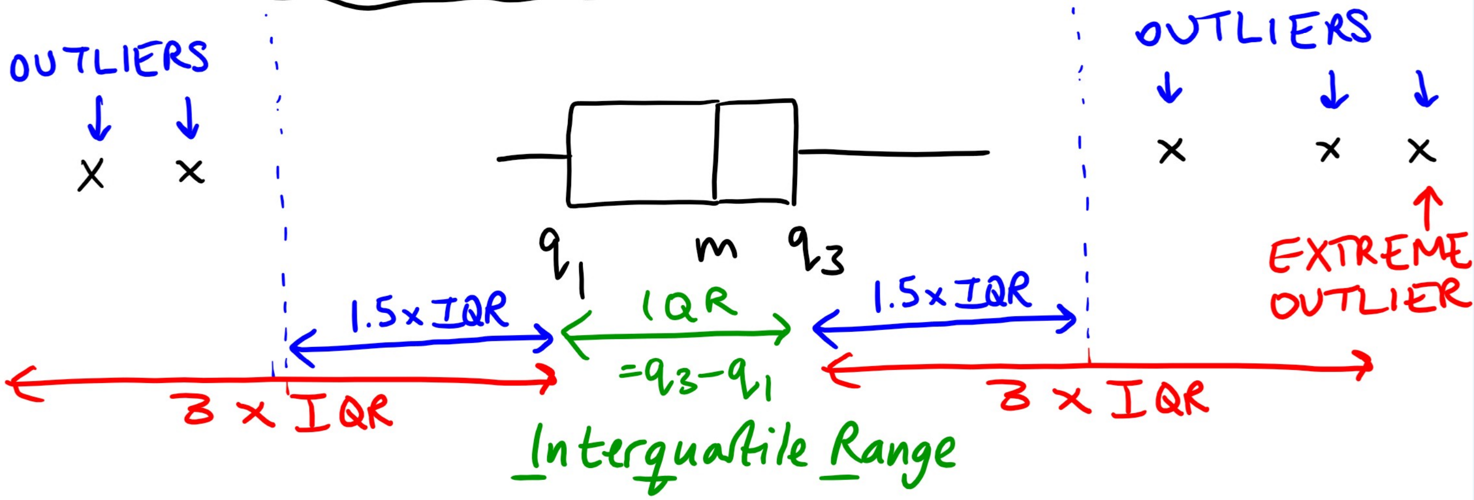
3Y03 – Probability and Statistics for Engineering

WS19 Lecture 19

Last Time

Box (& Whisker) Plots



m = median = middle value (n odd) or average of 2 middle values (n even)

q_1 = 1st quartile = median of the values below m

q_3 = 3rd quartile = median of the values above m.

Example Data Set 1 $n = 12$

$$10\ \ 11\ \ \underbrace{18\ \ 19}_{q_1 = 18.5}\ \ 23\ \ \underbrace{31\ \ 33}_{m=32}\ \ 39\ \ \overbrace{50\ \ 51}^{q_3 = 50.5}\ \ 72\ \ 105$$

$IQR = q_3 - q_1 = 50.5 - 18.5 = 32$

$1.5 \times IQR = 1.5 \times 32 = 48$

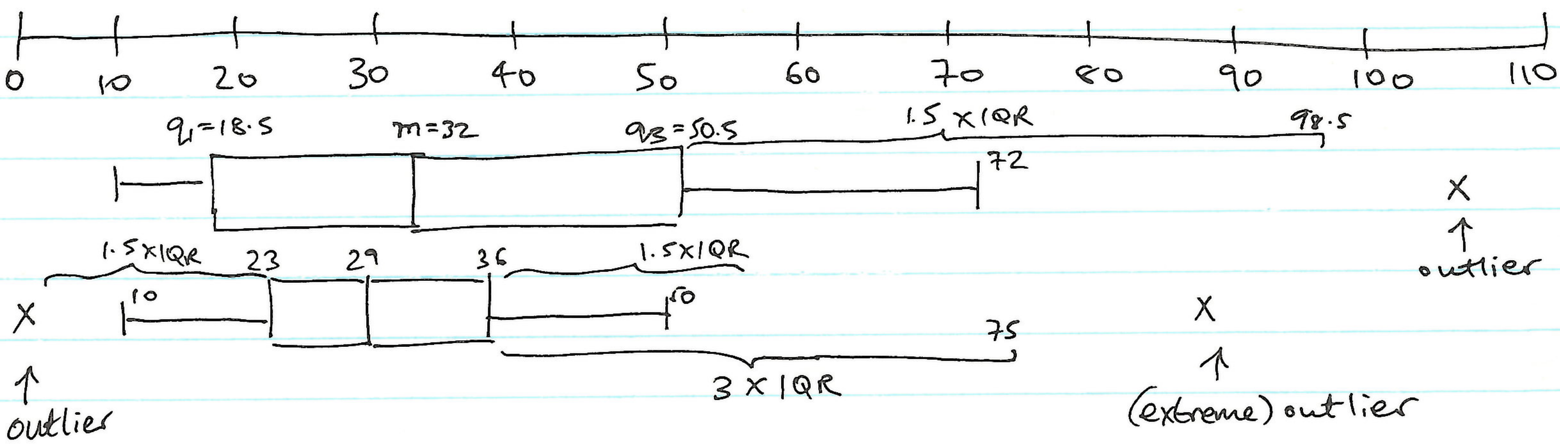
$3 \times IQR = 3 \times 32 = 96$

① 3Y03 L19 WS19

Data Set #2 $\quad n = 11 \quad$ 1 10 $\underset{q_1}{23}$ 25 27 $\underset{m}{29}$ 30 35 $\underset{q_3}{36}$ 50 89

$IQR = q_3 - q_1 = 36 - 23 = 13$

$1.5 \times IQR = 19.5 \qquad 3 \times IQR = 39$



6.7 Probability Plot

- histograms indicate underlying distribution
- in some situations you want to assume a certain distr.

We hypothesize a prob. distr. (with pdf $f(x)$ & cdf $F(x)$) & check hypothesis with a prob. plot.

Sample $\{x_1, \dots, x_n\}$

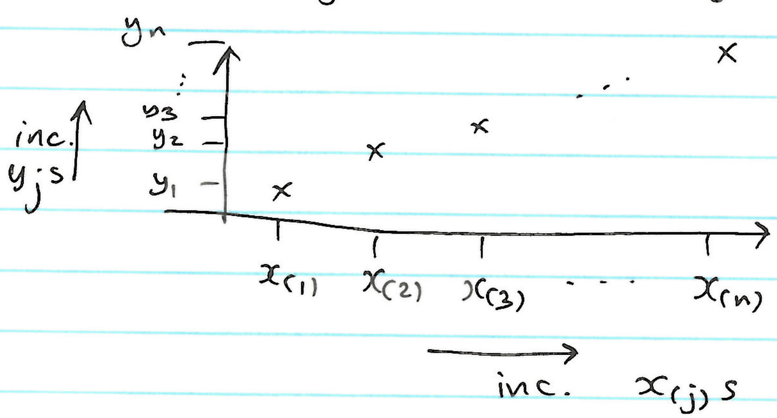
Rank / Order the sample & rename: $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$

<u>Idea</u> $x_{(j)}$ (jth observation in the list) should approximate the $100\left(\frac{j}{n}\right)$th percentile

Want y_j s.t.

$$P(X \leq y_j) = F(y_j) = \frac{j - 0.5}{n}$$

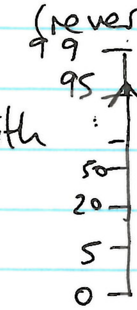
Plot y_j s against $x_{(j)}$ s — should be a straight line if our hypothesis is correct



Subjective judgment as to whether the line is straight or not!

Usually for us the hypothesis will be Normal distr.
i.e. $F(x) = \Phi(x)$ & values y_j are z_j -values from (reverse) normal table.

There is "normal prob. paper" with



— get to plot $\frac{j-0.5}{n}$ against $x_{(j)}$s.

Still the case that straight line = good hypothesis.

Even if line is not straight, plot can still indicate features of prob. distr. e.g. symmetry / skew — see textbook diagrams

7. Point Estimation of Parameters

Recall: sample $\{x_1, \ldots, x_n\}$ (data)
thought of as realizations of independent r.vs. with common underlying distribution $\{X_1, \ldots, X_n\}$ (also called random sample)

$\{x_1, \ldots, x_n\}$ ~~repre~~ is one set of possible values of $\{X_1, \ldots, X_n\}$ with some samples more likely ~~due~~ depending on prob. distr.

③ 3J03 L19 WS19

2 big concepts of statistical analysis/inference:

{ ① Point Estimation = parameter estimation
② Hypothesis Testing

Parameter : some feature/function of underlying distr.
e.g. pop. mean μ, pop. variance σ^2

usually little Greek letter: default is θ

A statistic is a function of $\{X_1, \dots, X_n\}$
(Remember – any function of r.v.s. is also a r.v.)

e.g. $\bar{X}$ = sample mean S^2 = sample variance

The prob. distr. of a statistic is called a sampling distribution

Estimat~~ion~~ Given a parameter θ

- an estimator is a statistic used to estimate θ
  ↳ notation often $\hat{\Theta}$ $\left[= h(X_1, \dots, X_n) \right]$ some h.

e.g. $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$ an estimator for μ
$S^2 = \frac{1}{n-1}\left(X_1^2 + \dots + X_n^2 - n\bar{X}^2\right)$ an estimator for σ^2

After we take a sample $\{x_1, \dots, x_n\}$ & calculate a value for $\hat{\Theta}$, ~~this~~ the value it takes is called the point estimate for θ, denoted $\hat{\theta}$

e.g. $\bar{x}$ is a point estimate for μ
s^2 σ^2