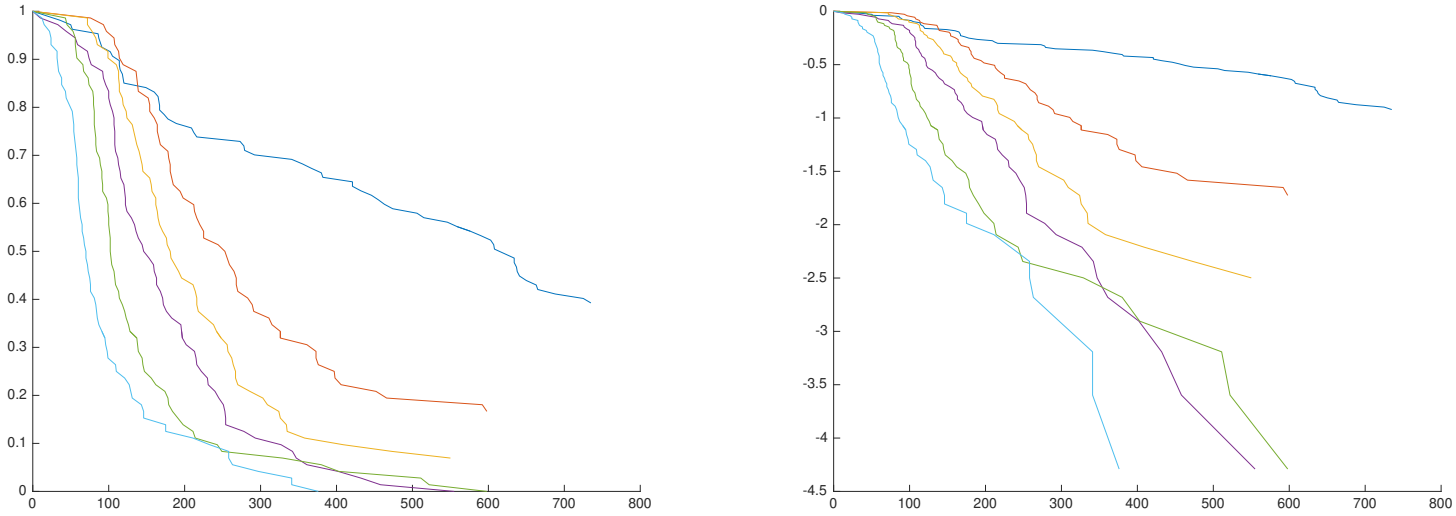
the hazard function is estimated by the negative slope of the log-survival empirical function.

Question. How would you find the yearly mortality rate for 64 years old men living in Sweden?

Case study: guinea pigs

Guinea pigs were randomly divided in 5 treatment groups of 72 each and one control group of 107 animals. The treatment groups were infected with increasing doses of tubercle bacilli. Lifetimes recorded (not all the animals died).

It is difficult to compare the groups just looking at numbers (next slide).



Left: the survival functions. Right: the log-survival functions. The negative slopes on the right give the hazard rates for different groups.

Control: 18 36 50 52 86 87 89 91 102 105 114 114 115 118 119 120 149 160 165 166 167 167 173 178 189 209 212 216 273 278 279 292 341 355 367 380 382 421 421 432 446 455 463 474 506 515 546 559 576 590 603 607 608 621 634 634 637 638 641 650 663 665 688 725 735

Dose I: 76 93 97 107 108 113 114 119 136 137 138 139 152 154 154 160 164 164 166 168 178 179 181 181 183 185 194 198 212 213 216 220 225 225 244 253 256 259 265 268 268 270 283 289 291 311 315 326 326 361 373 373 376 397 398 406 452 466 592 598

Dose II: 72 72 78 83 85 99 99 110 113 113 114 114 118 119 123 124 131 133 135 137 140 142 144 145 154 156 157 162 162 164 165 167 171 176 177 181 182 187 192 196 211 214 216 216 218 228 238 242 248 256 257 262 264 267 267 270 286 303 309 324 326 334 335 358 409 473 550

Dose III: 10 33 44 56 59 72 74 77 92 93 96 100 100 102 105 107 107 108 108 108 109 112 113 115 116 120 121 122 122 124 130 134 136 139 144 146 153 159 160 163 163 168 171 172 176 183 195 196 197 202 213 215 216 222 230 231 240 245 251 253 254 254 278 293 327 342 347 361 402 432 458 555

Dose IV: 43 45 53 56 56 57 58 66 67 73 74 79 80 80 81 81 81 82 83 83 84 88 89 91 91 92 92 97 99 99 100 100 101 102 102 102 103 104 107 108 109 113 114 118 121 123 126 128 137 138 139 144 145 147 156 162 174 178 179 184 191 198 211 214 243 249 329 380 403 511 522 598

Dose V: 12 15 22 24 24 32 32 33 34 38 38 43 44 48 52 53 54 54 55 56 57 58 58 59 60 60 60 60 61 62 63 65 65 67 68 70 70 72 73 75 76 76 81 83 84 85 87 91 95 96 98 99 109 110 121 127 129 131 143 146 146 175 175 211 233 258 258 263 297 341 341 376

Quantiles

The inverse of distribution function $F(x)$ is called quantile function

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

For a given distribution F and $0 \leq p \leq 1$, the p-quantile is $x_p = Q(p)$.

Special quantiles:

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

For a random variable X with $\mathrm{P}(X \leq x) = F(x)$,

$$\mathrm{P}(X \leq x_p) = F(x_p) = F(Q(p)) = p$$

so that quantile x_p cuts off proportion p of the smallest values of X.

The ordered sample values

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

are the jump points for the empirical distribution function.

Empirical quantiles

In the continuous case with $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$, we have

$$F_n(x_{(k)}) = \tfrac{k}{n}, \qquad F_n(x_{(k)} - \epsilon) = \tfrac{k-1}{n}.$$

This observation leads to the following definition of empirical quantiles

$$\boxed{x_{(k)} \text{ is called the empirical } (\tfrac{k-0.5}{n})\text{-quantile}}$$

QQ-plots

Two independent samples $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ are taken from two population distributions F_1 and F_2. The null hypothesis

$$H_0 : \{F_1(x) = F_2(a + bx) \text{for all } x \text{ and unspecified } a \text{ and } b\}$$

claimes the linear relation $Y = a + bX$ in distribution. It can be tested graphically using a QQ-plot.

$$\boxed{\text{QQ-plot is a scatter plot of } n \text{ dots with coordinates } (x_{(k)}, y_{(k)})}$$

If the QQ-plot is far from a straight line, we reject H_0.

The normality hypothesis H_{norm} states that a random sample $(x_1, \ldots, x_n)$ is drawn from $\mathrm{N}(\mu, \sigma)$ with unknown (μ, σ).

A special QQ-plot used for testing this hypothesis is called a normal probability plot (NPP). NPP is the scatter plot for

$$(x_{(1)}, y_1), \ldots, (x_{(n)}, y_n), \quad \text{where } y_k = \Phi^{-1}(\tfrac{k-0.5}{n}).$$

If NPP is far from a straight line, we reject H_{norm}.

If NPP is close to a straight line $y = a + bx$, so that

$$\Phi^{-1}(p) = a + bQ(p), \quad 0 < p < 1,$$

then after replacing p by $F(x)$, we get

$$\Phi^{-1}(F(x)) = a + bx, \quad -\infty < x < \infty$$

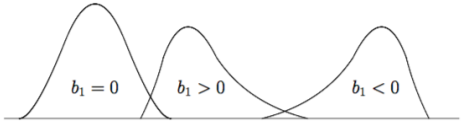
which gives $F(x) = \Phi(a + bx)$. This implies normality $F(x) \equiv \Phi(\frac{x-\mu}{\sigma})$.

Conclusion: NPP close to $y = a + bx$ supports the normal distribution model with point estimates $\hat{\mu} = -\frac{a}{b}$, $\hat{\sigma} = \frac{1}{b}$.

Population coefficients of skewness and kurtosis:

$$\beta_1 = \mathrm{E}[(\tfrac{X-\mu}{\sigma})^3], \quad \beta_2 = \mathrm{E}[(\tfrac{X-\mu}{\sigma})^4]$$

sample skewness and kurtosis:



$$b_1 = \frac{1}{n}\sum_{i=1}^{n}(\tfrac{x_i-\bar{x}}{s})^3, \quad b_2 = \frac{1}{n}\sum_{i=1}^{n}(\tfrac{x_i-\bar{x}}{s})^4$$

Symmetric $\beta_1 = 0$, skewed to the right $\beta_1 > 0$, and skewed to the left $\beta_1 < 0$ distributions.

Given that $b_1 \approx 0$, b_2 may inform on the curve shape. Normal distribution has $\beta_2 = 3$. Heavy tails: $\beta_2 > 3$. Light tails: $\beta_2 < 3$.

$\mathrm{Gam}(\alpha, \lambda)$ distribution has $\beta_1 = \frac{2}{\sqrt{\alpha}}$ and $\beta_2 = 3 + \frac{6}{\alpha}$. As $\alpha \to \infty$ the shape of the gamma distribution becomes normal.

Question. Draw a curve for the heights of adult males which is skewed to the right. Does your curve make clear that more than a half of the heights are below the average?

Boxplot

Box

upper edge of the box = upper quartile (UQ)

box center = median

lower edge of the box = lower quartile (LQ)

Wiskers

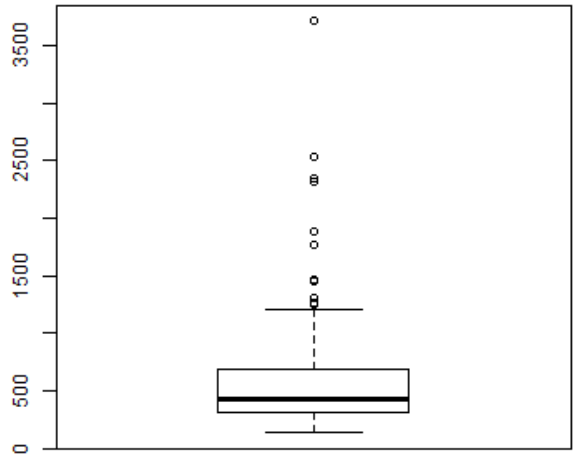
upper whisker end = {maximal data point $\leq$ UQ + 1.5 $\times$ IQR}

lower whisker end = {minimal data point $\geq$ LQ − 1.5 $\times$ IQR}

Outliers

upper dots = {data points $\geq$ UQ + 1.5 $\times$ IQR}

lower dots = {data points $\leq$ LQ − 1.5 $\times$ IQR}



Boxplots are convenient
for comparing several samples.

Boxplots of daily maximum
concentrations
of sulfur dioxide.