Serik Sagitov, Chalmers Tekniska Högskola, March 9, 2013

Chapter 10. Summarizing data

1 Empirical probability distribution

Population cumulative distribution function $F(x) = \mathrm{P}(X \le x)$. For an IID sample $(X_1, \dots, X_n)$ define

$$\boxed{\text{Empirical cdf } F_n(x) = \text{proportion of } X_i \le x}$$

For a fixed x the sample proportion $F_n(x)$ is an unbiased and consistent estimate of the population proportion $F(x)$.
After the sample is collected $F_n(x)$ is a cdf with mean $\bar{X}$ and variance $\frac{n-1}{n}s^2$.

Lifelength T with cdf $F(t) = \mathrm{P}(T \le t)$ and pdf $f(t) = F'(t)$.

$$\boxed{\text{Survival function } S(t) = \mathrm{P}(T > t) = 1 - F(t)}$$

Empirical survival function $S_n(t) = 1 - F_n(t)$ is the proportion of the data greater than t.

$$\boxed{\text{Hazard function } h(t) = f(t)/S(t)}$$

Mortality rate at age t: as δ tends to zero, $\mathrm{P}(t < T \le t + \delta | T \ge t) = \frac{F(t+\delta)-F(t)}{S(t)} \sim \delta \cdot h(t)$.
It is also 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 infected with tubercle bacillus, p. 349-353: 5 treatment and one control group. Fig 10.2: survival function. Fig 10.3: log-survival function.

The flat hazard function $h(t) = \lambda$ corresponds to the Exp(λ) model with $f(t) = \lambda e^{-\lambda t}$ and $S(t) = e^{-\lambda t}$.
Weibull(γ, λ) distribution, scale parameter $\lambda > 0$ and shape parameter $\gamma > 0$:
$f(t) = \lambda\gamma t^{\gamma-1}e^{-\lambda t^\gamma}$, $t \ge 0$, $S(t) = e^{-\lambda t^\gamma}$, $h(t) = \lambda\gamma t^{\gamma-1}$

2 Density estimation

Histogram: plot observed counts O_j for cells of width h. Small h - ragged histogram, large h - obscured histogram, find a balanced h.
Scaled histogram: plot $f_h(x) = \frac{1}{nh}O_j$ for x in cell j to ensure $\int f_h(x)dx = 1$.
Kernel density estimate with bandwidth h produces a smooth curve

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

Example. Male heights.
If hm is a column of 24 male heights, the for a given bandwidth h the following matstat code produces a plot for the 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)
```

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

17:056678899
18:0000112346
19:229

3 Q-Q plots

p-quantile of a distribution $x_p = F_{-1}(p)$, $0 \leq p \leq 1$

Quantile x_p cuts off proportion p of smallest values
$\mathrm{P}(X \leq x_p) = F(x_p) = F(F_{-1}(p)) = p$
Ordered sample $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$
$F_n(X_{(k)}) = \frac{k}{n}$ and $F_n(X_{(k)} - \epsilon) = \frac{k-1}{n}$

$X_{(k)}$ is the empirical $(\frac{k-0.5}{n})$-quantile

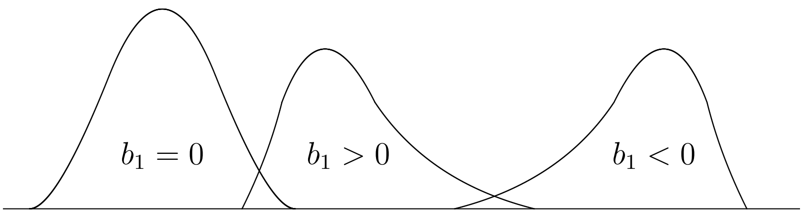
Two samples $(X_1, \ldots, X_n)$, $(Y_1, \ldots, Y_m)$
test H_0: two PDs are equal
by Q-Q plot = plot Y-quantiles against X-quantiles
Accept H_0 if the scatter plot is close to the bisector
equal quantiles = equal distributions

Linear model: $Y = a + b \cdot X$ in distribution

$\mathrm{P}(X \leq x) = \mathrm{P}(Y \leq a + bx)$
Linear model implies linear Q-Q plot $y_p = a + bx_p$

Normal probability plot. To test visually the normality hypothesis H_0: PD = $\mathrm{N}(\mu, \sigma^2)$ with unspecified parameters plot the normal quantiles $\Phi_{-1}(\frac{k-0.5}{n})$ against $X_{(k)}$.
Accept H_0 with $\mu = a$, $\sigma = b$, if the scatterplot is close to the straight line $x = a + by$.
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: $b_1 = \frac{1}{s^3 n} \sum (X_i - \bar{X})^3$



Kurtosis $b_2 = \frac{1}{s^4 n} \sum (X_i - \bar{X})^4$, normal data $b_2 = 3$

Leptokurtic distribution: $b_2 > 3$ (heavy tails). Platykurtic distribution: $b_2 < 3$ (light tails).

Example. Male heights. Summary statistics: $\bar{X} = 181.46$, $\hat{M} = 180$, $b_1 = 1.05$, $b_2 = 4.31$.
Heights of adult males are positively skewed: P(height of a random male < the average) > 50%.

4 Measures of location

Central point of a distribution: either population mean μ, or mode, or median M defined as $M = x_{0.5}$, if distribution is continuous.

Population median M: $\mathrm{P}(X < M) = \mathrm{P}(X > M)$

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 median $\hat{M}$ is a robust estimate, that is insensitive to outliers, while the sample mean $\bar{X}$ is sensitive to outliers.

Nonparametric sign test
Given an iid sample test H_0: $M = M_0$ against the two-sided alternative H_1: $M \neq M_0$. No parametric model is assumed. The sign test statistic $Y = \sum_{i=1}^{n} I(X_i \leq M_0)$ counts the number of observations below the null hypothesis value. Under the null hypothesis

$$\mathrm{P}(X_{(k)} < M_0 < X_{(n-k+1)}) = \mathrm{P}(k \leq Y \leq n - k),$$

and $Y \in \mathrm{Bin}(n, 0.5)$.

$(X_{(k)}, X_{(n-k+1)})$ = nonparametric $100 \cdot \mathrm{P}(k \leq Y \leq n - k)$% CI for the population median.

Reject H_0 if M_0 falls outside the corresponding confidence interval $(X_{(k)}, X_{(n-k+1)})$.

Example. For $Y \in \mathrm{Bin}(n, 0.5)$ with $n = 25$ we have

for $k =$	6	7	8	9	10	11	12
$\mathrm{P}(k \leq Y \leq n - k) =$	99.6	98.6	95.7	89.2	77.0	57.6	31.0

Thus $(X_{(8)}, X_{(16)})$ is a 95.7% CI for the median.

Trimmed means
Measures of location for the 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.

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

Nonparametric bootstrap
IID sampling from the empirical distribution = sampling with replacement from $x_1, \ldots, x_n$.
Simulate many new samples of size n to get an idea of the sampling distribution of an estimate like trimmed mean, sample median, s.

5 Measures of dispersion

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

interquartile range IQR $= x_{0.75} - x_{0.25}$
MAD = median of abs dev $|X_i - \hat{M}|$, $i = 1, \ldots, n$.

Three estimates of σ in $\mathrm{N}(\mu, \sigma^2)$: s, $\frac{\mathrm{IQR}}{1.35}$, $\frac{\mathrm{MAD}}{0.675}$

In the $\mathrm{N}(\mu, \sigma^2)$ case IQR $= (\mu + \sigma\Phi_{-1}(0.75)) - (\mu + \sigma\Phi_{-1}(0.25)) = 1.35\sigma$, because $\Phi_{-1}(0.75) = 0.675$.
Moreover, MAD $= 0.675\sigma$, since $\mathrm{P}(|X - \mu| \leq 0.675\sigma) = 0.5$.

Boxplot

box center = median
upper edge of the box = upper quartile (UQ)
lower edge of the box = lower quartile (LQ)
upper whisker end = {max data point ≤ UQ + 1.5 IQR}
lower whisker end = {min data point ≥ LQ – 1.5 IQR}
dots = {data ≥ UQ + 1.5 IQR} and {data ≤ LQ – 1.5 IQR}

Convenient to compare different samples. See for example Fig 10.14, p.374: daily SO_2 concentration data.