1 The Big Picture

1.1 Why Study Statistics?

Communication, Quality Control, Decision Making, Big Data, $\cdots$

Statistics = Art of finding meaningful Patterns in data

1.2 Descriptive and Inferential Statistics

Descriptive Statistics – quantifies characteristics of the distribution of data.

Behavior of data around centre

- Measures of central tendency mean, median, mode
- Measures of variation standard deviation, rang, variance
- Correlation, regression → curve fitting

Inferential Statistics – makes inferences about the population based on sample data.

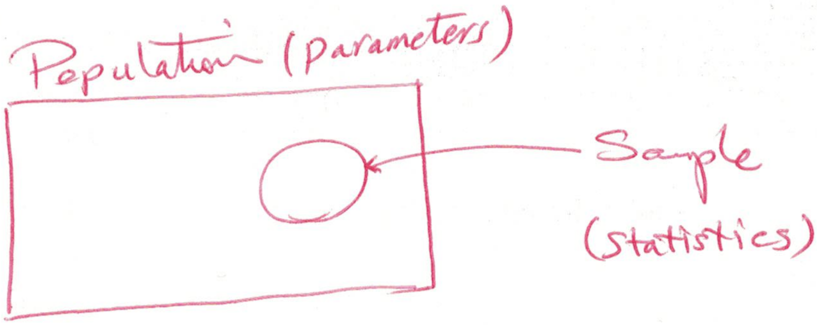
- Estimation
- Hypothesis testing
- Statistical comparison

1.3 Under the hood of Statistics: Probability Distributions

- Binomial distribution
- Poisson distribution
- Normal distribution Bell curve, Gussian distribution ★
- t-distribution
- Chi square distribution

1.4 Terminology: Population vs. Sample

Population: The complete collection of all elements (measurements, scores, etc.) to be studied.
Sample: A sub-collection of elements drawn from a population.
(simple random sampling, stratified sampling, etc.)



2 Distributions

2.1 Frequency Distribution: Histograms

Frequency v.s. observed quantity

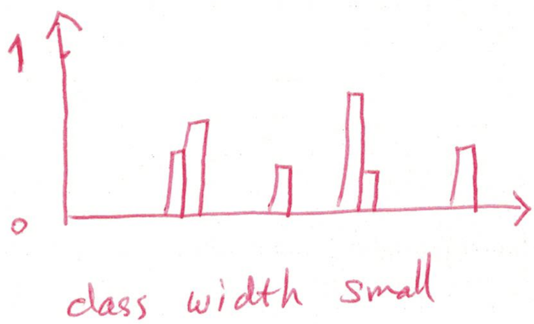
The *frequency* of a particular observation is the number of times the observation occurs in the data

Consider the following data of cosmic radiation counts under various depths of snow. [1]

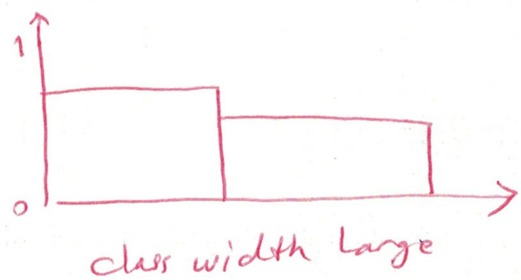
Depth x(cm)	Frequency (radiation count)	Relative Frequency	Mark (Label)
$0 \le x \le 3$	15	$\frac{15}{51} = 0.294 = 29.4\%$	1.5
$3 < x \le 6$	12	$\frac{12}{51} = 0.235 = 23.5\%$	4.5
$6 < x \le 9$	10	$\frac{10}{51} = 0.196 = 19.6\%$	7.5
$9 < x \le 12$	4	$\frac{4}{51} = 0.078 = 7.8\%$	10.5
$12 < x \le 15$	7	$\frac{7}{51} = 0.137 = 13.7\%$	13.5
$15 < x \le 18$	3	$\frac{3}{51} = 0.059 = 5.9\%$	16.5

- The range of measurements (radiation counts) is divided into intervals called **classes**. 6 class
- The size (width) of each class is called the **class width**. 3 cm
- **Class boundaries** can be set up so that there is no overlap between adjacent classes. (0, 3, 6, 9, ...)
- **Frequency** is the number of elements belong to a given class.
- **Relative frequency** is frequency / total number of elements in the sample.

Exercise 1: What would happen if class width is too small or too large?



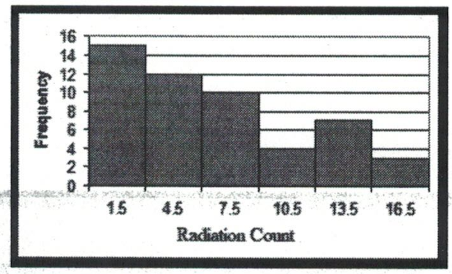
class width small

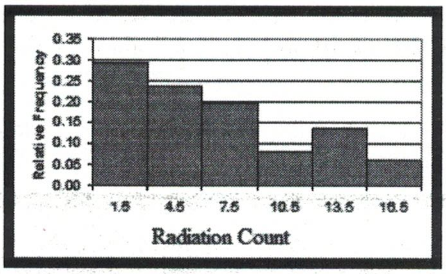


class width large

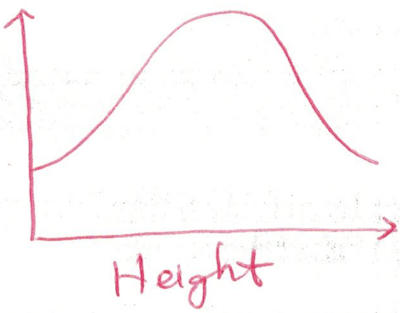
[1] "Deep snow measurements suggested using cosmic radiation" by Bissell and Burson, *Water Resources Research*, 1974, vol 10, no.6, p.1243.

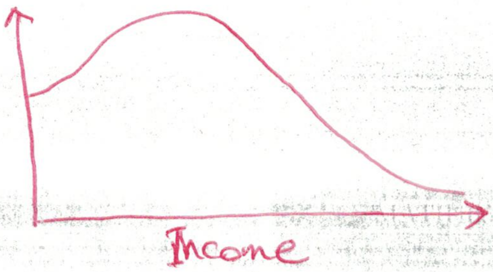
Histogram: Graphical representation of frequency distribution





Exercise 2: List a few examples of frequency distribution.





The most common distribution:

3 Measures of Central Tendency

Mean, Median, Mode, Mid-range

3.1 Mean (Arithmetic average)

For a *sample* with n elements $\{x_1, x_2, x_3, \cdots x_n\}$, the mean is defined to be

"x bar" $\leftarrow$ $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \textbf{(Sample mean)} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

N = size of population

- The mean of the whole *population* is denoted by $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$. "$\mu$ mu" population mean
- $\bar{x}$ is sensitive to extreme values in the data set.

Exercise 3: Calculate the mean of the GPA's of 3 students at BCIT. Their GPA's are $x_1 = 80, x_2 = 70, x_3 = 60$

$$\bar{x} = \frac{1}{3}(x_1 + x_2 + x_3) = \frac{1}{3}(80, 70, 60)$$

$$= 70$$

3.2 Median

e.g. $\{6,2,10,8,2\} \xrightarrow{\text{Sorted}} \{2,2,6,8,10\}$

The *median* is the "middle" of a sorted list of numbers. If the number of elements n is odd, then the median is defined to be

$$\tilde{x} = x_m, \quad \text{where } m = \text{ceil}(\tfrac{n}{2})$$

round up to nearest integer

"$\tilde{x}$ tilde"

If n is even

$$\tilde{x} = \frac{x_{\frac{n}{2}+1} + x_{\frac{n}{2}}}{2}$$

e.g., for $\{2,2,6,8\}$

$\tilde{x} = \frac{2+6}{2} = 4$

The median is less affected by extreme values than the mean.

Exercise 4: Consider the following data set:
$\{y_1 = 22, y_2 = 24, y_3 = 24, y_4 = 30, y_5 = 45, y_6 = 99\}$

1. Find the mean and the median of $\{y_1, y_2, y_3, y_4, y_5\}$

For $\{22, 24, 24, 30, 45\}$ we have

median $\tilde{y} = 24$

mean $\bar{y} = \frac{22+24+24+30+45}{5} = 29$

2. Find the mean and the median of $\{y_1, y_2, y_3, y_4, y_5, y_6\}$

For $\{22, 24, 24, 30, 45, 99\}$ we have

median $\tilde{y} = \frac{24+30}{2} = 27$

mean $\bar{y} = \frac{22+24+24+30+45+99}{6}$

$= 40.7$

mean is sensetive to extreme values

3.3 Mode

Position of peak(s)

The value that has the highest frequency or "peak" in the frequency distribution.

- A distribution with 2 peaks is called **bimodal distribution.** **Unimodal** for one peak, **multimodal** for many peaks.
- The class that has the greatest frequency is called the **modal class.**

Exercise 5: What is the mode of the data set given in §2.1?

mode = 1.5

3.4 Mid-range

The middle of the range of the data values. That is,

$$(\text{mid-range}) = \frac{(\text{minimum}) + (\text{maximum})}{2}$$

Exercise 6: Find all the measures of central tendency for the data set $\{2, 2, 2, 20, 34, 45, 210\}$.

mean $\bar{x} = \frac{1}{7}(2+2+20+34+45+210) = 45$

median $\tilde{x} = 20$

Mode $= 2$

mid-range $= \frac{2+210}{2} = 106$

Note: in Normal Distribution mean = median = mode = midrange

4 Measures of Variation (dispersion)

We often characterize the frequency distribution of given data using a representative value (e.g., mean, median) and a measure of variation, which is the "width" of the distribution. We will discuss two measures of variation; the range and the standard deviation.

4.1 Range

The range is the difference between the maximum value and the minimum value in the data set.

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

Exercise 7: The measured values of some quantity were $\{2.13, \underbrace{2.22}_{\text{max}}, \underbrace{2.09}_{\text{min}}, 2.11\}$. Find the range.

Range $= 2.22 - 2.09 = 0.13$

Exercise 8: Find the range of the following two distributions. Is the range a good measure of variation in this case?

(Exercise 9)

4.2 Standard deviation

The most commonly used measure of variation is the standard deviation. It is the *root mean square* (RMS) of ([data value] − [the mean]). That is,

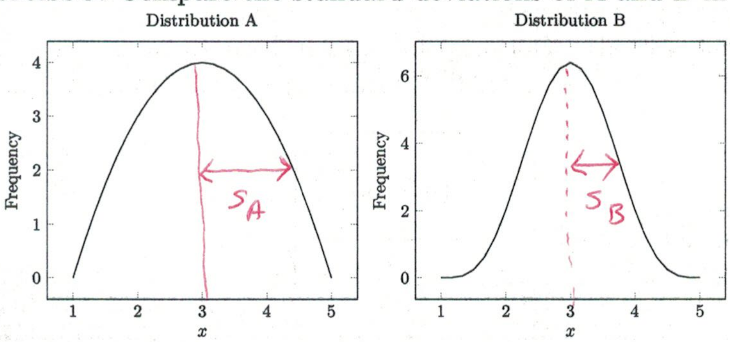
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}} \quad \dots \quad \text{for a population [population standard deviation]}$$

$$= \sqrt{\frac{1}{N}[(x_1-\mu)^2+(x_2-\mu)^2+\dots+(x_N-\mu)^2}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} \quad \dots \quad \text{for a sample [sample standard deviation]}$$

Note that in the standard deviation of a sample the sum is divided by $n-1$, instead of n, in order to compensate the tendency that s underestimate σ when the sample size is small.

Exercise 9: Compare the standard deviations of A and B in the figure above.



As we see B is more concentrated around Centre (mean)

Exercise 10: Calculate the standard deviation of the sample {1, 2, 2, 3} .

$$\bar{x} = \frac{1}{4}(1+2+2+3) = 2$$

$$s = \sqrt{\frac{(1-2)^2+(2-2)^2+(2-2)^2+(3-2)^2}{4-1}} = 0.816$$

4.3 Variance

s^2 and σ^2 are called the variance.