Lecture 6

May 27, 2019 6:51 PM

Population: The target group you want to analyse.

Sample: A small group of individuals from the population.

Some House hold — All cars in the world.

Sample $\Longrightarrow$ Population.

statistic: power of cars from these household $\Longrightarrow$ parameter: power of all cars in the world.

6.1 Numerical Summary:

Sample Mean:

Suppose we have a data set $x_1, x_2, x_3 \cdots x_n$.

The sample mean of them is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$ — $\bar{x}$ bar $\Longrightarrow$ μ

$\bar{x}$: statistic; μ: parameter

Population Mean:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$, N is size of population.

Sample Variance:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} \Longrightarrow \sigma^2$$

s^2: statistic; population variance σ^2: parameter

Example: 55, 63, 72, 41, 87, 75, 64, 60

$$\bar{x} = 64.625$$

$$s^2 = \frac{55^2 + 63^2 + \cdots 66^2 - 8 \cdot (64.625)^2}{8-1}$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

s = sample standard deviation

σ = population standard deviation.

Range:

Range = largest obs − smallest obs.

= 87 − 41 = 46

6.2 Stem and Leaf plots

Example: 55, 63, 72, 41, 87, 75, 64, 60.

Median:

is the middle value of an ordered data set.

- n is odd. $\frac{n+1}{2}$ th number
- n is even. average of middle two number.

41, 55, 60, 63, 64, 72, 75, 87

n = 8

median = $\frac{63+64}{2} = 63.5$

Stem	Leaf
1	4
2	5
(3)	6 | 0 3 4
3	7 | 2 5
1	8 | 7

first # = 1

first two # = 2

last two # = 3

last row # = 1

10 = interval length.

1 = leaf unit

[40, 50), [50, 60) ... ... [70, 80)

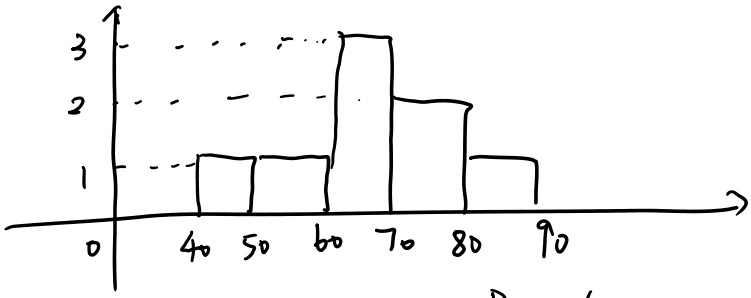
Convention: when the endpoints overlap the right endpoints is not included.

6.3 Frequency Distribution and Histogram:

Interval	Frequency	Cumulative Frequency	Relative Frequency
40 - 50	1	1	$\frac{1}{8}$
50 - 60	1	2	$\frac{1}{8}$
60 - 70	3	5	$\frac{3}{8}$
70 - 80	2	7	$\frac{2}{8}$
	1	8	1

60 - 70	2	7	$\frac{2}{8}$
70 - 80	1	8	$\frac{1}{8}$.
80 - 90			

Histogram.



6.4 Quartiles and Box plots

Quartile: the kth percentage P_k is the number that k% of the data is less than P_k.

median = 50% quartile, P_{50}.

Q_1 25% quartile $= \frac{n+1}{4}$ th number

Q_3 75% quartile $= \frac{3(n+1)}{4}$ th number

Example: 15, 13, 6, 5, 12, 50, 22, 18

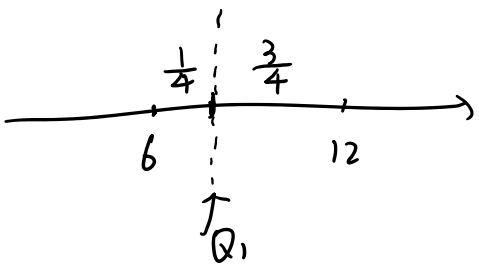
5, 6, 12, 13, 15, 18, 22, (50).

median: $\frac{13+15}{2} = 14$

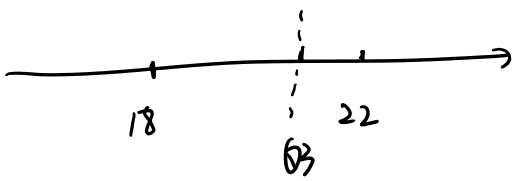
Q_1: $\frac{8+1}{4} = 2.25$ th $\Rightarrow$ 2nd 6, 3rd 12

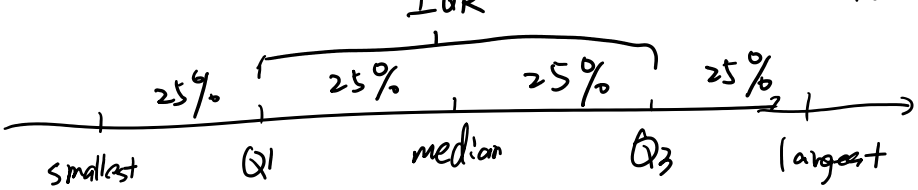
Q_3: $\frac{3(8+1)}{4} = 6.75$ th $\Rightarrow$

$Q_1 = 6 + \frac{1}{4}\cdot(12-6) = 7.5$



$Q_3 = 18 + \frac{3}{4}(22-18) = 21$





IQR: IQR = Inter quartile Range

$= Q3 - Q1$

Outlier: Extreme / unusual large or small value.

outlier $\begin{cases} > Q3 + 1.5\,IQR \\ < Q1 - 1.5\,IQR \end{cases}$

Mode: The most frequent number in the data set.

$Q3 + 1.5\,IQR = 21 + 1.5(21 - 7.5) = 41.25$

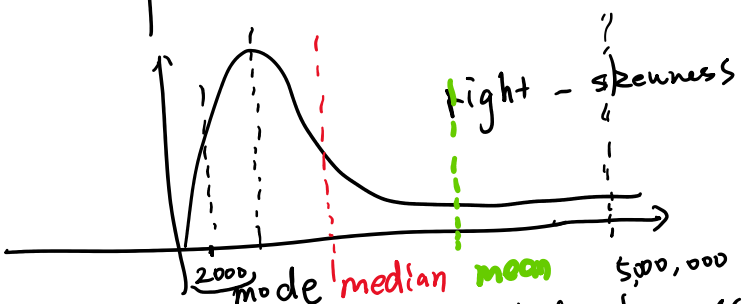
$Q1 - 1.5\,IQR = 7.5 - 1.5(21 - 7.5) = -12.75$

$\therefore 50 > 41.25$, 50 is an outlier

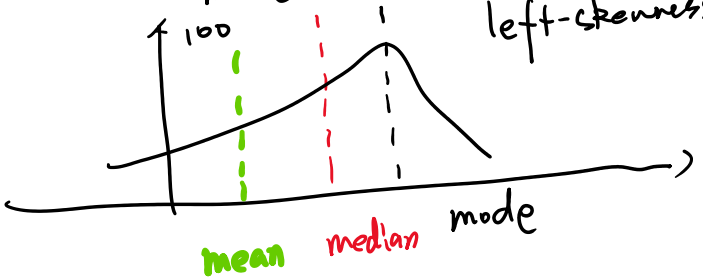
News: Average Salary of Toronto is $50,000.

mean = median = mode.

symmetric.



mode < median < mean

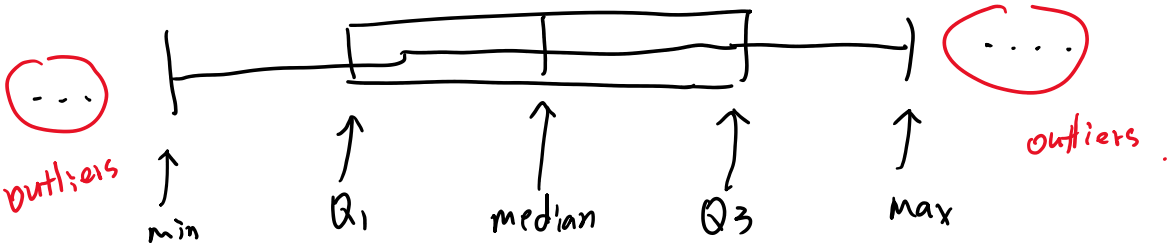


mode > median > mean.

Boxplot / five number summary.

min Q1 median Q3 max.

outliers ∴ ∴ outliers



6.7 Probability Plot.

Normal Probability Plot: Check whether sample follows Normal.

Given a sample, $X_1 \dots X_n$. Find.

$$Z_j = \Phi^{-1}\left(\frac{j-\frac{1}{2}}{n}\right), \quad j = 1, 2, \dots, n.$$

$\Phi(x) = F(x)$ for a standard normal.

A normal probability plot consists n pairs $(X_{(i)}, Z_i)$

where $X_{(1)} \le X_{(2)} \le \dots X_{(n)}$ are ordered data.

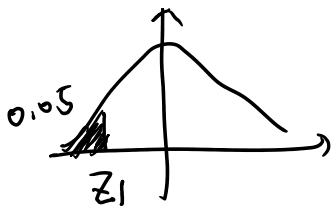
Example:

201, 214, 190, 176, 185, 205, 220, 183, 191, 192

$(n = 10)$

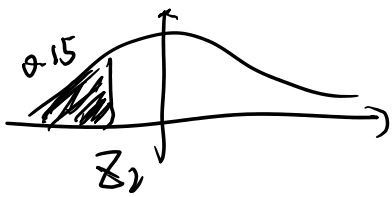
$$Z_1 = \Phi^{-1}\left(\frac{1-\frac{1}{2}}{10}\right) = \Phi^{-1}(0.05)$$

$$Z_1 = -1.645$$



$$Z_2 = \Phi^{-1}\left(\frac{2-\frac{1}{2}}{10}\right) = \Phi^{-1}(0.15)$$

$$Z_2 = -1.04$$

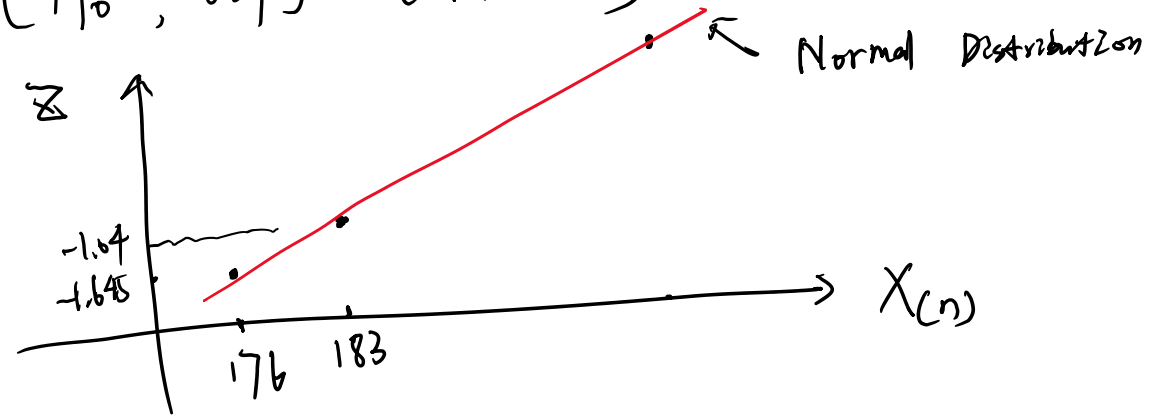


$X_{(1)}$ Z_1 $X_{(2)}$ Z_2 $X_{(3)}$ Z_3

(176, -1.645), (183, -1.04), (185, -0.67)

$X_{(4)}$ Z_4 $X_{(5)}$ Z_5 $X_{(10)}$ Z_{10}

(190, -0.39) (191, -0.13) ..... (220, 1.64)



Note: interpreting the plot:

If the data comes from a population that follows normal distribution, then the points on the plot should fall close to a straight line with no discernible curve pattern. The points should be randomly scattered about the line.

7.2 Central Limit Theorem.

Definition: $X_1, \ldots, X_n$ is a random sample of size n if they are independent and all have the same distribution

The statistic's distribution or the sampling distribution of $\bar{X}$ follows:

$$\bar{X} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$$

where $E(\bar{X}) = \mu$, $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$

Example:

An electronic company manufatures resisters that [illegible] and standard deviation 10Ω.

An electronic company manufactures ... have mean 100Ω and standard deviation 10Ω. Find the probability that a random sample of 25 the average of the sample is fewer than 95Ω.

$\bar{X} \Rightarrow \mu$

$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ ← $\mathrm{Var}(\bar{X})$

$$P(\bar{X} < 95) = P\left(Z < \frac{95 - 100}{10/\sqrt{25}}\right)$$

$$= P(Z < -2.5)$$

$$= 0.006210.$$

Example:

Let $X_1, X_2, \ldots X_{40}$ be a r.s. from the following distribution:

$$f(x) = \frac{1}{2}, \quad 4 \le x \le 6.$$

Find the sample mean ($\bar{X}$ instead of X) is greater than 5.1 exactly and use normal distribution to approximate it.

① $P(X \ge 5.1) = \int_{5.1}^{6} f(x)\,dx$ for individual.

individual

$$= \int_{5.1}^{6} \frac{1}{2}\,dx$$ not proper.

②

$$E(X) = \int_4^6 x \cdot \frac{1}{2}\,dx = 5$$

$$E(X^2) = \int_4^6 x^2 \cdot \frac{1}{2}\,dx$$

$$\mathrm{Var}(X) = E(X^2) - [E(X)]^2 = \frac{1}{3}.$$

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}(X) \cdot \frac{1}{40} = \frac{1}{120}.$$

Then

$\text{Var}(X) = \text{Var}(\bar{X}) = \frac{40}{} = 120$.

Then

$$P(\bar{X} > 5.1) = P\left(Z > \frac{5.1 - 5}{\sqrt{1/120}}\right) = P(Z > 1.095)$$

$$\simeq 0.1357$$

Result: If $\bar{X}_1$ and $\bar{X}_2$ are sample means of independent samples from population 1 μ_1, σ_1^2 population 2 μ_2, σ_2^2.

Then,

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

$$\left(Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \right)$$

$$\left(Z = \frac{X - \mu}{\sigma} \sim N(0,1) \right)$$

Example:

A group of bus drivers have salary as follow.

based on newspaper:

male : $\mu_1 = 5000$ $\sigma_1 = 40$

female : $\mu_2 = 5050$ $\sigma_2 = 30$

Then Find the probability that for a sample of $n_1 = 16$ and $n_2 = 25$. the male's average salary is 25 less than female's salary mean.

$$P(\bar{X}_1 - \bar{X}_2 < -25) = P\left(Z \leq \frac{-25 - (5000 - 5050)}{\sqrt{\frac{40^2}{16} + \frac{30^2}{25}}}\right)$$

$$P(\bar{X}_1 - \bar{X}_2 < -25) = P(Z < \frac{}{\sqrt{\frac{40^2}{16} + \frac{30^2}{25}}})$$

$$\approx P(Z \le -2.14)$$

$$= 1 - 0.9838.$$

$$= 0.1162$$

7.3 Concepts of Point Estimation:

If $X_1 \dots X_n$ is r.s from

$$f(x) = \frac{1}{\theta} e^{-x/\theta}, \quad x > 0$$

How to find θ?

sample	$\bar{X}$	S	$\hat{p}$	$\hat{\theta}$	estimators/statistic
polution	μ	σ	p	θ	true parameter
	mean	SD	proportion		

Point Estimator

Definition: A point estimator of parameter θ is sample statisc $\hat{\theta}$.

Unbiase Estimator

Definition: If $E(\hat{\theta}) = \theta$, then we call $\hat{\theta}$ an unbiased Estimator.

Example: Is $S^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1}$ an unbiased estimator of σ^2?

$$E(S^2) = \frac{1}{n-1} E\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$$

$$= \frac{1}{n-1} \left[E\left(\sum_{i=1}^{n} X_i^2\right) - E(n\bar{X}^2) \right]$$

$$E(S^2) = \frac{1}{n-1}\left[E\left(\sum_{i=1}^{n} X_i^2\right) - nE(\bar{X})\right]$$

$$= \frac{1}{n-1}\left[nE(X_i^2) - nE(\bar{X}^2)\right]$$

where $\begin{cases} E(\bar{X}^2) = Var(\bar{X}) + [E(\bar{X})]^2 = \frac{\sigma^2}{n} + \mu^2 \\ E(X^2) = Var(X) + [E(X)]^2 = \sigma^2 + \mu^2 \end{cases}$

$$= \frac{1}{n-1}\left[n(\sigma^2+\mu^2) - n\left[\frac{\sigma^2}{n} + \mu^2\right]\right]$$

$$= \frac{1}{n-1}\left[(n-1)\sigma^2\right]$$

$$= \sigma^2$$

$\therefore S^2$ is an unbiased estimator of σ^2

Compare estimators:

If two estimators are unbiased then the estimator with smaller variance is better.

$$SD(\hat{\theta}) = SE(\hat{\theta}) = \sigma_{\hat{\theta}} = \sqrt{Var(\hat{\theta})}$$

Example: Let p denote a population proportion. Then if I repeat the experiment n times and find x times success.

Then we use $\hat{p} = \frac{x}{n}$ as an estimator of p.

Find the bias and $SD(\hat{p})$.

$$\text{Bias} = E(\hat{\theta}) - \theta$$

$$= E(\hat{p}) - p$$

$$= E\left(\frac{x}{n}\right) - p$$

$$\left(X \sim \text{Binomial}(n, p), \quad E(X) = np\right)$$

$$= E\left(\frac{X}{n}\right) - p$$

$$= \frac{1}{n}E(X) - p \quad \left(E(X) = np\right)$$

$$= \left(\frac{1}{n} \cdot np\right) - p$$

$$= 0$$

$$SD(\hat{p}) = SE(\hat{p}) = SE\left(\frac{X}{n}\right)$$

$$= \sqrt{\frac{Var(X)}{n^2}}$$

$$= \sqrt{\frac{np(1-p)}{n^2}} \quad \left(\text{where } Var(X) = np(1-p)\right)$$

$$= \sqrt{\frac{p(1-p)}{n}}$$

$$n \to \infty$$

$$\to 0$$

The Mean Square Error of $\hat{\theta}$:

$$MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$$

$$= E\left[(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2\right]$$

$$= E\left[(\hat{\theta} - E(\hat{\theta}))^2\right] + E\left([E(\hat{\theta}) - \theta]^2\right) + \underbrace{E\left[2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta)\right]}_{0}$$

$$= Var(\hat{\theta}) + \left[Bias(\hat{\theta})\right]^2$$

$$\Longleftrightarrow E(X^2) = Var(X) + [E(X)]^2$$

$\Leftrightarrow \quad E(X^2) = Var(X) + [E(X)]^2$

$MSE(\hat{\theta}) = Var(\hat{\theta}) + [Bias(\hat{\theta})]^2$

Relative Efficiency of estimators $\hat{\theta}_1$, $\hat{\theta}_2$ for θ

$$A = \frac{MSE(\hat{\theta}_1)}{MSE(\hat{\theta}_2)}$$

If $A > 1$, $\hat{\theta}_1$ is worse than $\hat{\theta}_2$

wise versa.

(21-23) Sample Test 2.