3Y03 - 3J04 PROBABILITY & STATISTICS FOR (CIVIL) ENGINEERING

C01 - Lecture 19

Last time: VISUAL DATA DISPLAYS

In particular: HISTOGRAMS based on Frequency Distributions

height of blocks = (relative) frequency in corresponding bin

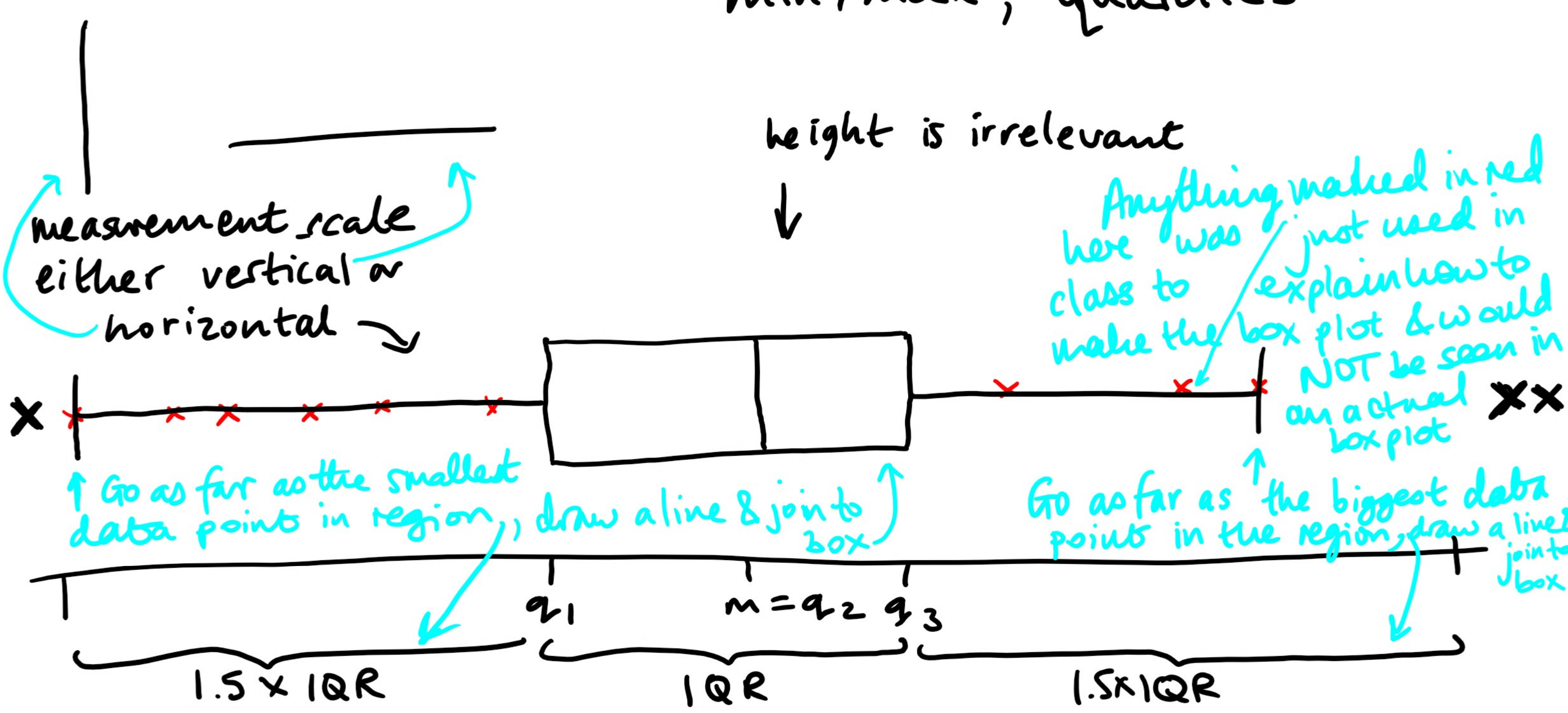
Range divided into (equal-sized) "bins"

If not equal, then height = $\dfrac{\text{frequency}}{\text{width}}$

ALSO
q_1 = 1st quartile: 25% of data points below
$m = q_2$ = 2nd quartile: 50% of data points below
q_3 = 3rd quartile: 75% of data points below.

6.4 Box (and Whisker) Plots

- combines different features of data set (sample) into one graph → min/max, quartiles



IQR = Interquartile Range = $q_3 - q_1$

In region either side of the box (from q_1 to q_3) draw "whiskers" out to most extreme data points lying in the region — these regions are $1.5 \times IQR$ either side

Then mark in all data points outside these two regions — these are called outliers

Definition of outlier: data points outside of $1.5 \times IQR$ from box

(1.5 — so that ~ 1% data points are outliers)

If a data point is $> 3 \times IQR$ outside the box, it is called an extreme outlier.

Can compare data sets (samples) by side-by-side box plots.

Data Set 1 $n = 12$.

$$10,\ 11,\ \underbrace{16,\ 19}_{q_1 = 17.5},\ 23,\ \underbrace{31,\ 33}_{m = \frac{31+33}{2} = 32 = q_2},\ 39,\ \underbrace{50,\ 51}_{q_3 = 50.5},\ 72,\ 105$$

$17.5 - 49.5 = -32$

$50.5 + 49.5 = 100$

$IQR = 50.5 - 17.5 = 33$

$1.5 \times IQR = 1.5 \times 33 = 49.5$

$3 \times IQR = 99$

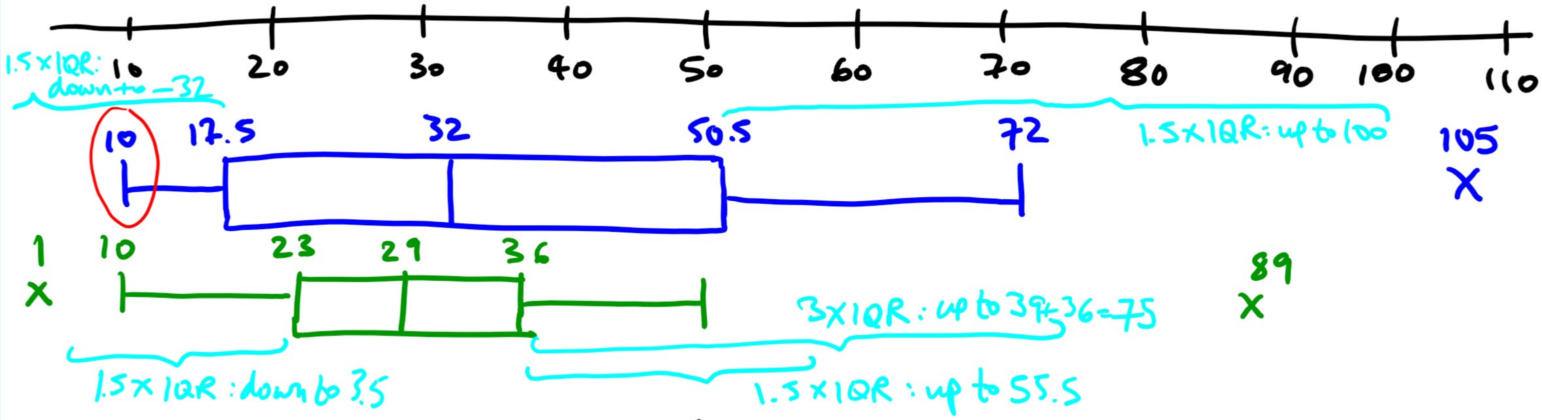
Data Set 2 $n = 11$

$$1,\ 10,\ \underset{q_1}{23},\ 25,\ 27,\ \underset{m}{29},\ 30,\ 35,\ \underset{q_3}{36},\ 50,\ 89$$

$23 - 19.5 = 3.5$

$IQR = 36 - 23 = 13$, $1.5 \times IQR = 19.5$, $3 \times IQR = 39$

$36 + 19.5 = 55.5$



Outliers: Data Set 1 : 105

Data Set 2 : 1, 89 ← extreme outlier $(36+33<89)$

6.7 Probability Plots

— indicator of underlying probability distribution

↓

Histograms help indicate underlying prob. distr. but only reliable for large sample size n

---

Idea : We hypothesize the prob. distr. with pdf $f(x)$ (and this gives us cdf $F(x)$)

We check our guess with a probability plot:

How to make: Sample is $\{x_1, \ldots, x_n\}$

→ Rank sample & rename so we have

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$$

So $x_{(j)}$ is the jth sample point in numerical order

Idea is that $x_{(j)}$ should approximate the

$(100)\left(\frac{j}{n}\right)$ th percentile

So find, for each j, the number y_j so that

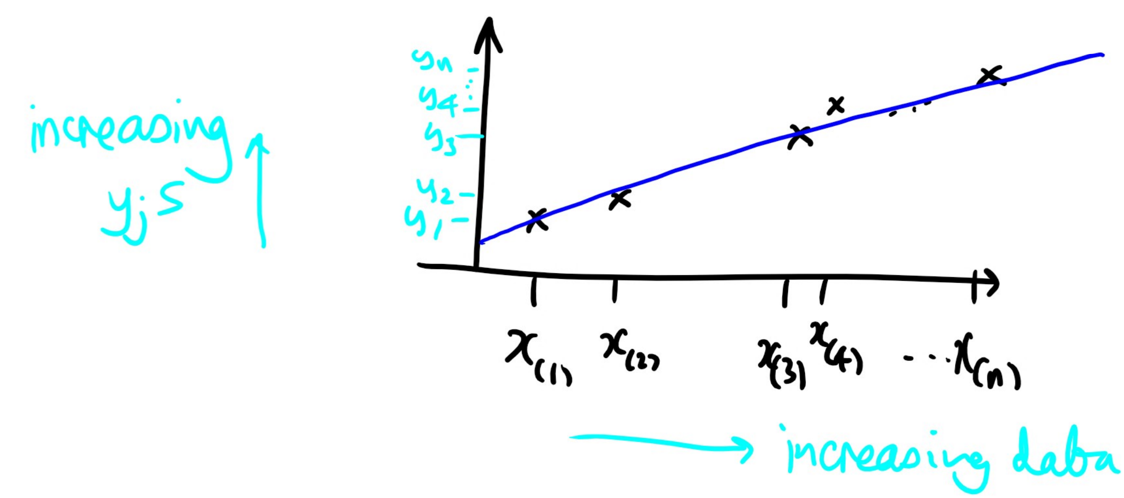
$$\underbrace{\frac{j-0.5}{n}}_{\text{correction factor as sample is finite}} = P(X \le y_j) = F(y_j)$$

X: underlying population r.v.

F: the hypothesized F

Then plot y_js against $x_{(j)}$s

— if guess is a good one then we get a straight line



Subjective judgment as to whether or not these points lie on a straight line.