WEEK-3 - STATISTICS

Dates

RANDOM PHENOMENON:

① Deterministic phenomenon:

phenomenon whose outcome can be predicted with a very high degree of confidence.

② Stochastic phenomenon:

phenomenon which can have many possible outcomes for same experimental conditions. Outcome can be predicted with limited confidence.

→ Sources of Errors:

① Model error - Lack of Knowledge of generating process

② Measurement error - Errors in sensors used for observing outcomes.

→ Types of random phenomena

(1) Discrete
(2) Continuous.

→ Sample Space :

Set of all possible outcomes of a random phenomena

→ Event :

Subset of a sample space.

→ Probability measure:

- $0 \leq P(A) \leq 1$
- $P(S) = 1$ S → sample space.

Interpretation of probability as a frequency:

- Conduct an experiment (coin toss) N times. If N_A is number of times outcome A occurs the $\boxed{P(A) = \frac{N_A}{N}}$

→ **Independent Events:**

$$P(A \cap B) = P(A) \times P(B)$$

→ **Mutually Exclusive events:**

Two events are mutually exclusive if occurence of one implies other event does not occur.

☆ $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

☆ if $B \subseteq A$, $P(B) \leq P(A)$

→ Conditional Probability:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad \text{if } P(A) > 0$$

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A) \quad \text{— Bayes Formula}$$

$$P(A) = P(A|B)\,P(B) + P(A|B')\,P(B')$$

RANDOM VARIABLE:

Random variable (RV) is a map from sample space to a real line such that there is a unique real number corresponding to every outcome of sample space.

→ Binomial Mass function

$$f(x=k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

for large n it tends to a Gaussian distribution.

→ Gaussian or Normal Density Function:

↳ used to chartegorize random errors in data.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

↳ ~~chatergo~~ characterized by 2 parameters μ and σ

↳ symmetric.

↳ Standard normal distribution:

$\mu = 0$

$\sigma = 1$

→ Chi-square density function

$$f(x) = \frac{1}{2^{n/2}\,\Gamma(n/2)}\, x^{n/2-1}\, e^{-x/2}$$

↳ characterised by parameter n (degrees of freedom)
↳ distribution of sum of squares of n independent standard normal rvs
↳ distribution of sample variance.

→ **Moments of a pdf :**

↳ Similar to describing a function using derivatives, a pdf can be described by its moments.

↳ for continuous distributions :

$$E[x^k] = \int_{-\infty}^{\infty} x^k f(x) \cdot dx$$

↳ for discrete

$$E[x^k] = \sum_{i=1}^{N} x_i^k p(x_i)$$

↳ Mean : $\boxed{\mu = E[x]}$

↳ Variance : $\sigma^2 = E[(x-\mu)^2] = E[x^2] - \mu^2$

↳ Standard Deviation = σ

⇒ Properties of ~~Guassian~~ Gaussian RVs

- Mean = $E[x] = \mu$
- Variance = $E[(x-\mu)^2] = \sigma^2$
- Symbolically $x \sim N(\mu, \sigma^2)$
- ~~Sanda~~ standard Gaussian RV $z \sim N(0,1)$

- if $x \sim N(\mu, \sigma^2)$ and $y = ax+b$ then

$$y \sim N(a\mu+b, a^2\sigma^2)$$

- Standardization :

if $x \sim N(\mu, \sigma^2)$, then $z = \left(\frac{x-\mu}{\sigma}\right) \sim N(0,1)$

⇒ Computation of probability using R:

① Function pnorm:

pnorm (X, mean, std, 'lower.tail' = TRUE/FALSE)

↳ can be replaced by chisq, exp, unif

② Function qnorm:

qnorm (p, mean, std, 'lower.tail' = TRUE/FALSE)

↳ to compute X given the probability p

↳ function dnorm to compute density function value.

↳ Function rnorm to generate random numbers from the distribution

→ **Joint pdf of two RVs:**

$f(x,y)$

- $P(x \le a, y \le b) = \int_{-\infty}^{b} \int_{-\infty}^{a} f(x,y)\, dx\, dy$
- Covariance b/w x and y : $\sigma_{xy} = E[(x-\mu_x)(y-\mu_y)]$
- Correlation b/w x and y : $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$
- Two RV x and y are uncorrelated if $\sigma_{xy} = 0$
- Two RVs x and y are independent if $f(x,y) = f(x) \cdot f(y)$

→ **Multivariate Normal Distribution:**

↳ a vector of RVs $x = [x_1\ x_2 \dots x_n]^T$

- Multivariate Gaussian Distribution : $x \sim N(\mu, \Sigma)$
  - $E[x] = \mu$: Mean vector
  - $E[(x-\mu)(x-\mu)^T] = \Sigma$: variance - covariance matrix
- $f(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$

- Structure of Σ :

$$\Sigma = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1 x_2} & \cdots & \sigma_{x_1 x_n} \\ \sigma_{x_2 x_1} & \sigma_{x_2}^2 & \cdots & \\ \vdots & & & \\ \sigma_{x_n x_1} & \cdots & \cdots & \sigma_{x_n}^2 \end{bmatrix}$$

STATISTICAL ANALYSIS:

(1) Descriptive Statistics:

- Graphical: organizing and presenting the data.
- Numerical: Summarizing the sample set.

② Inferential

- Estimation: estimate parameters of the pdf along with its confidence region.
- Hypothesis testing: Making judgements about $f(x)$ and its parameters.

→ Measures of central tendency:

① Mean:

$$\text{Mean (or average)} = \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- Best estimate in least squares criterion
- Unbiased estimate of population mean: $E[\bar{x}] = \mu$
- Affected by outliers.

② **Median :**

→ value of x_i such that 50% of the values are less than x_i and 50% of observations are greater than x_i.

*Robust WRT to outliers in data.
• best estimate in least absolute deviation sense.

③ **Mode :**

↳ Value the occurs most often (most probable value)

→ **Measures of Spread :**

→ Sample variance

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

- unbiased estimate of population variance : $E[s^2] = \sigma^2$

→ Mean absolute deviation

$$\bar{d} = \frac{1}{N} \sum_{i=1}^{N} |x_i - \bar{x}|$$

→ Distribution of sample mean and variance:

① Sample mean:

↳ unbiased estimate of population mean

if $x_i \sim N(\mu, \sigma^2)$ and all observations are mutually independent then $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{N}\right)$.

② Sample Variance

↳ unbiased estimate of population variance

if $x_i \sim N(\mu, \sigma^2)$ and all observations are mutually independent, then

$$\frac{(N-1)S^2}{\sigma^2} \sim \chi^2_{N-1}$$

⇒ Graphical Analysis:

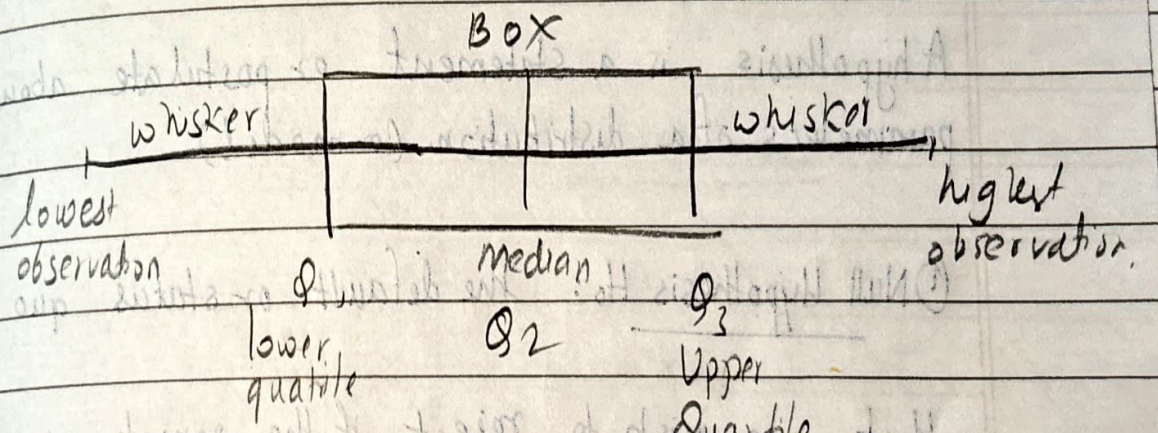
① Histogram

↳ divide the range of values in sample in sample set into small intervals and count how many observations fall within each interval.

↳ for each interval plot rectangle with width = interval size and height to number of observations in intervals.

② Box Plot → for stock prices,

- Find quartiles (Q1, Q2, and Q3), min and max values in range.
- Box is between Q1 and Q3, and whiskers is between min and max values.



③ Probability plot:

(p-p or q-q plot)

↳ determine diff quantile values from sample set. Plot computed quantiles vs theoritical quantile values from chosen distribution.

(4) Scatter Plot:

↳ Plot one RV(y) against another RV (x) to examine whether there is any dependence.

HYPOTHESIS TESTING:

A hypothesis is a statement or postulate about the parameters of a distribution (or model)

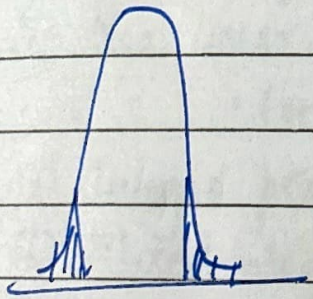
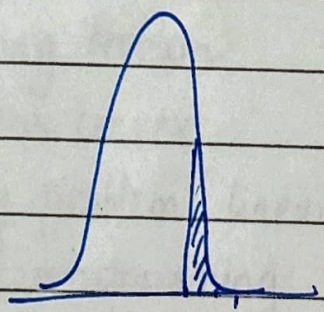
① Null Hypothesis H_0: the default or status quo postulate that we wish to reject if the sample set provides sufficient evidence. (eg. $n = n_0$)

② Alternative hypothesis H_1: the alternative postulate that is accepted if the null hypothesis is rejected. (eg $n < n_0$)

classmate
Date
Page

* the performance of a hypotheses test depends on
  - extent of variability in data.
  - Number of observations
  - Test statistic
  - Test criterion.

→ 2 sided and 1 sided test:

(1) Two sided test:	(2) One sided test:
$H_0 : \mu = 0$	$H_0 : \mu = 0$
$H_1 : \mu \neq 0$	$H_1 : \mu > 0$
	

⇒ **Errors in Hypothesis testing:**

Truth \ Decision	H_0 is not rejected	H_0 is rejected
H_0 is true	Correct decision	Type -1 error $-\alpha$
H_1 is true	Type II error β	Correct decision.

☆ If we decrease Type I error probability, the type-II probability will increase