Part A Statistics

Hilary Term and Trinity Term 2004

There are three kinds of lies: lies, damned lies, and statistics. (*Disraeli*)

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write. (*H.G. Wells)*

1. Exploratory Data Analysis

Statistical analysis:

understanding uncertainty:

produce an economical and informative description of data

model the data-generating mechanism

make predictions / decisions

Data: could be

- numerical
- counts
- ordinal
- categorical (categories, such as eye colour)

Here mostly: quantitative data; could be

- univariate: discrete (counts e.g.), continuous (measurement of speed, e.g.)
- multivariate: more than one observation per subject (weight and height, e.g.)

data could come from

- experiments
- observational studies

First: *visual display of the data*

Histogram

Partition space in which the data points lie into cells

blocks are erected over these cells such that the volume of each block is proportional to the number of data points in the cell

Example: Infants with SIRDS; see *Daly et al., p.4*

- grouping matters
- the scale is the area of the block, not the height of the block !
- related: pie charts, bar charts (categorical data)

Numerical summaries

Suppose that we have numerical data $x_1, \ldots, x_n$

Summaries for centre:

sample mean is the average

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

sample median is the middle observation if n is odd, and the the average of the middle two if n is even

Summaries for spread:

sample variance

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right\}
\end{aligned}
$$

(sample) quantiles: q_α, are defined for $0 \le \alpha \le 1$ so that a proportion of at least α of the data are less or equal to q_α and a proportion of at least $1-\alpha$ is greater or equal to q_α

There are many (at least 8) definitions of q_α if αn is not an integer; we shall use this one

Special quantiles: $q_{1/2}$ is the median

$q_{1/4} = q_L$ is the *lower sample quartile*, also called *lower hinge*

$q_{3/4} = q_R$ is the *upper sample quartile*, also called *upper hinge*

The *interquartile range* (IQR) is

$$IQR = q_{3/4} - q_{1/4}$$

Example: If $n = 50$, say, then order the data $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(50)}$;
median is $\frac{1}{2}(x_{(25)} + x_{(26)})$
$q_L = x_{(13)}$: have proportion $13/50 \geq 1/4$ of observation less or equal than $x_{(13)}$, and have proportion $38/50 \geq 3/4$ of observations bigger or equal than $x_{(13)}$
$q_R = x_{(38)}$

Box plots = *Box-and-whisker plots*

Central box:

- bounded below by the lower hinge

- bounded above by the upper hinge

- central line is the median

Whiskers:

- run to the observation that is nearest to $1.5\times$ the size of the box from the nearest hinge

- length is no larger than a *step* $= \frac{3}{2}IQR$

Observations that are more extreme are shown separately; these are also called *outliers*

Example: Rayleigh's nitrogen data: can distinguish the two groups

Quantile plots= Q-Q plots

Empirical distribution function (ecdf)

$F_n(x) = \frac{1}{n}\times$ the number of observations $\leq x$

jumps $1/n$ at each of the observations

would be close to a straight line for a uniform distribution

the quantiles can be read off by "inverting" an e.c.d.f.plot

Q-Q plots compare two sets of data by plotting the quantiles of one against the other

often one set of data is replaces by the quantiles of a theoretical distribution; then also called *probability plot*

Assume that X is a continuous random variable with c.d.f. F, density f; Let $0 < p < 1$

The pth quantile of f, denoted by $Q(p)$, is a value such that

$$F(Q(p)) = \int_{-\infty}^{Q(p)} f(x)dx = p$$

For $U_1, \ldots, U_n \sim U(0,1)$ independent, order $U_{1:n} < U_{2:n} < \cdots < U_{n:n}$; we have

$$EU_{k:n} = \frac{k}{n+1}$$

and $Var(U_{k:n})$ is small (Exercise)

Recall: if $U \sim U(0,1)$ then $X = Q(U)$ has density f

so $Q\left(\frac{k}{n+1}\right)$ should be a good approximation for $EX_{k:n}$

Data: order $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$

Probability plot: plot $x_{(k)}$ against $Q\left(\frac{k}{n+1}\right)$

Example: $F(x) = 1 - e^{-x}, \quad x \geq 0$ c.d.f. of exponential (1) - distribution

$f(x) = e^{-x}, \quad x \geq 0$

calculate $Q(p) = -\log(1-p)$

exponential quantile plot:

plot $x_{(k)}$ against $-\log(1 - k/(n+1))$

Normal approximation for e.c.d.f.

Suppose that $x_1, \ldots, x_n$ are realizations of i.i.d. random variables $X_1, \ldots, X_n$ with c.d.f. F

Fix a; let $I_k = \mathbf{1}(X_k \leq a)$, the indicator function which takes the value one if $X_k \leq a$, and 0 otherwise, then

$$EI_k = P(X_k \leq a) = F(a)$$

Put $W = \frac{1}{n}\sum_{k=1}^{n} I_k$, then $EW = F(a)$ and

$$VarW = \frac{1}{n}F(a)(1 - F(a))$$

from Central Limit Theorem: W is approximately $\mathcal{N}(F(a), \frac{1}{n}F(a)(1 - F(a)))$

and $F_n(a)$ is a realization of W

Distribution of order statistics

Let $X_1, \ldots, X_n$ be i.i.d. with continuous distribution function F and density f

order them $X_{1:n} < X_{2:n} < \cdots < X_{n:n}$

$X_{k:n}$ is called the *kth order statistic*

The uniform case: Suppose that $U_1, \ldots, U_n$ are independent $U[0,1]$-random variables

By symmetry, there are $n!$ different possible orderings, and they are all equally likely, so:

The joint density of $(U_{1:n}, U_{2:n}, \ldots, U_{n:n})$ is given by

$$f(u_{(1)}, u_{(2)}, \ldots, u_{(n)}) = n!,$$

for $0 < u_{(1)} < u_{(2)} < \cdots < u_{(n)} < 1$

Theorem 1 *1. The density of* $U_{k:n}$ *is given by*

$$f_{(k)}(x) = \frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k}$$

for $0 < x < 1$

2. For $j < k$ *the joint density of* $(U_{j:n}, U_{k:n})$ *is given by*

$$f_{(j,k)}(x,y) = \frac{n!}{(j-1)!(k-j-1)!(n-k)!} x^{j-1}(y-x)^{k-j-1}(1-y)^{n-k},$$

for $0 < x < y < 1$

Proof:

1. Fix $0 < x < 1$ and let $N(x)$ denote the number of $U's$ that are less or equal to x then $N(x) \sim Binomial(n, x)$ as, for the uniform, $F(x) = x$

   We obtain as c.d.f. of $U_{k:n}$

$$\begin{aligned} F_{(k)}(x) &= P(U_{k:n} \leq x) \\ &= P(N(x) \geq k) \\ &= \sum_{j=k}^{n} \binom{n}{j} x^j (1-x)^{n-j} \end{aligned}$$

Differentiate:

$$
\begin{aligned}
f_{(k)}(x) &= \sum_{j=k}^{n} \binom{n}{j} \Big\{ j x^{j-1} (1-x)^{n-j} \\
&\qquad - (n-j) x^{j} (1-x)^{n-j-1} \Big\}
\end{aligned}
$$

Put

$$T_j = \binom{n}{j} (n-j) x^{j} (1-x)^{n-j-1}$$

then have telescope sum

$$f_{(k)}(x) = \sum_{j=k}^{n} (T_{j-1} - T_j)$$

wiht $T_n = 0$, giving $f_{(k)}(x) = T_{k-1}$, which is the first assertion.

2. Informal argument:

Suppose $u_{(j)} \in (x, x + dx)$ and

$u_{(k)} \in (y, y + dy)$

then there are $j - 1$ of the $U's$ in $(0, x)$,

one in $(x, x + dx)$,

$k - j - 1$ in (x, y),

one in $(y, y + dy)$,

$n - k$ in $(y, 1)$

now multiply the probabilities to obtain the assertion

General case

Use $X_k = F^{-1}(U_k)$ to show:

Theorem 2 *1. The joint density of*

$(X_{1:n}, X_{2:n}, \ldots, X_{n:n})$ *is given by*

$$f(x_{(1)}, x_{(2)}, \ldots, x_{(n)}) = n! \prod_{j=1}^{n} f(x_{(j)}),$$

for $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$

2. The density of $X_{k:n}$ *is given by*

$$f_{(k)}(x) = \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1} f(x)(1 - F(x))^{n-k}$$

3. *For* $j < k$ *the joint density of* $(X_{j:n}, X_{k:n})$ *is given by*

$$\begin{aligned}
& f_{(j,k)}(x,y) \\
&= \frac{n!}{(j-1)!(k-j-1)!(n-k)!} \\
& \quad F(x)^{j-1} f(x) (F(y) - F(x))^{k-j-1} \\
& \quad (1 - F(y))^{n-k} f(y)
\end{aligned}$$

Further reading:

B.D. Ripley, What is Statistics? Simple Summaries and Plots, at

`http://www.stats.ox.ac.uk/ ripley/StatMethods/Lect1.pdf`