# 3Y03 - Probability and Statistics for Engineering

WS19 Lecture 18

Yesterday we started on **Visual Representations of Data**

## Stem & Leaf Diagram ; can illustrate one data set or can compare 2 data sets by putting diagrams back to back with same stems.

↑ All but last digit, unless splitting up the arising categories further would be more meaningful.

↑ Last digit

| BC | | ON |
|---|---|---|
| 6 | 7. | |
| 6 4 4 3 2 1 0 | 8. | 0 6 |
| | 9. | 1 1 2 3 |
| | 10. | 2 4 |
| ↑ leafs | ↑ stems | ↑ leafs |

- Shows general shape of distribution
- Not practical with large amounts of data (by hand)

- Allows us to read off:

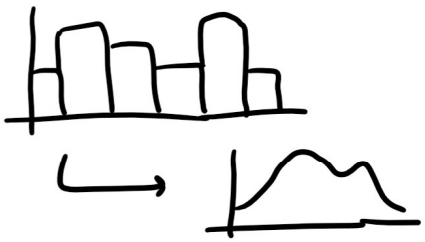
2nd quartile → • Median m: ~ 50% data points above m, ~50% below

- 1st quartile $q_1$: ~75% " " " , ~25% below
- 3rd quartile $q_3$: ~25% " " " , ~75% below
- nth percentile : ~(100-n)% " " " ~ n% below

## 6-3 Frequency Distributions & Histograms

- group data into "bins" / "class intervals" / "cells"
  - usually of equal width
- count frequency in each bin

How many bins? — Too many: lose shape
— Too few: lose detail



If $n$ data points, good rule is $\sqrt{n}$ bins.

As $n$ increases, $\sqrt{n}$ increases so bin width decreases

In limit, as $n \longrightarrow \infty$, frequency distribution → underlying pdf $f(x)$.

**Example** Building permits / year.

20 years = $n$

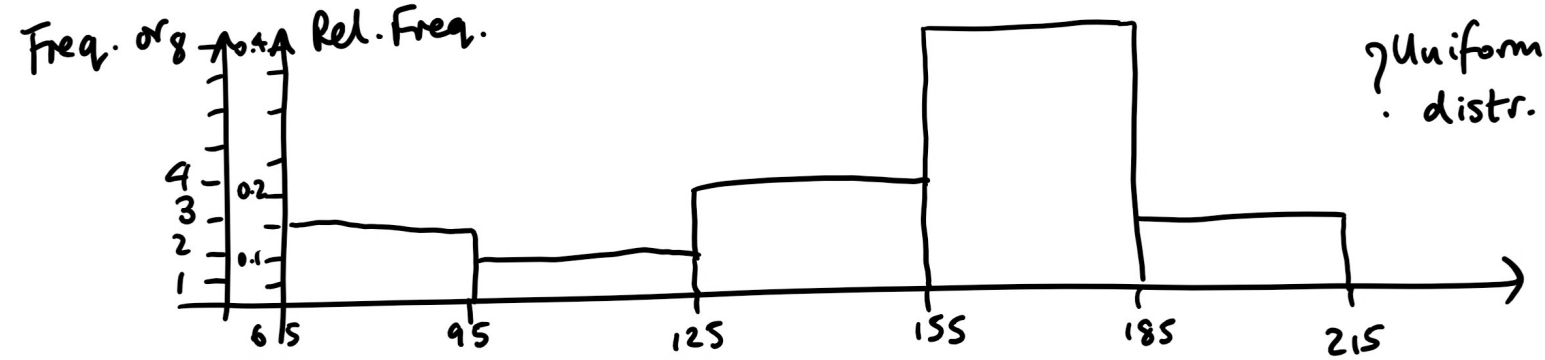
$\#\text{ bins} \sim \sqrt{n} = 4.47$

So, say, choose $\#\text{ bins} = 5$.

Range of data: max = 213, min = 73 ] ~ bins of width 30

— See separate pdf file on course website or the table at the bottom of this document.

| Bins | $65 \leq x < 95$ | $95 \leq x < 125$ | $125 \leq x < 155$ | $155 \leq x < 185$ | $185 \leq x < 215$ |
|---|---|---|---|---|---|
| Frequency | 3 | 2 | 4 | 8 | 3 |
| Relative Frequency | $3/20 = 0.15$ | $2/20 = 0.1$ | 0.2 | 0.4 | 0.15 |
| Cumulative Frequency | 3 | 5 | 9 | 17 | 20 |



↳ Rel. Freq. Distribution : total area of blocks equals 1

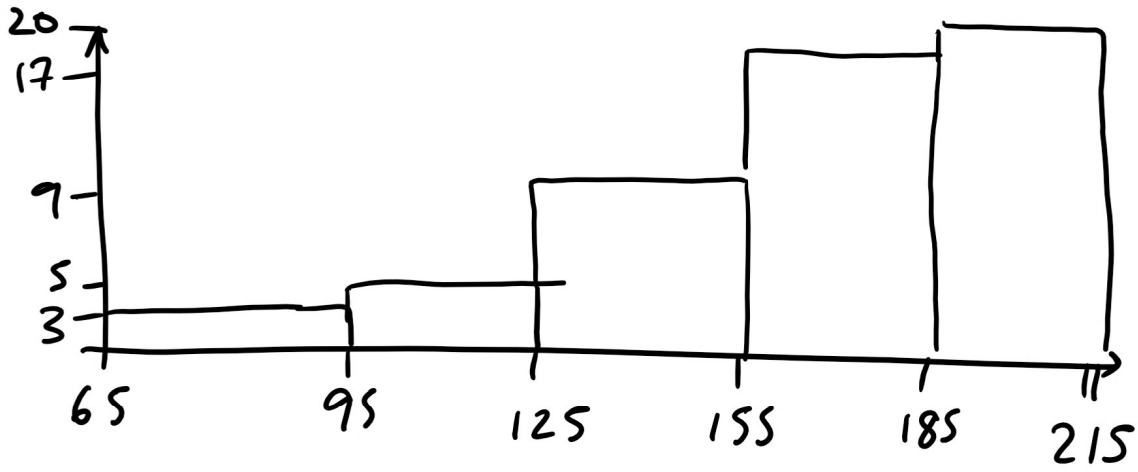
↑ Equal width

→ converges to $f(x)$ as $n \to \infty$

If unequal width make the area of the blocks = frequency

$$\text{So height} = \frac{\text{frequency}}{\text{width}}.$$

We can also plot an analogue to the cdf $F(x)$ by plotting cumulative frequency against bins:



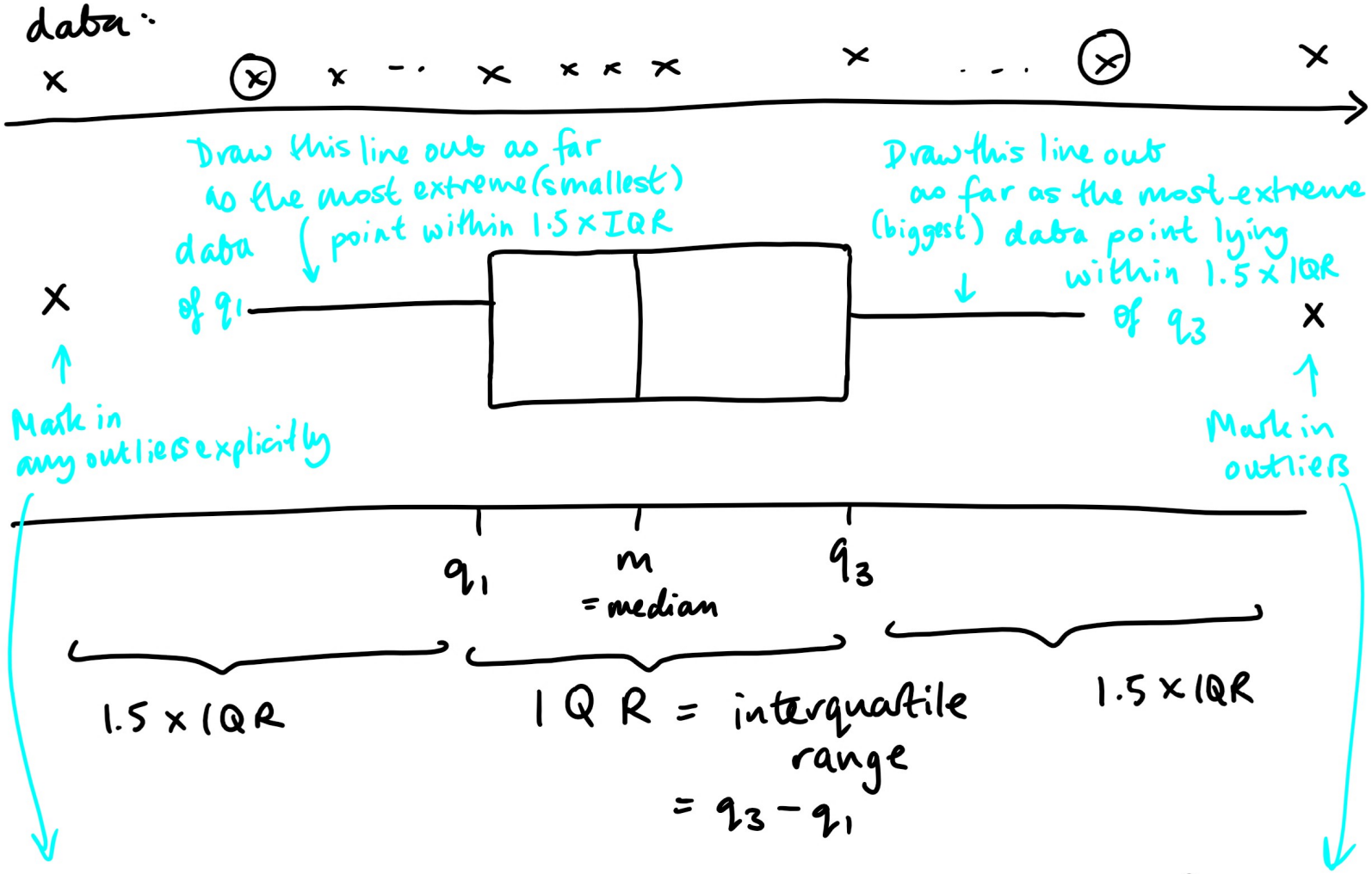

As $n \to \infty$, this should converge to the cdf $F(x)$.

Also a perfectly valid representation of the data: to form a histogram using other categories eg. year v. # permits

# 6.4 Box Plots (Box & Whisker Plots)

- combines different features of data into 1 graphic
(e.g. min/max, & quartiles)

(height of what we do not important if horizontal)




Outlier: Any data point lying $> 1.5 \times IQR$ from box ends i.e. $q_1, q_3$

Extreme Outlier: Any data point lying $> 3 \times IQR$ from box ends

Can compare data sets using side-by-side box plots

Data Set 1 $n = 12$

10 11 18 19 23 31 | 33 39 50 51 72 105

$q_1 = 18.5$ $m = 32$ $q_3 = 50.5$

To compute $q_1$ (& $q_3$), repeat process for finding median, i.e. count # data points below (& above) $q_3$ / the value of $m$ & if odd, take $q_1$ middle value, if even average 2 middle values. T.B.C.

Here is the table for the Example about building permits under 6-3 above.

| Industrial Building Permits Issued in Hamilton by Year | | | | |
|---|---|---|---|---|
| | | | | |
| YEAR | PERMITS ISSUED | | | |
| 1998 | 86 | | | |
| 1999 | 90 | | | |
| 2000 | 73 | | | |
| 2001 | 170 | | | |
| 2002 | 128 | | | |
| 2003 | 140 | | | |
| 2004 | 112 | | | |
| 2005 | 122 | | | |
| 2006 | 188 | | | |
| 2007 | 158 | | | |
| 2008 | 142 | | | |
| 2009 | 172 | | | |
| 2010 | 157 | | | |
| 2011 | 213 | | | |
| 2012 | 146 | | | |
| 2013 | 178 | | | |
| 2014 | 183 | | | |
| 2015 | 183 | | | |
| 2016 | 172 | | | |
| 2017 | 193 | | | |