Chapter 1

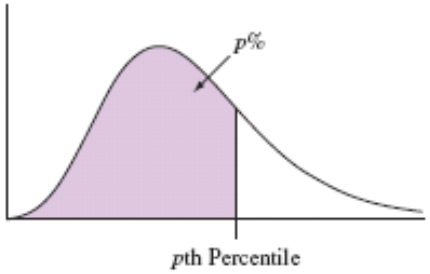
Summary and Display of Univariate Data

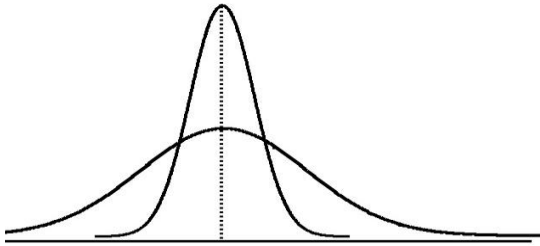
(contd.)

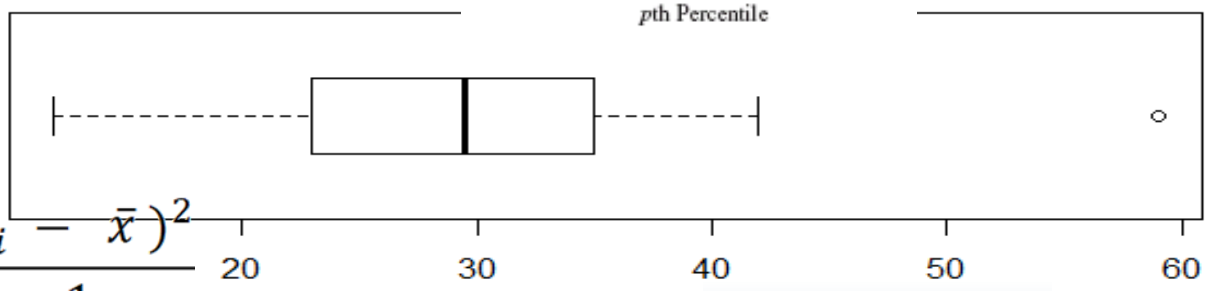
Lecture 4

Measure of Variability
Percentiles, Quartiles
Box plots

Dr. Lasantha Premarathna





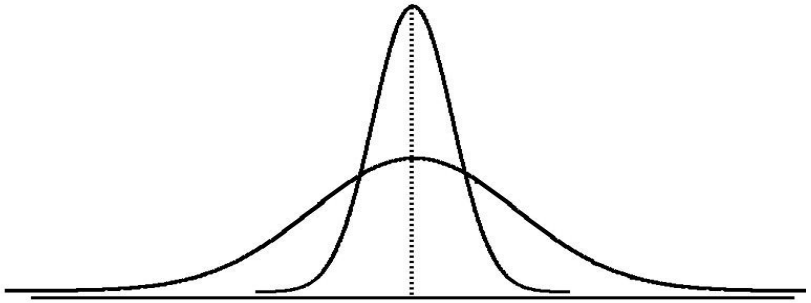


$$s^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1}$$

IQR = Q3 - Q1

Measures of variability

Measures of variation give information on the **spread** or **variability** or **dispersion** of the data values



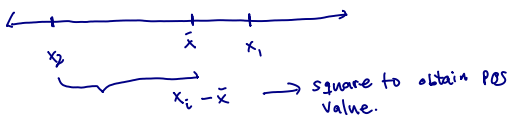
Same center, different variation

- **Range**
  - Difference between the largest and the smallest values

$$\textbf{Range} = X_{largest} - X_{smallest}$$

  - The range is strongly affected by outliers

Measures of variability



- **Variance and Standard Deviation**

for pos.

Sample variance = $s^2 = \dfrac{\sum_i^n (x_i - \bar{x})^2}{n-1} = \dfrac{\sum_i^n x_i^2 - n\bar{x}^2}{n-1} = \dfrac{sum\ of\ squared\ deviations}{sample\ size - 1}$

- Decrease in freedom (indp. Vd) mean known

Sample standard deviation = $s = \sqrt{\dfrac{\sum_i^n (x_i - \bar{x})^2}{n-1}} = \sqrt{\dfrac{\sum_i^n x_i^2 - n\bar{x}^2}{n-1}}$

english for sample

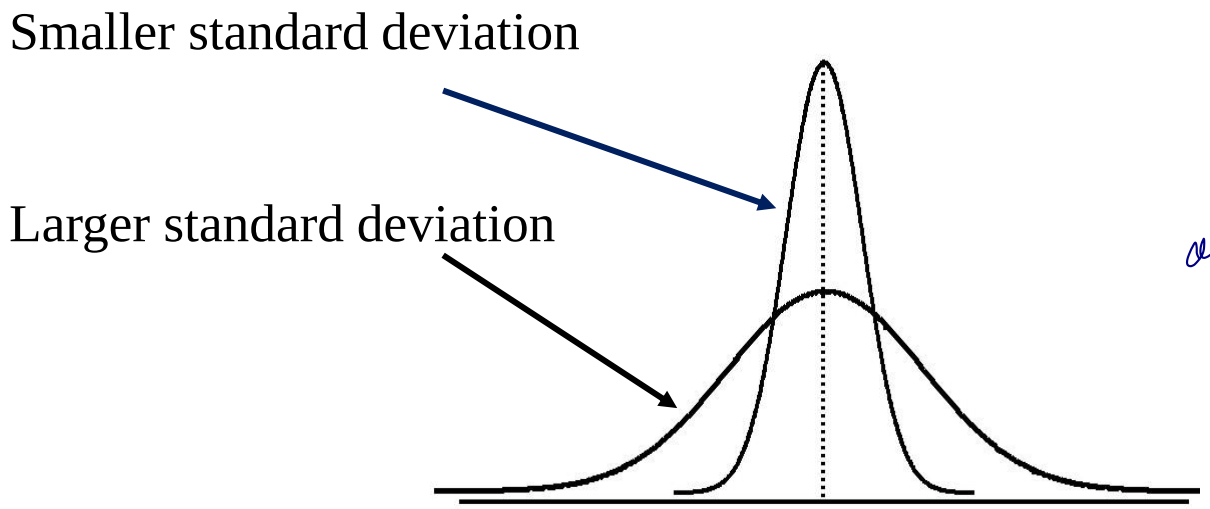
e.g. Number of hours spent studying per week for 5 students are 4, 6, 8, 7, 5.
Find the standard deviation of number of hours spent studying

$$s^2 = \frac{(4-6)^2 + (6-6)^2 + (8-6)^2 + (7-6)^2 + (5-6)^2}{5-1}$$

$$= 2.5\ hrs^2$$

$$\bar{x} = \frac{4+6+8+7+5}{5} = 6$$

$$s = \sqrt{2.5}$$



affected by outliers.

Properties of the Standard Deviation

- consider all data pts

- s measures the spread of the data
- $s = 0$ only when all observations have the same value, otherwise $s > 0$. As the spread of the data increases, s gets larger.
- s has the same units of measurement as the original observations. The variance = s^2 has units that are squared → Must report units.
- s is not resistant. Strong skewness or a few outliers can greatly increase s. → identify outliers

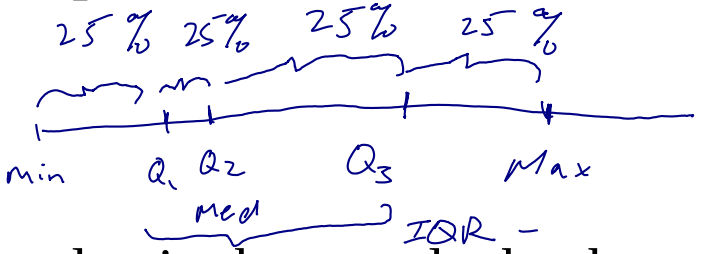
Measures of variability

Non affected by outliers.

Interquartile Range (IQR):

length of the range of an interval that captures the middle 50% of the data

IQR = Q3 - Q1



- ***Q1 = First quartile*** = 25th percentile is the value in the sample that has 25% of the data below it.
- ***Q3 = Third quartile*** = 75th percentile is the value in the sample that has 75% of the data below it.

If we were to split the data in half, the *first quartile is the median of the lower half* and *the third quartile is the median of the upper half* of the data. Notice the ***median*** is also called the ***50th percentile*** or ***second quartile*** (***Q2***) since 50% of data falls below it.

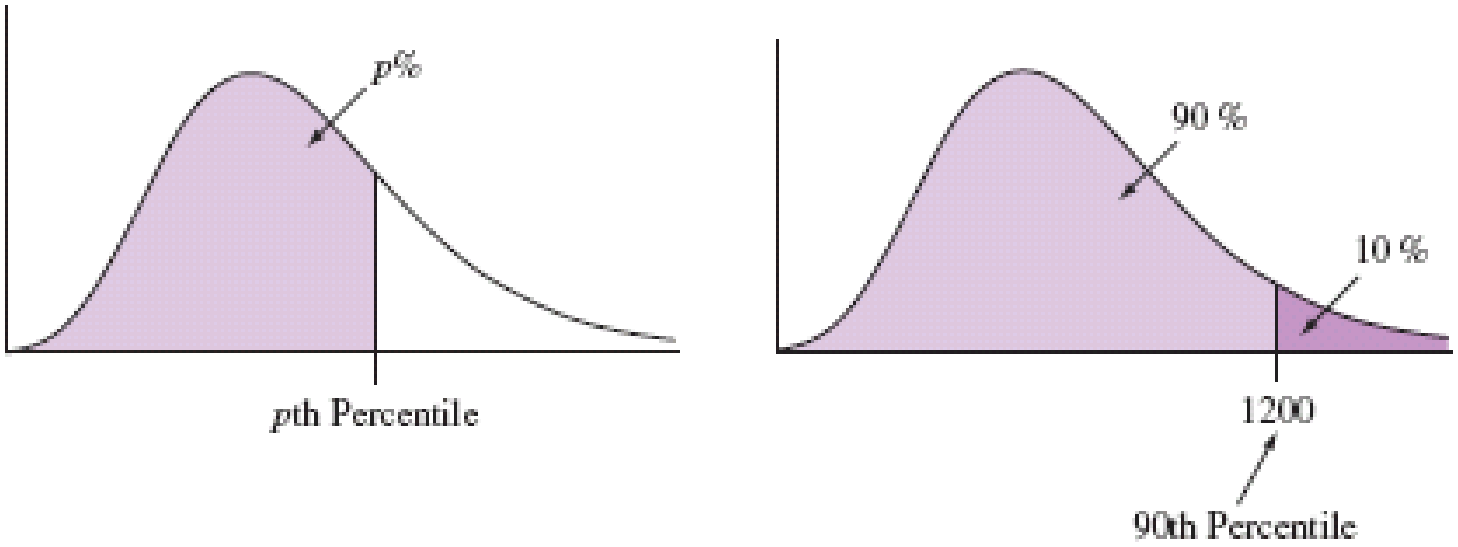
Note: Some software packages or textbooks use slightly different rules to find quartiles thus different sources may give varying results.

Identifying an outlier

An observation is an outlier if it falls more than *1.5* x *IQR* below the first quartile or more than *1.5* x *IQR* above the third quartile

Percentile

- The p^{th} percentile is a value such that p percent of the observations fall below or at that value



Sample Quantiles – general

Let $0 < p < 1$ be fixed. The sample quantile or order p, $Q_{(p)}$ is a number with the property that approximately $p100\%$ of the data points are smaller than it.

To compute $Q_{(p)}$ we must follow the following steps

- Sort the data from smallest to largest

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

smallest (↑ $x_{(1)}$) largest (↑ $x_{(n)}$)

The i^{th} order statistic is denoted by $x_{(i)}$. First order statistic (smallest order statistic) is the minimum and the n^{th} order statistic (largest order statistic) is the maximum.

- Compute the number $np + 0.5$ ← # of data pts.
  - If this number is an integer, m, then

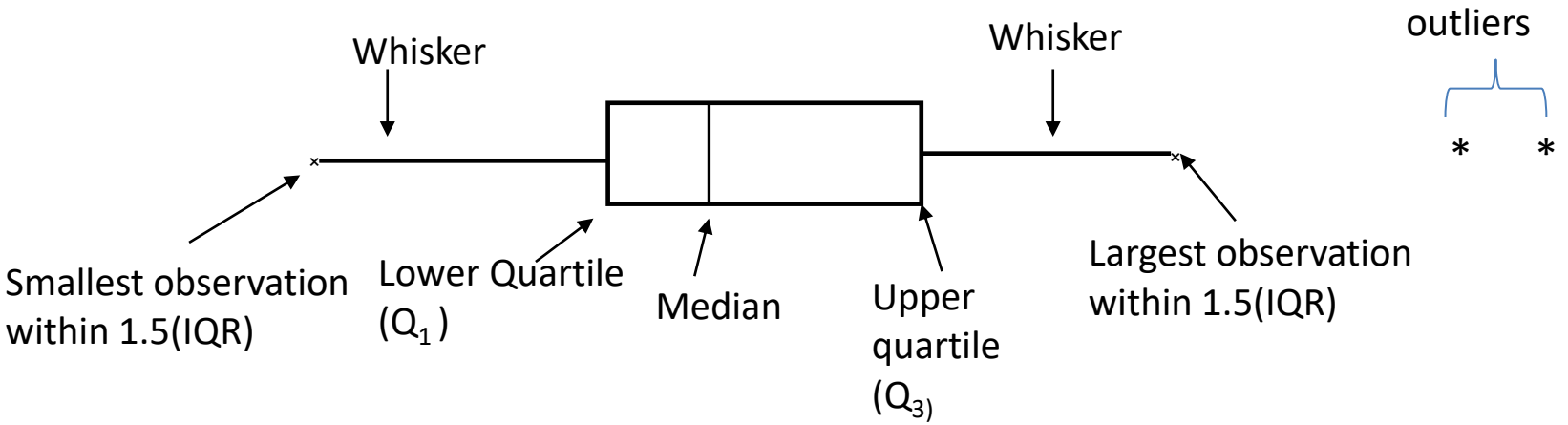
$$Q_{(p)} = x_{(m)}$$

  - If $np + 0.5$ is not an integer and $m < np + 0.5 < m + 1$ for some integer m then

$$Q_{(p)} = \frac{x_{(m)} + x_{(m+1)}}{2}$$

Box plot

- A box plot is another way of looking at a data set in an effort to determine its central tendency, spread, skewness, and the existence of outliers. We use box plots for Quantitative data.
- A box plot is a graph of a set of five summary measures of the distribution of the data.
    1. The smallest observation
    2. The lower quartile, Q1 (25th percentile)
    3. The median (Q2) of the data (50th percentile)
    4. The upper quartile, Q3 (75th percentile)
    5. The largest observation

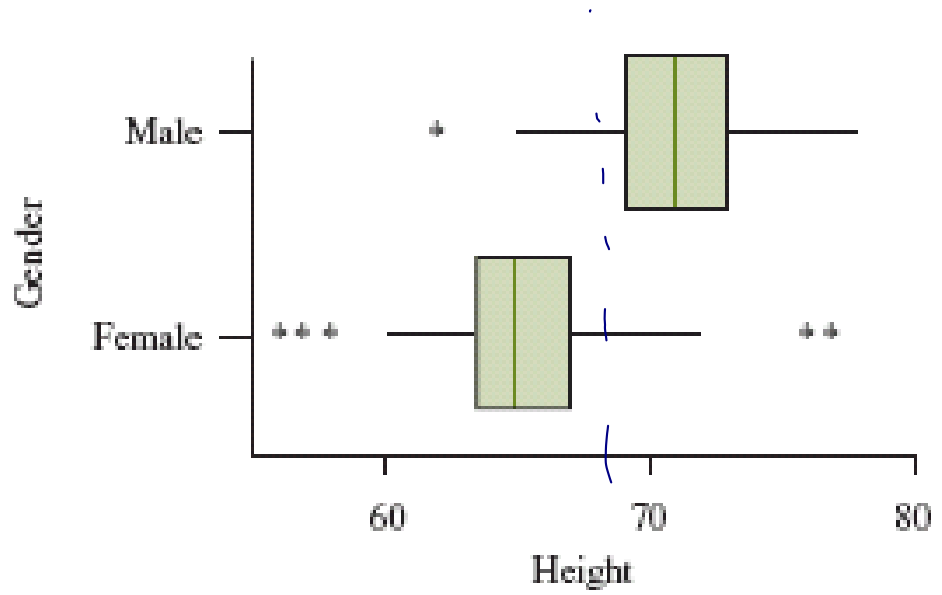


Comparison using side by side box plots

Box Plots do not display the shape of the distribution as clearly as histograms, but are useful for making graphical comparisons of two or more distributions.

75% of males taller than 75% of females.

Eg:



e.g. Construct the box plot for the following data.

12	14	17	22	22	24	25	26	27	29
30	31	33	34	35	35	39	40	42	59
1	2	3	4	5	6	7	8	9	10

$n = 20$ items

$p = 0.25$ or 0.75 percentile.

Q_1

$np = 20 \times 0.25 = 5$

$np + 0.5 = 5.5$

$= \frac{5^{th} + 6^{th}}{2}$

$= \frac{22 + 24}{2} = 23$

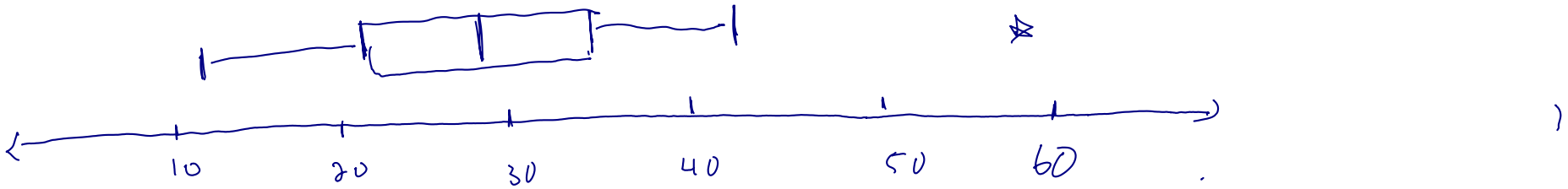
Q_3

$np = 20 \times 0.75 = 15$

$np + 0.5 = 15.5$

$= \frac{15^{th} + 16^{th}}{2}$

$= \frac{35 + 35}{2} = 35$



12 23 28.5 35 42 59

$1.5 \times IQR = 12 \times \frac{3}{2} = 18$

lower outlier range,

$= 23 - 18 = 5$

$35 + 18 = 52$

upper outlier range

$= 35 + 18 = 52$.

How do location/scale changes affect mean and variance

i.e. Changing Celsius data to Fahrenheit

x_i = i^{th} measurement in ^{o}C

y_i's are in Fahrenheit

sample mean and variance of $x_i's$ are $\bar{x}$ and s_X^2

$$y_i = 32 + \frac{9}{5} x_i$$

a, b

$$y_i = a + b\, x_i$$

Mean of y_i 's

$$\bar{y} = \frac{1}{n}\sum_i^n y_i = \frac{1}{n}\sum_i^n (a + bx_i)$$

$$= \frac{1}{n}\left(na + b\sum_i^n x_i\right)$$

$$= \frac{1}{n}(na + bn\bar{x})$$

$$\bar{y} = a + b\,\bar{x}$$

→ Mean will shift
Mean will scale

Variance of y_i's

$$s_Y^2 = \frac{1}{n-1}\sum_i^n (y_i - \bar{y})^2$$

$$= \ldots.$$

$$= \ldots.$$

$$= \ldots.$$

$$= b^2\, s_X^2$$

→ Variance will scale by b^2

→ No shift

Summary

- Measure of Variability (Range, variance, standard deviation, IQR)
- Percentiles, Quartiles
- Box plots

Before the next class

- Review the lecture 4 and related sections in the text book

Next Class:

- Chapter 3 : Sets and Probability