3Y03 – 3J04 Probability & Statistics for (Civil) Engineering

C01 – Lecture 18

Last time: Numerical Summaries of Data

Sample: $\{x_1, \ldots, x_n\}$

Sample... Mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

... Median m = middle value (interpolate if necessary)

... Mode most commonly occurring value

... Variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$

... Range $\max\{x_i\} - \min\{x_i\}$ ... Standard Deviation $s = +\sqrt{s^2}$

Example Prices of Cannabis sold for medical useage in ON, 2010–2017.

year	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	x_i^2	x_i
2010	9.07	-0.1625	0.02640625	82.2649	10.37
2011	9.16	-0.0725	0.00525625	83.9056	10.18
2012	9.31	0.0775	0.00600625	86.6761	9.31
2013	10.37	1.1375	1.29390625	107.5369	9.16
2014	10.18	0.9475	0.89775625	103.6324	9.11
2015	9.11	-0.1225	0.01500625	82.9921	9.07
2016	8.64	-0.5925	0.35105625	74.6496	8.64
2017	8.02	-1.2125	1.47015625	64.3204	8.02
	73.86		4.06555	$\frac{1}{7}\left(685.978 - 8\cdot(9.23)^2\right)$	
$n = 8$					
	$\boxed{9.2325}$		0.580792857	0.580792857	9.135
	=Mean $\bar{x}$		=Var s^2	from shortcut method	= Median m
			0.762097669		
			=Std Dev s		

To find median: $\frac{9.16 + 9.11}{2}$

Mode = every data point

Range = $10.37 - 8.02 = 2.35$.

Stem & Leaf Diagram

→ Divide up the numerical values of the data into 2 parts :

stem + leaf

all but last digit

last digit

<u>Example</u> Price (from above) to nearest 10c

\$9.0, \$9.2, \$9.3, \$10.4, \$10.2, \$9.1, \$8.6, \$8.0

Stem	Leaf
\$8.	0 6
\$9.	0 1 2 3
\$10.	2 4

We can also make back-to-back stem & leaf diagrams

<u>Example</u> Compare above with data set from BC:

\$8.2, \$8.3, \$8.4, \$8.4, \$8.0, \$8.1, \$8.6, \$7.6

BC Leaf	Stem	ON Leaf
6	7.	
6 4 4 3 2 1 0	8.	0 6
	9.	0 1 2 3
	10.	2 4

Also can have split stems e.g.:

"Upper" & "Lower" segments of each stem

BC		ON
6	7.U	
4 4 3 2 1 0	8.L	0
6	8.U	6
	9.L	0 1 2 3
	9.U	
	10.L	2 4

→ Not useful/practical with large amounts of data

→ Shows general shape of distribution

→ Also allows us to read off <u>quartiles</u> in data or <u>percentiles</u> in data:

1st quartile q_1 : ~25% data points below here

2nd quartile q_2 : ~50% " " " "

= median m

3rd quartile q_3 : ~ 75% " " " "

nth percentile : ~ n % " " " "

Frequency Distributions & Histograms

Group data into "class intervals" or "cells" or "bins"

↳ usually of equal width

& count the frequency of data in each bin

Example

industrial building permits issued by City of Hamilton every year since 1998

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

- Free to choose interval length & # bins (bin size) ← must cover all data points
- Usually good rule in $\sim \sqrt{n}$ bins

Here $n = 20$, so # bins $\sim \sqrt{20} \approxeq 4.5$

So let's make 5 bins

Want to cover range 73 to 213 $140/5 = 28$

So here's one possibility: with bin width $= 30 \approx 28$

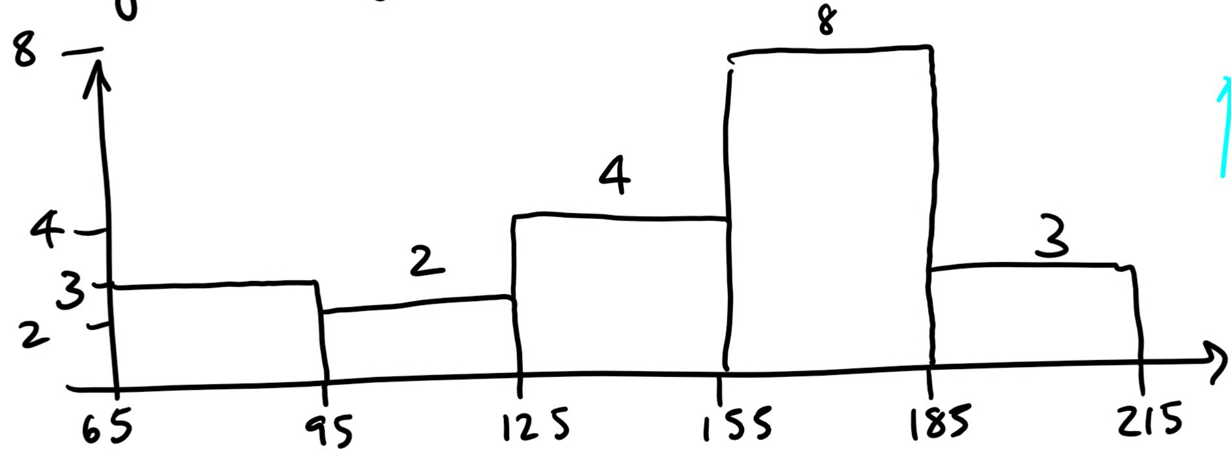
data points from list in each bin

Freq. as proportion of total

data points in all the bins up to & including current one

Bin	$65 \le x < 95$	$95 \le x < 125$	$125 \le x < 155$	$155 \le x < 185$	$185 \le x$ (<215)
Freq.	3	2	4	8	3
Relative Freq.	$\frac{3}{20} = 0.15$	$\frac{2}{20} = 0.1$	0.2	0.4	0.15
Cumulative Freq.	3	5	9	17	20

Histogram : graphical representation



height = frequency (can also do the same with height = relative frequency)

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

As n increases, $\sqrt{n}$ = # bins increases & the histogram

of rel. freq.
converges as $n \to \infty$ ~~is~~ to the underlying probability density function $f(x)$

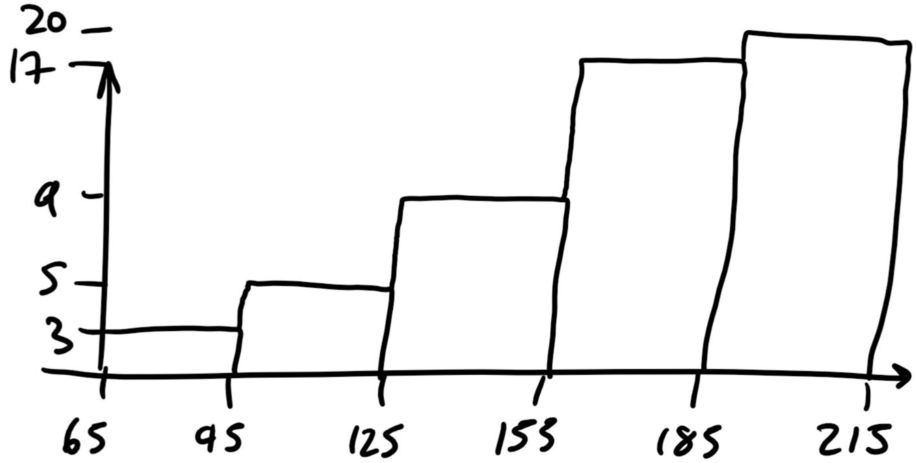
(With rel. freq histogram, total area of blocks = 1, which should be total area under $f(x)$)

If bins are NOT equal width, make AREA of blocks = frequency so now height of blocks is

$$\frac{\text{frequency}}{\text{width}}.$$

(AREA represents relative frequency — happens by accident plotting rel. freq. histogram if all bins same width)

We also have an analogue to cumulative distr. function $F(x)$: plot cumulative frequency against bins:



↑ height = cumulative freq.

Also get histograms using categories; here could make a different graphical representation of data by plotting # applications against year: