Statistics, the exploration and Analysis of Data
- Roxy Peck & Joy Devore

Schaum's Series.

ABE 302: Engineering Statistics (3)

Course Content

(1) Probability & Statistics

- Probability Space theory
- ~~Theory~~
- Conditional probability & independence
- Random variables, discrete and continuous distributions
- Mean and Variance
- Binomial, Poisson, hypergeometry
- Exponential and Normal distributions and their characteristics
- Centre limit theory
- Elementary Sample theory for normal population.
- Statistical Inference on mean and Variance
- Simple linear regression and general applications
- Chi-Square test

Probability & Statistics

Probability is the measure of the likelihood that a particular event will occur in any one trial or experiment carried out in a prescribed condition.

Notations

The probability that a certain event A will occur is denoted by $P(A)$ or $p(A)$ It is also denoted by (P success)

Success or Failure

When an event occurs in any one trial, it is called Success, but when it fails to occur, it is called failure.

In any N trials, there are x successes, there will also be $(N-x)$ failures

$$P(\text{success}) = \frac{x}{N}$$

$$P(\text{failure}) = \frac{(N-x)}{N}$$

$P(A) + P(\text{Not } A) = 1$

or $P(A) + P(\bar{A}) = 1$

Therefore, $\frac{x}{N} + \frac{N-x}{N} = 1$

Types of Probabilities:

Determination of probability can be undertaken from two perspectives.

(1) Emperical or Experimental

(2) Classical or Theoretical

(1) Emperical is based on the previous known result

Expectation

This is defined as the product of the number of trials and probability that event A will occur in a number of trials

(2) The classical approach to probability is based on the application of theoretical ~~knowledge~~ number of ways in which it is possible for an event to occur.

Classical probability P of an event (A) occurring is defined by

$$P(A) = \frac{\text{number of ways in which event A occurs}}{\text{Total number of all possible outcomes}}$$

Mutually Exclusive and Mutually Non-Exclusive Events

Mutually exclusive events are events which cannot occur together

Mutually non-exclusive events are those which happen simultaneously, e.g a pair of events like 2 dice casted.

Additional law of probabilities

For a given event, A and B are mutually exclusive if either A or B occurs, but

or = Addition
but = Multiplication

not both.

$P(A \text{ or } B) = P(A) + P(B)$

Independent Events and Dependent Events

Events are independent when the occurrence of one event does not affect the probability of occurrence of the second event.

Events are dependent when one event just affects the probability of occurrence of a second event.

Independent — Replacement of events
Dependent - Non-replacement.

red orange = 12
blue orange = 10
white orange = 8

~~P(white) P(white orange~~

$$\left.\begin{aligned} P(\text{picking 2 white orange}) &= \frac{2}{30} \quad \text{A} \\ P(\text{picking 5 blue orange}) &= \frac{5}{30} \quad \text{B} \end{aligned}\right\} \text{Independen}$$

$$P(\text{picking 5 red orange}) = \frac{5}{28} \Big\} \text{Dependent}$$

Conditional probability

The conditional probability of an event B occurring when given an event A has already taken place vice-versa
In case of the event B above

$$P\left(\frac{B}{A}\right) = \frac{5/30}{2/30}$$

19-05-2021

Binomial Distribution

This can be used when the probability of success of an event is very high.

The Binomial Formula = $b(x; n, p) = {}^nC_x \cdot p^x \cdot (1-p)^{n-x}$

$$\boxed{b(x; n, p) = {}^nC_x \cdot p^x \cdot (1-p)^{n-x}}$$

Where b = binomial probability

x = total number of successes

p = probability of successes in an individual trial.

n = number of trials.

$${}^nC_x = \frac{n!}{(n-x)!\,x!}$$

Example : A coin is tossed 10 times. What is the probability of getting exactly 6 heads

$p = 0.5$

$q = 1 - 0.5$

$n = 10$

$x = 6$

$$P_6 = {}^nC_x \; p^x \cdot (1-p)^{n-x}$$

$$= {}^{10}C_6 \cdot 0.5^6 \cdot (1-0.5)^{10-6}$$

$$= \frac{10!}{(10-6)!\,6!} \times 0.5^6 \times 0.5^4$$

$$= 210 \times \frac{1}{64} \times \frac{1}{16}$$

$$P_6 = \underline{0.2051}$$

μ - population mean $\quad$ s - SD of sample

$\bar{x}$ - Sample mean $\quad$ σ = SD of population.

Mean = np

Variance = npq

SD = $\sqrt{npq}$

When we have a rare event, the probability is very low

Poisson Distribution.

$$P(x : \mu) = \frac{(e^{-\mu})(\mu^x)}{x!}$$

Where x = Actual number of successes that result from a no of trials

$e = 2.71828$

μ = mean of population.

Properties of Poisson Distribution.

Mean = μ

Variance = μ

SD = $\sqrt{\mu}$

(μ) mean = np

Example

The average number of homes ~~[illegible]~~ sold by a company is 2 homes per day. What is the probability that exactly 3 homes will be sold tomorrow?

Solution

$\mu = 2$ since 2 homes are sold on average per day

$X = 3$

$$P(x:\mu) = \frac{(e^{-\mu})(\mu^{x})}{X!}$$

$$P(x:\mu) = \frac{(e^{-2})(2^{3})}{3!}$$

$$p(x:\mu) = \frac{1.083}{6} = \underline{0.180}$$

Hyper geometric Distribution.

When taking an exact number from a sample of a population.

If a population of ~~tites~~ size "N" contains "K" items of success, failure = $(N-K)$,

then the probability of the hypergeometric random variable (X),

The number of successes in a random sample size (n) is

$$P(X=k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

$$P(x=k) = \frac{{}^{K}C_{k}\ {}^{(N-K)}C_{(n-k)}}{{}^{N}C_{n}}$$

K = success picked from population

k = success picked from sample

Example

A deck of cards contains 20 cards — 6 red cards and 14 black cards. 5 cards are drawn at random, without replacement. What is the probability that exactly 4 red cards are drawn?

Solution

Exactly $= 4 \Rightarrow k$

$N = 20$ $\quad\quad$ $k = 6$

$n = 5$

$N - k = 20 - 6 = 14$

$n - k = 5 - 4 = 1$

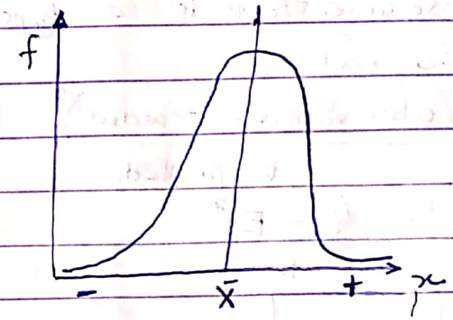
$$\frac{\binom{6}{4}\binom{14}{1}}{\binom{20}{5}}$$

$$= \frac{({}^{6}C_4)({}^{14}C_1)}{{}^{20}C_5}$$

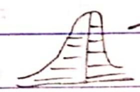
$$= \frac{15 \times 14}{15504}$$

$$= 0.0135$$

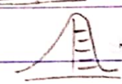
Normal Distribution



$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$



→ 2 tail
no subtraction



→ one tail
Subtract answer from 0.5

$Z =$ Size

$\bar{x} =$ Mean of sample

$\mu =$ mean of population

$\sigma =$ Standard deviation of population

$n =$ size of sample

non-parametric tests – based on observations

26-05-2021

Chi-Square (x^2)

The chi-square test is a type of non-parametric test

$$x^2 = \frac{(\text{Observation} - \text{Expected})^2}{\text{Expected}}$$

$$x^2 = \frac{(O-E)^2}{E}$$

	O	E
ABE	5	5
CVE	6	5
EEE	7	5

5%

Level of significance 0.05

$$\Sigma x^2 = \frac{(5-5)^2}{5} + \frac{(6-5)^2}{5} + \frac{(7-5)^2}{5}$$

Degree of freedom = $n-1$ = 3-1=2

no of ~~observations~~ samples $n=3$

$$0 + \frac{1^2}{5} + \frac{2^2}{5}$$

$$= 0 + 0.2 + 0.8 = 1$$

$$x^2 = 1$$

0

1

3

Observation

	Small	Medium	Large	Σ
White	10	12	8	30
Red	5	13	10	28
Black	2	5	~~12~~	19
Σ	17	30	30	77

Expected

	Small	Medium	Large
White	$\frac{30 \times 17}{77}$	$\frac{30 \times 30}{77}$	$\frac{30 \times 30}{77}$
Red	$\frac{28 \times 17}{77}$	$\frac{28 \times 30}{77}$	$\frac{28 \times 30}{77}$
Black	$\frac{19 \times 17}{77}$	$\frac{19 \times 30}{77}$	$\frac{19 \times 30}{77}$

White	6.62	11.68	11.68
Red	6.18	10.91	10.91
Black	4.19	7.40	7.40

7	12	12
6	11	11
4	7	7

$$x^2 = \Sigma \frac{(O-E)^2}{E} = \frac{(10-7)^2}{7} + \frac{(12-12)^2}{12} + \frac{(30-12)^2}{12}$$

$+ \ldots$

White, Red, Black

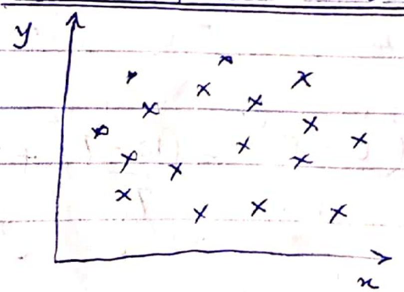
Small, Medium, Big.

degree of freedom = $(3-1)(3-1) = 4$

Measures of Association

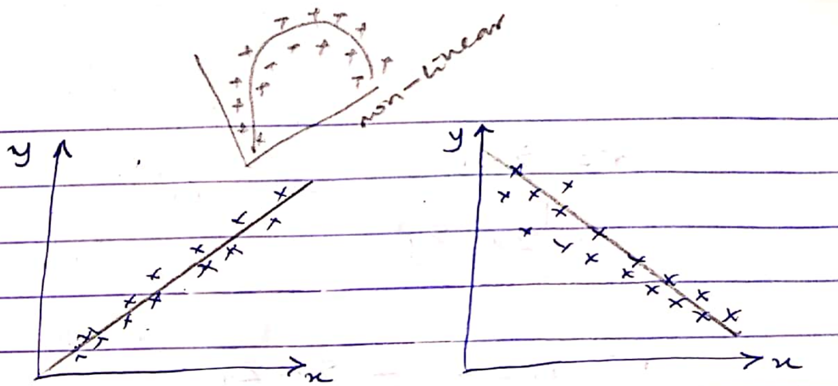
The study of association has to do with bivariance (2 variables), whereby 2 or more variables are considered as a means of viewing the relationship existing between them, When an increase in variable leads to an increase in another variable, we have a positive association of the relations. However when an increase in a variable leads to a decrease in the other, then we have a negative association

Types of Statistical Relationship



Scatter diagram

no correlation



Positive

Positive / Direct correlation

Negative

Negative / Inverse correlation

Karl Pearson's Product - Moment Correlation Coefficient.

This is the most popular mathematical method of quantifying or measuring correlation. It is always denoted by the symbol r, and is given by

$$r = \frac{\text{Covariance } (x, y)}{\sqrt{x} \cdot \sqrt{y}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$\frac{Cov(x,y)}{\sqrt{Var(x) \cdot Var(y)}}$$

$$= \frac{\Sigma xy}{M \sigma_x \cdot \sigma_y}$$

where σ_{xy} = standard deviation of ~~xy~~ x and y respectively.

M = population size

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}} \quad — (1)$$

$$r = \frac{N\Sigma xy - \Sigma x \Sigma y}{\sqrt{\left(N\Sigma x^2 - (\Sigma x)^2\right)\left(N\Sigma y^2 - (\Sigma y)^2\right)}}$$

σ_x σ_y

02/06/2021

<u>Linear Regression</u>

It is used for prediction of dependent variable X and for testing the strength of independent variable Y.

$r = a + bx \quad — (1)$

$\Sigma y = an + b\Sigma x \quad — (2)$

$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad — (3)$

where

a & b = constants

Find the regression line

n = number of sample

$y = a + bx_1 + cx_2 \quad — (1)$

$y = an + b\Sigma x_1 + c\Sigma x_2 \quad — (2)$

$\Sigma x_1 y = a\Sigma x_1 + b\Sigma x_1^2 + c\Sigma x_1 x_2 \quad — (3)$

$\Sigma x_2 y = a\Sigma x_2 + b\Sigma x_1 x_2 + c\Sigma x_2^2 \quad — (4)$

x	y	xy	x^2
4	5	20	16
6	4	24	36
7	5	35	49
8	6	48	64

y^2 44 35 79 78 127

$n = 4$

$\Sigma x = 25$

$\Sigma y = 20$

$\Sigma xy = 127$

$\Sigma x^2 = 165$

$\Sigma y^2 =$

$$b = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

$a = \bar{y} - b\bar{x}$

$$b = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

$y = 5 - 2x$ when $x = 40$, what is y?

$y = 3.6 + 0.2x$

$\Sigma y = an + b\Sigma x$

$20 = 4a + 25b$ —— (1)

$\Sigma xy = a\Sigma x + bx^2$

$127 = 25a + 165b$ —— (2)

Solving simultaneously,

$a = 3.6$

$b = 0.2$

Mrs Fayose

Analysis of Variance : One Way Test

Let us consider sample datas presented below.

2, 4, 4, 3, 4, 8, 7, 8, 9, 9, 2

t test = $\dfrac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

μ = mean of population (given data).

S = SD of sample

n = no of sample

$\bar{x}$ = mean

degree of freedom = $11 - 1 = 10$

μ = mean of entire population

Statistics S1 A-level

i t-table

female - 4, 3, 4, 3, 5, 7, 6, 9, 2, 5, 3

f ratio = $\dfrac{\text{larger variance}}{\text{smaller variance}}$

$\dfrac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \pm \dfrac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

male female

Engr. Dr. Mrs Fayose

Statistical Inference

2 types of statistics

1. Descriptive
2. Inferential

There are 3 main underlying ideas in inference

1. Sample is likely to be a good representative of the population, if the sample is well collected.

Types of Sampling

(1) Random Sampling

Every member of the population is given the equal chance to be selected in a situation. e.g to determine the quality of palm kernels for sale.

Random Sampling is used to eliminate bias, and most theorems in statistics are based on random sampling.

Disadvantage - Random Sampling may not be a good representation of the population.

(2) Stratified Sampling

The population is first divided into 2 or more groups (called strata) and random sampling is then conducted.

Advantage - ~~But~~ They are usually more representative of the population.

Sampling distribution
Confidence Interval
Hypothesis testing.

(3) Cluster Sampling - dividing the population of interest into non-overlapping subgroups, called clusters

Units of population exists already in particular groups. Samples are usually representative of the population

Disadvantage - It may be difficult to locate cluster values

(4) ~~Schematic~~ Systematic Sampling

Elements are selected from a pre-determined interval in population frame

Sampling/Samples collected may or may not be a good representation of population.

To determine the extent of uncertainty in a sample

09-06-2021

SAMPLING DISTRIBUTION

Parameter $\longrightarrow$ Statistic

A parameter is a quantity that measures or describes the population we want to study

A statistic is a portion or part of the sample

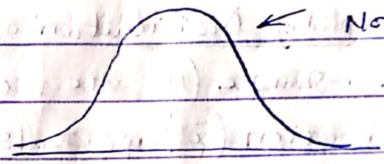
Population (parameter)	Sample (statistic)
Population mean μ	Sample mean $\bar{x}$
Population variance σ	Sample variance
Population Standard deviation	Sample Standard deviation
Population proportion	Sample proportion

$$\begin{array}{r} 17 \\ 21 \\ 21 \\ \hline 60 \end{array} \qquad \begin{array}{r} 18 \\ 19 \\ 25 \\ \hline 62 \end{array}$$

$\bar{x}_1 = \frac{60}{3} = 20$ $\qquad$ $\bar{x}_2 = \frac{62}{3} = 20.6$

Picking 3 samples from a population of 10

$\binom{10}{3} = {}^{10}C_3 = 120$



Normal Distribution

The more the number of samples, the mor your distribution tends to a normal distrib

If $n \geq 30$, it tends to a normal distributio

For Normal distribution,

$$\text{Mean} = \frac{\Sigma x}{n}$$

$$\text{Variance} = \frac{\Sigma (x-\bar{x})^2}{n}$$

$$S.D = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n}} = \sqrt{V}$$

For Sampling distribution

$$\text{Variance} = \frac{\sigma}{\sqrt{n}}$$ $\Rightarrow$ standard error of sampling distribution.

$$\text{Proportion } (\pi) = \sqrt{\frac{\pi(1-\pi)}{n}}$$

$$\sqrt{\frac{P(1-P)}{n}}$$

(i) $\mu_1 = \mu$
(ii) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Properties of Sampling Distribution of Mean

1. For any number of Sample (N), the center of the mean distribution ($\bar{x}$) will also coincide with the mean population being sampled by the spread of the mean (S.D).
2. Distribution will decrease as N increases

Conditions for Properties of Sampling distribution of Sample mean.

$\bar{x}$ = Mean of the observation in a random Sample of size n.

From a population having mean (μ) and standard deviation σ, the mean value of the $\bar{x}$ distribution of (μ) and the Standard Deviation of the $\bar{x}$ distribution by $\sigma_{\bar{x}}$. The following rules hold

Rule 1 : $\mu_1 = \mu$

Rule 2 : $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

=>

Rule 3

When the population distribution is normal, the Sampling distribution is also normal for any Sample size n

Rule 4

Central Limit Theory states that when 'n' is sufficiently large, the sampling distribution is well approximated by a normal curve even when the population distribution is not itself normal.

Sample Distribution for Sample Proportion

Parameter	Statistic
π = Population proportion	$P \Rightarrow$ Sample Proportion

$$P = \frac{\text{No of Successes in the Sample}}{n}$$

Properties of Sample Proportion in Sample Distribution

(1) The sampling distribution of P is centered at π. i.e $\mu_p = \pi$. Therefore, P is an unbiased

statistic for estimating $\bar{\pi}$

(2) The standard deviation of $\sigma_p = \sqrt{\frac{\bar{\pi}(1-\bar{\pi})}{n}}$

as long as n is large and $n\pi \geq 10$

n = no of sample $\quad n(1-\bar{\pi}) \geq 10$

π = proportion

16/06/2021

Confidence Interval

Introduction to confidence Interval for one mean (sigma(σ) known)

Standard deviat

ostatistics

When sampling from a normally distributed population with a known value of σ

$\bar{x}$ is a point estimator of μ - population mean

Sample mean

~~How close is the~~

The confidence interval shows or relates how close the sample mean is to the population mean

->

Point estimate - single value.

Confidence interval is of the form $\bar{x} \pm$ Margin of Error.

Some assumptions will be made, and then an appropriate margin of error is determined.

(1) A simple random sample from a population of interest

(2) A normally distributed population.

(3) The population standard deviation (σ) must be known

Under these assumptions, the confidence interval formula

$(1-\alpha)100\%$ confidence interval for μ

$$\bar{x} \pm Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

where $Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ is the margin of error.

The $(1-\alpha)100\%$ is the confidence level, often chosen to be 95%, ~~99%~~ or ~~99.7%~~

95%
90%
~~[illegible]~~
99.9%

} Standard Intervals

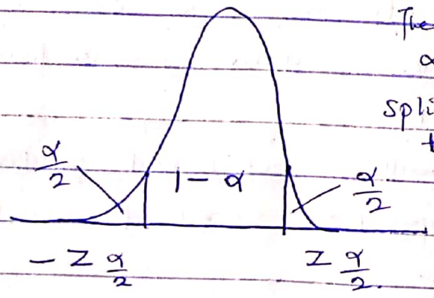
The appropriate Z value is based on the confidence level.

for $(1-\alpha)100\%$

The area under the entire curve is 1

~~The~~ remaining area is α.

split it evenly into both tails



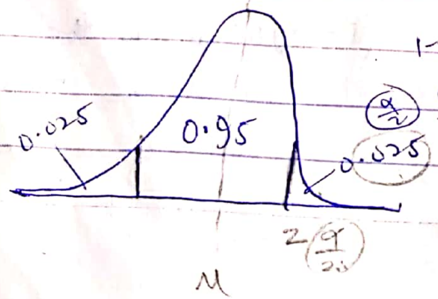
The appropriate Z value of any given confidence level will have to be found from the standard normal distribution table.

For a 95% confidence interval,

$(1-\alpha)100\% = 95\%$

$1-0.95 = 0.05$

$\frac{\alpha}{2}$ $\frac{0.05}{2} = 0.025$



$Z\left(\frac{\alpha}{2}\right)$

$= Z(0.025) = 1.96$ from normal dist. table.

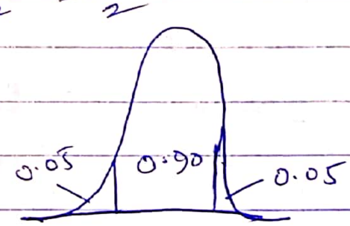
$\therefore \quad \bar{X} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$ $\quad \pm \; \mp$

for a 90% confidence interval,

$(1-\alpha)100\% = 90\%$

$1-0.90 = 0.10$

$\frac{\alpha}{2} = \frac{0.10}{2} = 0.05$



$Z\left(\frac{\alpha}{2}\right) = Z(0.05) = 1.645$

Example $n = 135$

$\bar{x} = 0.988$

find 95% confidence interval for the population mean

(suppose it is known that $\sigma = 0.028$)

$n = 135$

$\bar{x} = 0.988$ $\qquad \alpha = 1 - 0.95 = 0.05$

$\sigma = 0.028$ $\qquad \frac{\alpha}{2} = \frac{0.05}{2} = 0.025$

$\bar{x} \pm Z\left(\frac{\alpha}{2}\right) \times \frac{\sigma}{\sqrt{n}}$ $\qquad Z(0.025) = 1.96$

$= 0.988 \pm Z(0.025) \times \frac{0.028}{\sqrt{135}}$

$= 0.988 \pm 1.96 \times \frac{0.028}{\sqrt{135}}$

$= 0.988 \pm 0.0047$

$\simeq [0.988 + 0.0047, \; 0.988 - 0.0047]$

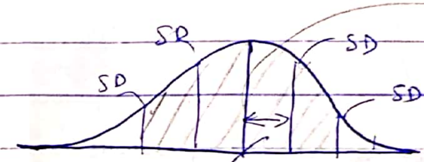
$= (0.983, 0.993)$

$(1-\alpha)100 = 95\%$

$1 - 0.95 = 0.05$

Therefore we are 95% confident that the population mean lies in this interval.

Video 2 – Confidence Intervals and margin of error – Khan Academy.



P = population proportion.

$\hat{p}$ = sampling proportion

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

There is a 95% probability that ~~is~~ the ~~sample~~ population proportions P is within 2 standard deviations of $\hat{p}$

Sample proportion $\hat{p} = 0.54$, $n = 100$

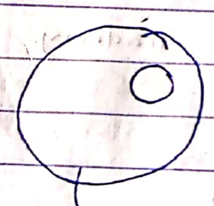
~~Sample~~ Standard Error of sample proportion $\sigma_{\hat{p}}$

$\sigma \; SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} =$

$0.54 - (0.10)$, $0.54 + (0.10)$

$= \sqrt{\frac{0.54 \cdot 0.46}{100}}$

≈ 0.05



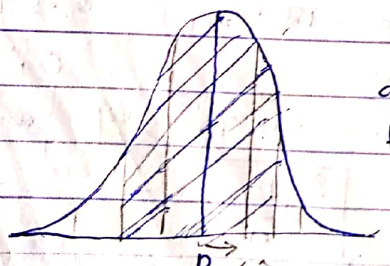
Population of likely voters

P = proportion that supports candidate A

Sample 1 : n = 100

sample proportion $\hat{p} = 0.54$

Sample distribution of the sample proportion for $n = 100$



The mean of the sampling distribution = Population of the proportion

$P \quad \sigma\hat{p}$ = standard deviation of the sample proportion.

23/06/2021 **SAMPLING DISTRIBUTION**

The average male drinks 2L of water when active outdoors (with a standard deviation of 0.7L). You are planning a full day nature trip for 50 men and will bring 110L of water. What is the probability that you will run out

(average / mean) population mean μ = 2L

$\sigma = 0.7L$

$n = 50$

$P > 110L$

$n > 30$ for normal distribution

Sample mean $= \bar{x} = \frac{110}{50} = 2.2L$ (sample mean)

Population mean $\mu = 2L$

Sample mean $= \bar{x} = 2.2L$

Variance $(\sigma_{\bar{x}}^2) = \frac{\sigma^2}{n}$

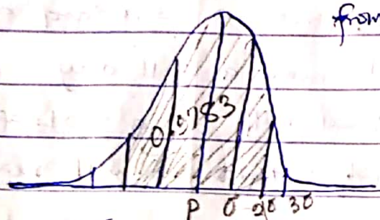
$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.7}{\sqrt{50}} = 0.099$

$$Z = \frac{\bar{x} - \mu}{\sigma} = \frac{2.2 - 2}{0.99} = 2.02$$

from Z table, $Z(2.02)$

$= 0.9783$



$P(\text{run out of water}) = 1 - 0.9783$

$= 0.0217$

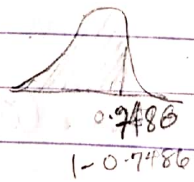
$= \underline{2\%}$

② Suppose samples of 100 items were drawn from a population with a mean of 200 and standard deviation 30. What proportion of the samples will have mean

(a) greater than 202

(b) < 199

(c) Between 198.5 and 203

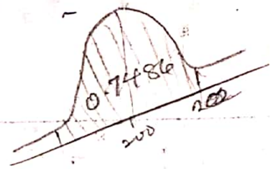


$n = 100$

$\mu = 200$ (population mean)

$\sigma = 30$

(i) $\bar{x} = 202$



$$Z = \frac{x - \mu}{\sigma_{\bar{x}}} \qquad x = 202$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{30}{\sqrt{100}} = 3.$$

$$Z = \frac{202 - 200}{3}$$

$$Z = \frac{2}{3} = 0.67$$

$P(z = 0.67)$ positive tables

$z = 0.7486$

$= 1 - 0.7486 = (0.2514) \Rightarrow 25.14\%$

(ii) < 199

$$Z = \frac{x - \mu}{\sigma_{\bar{x}}}$$

$$z = \frac{199 - 200}{3} = -\frac{1}{3} = -0.33$$ negative tables.

(negative tables) $= 0.3707 = 37.07\%$

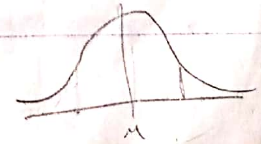
(iii) Between 198.5 and 203.

$$z = \frac{198.5 - 200}{3}$$

$$= -\frac{1.5}{3} = -0.50$$

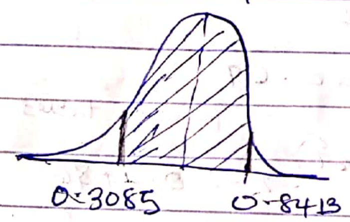
$\le$

$Z(-0.5) = 0.3085$



$$Z = \frac{203-200}{3} = \frac{3}{3} = 1.00$$

$Z_{1.00} = 0.8413$



$\bar{x}$ is a point estimator of μ. $\bar{x}$ takes a single value.

$= 0.8413 - 0.3085$

$= 0.5328$

$= 53.28\%$

Confidence Interval

A sample of 135 women yielded an average of 2D/4D ratio of 0.988. What is a 95% confidence interval for the population mean? (Suppose it is known that $\sigma = 0.028$)

$0.95 = 1.6$

$0.90 = 2.58$

30-06-2021

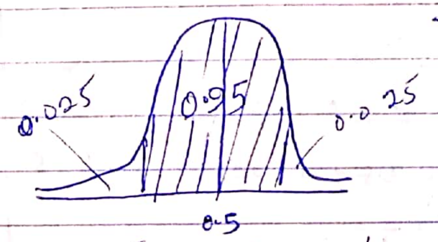
Confidence Interval

Let P = Sample proportion

Let π = the population proportion of the defective nails

P = sample proportion = $\frac{\text{no of defective in the sample}}{n}$

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



$2(0.025) \Rightarrow (0.025 + 0.95)$

$= 0.9750$

The variance of bolts ~~joined~~ heights produced in an industry is 100 while the mean height of ~~samples it~~ 16 bolts collected from the

industry is 185
find a 95% confidence interval for the population mean and make a confidence statement.

Solution

Bolt height (σ^2) - population variance $= 100$

$S.D \doteq \sigma = \sqrt{100} = 10$, $n=16$, $\bar{x} = 185$.

$$\bar{x} \pm z\left(\frac{\alpha}{2}\right) \frac{\sigma}{\sqrt{n}}$$

$$1 - 0.95 = 0.05$$

$$\frac{\alpha}{2} = 0.025$$

$$z(0.95 + 0.025) = 0.975$$

$$\simeq 1.96$$

$$185 \pm 1.96 \times \frac{10}{\sqrt{16}}$$

Upper limit $= 189.15$

lower limit $= 180.10$.

Confidence statement - The result shows that at 95% confidence level (~~of~~ probability), the interval 180.10 and 189.5 contains the mean height of the bolt population.

~~One of the purposes of the survey on affirmative action described~~

Example 9.1

$$P = \frac{\text{no of successes in the sample}}{n}$$

$$P = \frac{537}{1013} = 0.530$$ - Point estimate of proportion.

1. Assume a confidence level first. (95%)

$$\bar{\pi} = P \pm Z\left(\frac{\alpha}{2}\right) \sqrt{\frac{P(1-P)}{n}} \qquad nP \geq 10$$

$$= 0.530 \pm 1.96 \sqrt{\frac{0.530(1-0.530)}{n}}$$

$$= (0.499, 0.561)$$

confidence statement - ...

Confidence interval ~~for~~ for the population mean (μ) when σ is unknown.

σ — population S.D

$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$-1.96 < \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1.96$$

$$\bar{x} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{x} + 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

If S.D is unknown, use the sample data to determine it.

Then t distribution

$$t = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

* sigma unknown
* sample size small

↓

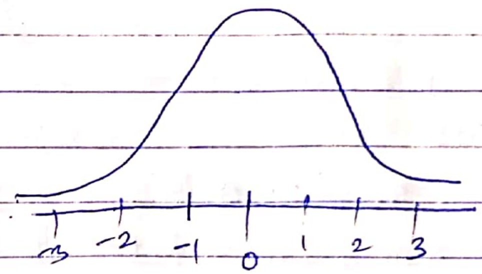
Use T-table

estimate from the sample S.D

Use t-table

Curve for t table

Similar to the z table



Properties

The t-value corresponding to any number of degree of freedom is

Degree of freedom $= n - 1$

If $n = 30$, $d = (30 - 1) = 29$.

Example (Population SD unknown).

From a sample of 49 students, the mean mathematics score in a school was found to be 55%, with a SD of 15.

Find the 99% confidence interval for the real mean score of mathematics in the school.

$\bar{x} = 55\%$

$n = 49$

$d = 49 - 1 = 48$

confidence level = 99%

$S = 15$

$t(df) = t_{95\%}^{(48)} = 2.70$

A solved problem from youtube on t-distribution

$$x = a'y + b$$

$$a' = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum y^2 - (\sum y)^2}$$

$$b' = \bar{x} - a'\bar{y}$$

SUMMARY NOTES

Types of Statistics

(1) **Descriptive Statistics**

This summarizes data using graphs and summary values like means and interquartile range. Descriptive statistics helps to realize relationships and patterns but do not draw conclusions beyond the data we have at hand

(2) **Inferential Statistics**

This analysis helps us to draw conclusions beyond the data we have from the population from which it was drawn.

Statistical Inference is the process of drawing conclusions about population parameters based on a sample taken from the population. The data we collect from the population is called a Sample

* The Sampling distribution of a statistic is the probability distribution of that statistic
* Statistic comes from sample
* Parameter comes from a population.
* A parameter is a measurement or quantity that ~~comes from~~ describes a population
* A statistic is the measurement or quantity that helps to describe the sample.

Examples of parameter

1. Population mean
2. population Variance
3. Population Standard Deviation

Examples of Statistic

1. Sample mean
2. Sample variance
3. Sample standard deviation

* The Sampling distribution of the statistic is the subject of statistical Inference. It is the distribution of the statistic if we are to repeatedly draw samples from the population.

A large Sample is a sample with more than 30 items i.e $n > 30$

The Standard Error - Standard deviation of a sampling distribution.

Standard Error of Sample mean $= \frac{\sigma}{\sqrt{n}}$

Standard Error of proportions $\sqrt{\frac{p(1-p)}{n}}$

* Statistic - Any quantity computed from values in a sample is called a statistic.

* Sample Proportion - The fraction of individuals or objects in a sample that have some characteristics of interest.

OR - The ~~population~~ proportion of individuals in a sample that possesses a population property.

* **Unbiased Statistic**

A statistic whose mean value is equal to the value of the population characteristic being estimated is said to be an unbiased Statistic

A statistic that is not unbiased is said to be biased.

* Among several unbiased statistics, the best statistic to use is the one with the smallest standard deviation.

* For unbiased statistic, the best statistic for obtaining a point estimate of the population mean μ is the sample mean $\bar{x}$

* **Point Estimate:** The point Estimate of a population characteristic is a single number that is based on sample data and represents a plausible value of the characteristic.

Population – The entire collection of individuals or objects about which an information is desired

Sample – A subset of the population selected for study

A bowl contains 5 red balls, 4 white and 3 green balls. 2 balls are drawn one after the other without replacement. What is the probability of drawing a red ball first ? ~~(ii) that a white ball is drawn the first~~

Soln

(i) $P(RW) + P(RR) + P(RG)$

$\because$ Total $= 5 + 4 + 3 = 12$ balls

$P(R,W) = \frac{5}{12} \times \frac{4}{11} = \frac{5}{33}$

$P(R,R) = \frac{5}{12} \times \frac{4}{11} = \frac{5}{33}$

$P(R,G) = \frac{5}{12} \times \frac{3}{11} = \frac{5}{44}$

$= \frac{5}{33} + \frac{5}{33} + \frac{5}{44} = \frac{5}{12}$

(ii) The probability of picking a white ball the second time, given ~~that~~ a red ball was drawn the first time.

$P(R,W) = \frac{5}{12} \times \frac{4}{11} = \frac{5}{33}$

(iii) Drawing a green ball and white ball in that order

$P(G,W) = \frac{3}{12} \times \frac{4}{11} = \frac{1}{11}$

- 2 types of t-distribution
- When to use z-tables and t-tables
- Notes, Examples, Assignments

28-07-2021

A statement made/claim made that is to be tested – Null Hypothesis e.g The average height of FUOYE Students ~~statem~~ is 1.7m.

Alternative Hypothesis – Hypothesis against the null-hypothesis

* When the population standard deviation is not known, the t-distribution is used.
* When the standard deviation is ~~not~~ known, use z-table $z = \frac{x - \mu}{\frac{\sigma}{\sqrt{n}}}$ — Standard error

$t_0 = \frac{x = \mu}{\frac{s}{\sqrt{n}}}$ $s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n-1}}$

Example

$\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$ The breaking strength of cables produced by a manufacturer has a mean of $\mu = 1800lb$ and standard deviation σ of $100 lb$. By a new

90% − 1.645
95% − 1.96
99% — 2.58
~~99.73 − 3~~

technique in manufacturing, it is claimed that the strength is increased. To test this claim, a sample of 50 cables is tested. It was found out that the mean is 1850 Lb ($\bar{x}$). Can we support this claim?

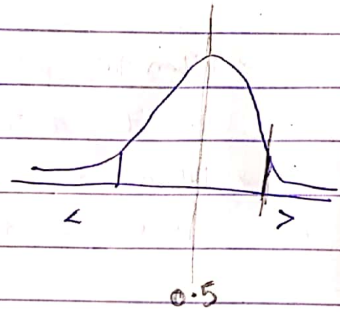
Solution

$\mu = 1800$ lb

$\sigma = 100$ lb

$n = 50$

$\bar{x} = 1850$ lb



$H_0 : \mu = 1800$ lb
$H_a : \mu > 1800$ lb

This test is a one-tailed test.

Select a significance level

Significance level = 1 − Confidence level.

Use significance level of $\alpha = 0.01$ (99%) for this question.

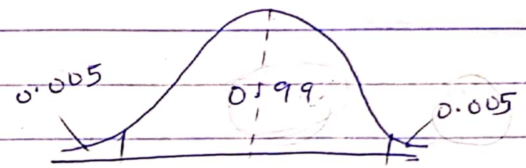
test statistic = z

$$\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{1850 - 1800}{\frac{100}{\sqrt{50}}} = 3.54$$

95% 0.95
$1 - 0.95 = 0.05$ α
0.025 0.95 0.025

Based on the significance level, go to the z table and check the p-value

$0.01 \Rightarrow (1 - 0.01) = 0.99$



$$\begin{array}{r} 0.99 \\ +\,0.005 \\ \hline 0.995 \end{array}$$

$= 2.58$

Using the 2-tailed table ⇒ 0.35

$\boxed{2.33}$

For a right-hand tailed test, use 2.33 for 0.01 significant level (99%).

~~Since p-value = 2.33, is less than 3.54, the Null hypothesis is accepted~~

If z-score is > the value of the p-value, the H_0 is not correct; H_a is accepted so the claim is supported.