3Y03 - Probability and Statistics for Engineering

WS19 Lecture 17

Last time — Linear Functions of Random Variables

If $X_1, \dots, X_n$ are independent, then $Y = c_1X_1 + \dots + c_nX_n$ $(c_i \in \mathbb{R})$ has $E(Y) = c_1E(X_1) + \dots + c_nE(X_n)$ & $V(Y) = c_1^2V(X_1) + \dots + c_n^2V(X_n)$

If $X_1, \dots, X_n$ are normally distributed, then so is Y.

Today — Statistics & Data

↓

- real #s, observations, measurements

  e.g. different volume measurements of soda bottles

- take sample $\{x_1, \dots, x_n\}$ (measurements)

  from wider population (e.g. all bottles in a certain factory)

- want to infer information about the pop. from sample
- interested in some feature of setup e.g. average

  + using data to estimate this.

- each x_i is assumed to be a realization of a random variable X_i, where $X_1, \dots, X_n$ are independent and have same distribution

What kinds of information can we extract from data?

Measures of Location

Sample Mean $= \bar{x} = \frac{1}{n}(x_1 + \ldots + x_n)$

- estimates μ the underlying population mean i.e. the mean of each X_i

(Sample) Median : arrange data from smallest to biggest

↳ $m = \begin{cases} \text{middle value} & \text{if } n \text{ odd} \\ \text{average of two middle values} & \text{if } n \text{ even} \end{cases}$

- less vulnerable to extreme observations than the mean esp. if underlying distribution is not symmetric

(Sample) Mode = most commonly occurring data value. (can be multiple modes)

Measures of Variation

Sample Variance $s^2 = \frac{1}{n-1}\left[(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2\right]$

↑ $\frac{1}{n-1}$ not $\frac{1}{n}$: reason has to do with called "bias"

Sample standard deviation : $s = +\sqrt{s^2}$

Tedious to compute so have shortcut:

$$s^2 = \frac{1}{n-1}\left[(x_1^2 + \ldots + x_n^2) - n\bar{x}^2\right].$$

(Sample) Range $= r = \max(x_i) - \min(x_i)$

– very crude measure of spread esp. in n large

Example Maximum daily temperatures in Hamilton 1st Feb – 12th Feb 2019.

Date	Temp x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	x_i^2	x_i
01/02/2019	-9.2	-12.1167	146.8136	84.64	-9.2
02/02/2019	1.6	-1.31667	1.733611	2.56	-3.6
03/02/2019	9.5	6.583333	43.34028	90.25	-3
04/02/2019	14.3	11.38333	129.5803	204.49	-1.9
05/02/2019	10.2	7.283333	53.04694	104.04	0.3
06/02/2019	0.3	-2.61667	6.846944	0.09	1.4
07/02/2019	7.3	4.383333	19.21361	53.29	1.6
08/02/2019	8.1	5.183333	26.86694	65.61	7.3
09/02/2019	-3.6	-6.51667	42.46694	12.96	8.1
10/02/2019	-3	-5.91667	35.00694	9	9.5
11/02/2019	-1.9	-4.81667	23.20028	3.61	10.2
12/02/2019	1.4	-1.51667	2.300278	1.96	14.3
Sum	35		530.4167	632.5	1.5 = Median m
	= Sample Mean $\bar{x}$		48.2197	48.2197	
			=Sample Var s^2		
			6.94404		
			=Sample Std Dev s		

$n = 12$

$x_1 + \ldots + x_{12} \Rightarrow$ Sum $= 35$

$(x_1 - \bar{x})^2 + \ldots + (x_{12} - \bar{x})^2 = 530.4167$

$= x_1^2 + \ldots + x_{12}^2 = 632.5$

Average the values

$\frac{1.6 + 1.4}{2}$

$\frac{x_1 + \ldots + x_{12}}{12} = \frac{35}{12} = 2.916666667$ = Sample Mean $\bar{x}$

$\frac{530.4167}{11} = 48.2197$

$= 632.5 - 12(2.916)^2$

$= x_1^2 + \ldots + x_{12}^2 - 12\bar{x}^2$

$= \sqrt{48.2197} = +\sqrt{s^2}$.

Representing Data Graphically

Stem & Leaf Diagram

↑ Arrange data in a visually informative way

- Table

- Divide the numerical values of the data into 2 parts : stem + leaf

  stem ↑ all but last digit

  leaf ↑ last digit

Example

Cannabis price (medical useage) in ON

$/gram to nearest 10c	$/gram	Year
9.10	9.07	2010
9.20	9.16	2011
9.30	9.31	⋮
10.40	10.37	
10.20	10.18	⋮
9.10	9.11	⋮
8.60	8.64	2017
8.00	8.02	

Sorry, I wrote this down & then we didn't use it in class.

Stem & Leaf plot:

8 . 0 6

9 . 1 1 2 3

10 . 2 4

We can make back-to-back stem & leaf plots to compare data sets eg. prices of cannabis (medicated useage) in BC over the same timeframe (2010–2017) were (to the nearest 10c) : $8.20, $8.30, $8.40, $8.40, $8.00, $8.10, $8.60, $7.60

We get:

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

We compare shapes to compare the distributions.

This is not very refined. We can also split "stems" into categories e.g. Upper $_{(0,\ldots,4)}$ & Lower $_{(5,\ldots,9)}$ to refine our presentation of the data e.g.

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

Then we see here that, even with this refinement, the data from BC is still clustered — highlights this feature.