Serik Sagitov, Chalmers and GU, January 27, 2016

Chapter 10. Summarising data

1 Empirical probability distribution

Consider an IID sample $(X_1, \ldots, X_n)$ from the population distribution $F(x) = \mathrm{P}(X \le x)$.

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

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 $F(t) = \mathrm{P}(T \le t)$, then its survival function is $S(t) = \mathrm{P}(T > t) = 1 - F(t)$.

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

The hazard function is the mortality rate at age t:

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

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.

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

2 Density estimation

A histogram displays the observed counts $O_j = \sum_{i=1}^n 1_{\{X_i \in \mathrm{cell}_j\}}$ 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

$f_h(x) = \frac{1}{nh}\sum \phi(\frac{x - X_i}{h})$, 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 $\mathrm{P}(X \leq x) = F(x)$:
$\mathrm{P}(X \leq x_p) = F(x_p) = F(Q(p)) = p$.
The ordered sample values $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ are the jump points for the empirical distribution function. Since
$F_n(X_{(k)}) = \frac{k}{n}$ and $F_n(X_{(k)} - \epsilon) = \frac{k-1}{n}$, $X_{(k)}$ is called the empirical $(\frac{k-0.5}{n})$-quantile.

Suppose we have two independent samples $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ with population distribution functions F_1 and F_2. A relevant null hypothesis H_0: $F_1 \equiv F_2$ is equivalent to H_0: $Q_1 \equiv Q_2$.
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 $\mathrm{P}(X \leq x) = \mathrm{P}(Y \leq a + bx)$, in other words, $Y = a + b \cdot X$ in distribution, then under $Q_2(p) = a + bQ_1(p)$, and the Q-Q plot should approximate the straight line $y = a + bx$. Indeed,
$F_1(x) = F_2(a + bx)$ implies $Q_2(F_1(x)) = a + bx$, and therefore $Q_2(p) = a + bQ_1(p)$.

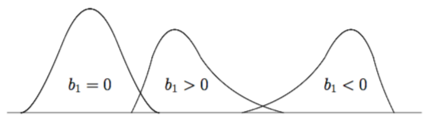
4 Testing normality

The normality hypothesis H_0 states that the population distribution for the sample $(X_1, \ldots, X_n)$ is normal $\mathrm{N}(\mu, \sigma^2)$ 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, $\mathrm{P}(X < \mu) > 0.50$.

The gamma distribution Gamma(α, λ) 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 $(X_1, \ldots, X_n)$ 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

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

can be computed from the symmetric binomial distribution $Y \sim \mathrm{Bin}(n, 0.5)$.
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 $Y_0 = \sum_{i=1}^{n} 1_{\{X_i \le M_0\}}$ counts the number of observations below the null hypothesis value. It has a simple null distribution $Y_0 \overset{H_0}{\sim} \text{Bin}(n, 0.5)$. 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 $\bar{X}_\alpha$ = 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 $x_1, \ldots, x_n$.
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 IQR $= Q(0.75) - Q(0.25)$ is the difference between the upper and lower quartiles,
MAD = Median of Absolute values of Deviations from the sample median $|X_i - \hat{M}|$, $i = 1, \ldots, n$.

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

Under the normality assumption

IQR $= (\mu + \sigma\Phi_{-1}(0.75)) - (\mu + \sigma\Phi_{-1}(0.25)) = 2\sigma\Phi_{-1}(0.75) = 1.35\sigma$, because $\Phi_{-1}(0.75) = 0.675$.
MAD $= 0.675\sigma$, since $\text{P}(|X - \mu| \le 0.675\sigma) = (\Phi(0.675) - 0.5) \cdot 2 = 0.5$.

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 $\ge$ UQ + 1.5 IQR}
upper whisker end = {max data point $\le$ 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 $\ge$ LQ – 1.5 IQR}
lower dots = {data $\le$ LQ – 1.5 IQR}