Serik Sagitov, Chalmers Tekniska Högskola, April 13, 2005

Chapter 10. Summarizing data

1. Empirical probability distribution

IID sample $(X_1, \dots, X_n)$ with population cdf $F(x)$

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

For fixed x sample proportion $F_n(x)$ is an unbiased
and consisitent estimate of pop. proportion $F(x)$

After the sample is collected
$F_n(x)$ is a cdf with mean $\bar{X}$ and variance $\frac{n-1}{n}s^2$

Lifelengths

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

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

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

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

Mortality rate at age t
$\mathrm{P}(t < T \le t + \delta | T \ge t) \approx \delta \cdot h(t)$

The negative of the slope of the log survival function
$h(t) = -\frac{d}{dt} \log S(t)$

Exp(λ): flat hazard function $h(t) = \lambda$

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

Weibull(γ, λ) distribution on $[0, \infty)$

$f(t) = \lambda\gamma t^{\gamma-1}e^{-\lambda t^\gamma}$, $S(t) = e^{-\lambda t^\gamma}$, $h(t) = \lambda\gamma t^{\gamma-1}$

scale parameter $\lambda > 0$ and shape par. $\gamma > 0$

Ex 1: 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

Density estimation

Histogram: observed counts O_j for cells of width h

small h - ragged histogram

large h - obscured histogram, find a balanced h

Scaled histogram

$f_h(x) = \frac{1}{nh} O_j$ for x in cell j to ensure $\int f_h(x)dx = 1$

Kernel density estimate with bandwidth h

produces a smooth curve

$$\boxed{f_h(x) = \frac{1}{nh} \Sigma\, \phi(\tfrac{x-X_i}{h}), \text{ where } \phi(x) \text{ is the N(0,1) pdf}}$$

Ex 2: male heights

If hm is a column of 24 male heights

```
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

17:056678899

18:0000112346

19:229

distribution shape plus the numerical information

2. Q-Q plots

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

Quantile x_p cuts off proportion p of smallest values

$\mathrm{P}(X \le x_p) = F(x_p) = F(F_{-1}(p)) = p$

Ordered sample $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$

$F_n(X_{(k)}) = \frac{k}{n}$ and $F_n(X_{(k)} - \epsilon) = \frac{k-1}{n}$

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

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

$$\boxed{\text{Linear model: } Y = a + b \cdot X \text{ in distribution}}$$

$\mathrm{P}(X \le x) = \mathrm{P}(Y \le a + bx)$

Linear model implies linear Q-Q plot $y_p = a + bx_p$

Normal probability plot

To test H_0: PD = N(μ, σ^2) with unspecified parameters

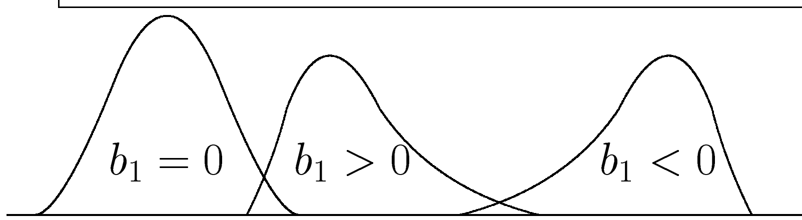
plot the normal quantiles $\Phi_{-1}(\frac{k-0.5}{n})$ against $X_{(k)}$

Accept H_0 with $\mu = a$, $\sigma = b$

if the scatterplot is close to the line $x = a + by$

Light tails profile and heavy tails profile

$$\boxed{\text{Coefficient of skewness: } b_1 = \frac{1}{s^3 n}\Sigma(X_i - \bar{X})^3}$$



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

Leptokurtic distribution: $b_2 > 3$ heavy tails

platykurtic distribution: $b_2 < 3$ light tails

Ex 2: male heights

$\bar{X} = 181.46$, $\hat{M} = 180$, $b_1 = 1.05$, $b_2 = 4.31$

Heights of adult males are positively skewed

P(height of a random male $<$ the average) $>$ 50%

3. Measures of location

Central point of a distribution:

population mean μ, mode or median M

$M = x_{0.5}$ if distribution is continuous

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

Sample median $\hat{M} = X_{(k)}$, if $n = 2k - 1$

$\hat{M} = \frac{X_{(k)} + X_{(k+1)}}{2}$, if $n = 2k$

$\hat{M}$ is a robust estimate = insensitive to outliers

sample mean $\bar{X}$ is sensitive to outliers

Nonparametric sign test

Test H_0: $M = M_0$ against two-sided H_1: $M \neq M_0$

sign test statistic: $Y = \sum I(X_i \leq M_0)$

null distribution $Y \in \mathrm{Bin}(n, 0.5)$

Reject H_0 if M_0 falls outside $(X_{(k)}, X_{(n-k+1)})$

where k is such that $\mathrm{P}(Y < k) = \frac{\alpha}{2}$, $Y \in \mathrm{Bin}(n, 0.5)$

$$\boxed{(X_{(k)}, X_{(n-k+1)}) = \text{nonparametric CI for } M}$$

$n = 25$ and $k =$	6	7	8	9	10	11	12
$100(1 - \alpha)\%$	99.6	98.6	95.6	89.2	77.0	57.6	31.0

Trimmed means

Measures of location for the central portion of the data

$$\boxed{\alpha\text{-trimmed mean } \bar{X}_\alpha = \text{sample mean without } \tfrac{n\alpha}{2} \text{ smallest and } \tfrac{n\alpha}{2} \text{ largest observations}}$$

Ex 2: male heights $\bar{X}_{0.4}$=180.36

$$\boxed{\text{When summarizing data compute several measures of location and compare the results}}$$

Nonparametric bootstrap

IID sampling from the empirical distribution

= sampling with replacement from $x_1, \ldots, x_n$

simulate many new samples of size n

Used to view the sampling distribution of an estimate

like trimmed mean, sample median, s

4. 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 $|X_i - \hat{M}|$, $i = 1, \ldots, n$

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

$\mathrm{IQR} = (\mu + \sigma\Phi_{-1}(0.75)) - (\mu + \sigma\Phi_{-1}(0.25)) = 1.35\sigma$

$\mathrm{P}(\frac{|X-\mu|}{\sigma} \leq z) = 0.5$ for $z = \Phi_{-1}(0.75) = 0.675$

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 $\leq$ UQ + 1.5 IQR}

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

dots = {data $\geq$ UQ + 1.5 IQR} and {data $\leq$ LQ – 1.5 IQR}

Convenient to compair different samples

Fig 10.14, p.374: daily SO_2 concentration data