

Chapter 10. Summarising data

1 Empirical probability distribution

Consider an IID sample (X_1, \dots, X_n) from the population distribution $F(x) = P(X \leq x)$.

$$\text{Empirical distribution function } F_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{\{X_i \leq x\}}.$$

For a fixed x , $F_n(x)$ is the sample proportion estimating the population proportion $F(x)$. $F_n(\cdot)$ is a cumulative distribution function with mean \bar{X} and variance $\frac{n-1}{n} s^2$.

If the data describes life lengths, then it is more convenient to use the empirical survival function $S_n(x) = 1 - F_n(x)$, the proportion of the data greater than x . If the lifelength T has distribution function $F(t) = P(T \leq t)$, then its survival function is $S(t) = P(T > t) = 1 - F(t)$.

$$\text{Hazard function } h(t) = \frac{f(t)}{S(t)}, \text{ where } f(t) = F'(t) \text{ is the probability density function.}$$

The hazard function is the mortality rate at age t :

$$P(t < T \leq t + \delta | T \geq t) = \frac{F(t + \delta) - F(t)}{S(t)} \sim \delta \cdot h(t), \quad \delta \rightarrow 0.$$

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

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

Example (Guinea pigs)

Guinea pigs were infected with tubercle bacillus, then divided in 5 treatment groups and one control group. The survival times were recorded. The data is illustrated by two graphs: one for the survival functions and the other for the log-survival functions.

$$\text{A constant hazard rate } h(t) = \lambda \text{ corresponds to the exponential distribution } \text{Exp}(\lambda).$$

2 Density estimation

A histogram displays the observed counts $O_j = \sum_{i=1}^n 1_{\{X_i \in \text{cell}_j\}}$ over the adjacent cells of width h . The choice of a balanced width h is important: smaller h give ragged profiles, larger h give obscured profiles.

Put $f_h(x) = \frac{1}{nh} O_j$ for x belonging to the cell j , and notice that $\int f_h(x) dx = \frac{1}{nh} \sum_j O_j = 1$. The scaled histogram given by the graph of $f_h(x)$ is a density estimate.

Kernel density estimate with bandwidth h produces a smooth curve

$$f_h(x) = \frac{1}{nh} \sum \phi\left(\frac{x - X_i}{h}\right), \text{ where } \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Example (male heights)

Let = column of 24 male heights. For a given bandwidth h , the following matlab code produces a plot for a kernel density estimate

```
x=160:0.1:210; L=length(x);
f=normpdf((ones(24,1)*x - hm*ones(1,L))/h);
fh=sum(f)/(24*h); plot(x,fh)
```

The stem-and-leaf plot for the 24 male heights indicates the distribution shape plus gives the full numerical information:

```
17:056678899
18:0000112346
19:229
```

3 Q-Q plots

The inverse of the cumulative distribution function F is called the quantile function $Q = F_{-1}$. The quantile function Φ_{-1} for the standard normal distribution Φ is called the profit function (from PROBABility unit).

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

Special quantiles:

median $M = Q(0.5)$, lower quartile $Q(0.25)$, upper quartile $Q(0.75)$.

Quantile x_p cuts off proportion p of smallest values of a random variable X with $P(X \leq x) = F(x)$:

$$P(X \leq x_p) = F(x_p) = F(Q(p)) = p.$$

The ordered sample values $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ are the jump points for the empirical distribution function. Since

$$F_n(X_{(k)}) = \frac{k}{n} \text{ and } F_n(X_{(k)} - \epsilon) = \frac{k-1}{n}, \quad \boxed{X_{(k)} \text{ is called the empirical } \left(\frac{k-0.5}{n}\right)\text{-quantile.}}$$

Suppose we have two independent samples (X_1, \dots, X_n) and (Y_1, \dots, Y_n) with population distribution functions F_X and F_Y . A relevant null hypothesis $H_0: F_X \equiv F_Y$ is equivalent to $H_0: Q_X \equiv Q_Y$.

It can be tested graphically using a Q-Q plot.

The Q-Q plot is a scatter plot of n dots with coordinates $(X_{(k)}, Y_{(k)})$.

We accept the H_0 of equal distributions if the scatter plot is close to the bisector, that is when we have almost equal quantiles.

More generally, if $Y = a + bX$ in distribution, then $F_X(x) = F_Y(a + bx)$, then for under $Q_X(p) = x$, we get $Q_Y(p) = a + bx$. Thus $Q_Y(p) = a + bQ_X(p)$ and the Q-Q plot should approximate the straight line $y = a + bx$.

4 Testing normality

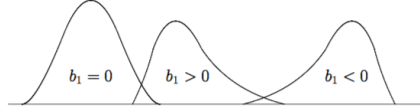
The normality hypothesis H_0 states that the population distribution for the sample (X_1, \dots, X_n) is normal $N(\mu, \sigma^2)$ with unspecified parameter values. A Q-Q plot used for testing this hypothesis is called normal probability plot.

If the normal probability plot is close to a straight line $y = a + bx$, then we accept H_0 and use the point estimates $\hat{\mu} = a$, $\hat{\sigma} = b$.

Normal probability plot is the scatter plot for (x_k, y_k) , where $x_k = \Phi_{-1}(\frac{k-0.5}{n})$ and $y_k = X_{(k)}$.

If normality does not hold, draw a straight line via empirical lower and upper quartiles to detect a light tails profile or heavy tails profile.

Coefficient of skewness: $\beta_1 = \frac{E(X-\mu)^3}{\sigma^3}$, sample skewness: $b_1 = \frac{1}{s^3 n} \sum (X_i - \bar{X})^3$



Kurtosis $\beta_2 = \frac{E(X-\mu)^4}{\sigma^4}$, sample kurtosis: $b_2 = \frac{1}{s^4 n} \sum (X_i - \bar{X})^4$

For the normal distribution $\beta_2 = 3$. Leptokurtic distribution: $b_2 > 3$ (heavy tails). Platykurtic distribution: $b_2 < 3$ (light tails).

Example (male heights)

Summary statistics: $\bar{X} = 181.46$, $\hat{M} = 180$, $b_1 = 1.05$, $b_2 = 4.31$. Good to know: the distribution of the heights of adult males is positively skewed, so that $M < \mu$, or in other terms, $P(X < \mu) > 0.50$.

The gamma distribution $\text{Gamma}(\alpha, \lambda)$ is positively skewed $\beta_1 = \frac{2}{\sqrt{\alpha}}$, and leptokurtic $\beta_2 = 3 + \frac{6}{\alpha}$.

5 Measures of location

The central point of a distribution can be defined in terms of various measures of location, for example, as the population mean μ or the median M . The population median M is estimated by the sample median.

Sample median: $\hat{M} = X_{(k)}$, if $n = 2k - 1$ and $\hat{M} = \frac{X_{(k)} + X_{(k+1)}}{2}$, if $n = 2k$.

The sample mean \bar{X} is sensitive to outliers while the sample median \hat{M} is not, \hat{M} is a robust estimator.

Confidence interval for the median

Consider an IID sample (X_1, \dots, X_n) without assuming any parametric model for the unknown population distribution. Let $Y = \sum_{i=1}^n 1_{\{X_i \leq M\}}$ be the number of observations below the median, then

$$p_k = P(X_{(k)} < M < X_{(n-k+1)}) = P(k \leq Y \leq n - k)$$

can be computed from the symmetric binomial distribution $Y \sim \text{Bin}(n, 0.5)$.

This yields the following non-parametric formula for an exact confidence interval for the median.

$(X_{(k)}, X_{(n-k+1)})$ is a $100 \cdot p_k\%$ CI for the population median M .

Example. For $n = 25$, from the table below we find that $(X_{(8)}, X_{(18)})$ gives a 95.7% CI for the median.

k	6	7	8	9	10	11	12
p_k	99.6	98.6	95.7	89.2	77.0	57.6	31.0

Sign test

The sign test is a non-parametric test of $H_0: M = M_0$ against the two-sided alternative $H_1: M \neq M_0$. The sign test statistic $Y_0 = \sum_{i=1}^n 1_{\{X_i \leq M_0\}}$ counts the number of observations below the null hypothesis value. It has a simple null distribution $Y_0 \stackrel{H_0}{\sim} \text{Bin}(n, 0.5)$. Connection to the above CI formula: reject H_0 if M_0 falls outside the corresponding confidence interval $(X_{(k)}, X_{(n-k+1)})$.

Trimmed means

A trimmed mean is a robust measure of location computed from a central portion of the data.

α -trimmed mean \bar{X}_α = sample mean without $\frac{n\alpha}{2}$ smallest and $\frac{n\alpha}{2}$ largest observations

Example (male heights)

Ignoring 20% of largest and 20% of smallest observations we compute $\bar{X}_{0.4}=180.36$. The trimmed mean is between $\bar{X} = 181.46$ and $\hat{M} = 180$.

When summarizing data compute several measures of location and compare the results.

Nonparametric bootstrap

Substitute the population distribution by the empirical distribution. Then a bootstrap sample is obtained by resampling with replacement from the original sample x_1, \dots, x_n .

Generate many bootstrap samples of size n to approximate the sampling distribution for an estimator like trimmed mean, sample median, or s .

6 Measures of dispersion

Sample variance s^2 and sample range $R = X_{(n)} - X_{(1)}$ are sensitive to outliers. Two robust measures of dispersion:

interquartile range $\text{IQR} = Q(0.75) - Q(0.25)$ is the difference between the upper and lower quartiles,

$\text{MAD} = \text{Median of Absolute values of Deviations from the sample median } |X_i - \hat{M}|, i = 1, \dots, n$.

Three estimates of σ for the normal distribution $N(\mu, \sigma^2)$ model: $s, \frac{\text{IQR}}{1.35}, \frac{\text{MAD}}{0.675}$

Under the normality assumption

$\text{IQR} = (\mu + \sigma\Phi_{-1}(0.75)) - (\mu + \sigma\Phi_{-1}(0.25)) = 2\sigma\Phi_{-1}(0.75) = 1.35\sigma$, because $\Phi_{-1}(0.75) = 0.675$.

$\text{MAD} = 0.675\sigma$, since $P(|X - \mu| \leq 0.675\sigma) = (\Phi(0.675) - 0.5) \cdot 2 = 0.5$.

Box plot

The box plots are convenient to use for comparing different samples (illustrate using the daily SO_2 concentration data). A box plot is built of the following components

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

upper whisker end = {max data point \leq UQ + 1.5 IQR}

upper edge of the box = upper quartile (UQ)

box center = median

lower edge of the box = lower quartile (LQ)

lower whisker end = {min data point \geq LQ - 1.5 IQR}

lower dots = {data \leq LQ - 1.5 IQR}