Probability and Statistics

ishma hafeez
notes

Osama.eifaz @nu

1. Walpole
2. Prem S
3. Neil A Weiss

Ishma Hafeez notes

Mid 1	15%
Mid 2	15%
final	50%
Qs/As	20%

- Introduction, variables, momentum of scalars, types of data
- mean, median, mode, quartiles, variants, standard deviation, co.efficience of variant } Mid 1
- Histogram, Bar plot, Box plot, dot plot, frequency, polygon, stan & leaf
- Probability, conting technique, Permutation, combination
- venn digram, addition rule, multiplicative rule
- Conditional Probability and Bayers rule

Statistics

The science of collection, Presentation, analysis and interpretation of numerical facts/data

OR

its the collection of procedures and principals for gathering data and analysing information to help people make decisions when found with uncertainty

STATISTICS

Description
- Data summary
- Graphs tables
- Mean, median, mode

Information
- Drawing conclusions about population using samples

Observational study

↳ sample survey

1. researchers observe characteristics
2. take measurements

Designed Eperiment

1. researchers impose treatments and controls
2. then observe characteristics
3. take measures

Population:
The collection of all individuals/objects under consideration in a statistical study

Parameter:
characteristics of population
- mean (μ), variant (σ^2) etc

Sample:
subset of population

Statistic:
characteristic of sample
- mean ($\bar{x}$), variant (s^2) etc

Random sample:
It is a sample selected in such a way that every member of the population has an equal chance of being selected

VARIABLE

→ THIS WHOLE COURCE

Qualitative
- ↳ non numerical
- ↳ categorical
- ↳ Gender
- ↳ Religion
- ↳ Races

LONG STATEMENTS
FEELINGS
JOB INTERVIEWS

Quantitative
- ↳ numerical
- ↳ A, W, H

QUESTIONARIES

Discrete
- ↳ whole numbers
- ↳ data values consists of gaps
- ↳ no of students

Continous
- ↳ w/o gap
- ↳ intervals
- ↳ 64"
  63 — 64.1"

DATA

PRIMARY
- ↳ raw data
- ↳ first hand information
- ↳ survery questionnaries

SECONDARY
- ↳ already collected by someone
- ↳ Published reports of research organisations

Cross section data

data collected on different elements at the same point in time

↳ objects are different

↳ Time is constant

University	Annual Cost (dollars)
Harvard University	69,600
Princeton University	66,150
Stanford University	69,100
University of California, Berkeley	65,003
University of Chicago	75,735
University of Pennsylvania	71,715
Yale University	71,290
MIT	67,430

Time series data

Data collected on same element for same element at different intervals in time

↳ 1 variable

↳ different time intervals

Years	Total Population
2015	320,878,310
2016	323,015,995
2017	325,084,756
2018	327,096,265
2019	329,064,917

Panel

↳ different cross sections

↳ different time period

MEASUREMENTS OF SCALES

Qualitative → String

if you can't take avg

Quantitative

Nominal scales

↳ non numerical data

↳ categorial data

↳ classification

↳ Gender / Blood group

↳ Religion

↳ Proffesion

↳ zip code

Ordinal scale

↳ ranking

↳ ordered data

↳ customer rating (strongly agree, agree, neutral, disagree)

↳ Likert scale

↳ Performance of students

Interval scale

↳ +, −

↳ absolute zero does not exist

↳ Temprature

↳ IQ scores

↳ variables for which '0' isn't meaningful

Ratio Scale

↳ +, −, *, ÷

↳ absolute zero exist (Show something absence)

↳ Area / volume / lenght

↳ Distance / Weight / Height

↳ Time

↳ salary

↳ Age

8)

variable	Scale
ranking of golfers	ordial
temp	interval
weight	ratio
Salries	ratio
ratings	Ordinal
categories	nominal

Frequency Distribution

↳ Histogram

↳ Polygon

↳ Ogive

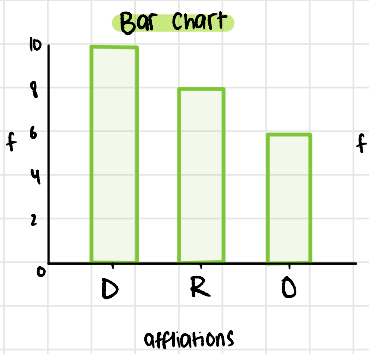
Ordered Array

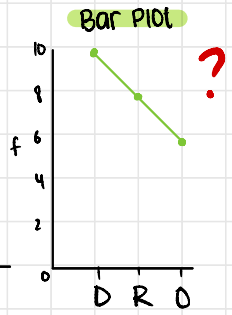
↳ Stem and leaf

Affiliations	Tally	f
D	## ##	10
R	## ///	8
O	## /	6

↗ qualitative

D	R	O	R
D	D	R	R
O	O	D	D
R	R	R	R
D	D	D	O
O	D	O	D





Stem and leaf Plot

leading digits ↓ ↘ trailing digits

70	64	99	55	64	89	87	65
62	38	67	70	60	69	78	39
75	56	71	51	99	68	95	86
57	53	47	50	55	81	80	98
51	36	63	66	85	79	83	70

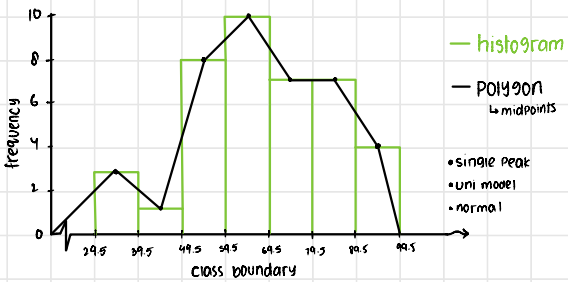
3	8,6,9
4	7
5	7,1,6,3,5,1,0,5
6	2,4,7,3,6,4,0,7,5
7	0,5,1,0,9,8
8	5,9,1,7,0,3,6
9	9,9,5,8

SORT IT

Stem	leaf
3	6 8 9
4	7
5	0 1 1 3 5 5 6 7
6	0 2 3 4 4 5 6 7 7
7	0 0 1 5 8 9
8	0 1 3 5 6 7 9
9	5 8 9 9

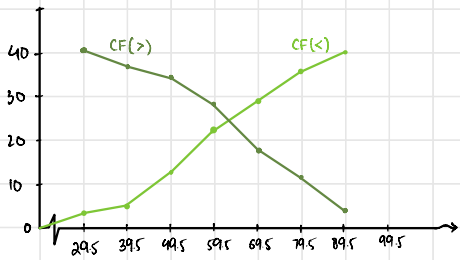
Class interval (C.I)	Tally	frequency (f)	cumulative frequency CF(<)	CF(>)	R.F ($\frac{f}{\text{total}}$) rounded up	-0.5 / +0.5 class boundary
30 - 39		3	3	40	$3/40 = 0.075$	29.5 - 39.5
40 - 49		1	4	37	$1/40$	39.5 — 49.5
50 - 59		8	12	36		49.5 — 59.5
60 - 69		10	22	28		59.5 — 69.5
70 - 79		7	29	18		69.5 — 79.5
80 - 89		7	36	11		79.5 — 89.5
90 - 99		4	40	4	$4/40$	89.5 — 99.5
Total		40				

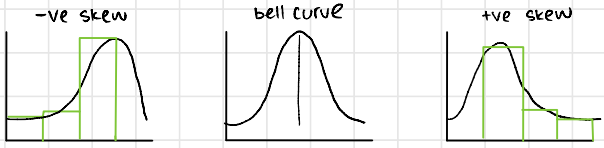
Histogram / Polygon



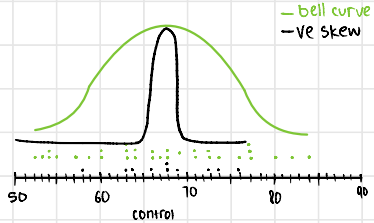
Ogive

↳ CF





Dot Plot



Intervention		Control						
68	66	74	52	67	63	77	57	80
74	58	77	53	76	54	73	54	
69	63	60	77	63	60	68	64	
68	73	66	71	66	55	71	84	
64	76	63	73	59	68	64	82	

Pareto Charts

When the variable displayed on the horizontal axis is qualitative or categorical, a Pareto chart can also be used to represent the data.

A **Pareto chart** is used to represent a frequency distribution for a categorical variable, and the frequencies are displayed by the heights of vertical bars, which are arranged in order from highest to lowest.

Homeless People

The data shown here consist of the number of homeless people for a sample of selected cities. Construct and analyze a Pareto chart for the data.

City	Number
Atlanta	6832
Baltimore	2904
Chicago	6680
St. Louis	1685
Washington	5518

BAR GRAPH but arranged by their heights in descending order

Solution

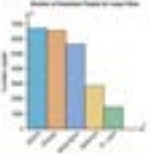
Step 1 Arrange the data from the largest to smallest according to frequency.

City	Number
Atlanta	6832
Chicago	6680
Washington	5518
Baltimore	2904
St. Louis	1685

Step 2 Draw and label the x and y axes.

Step 3 Draw the bars corresponding to the frequencies.

The graph shows that the number of homeless people is about the same for Atlanta and Chicago and a lot less for Baltimore and St. Louis.



Suggestions for Drawing Pareto Charts

1. Make the bars the same width.
2. Arrange the data from largest to smallest according to frequency.
3. Make the units that are used for the frequency equal in size.

The Time Series Graph

When data are collected over a period of time, they can be represented by a time series graph. used to study patterns

A **time series graph** represents data that occur over a specific period of time.

Workplace Homicides

The number of homicides that occurred in the workplace for the years 2007 to 2008 is shown. Draw and analyze a time series graph for the data.

Year	'02	'04	'05	'06	'07	'08
Number	632	559	567	540	628	517

Source: Bureau of Labor Statistics

Solution

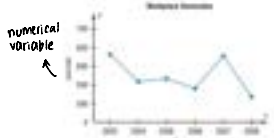
Step 1 Draw and label the x and y axes.

Step 2 Label the x axis for years and the y axis for the number.

Step 3 Plot each point according to the table.

Step 4 Draw line segments connecting adjacent points. Do not try to fit a smooth curve through the data points.

There was a slight decrease in the years '04, '05, and '06, compared to '03, and again an increase in '07. The largest decrease occurred in '08.



numerical variable

time period

The Pie Graph

Pie graphs are used extensively in statistics. The purpose of the pie graph is to show the relationship of the parts to the whole by visually comparing the sizes of the sections. Percentages or proportions can be used. The variable is nominal or categorical.

A **pie graph** is a circle that is divided into sections or wedges according to the percentage of frequencies in each category of the distribution.

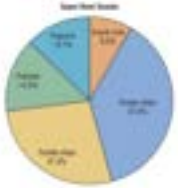
Super Bowl Snack Foods

This frequency distribution shows the number of pounds of each snack food eaten during the Super Bowl. Construct a pie graph for the data.

Snack	Pounds (frequency)	Percentages Step 1	Angles Step 2
Potato chips	11.2 million	(11.2/30)*100 = 37.3	(11.2/30)*360 = 134°
Tortilla chips	8.2 million	(8.2/30*100 = 27.3	(8.2/30)*360 = 98°
Pretzels	4.3 million	14.3	52
Popcorn	3.8 million	12.7	46
Snack nuts	2.5 million	8.3	30
Total	30.0 million	99.9	360

Step #3

Next, using a protractor and a compass, draw the graph using the appropriate degree measures found in step 2, and label each section with the name and percentages



C.I (Class interval)	Tally	f (Frequency)	Mid Point u	fu	c.f
30 – 39		3	$\frac{30+39}{2} = 34.5$	$f \times u =$ 103.5	3
40 – 49		1	44.5		4
50 – 59		8	54.5		12 → Preceding value
60 – 69		10	64.5		22 → closest to 20
70 – 79		7	74.5		29
80 – 89		7	84.5		36
90 – 99		4	94.5	377.8	40
Total		40		2720	

$$\text{mean}(\mu) = \frac{\Sigma u}{n} \qquad \text{coeff of variance} = \frac{\sigma \times 100}{\mu}$$

$$\text{median} = \left(\frac{n+1}{2}\right)^{th} \qquad \text{Standard deviation } (\sigma)$$

$$\tilde{x} = l + \frac{h}{f}\left[\frac{\Sigma f}{2} - Cf_{\text{preceding}}\right]$$

(l = (L.C.B) lower class boundary median class; h = class width; f = median class frequency; $Cf_{\text{preceding}}$ = cumulative frequency previous median class)

LCB = −0.5
UCB = +0.5

$$\text{mode} = l + \left[\frac{f_m - f_1}{2f_m - f_1 - f_2}\right] h$$

(l = (L.C.B) lower class boundary modal class; f_m = max frequency; f_1 = preceding; f_2 = following; h = class size)

grouped ← variance → ungrouped

$$\sigma^2 = \frac{\Sigma[f(u-\mu)^2]}{\Sigma f} \qquad \sigma^2 = \frac{\Sigma(u-\mu)^2}{n} = \frac{\text{Sum of squared deviations from mean}}{\text{no of observations}}$$

GROUP

Q) find mean

$\mu = \frac{2720}{40} = 68$

Q) find mode

max frequency is at 60–69 → $59.5 + \frac{10-8}{2(10)-8-7} \times 10$

$= 63.5$

Q) $\tilde{x} = 59.5 + \frac{10}{10}\left[\frac{40}{2} - 12\right]$ (60-0.5; 40/2 = 20; choose preceding to 22)

$= 67.5$

Q) find standard deviation

$\mu = 68$ → found above

$$\sigma^2 = \frac{3(34.5-68)^2 + 1(44.5-68)^2 + 8(54.5-68)^2 + 10(64.5-68)^2 + 7(74.5-68)^2 + 7(84.5-68)^2 + 4(94.5-68)^2}{40}$$

$$= \sqrt{\frac{10\,510}{40}} = 16.209$$

Q) find coeff of variance

$= \frac{16.209}{68} \times 100 = 238\%$

68 > 67.5 > 63.5

$\bar{x}$ mean

$\tilde{x}$ median

$\hat{x}$ mode

$\bar{x} > \tilde{x} > \hat{x}$

+ve skewed

UNGROUP

Q) $u = 2, 4, 6, 8$, find mean → no mode as all values occur exactly once

$\mu = \frac{2+4+6+8}{4} = 5$

Q) 1, 2, 3, find medium

$= \frac{3+1}{2} = 2^{nd}$ position → has 2

Q) 2, 4, 6, 8, find medium

$\frac{4+1}{2} = 2.5^{th} \rightarrow \frac{2^{nd} + 3^{rd}}{2} \rightarrow \frac{4+6}{2} = 5$

Q) 2, 3, 5, 8, find variance

$\mu = \frac{18}{4} = 4.5$

$$\sigma^2 = \frac{(2-4.5)^2 + (3-4.5)^2 + (5-4.5)^2 + (8-4.5)^2}{4}$$

$= \sqrt{5.25}$

$= 2.3$

APPLICATIONS

3.49 The following table gives the frequency distribution of the number of hours spent last week on cell phones (making phone calls and texting) by all 100 students in the tenth grade at a school.

Hours	Number of Students (f)
0 to less than 4	14
4 to less than 8	18
8 to less than 12	25
12 to less than 16	18
16 to less than 20	16
20 to less than 24	9
	100

Find the mean, variance, and standard deviation

x	fx	cf
2	28	14
6	108	32
10	250	57
14	252	75
18	288	91
22	198	100
	1124	

$$\mu = \frac{1124}{100} = 11.24$$

$$\text{mode} = 8.5\left[\frac{25-18}{2(25)-18-18}\right] \times 4$$

$$\sigma^2 =$$

$\bar{x} \pm s$ → Approx. 68% observations

$\bar{x} \pm 2s$ → Approx. 95% observations

$\bar{x} \pm 3s$ → Approx. 99.7 observations

• Consider the following sample of exam scores, arranged in increasing order. The sample mean and sample standard deviation of these exam scores are, respectively, 85 & 16.1.

39	57	58	64	69	74
79	80	81	85	85	87
87	89	89	90	92	93
94	94	95	96	96	97
97	97	97	98	100	100

Use the data to obtain the exact percentage of observations that lie within two standard deviations to either side of the mean. Compare your answer

Empirical rule is fixed if the data is normally distributed or symmetric.

(1s = 85 ± 16.1 → 68% observation will fall in this interval

(1s2s = 85 ± 2*16.1 → 95 observation will fall in this interval

(1s3s = 85 ± 3*16.1 → 99.73 observations will fall in this interval.

[do this if you are asked to check the upper empirical rule otherwise not needed!]

Rule	Lower to Upper	Frequency
85 ± 16.1	68.9 to 101.1	
85 ± 2*16.1	52.8 to 117.2	
85 ± 3*16.1	36.7 to 133.3	

Quartiles

5, 15, 16, 20, 21, 25, 26, 27, 30, 30

31, 32, 32, 34, 35, 38, 38, 41, 43, 66 → outlier > outside max

data set represent

no. of observations

25% $Q_1 = \left(\frac{n+1}{4}\right)^{th}$ lower data set $\left(\frac{20+1}{4}\right)^{th} \Rightarrow 5.25^{th} \rightarrow \frac{5^{th}+6^{th}}{2} = \frac{21+25}{2} = 23$

50% $Q_2 = \left(\frac{2(n+1)}{4}\right)^{th}$ median $2(5.25) \Rightarrow 10.5^{th} \rightarrow \frac{10^{th}+11^{th}}{2} = \frac{30+31}{2} = 30.5$

75% $Q_3 = \left(\frac{3(n+1)}{4}\right)^{th}$ upper data set $3(5.25) \Rightarrow 15.75 \rightarrow \frac{15^{th}+16^{th}}{2} = \frac{35+38}{2} = 36.5$

Box and Whiskers Plot

↳ Data shape (skew)

↳ Outliers

↳ variation (box length) (IQR)

$IQR = Q_3 - Q_1$

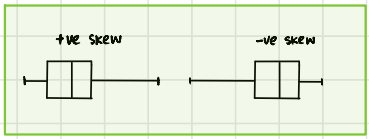
min, Q_1, Q_2, Q_3, max

min ↓ L

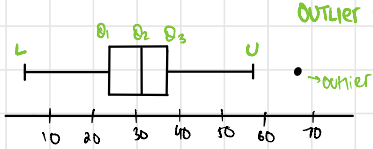
max ↓ U

$Q_1 - 1.5(IQR) = 2.75$

$56.75 = Q_3 + 1.5(IQR)$



OUTLIER SO USE L U



NO outlier so use max min

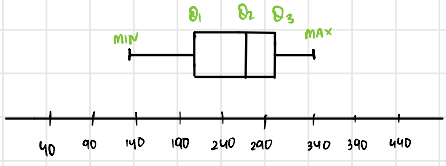
	1	2	3	4	5	6	7	8
Real cheese	$x_1 \rightarrow$	40	45	90	180	220	240	310	420
Cheese substitute	$x_2 \rightarrow$	130	180	250	260	270	290	310	340

2.25	4.50	6.75				no outliers
Q_1	Q_2	Q_3	IQR	L		U
67.5	200	275	207.5	-243.75		586.25
215	265	300	85	87.5		431.25

REAL CHEESE +ve skew

has more sodium content

greater variation



CHEESE SUBSTITUTE -VE SKEW

Probability

Statistical Experiment

Any activity/process that generates data

↳ Tossing a coin

↳ Rolling a dice

↳ Playing cards

↳ survery/opinion of voters

Sample space

List of all possible outcomes of a statistical experiment

↳ Tossing a coin 2 times: {HH, HT, TH, TT}

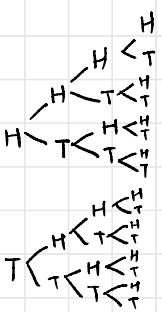
↳ Rolling a device: {1,2,3,4,5,6}

event: subset of a sample space

* Toss a coin 4 times / 4 coins tossed together

possibilities 2^{4} ← tosses $= 6$

Tree Digram



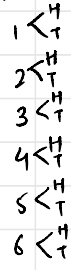
list of elements

S = { HHHH, HHHT, HHTH, HHTT, HTHH, HTHT, HTTH, HTTT, THHH, THHT, THTH, THTT, TTHH, TTHT, TTTH, TTTT }

* Tossing a dice & coin together

$6 \times 2 = 12$?

Tree Digram



list of elements

(1,H)(1,T)

(2,H)(2,T)

(3,H)(3,T)

(4,H)(4,T)

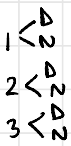
(5,H)(5,T)

(6,H)(6,T)

* 3 items < detected / non detected

$2^{3} = 8$

Tree Digram



list

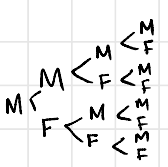
(1,D)(1,N)

(2,D)(2,N)

(3,D)(3,N)

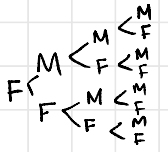
* 4 students as either M or F

1st sample space



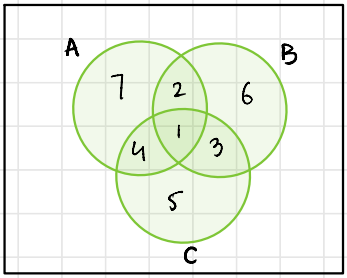
2nd sample space

$S_2 = \{0,1,2,3,4\}$



$S_1 = \{$ MMMM, MMFM ... $\}$

VENN DIGRAM



$A \cap B = 1, 2$

$B \cap C = 1, 3$

$A \cup C = 1, 2, 3, 4, 5, 7$

$B' \cap A = 4, 7$

$A \cap B \cap C = 1$

$(A \cup B) \cap C' = 7, 2, 6$

Permutations

↳ order matters

↳ organise

↳ ways

$${}^{n}P_{r} = \frac{n!}{(n-r)!}$$

Combinations

↳ order doesn't matter

↳ different types

↳ select

$${}^{n}C_{r} = \frac{n!}{(n-r)!\,r!}$$

Empirical formula

Probability

It is a measure of chance that an uncertain event will occur

Subjective

↳ personal experience

↳ judgement

Objective

↳ classical approach

$P(\text{Event : A}) = \frac{\text{favourable case of A}}{\text{all possible cases}}$

$\leq$ Prob $= 1$, $0 \leq P(A) \leq 1$

1) A coin is tossed twice. Prob atleast 1 head

$S = \{HH, TH, HT, TT\}$ $= 3/4$

3) A dice drawn such that an even no is twice likely to occur than odd

$$\begin{Bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ w & 2w & w & 2w & w & 2w \end{Bmatrix}$$

$P(X<4) = \frac{1}{9} + \frac{2}{9} + \frac{1}{9} = \frac{4}{9}$

5) A card is drawn at random from 52 cards

a) $P(I) = \frac{{}^4C_1}{52C_1} = \frac{4}{52}$

b) In a poker hand consisting of 3 cards Prob of 2 Aces and 3 Jacks

$$\frac{({}^4C_2)({}^4C_3)}{{}^{52}C_5} =$$

(2 Aces, 3 Jacks; total; need now anyway)

8) 10 E 10 M

25 I 8 C

a) $P(I) = 25/53$

b) $P(C \text{ or } E) = P(C) + P(E) = \frac{8}{53} + \frac{10}{53}$

9) Dice rolled. Sum 7 or 11 → exclusive

$P(7) = \frac{16}{43}\ \frac{25}{52}\ \frac{36}{61}$ $P(11) = \frac{65}{56}$

$P(7 \text{ or } 11) = \frac{6}{36} + \frac{2}{36}$

OR → has + → only one event from choice

AND → has * → more than 1 events

Mutually exclusive events

↳ nothing in common

$P(A \cup B) = P(A) + P(B)$

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

Non Mutually exclusive events

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

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

Independent events

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

if $P(A|B) = P(A)$ then independent

else dependent

Dependent events

$P(A \cap B) = P(A)\,P(B|A)$

BAYES RULE

$$P(A|B) = \frac{P(B|A_i)\,P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

A will occur given that B already

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

↳ $P(A|B) = \frac{P(A \cap B)}{P(B)}$ $\quad P(B|A) = \frac{P(A \cap B)}{P(A)}$

↳ $P(A|B)\,P(B) = P(B|A)\,P(A)$

CONDITIONAL PROBABILITY

8)

	non smokers	moderate smokers	heavy smokers	Total
H	21	36	30	87
NH	48	26	19	93
Total	69	62	49	180

a) $P(H/H_3) = \frac{P(H \cap H_3)}{P(H_3)} = \frac{30}{49}$

b) $P(N_3/NH) = \frac{P(N_3/NH)}{P(NH)} = \frac{48}{93}$

8)

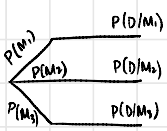
	P of choosing M		P of D of M		P of it being M and D
$P(M_1)$	0.30	x	$P(D/M)$ 0.02	=	0.006
$P(M_2)$	0.45	x	$P(D/M_2)$ 0.03	=	0.0135
$P(M_3)$	0.25	x	$P(D/M_3)$ 0.02	=	0.005
$P(D)$	?				0.0245

$P(D) = P(D \cap M_1) + P(D \cap M_2) + P(D \cap M_3)$

$= P(M_1)P(D/M_1) + P(M_2)(D/M_2) + P(M_3)(D/M_3)$

$= 0.0245$

$P(M_1/D) = \frac{P(M_1 \cap D)}{P(D)} = \frac{0.006}{0.024}$



1)

Station	A	B	C	Total
Electric	2	1	1	
Computers	4	3	2	
Equipment	5	4	2	
Human Error	7	7	5	
Total				

$$P(C/HE) = \frac{P(C \cap HE)}{P(HE)} = \frac{5}{19}$$

2)

	C	D	
Male	10	2	
Female	5	3	
			20

P(diploma holder is a male)

$$P(M/D) = \frac{P(M \cap D)}{P(D)} = \frac{2}{5}$$

3)

	1	2	3	Total
G	1	3	4	8
Y	2	2	3	2
R	4	5	1	10
Total	7	10	8	25

a) $P(Y) = P(Y \cap B_1) + (Y \cap B_2) + P(Y \cap B_3)$

$$\frac{1}{3} \times \frac{2}{7} + \frac{1}{3} \times \frac{2}{10} + \frac{1}{3} \times \frac{3}{8} = \frac{241}{840}$$

b) $P(B_2/Y) = \frac{P(B_2 \cap Y)}{P(Y)}$

$$= \frac{\frac{1}{3} \times \frac{2}{10}}{241/840}$$

4) M_1 M_2 M_3 → chances M hits

$\frac{1}{6}$ $\frac{1}{4}$ $\frac{1}{3}$

P(target was hit by M_1) =

$$P(M_1/H) = \frac{P(M_1 \cap H)}{P(H)} = \frac{\frac{1}{3} \times \frac{1}{6}}{\frac{1}{3}(\frac{1}{6}) + \frac{1}{3}(\frac{1}{4}) + \frac{1}{3}(\frac{1}{3})} = \frac{2}{9}$$?

5) H_1 H_2 H_3 → chances

$\frac{1}{6}$ $\frac{1}{4}$ $\frac{1}{3}$

$P(M_1) = \frac{1}{6}(1 - \frac{2}{4})(1 - \frac{1}{3})$

$P(M_2) = (1 - \frac{1}{6})\frac{2}{4}(1 - \frac{1}{3})$

$P(M_3) = (1 - \frac{1}{6})(1 - \frac{2}{4})\frac{1}{3}$

P(target was hit by M_1) =

P( ) =

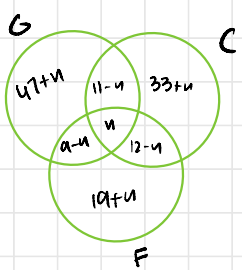
a) How many diff ways can student check of
1 ans to each question. 5 questions 4 option
4^5 .

b) get all wrong
3^5 → 5 questions
→ 3 wrong

2.37) 4B, 5G
if B, G must alternate
$5 \times 4 \times 4 \times 3 \times 3 \times 2 \times 2 \times 1 \times 1 = 2880$
G B G B G B G B G

e) How many ways no 2 students have same bday. 60 students
$365 P_{60}$ ← days, → students

θ) In a survey, B=136, G=67, C=56, F=40, G∩C=11, C∩F=12, G∩F=9, each boy played atleast 1
How many played all games



0, 1 2 3 4 5 6 NO REPETITION, in 3 DIGITS

a) 

6 × 6 × 5 = 180

↳ for 0

b) How many of these are odd numbers

5 × 5 × 3 = 75

c) How many > 330

3 × 6 × 5 + 1 × 3 × 5 = 105

4007 (330)

a) How many no each of 5 digits constructed

0 4 5 6 7

4 and 5 next to each other

3 × 3! × 2! = 36

b) not to be together 4 and 5

5 digits. 4 × 4! = 96

96 - 36 = 60

APPLICATIONS

2.11 A local gas station collected data from the day's receipts, recording the gallons of gasoline each customer purchased. The following table lists the frequency distribution of the gallons of gas purchased by all customers on this one day at this gas station.

Gallons of Gas	Number of Customers	mid point $\frac{n}{2}$	R.f	P.D
0 to less than 4	31		0.084	8.4
4 to less than 8	78	6	0.211	21.1
8 to less than 12	49	10	0.133	13.3
12 to less than 16	81	14	0.220	22
16 to less than 20	117	18	0.317	31.7
20 to less than 24	13	22	0.035	3.5
	369			

a. How many customers were served on this day at this gas station? 369

b. Find the class midpoints. Do all of the classes have the same width? If so, what is this width? If not, what are the different class widths? all have same widths $n=4$

c. Prepare the relative frequency and percentage distribution columns.

d. What percentage of the customers purchased 12 gallons or more? 57.2

e. Explain why you cannot determine exactly how many customers purchased 10 gallons or less. not specified

f. Prepare the cumulative frequency, cumulative relative frequency, and cumulative percentage distributions using the given table.

Make a dotplot for these data.

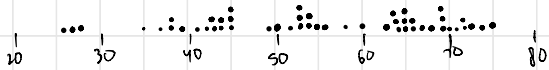
2.35 The following data give the times (in minutes) taken by 50 students to complete a statistics examination that was given a maximum time of 75 minutes to finish.

41	28	65	69	51	69	70	53	65	68
37	44	42	38	74	53	66	65	52	64
26	45	66	35	43	44	39	55	64	54
38	52	58	72	67	65	43	65	68	27
64	49	71	75	45	69	56	73	53	72

20 80

Create a dotplot for these data.

2.36 The following data give the one-way commuting times (in



EMPERICAL?
SETS DIGRAM?
PERMUTATIONS FOR STRING?

each stem in increasing order.

2.27 The following data give the times (in minutes) taken by 50 students to complete a statistics examination that was given a maximum time of 75 minutes to finish.

41	28	65	69	51	69	70	53	65	68
37	44	42	38	74	53	66	65	52	64
26	45	66	35	43	44	39	55	64	54
38	52	58	72	67	65	43	65	68	27
64	49	71	75	45	69	56	73	53	72

a. Prepare a stem-and-leaf display for these data. Arrange the leaves for each stem in increasing order.

b. Prepare a split stem-and-leaf display for the data. Split each stem into two parts. The first part should contains the leaves 0,

2 8 6 7
3 7 8 8 5 9
4 1 5 4 5 9 2 3 5 4
5 3 0 2 8 3 3 6 5 2 3 4
6 0 9 3 8 4 6 7 5 9 6 5 5 4 8 4
7 0 1 2 5 4 3 2

RANDOM VARIABLE (RV)

A r.v is a function that associates a real no. with each element in the sample space

Discrete random variable: possible outcomes countable

Continous random variable: values on a continous scale

e) Let X_0 = no. of heads in toss of 3 coins

x	$2^3=8$	P(x)	F(x)
0 ← TTT		1/8	1/8
1 ← HTT, THT, TTH		3/8	4/8
2 ← HHT, THH, HTH		3/8	7/8
3 ← HHH		1/8	8/8

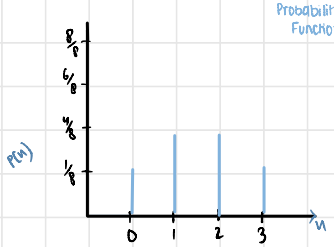
Cumulative Distribution function (CDF)

$F(X) = P(X \leq n)$

$$F(X) = \begin{cases} 0 & n < 0 \quad P_{<0}\ P(n<0) \\ 1/8 & 0 \leq n < 1/8 \quad P_{\leq 0}\ P(n=0) \\ 4/8 & 1/8 \leq n < 4/8 \quad P_{\leq 1}\ P(n=0)+P(n=1) \\ 7/8 & 4/8 \leq n < 7/8 \quad P_{\leq 2}\ P(n=0)+P(n=1)+P(n=2) \\ 1 & 7/8 \leq n < 1 \quad P_{\leq 3}\ P(n=0)+P(n=1)+P(n=2)+P(n=3) \end{cases}$$

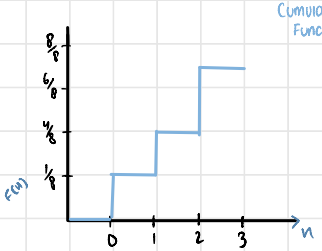
Line Graph → PMF → uses P(n)

Probability Mass Function



Line Graph → CDF → uses F(n)

Cumulative Distribution Function



e) Let X_0 = no. of heads in toss of 4 coins

x	$2^4=16$	P(x)	F(x)
0 ← TTTT		1/16	1/16
1 ← HTTT, THTT, TTHT, TTTH		4/16	5/16
2 ← HHTT, TTHH, HTHT, THTH		4/16	9/16
3 ← THHH, HTHH, HHTH, HHHT		4/16	13/16
4 ← HHHH		1/16	14/16

Cumulative Distribution Function (CDF)

↳ for continuous RV

↳ $F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)\,dt$

↳ $F(x) = P(X \le x) = \sum P(x)$ (CDF = $P(X \le x)$, PMF = $\sum P(x)$)

↳ $P(a < x < b) = F(b) - F(a)$

$f(x) = \frac{dF(x)}{dx}$ $\qquad F(n) = \sum_{x=1}^{n} (x+1)^n$

Probability mass Function (PMF)

1. $f(x) \ge 0$, for all $x \in R$

↳ $\sum f(x)\,dx = 1$ → verify PMF

↳ $P(X=x) = F(x) - F(x-1)$

↳ $P(a < x < b) = F(b) - F(a)$

$P(n) = (n+1)^n$

Example 1

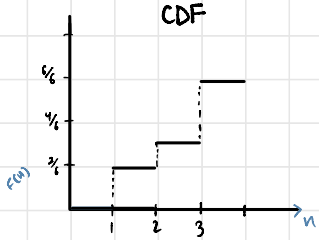
- Consider the following pmf: f(x) = (x/6), x = 1, 2, 3, zero elsewhere
(i) Find distribution function and its graph.
(ii) Calculate P(1.5 < x ≤ 4.5)

i) $f(x) = x/6$ → Distribution function

x	f(x)	F(x)
1	1/6	1/6
2	2/6	3/6
3	3/6	6/6

$F(x) = \begin{cases} 0 & x < 1 \\ 1/6 & 1 \le x < 2 \\ 3/6 & 2 \le x < 3 \\ 6/6 & x \ge 3 \end{cases}$

graph



ii) $P(1.5 < x \le 4.5)$

$= F(4.5) - F(1.5)$

$= x \ge 3 - (1 \le x < 2)$

$= 1 - 1/6$

$= 5/6$

Example 3

(3) Consider the following functions
(i) f(x) = (x + 2) / 5 for x = 1, 2, 3, 4, 5.
(ii) f(x) = (4Cx) / (2^5), for x = 0, 1, 2, 3, and 4.
and check whether the functions can serve as a pmf ?

i) $f(x) = \frac{(x+2)}{5}$

Since $0 \le x \le 1$ → what

$\sum f(x) \ne 1$

So not PMF

x	f(x)
1	3/5
2	4/5
3	1
4	6/5
5	7/5

$\sum f(x) = 5$

ii) $f(x) = \frac{{}^4C_x}{2^5}$

not PMF as $\sum f(x) \ne 1$

x	f(x)
1	1/32
2	1/8
3	3/16
4	1/8
5	1/32

$\sum f(x) = 1/2$

Example 4

(4) A coin is biased so that a head occurs 3 times of tail. If the coin is tossed 3 times, find the probability distribution for the number of heads and also find P (1 ≤ x ≤ 3).

H 3/4 T 1/4

0 { TTT $\frac{1}{4}\times\frac{1}{4}\times\frac{1}{4}\times 1$

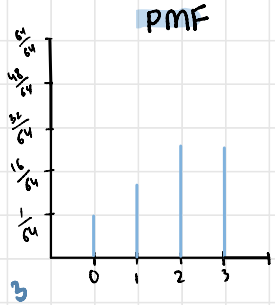
1 { HTT, THT, TTH $\frac{3}{4}\times\frac{1}{4}\times\frac{1}{4}\times 3$

2 { HHT, HTH, THH $\frac{3}{4}\times\frac{3}{4}\times\frac{1}{4}\times 3$

3 { HHH $\frac{3}{4}\times\frac{3}{4}\times\frac{3}{4}\times 1$

X	P(x)
0	1/64
1	9/64
2	27/64
3	27/64

PMF



$P(1 \le x \le 3) = \overset{1}{9/64} + \overset{2}{27/64} + \overset{3}{27/64}$

$= 63/64$

Example 7

- The distribution function for a discrete random variable x is given as:

$F(x) = 1 - (1/2)^{(x+1)}$, for x = 0, 1, 2, ...

Find:

(a) P (X = 3)

(b) P (7 ≤ x < 10)

(c) Probability Mass Function.

$f(x) = 1 - (1/2)^{(x+1)}$

a) $P(x=3) = F(3) - F(2)$

$\therefore P(X=x) = F(x) - F(x-1) = \left[1 - \left(\frac{1}{2}\right)^{3+1}\right] - \left[1 - \left(\frac{1}{2}\right)^{2+1}\right]$

$= 1/16$

b) $P(7 \le x < 10) = F(9) - F(6)$ what $F(10) - F(6)$

$= 7/1024$ $= 15/2048$

c) PMF

$= F(x) - F(x-1)$

$= 1 - \frac{1}{2}^{x+1} - 1 + \frac{1}{2}^{x-1+1}$

$= \left(-\frac{1}{2}\right)^{x+1} + \left(\frac{1}{2}\right)^{x}$

$= \frac{1}{2}^{x+1}$ → WHAT

$\left(-\frac{1}{2}\right)^{x} \cdot \left(-\frac{1}{2}\right)^{1} + \left(\frac{1}{2}\right)^{x}$

$-\frac{1}{2}\left(-\frac{1}{2}\right)^{x} + \left(\frac{1}{2}\right)^{x}$

Example 8

3.5 Determine the value c so that each of the following functions can serve as a probability distribution of the discrete random variable X:

(a) $f(x) = c(x^2 + 4)$, for $x = 0, 1, 2, 3$;

a) $f(x) = c(x^2 + 4)$

$\sum_{x=0}^{3} f(x) = 1$

$c(4) + c(5) + c(8) + c(13) = 1$ (0, 1, 2, 3)

$30c = 1$

$c = \frac{1}{30}$

Q) find $P(n=2) = f(n=2)$ using CDF

$P(2) = F(2) - F(1)$

$= P(n \le 2) - P(n \le 1)$

$= 11/16 - 5/16$

$= 3/8$

Q1) A fair coin is tossed until H appears for the first time. Find

a) PMF :- $(1/2)^n \quad \therefore n = 1, 2, 3, \ldots$

b) CDF = $\sum_{n=1}^{n} (1/2)^n$

c) $F(4) = \sum_{n=1}^{n=4} (1/2)^n$

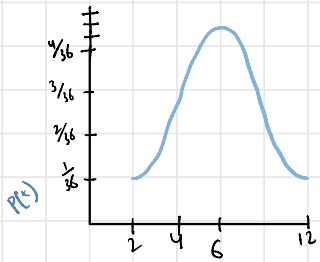
Q) if two dice are rolled once, find the PMF of the sum of points on dice, CDF and their graph

let $t = X + Y$

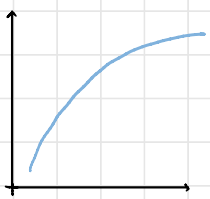
z	2	3	4	5	6	7	8	9	...	12
f(t): P(t):	1/36	2/36	3/36	4/36	5/36	6/36	7/36	8/36	.....	
F(t):	1/36	3/36	6/36	....						36/36

$$F(t) = \begin{cases} 0 & z < 2 \\ 1/36 & 2 \le z \le 3 \\ 3/36 & 3 \le z \le 4 \\ \vdots & \\ 36/36 & z \ge 12 \end{cases}$$

PMF



CDF



* Expected Value

let suppose a coin tossed 2 times

X = Head calculate

E(X) = Mass of a row x

X	P(x)	x.P(x)	x^2.P(x)
0	1/4	0	0
1	2/4	2/4	2/4
2	1/4	2/4	4/4
Sum	1	E(x) = 1	$E(x^2) = 3/2$

$E(x) = \sum (x) P(x)$ → Expected value

$\sigma^2 X = V(x) = E(x^2) - [E(x)]^2$ → Variance

Joint Probability Distribution

1. $f(x,y) \geq 0$
2. $\sum_x \sum_y f(x,y) = 1$
3. $P(X=x, Y=y) = f(x,y)$
   ↳ $P[(X,Y) \in A] = \sum_A \sum f(x,y)$

Example 9

- Two ballpoint pens are selected at random from a box that contains 3 blue pens, 2 red pens, and 3 green pens. If X is the number of blue pens selected and Y is the number of red pens selected, find
- (a) the joint probability function $f(x,y)$
- (b) $P[(X,Y) \in A]$, where A is the region $\{(x,y) \mid x + y \leq 1\}$ → WHAT

$B=3, R=2, G=3 \Rightarrow 8$ → select 2 w/o replacement

JOINT PROB MASS FUNCTION

Marginal dist $g(x)$

Blue ↓ x \ Red → y	0	1	2	$g(x)$	$x\,g(x)$	$x^2 g(x)$
0	$\frac{{}^3C_0\,{}^2C_2}{{}^8C_2} = 3/28$	$\frac{{}^3C_0\,{}^2C_1\,{}^3C_1}{{}^8C_2} = 6/28$	$\frac{{}^3C_0\,{}^2C_2\,{}^3C_0}{{}^8C_2} = 1/28$	$10/28$	0	0
1	$\frac{{}^3C_1\,{}^2C_1\,{}^3C_0}{{}^8C_2} = 9/28$	$\frac{{}^3C_1\,{}^2C_1\,{}^3C_1}{{}^8C_2} = 6/28$	—	$15/28$	$15/28$	$15/28$
2	$\frac{{}^3C_2\,{}^2C_0\,{}^3C_0}{{}^8C_2} = 3/28$	—	—	$3/28$	$6/28$	$12/28$
$h(y)$	$15/28$	$12/28$	$1/28$	1	$E(x) = 21/28$	$E(x^2) = 27/28$
$y\,h(y)$	0	$12/28$	$2/28$	$E(y) = 14/28$		
$y^2 h(y)$	0	$12/28$	$4/28$	$E(y^2) = 16/28$		

Marginal dist of y

$E(x^4)$

$E(\sqrt{x})$

$E[3x] = 3E(x)$

$\sigma^2 x = V(x) = E(x^2) - [E(x)]^2$ → variance

$= \frac{27}{28} - \left(\frac{21}{28}\right)^2$

a) $P[X \leq 2, Y = 1] = f(0,1) + f(1,1) + f(2,1)$
$= 6/28 + 6/28 + 3/28$
$= 12/28$

b) $P[X \geq 2, Y \leq 1] = f(2,1) + f(2,0)$

c) $P[X > Y] = f(1,0) + f(2,0)$

d) $P[X + Y = 4] = 0$

e) Conditional Probability
$P[X=0 / Y=1] = \frac{f(x=0, y=1)}{h(y=1)}$
$= \frac{6/28}{12/28} = \frac{6}{12}$

f) $P[Y=1 / X=0] = \frac{f(y=1, x=0)}{g(x=0)} = \frac{6/28}{10/28}$

g) $E[XY] = \sum\sum xy f(x,y) = 6/28$

0	0	×	3/28
0	1	×	6/28
0	2	×	1/28
1	0	×	9/28
1	1	×	6/28
1	2		
2	0	×	3/28
2	1		
2	2		
			6/28

h) Covariance $(x,y) = E(XY) - E(x)\,E(y)$
$= 6/28 - \left(\frac{21}{28}\right)\left(\frac{14}{28}\right)$
$= -9/56$

$-1 \leq$ Correlation $(x,y) = \frac{\text{covariance}(x,y)}{\sqrt{V(x)}\sqrt{V(y)}} \leq +1$

$= \frac{-9/56}{\sqrt{\frac{45}{112}}\sqrt{\frac{9}{28}}}$

$= -0.447$ moderate -ve relationship

MATHAMETICAL EXPECTATION

$V(x) = E(x^2) - [E(x)]^2$ (σ^2 → variance)

$\mu = E(x) = \sum x f(x)$ (mean)

$V(y) = E(y^2) - [E(y)]^2$

$E(xy) = \sum\sum xy\, f(x,y)$

Covariance $(x,y) = E(xy) - E(x)\,E(y)$

Correlation $(x,y) = \frac{\text{Covariance}(x,y)}{\sqrt{V(x)}\sqrt{V(y)}}$

± 1 → strong
± 0.5 → moderate
0 → negligible

JPMF

$f(x,y) = \frac{x+y}{30}$ for $x = 0,1,2,3$ $y = 0,1,2$

Calculated

a) $P[X \le 2, Y = 1] = F(0,1) + F(1,1) + F(2,1)$
$= \frac{1}{30} + \frac{2}{30} + \frac{3}{30} = \frac{6}{30}$

b) $P[X > 2, Y \le 1] = F(3,1) + F(3,0)$
$= \frac{3}{30} + \frac{4}{30} = \frac{7}{30}$

c) $P[X > Y] = F(3,2) + F(3,1) + F(3,0) + F(2,1) + F(2,0) + F(1,0)$
$\frac{5}{30} + \frac{4}{30} + \frac{3}{30} + \frac{3}{30} + \frac{2}{30} + \frac{1}{30} = \frac{18}{30}$

d) $P[X + Y = 4] = F(3,1) + F(2,2)$
$= \frac{4}{30} + \frac{4}{30} = \frac{8}{30}$

e) Calculate

marginal Dist X

x \ y	0	1	2	g(x)	xg(x)	$x^2 g(x)$
0	0	1/30	2/30	3/30	0	0
1	1/30	2/30	3/30	6/30	6/30	6/30
2	2/30	3/30	4/30	9/30	18/30	36/30
3	3/30	4/30	5/30	12/30	36/30	108/30
h(y)	6/10	10/30	14/30	1	E(x) = 2	E(x²) = 5
yh(y)	0	10/30	28/30	E(y) = 38/30		
$y^2 h(y)$	0	10/30	56/30	E(y²) = 66/30		

$(X-\mu)^2 f(x)$
$(X-2)^2 f(x)$

$E(X) = \frac{60}{30} = 2$
$E(X^2) = \frac{150}{30} = 5$

$V(X) = E(X^2) - [E(X)]^2 = 5 - (2)^2 = 1$
$V(Y) = E(Y^2) - [E(Y)]^2 = \frac{66}{30} - (\frac{38}{30})^2 = 0.59$

$E(XY) = \sum\sum xy\, f(x,y) = 2.4$

$E(XY) = \sum\sum xy\, f(x,y) = 2.4$

00 = 0	0
01 = 0	1/30 × 0
02 = 0	2/30 × 0
10 = 0	1/30 × 0
11 = 1	2/30 × 1
12 = 2	3/30 × 2
20 = 0	2/30 × 0
21 = 2	3/30 × 2
22 = 4	4/30 × 4
30 = 0	3/30 × 0
31 = 3	4/30 × 3
32 = 6	5/30 × 6

Covariance $(X,Y) = E(XY) - E(X)\,E(Y) = 2.4 - 2(\frac{38}{30}) \rightarrow \frac{-2}{15} \rightarrow -0.133$

Correlation $(X,Y) = \frac{\text{Covariance}(X,Y)}{\sqrt{V(X)}\ \sqrt{V(Y)}} = -0.17277$

↳ weak relation

↳ inverse relation as -ve sign

$= \frac{-0.133}{\sqrt{1}\ \sqrt{0.59}}$

Conditional Distribution of X

$$f(x|y) = \frac{f(x,y)}{h(y)} \quad \therefore h(y) > 0$$

Conditional Distribution of Y

$$f(y|x) = \frac{f(x,y)}{g(x)} \quad \therefore g(x) > 0$$

Slide 20

3.51 Three cards are drawn without replacement from the 12 face cards (jacks, queens, and kings) of an ordinary deck of 52 playing cards. Let X be the number of kings selected and Y the number of jacks. Find

(a) the joint probability distribution of X and Y;

(b) $P[(X, Y) \in A]$, where A is the region given by $\{(x, y) \mid x + y \geq 2\}$.

X = no of kings Y = no of Jacks total facecards = 12 total cards = 52

a)

y \ x	0	1	2	3	Marginal dist g(x) Total g(x)
0	${}^4C_0\,{}^8C_3 / {}^{12}C_3 = \frac{56}{220} = \frac{14}{55}$	${}^4C_1\,{}^8C_2 / {}^{12}C_3 = \frac{112}{220} = \frac{28}{55}$	${}^4C_2\,{}^8C_1 / {}^{12}C_3 = \frac{48}{220} = \frac{12}{55}$	${}^4C_3\,{}^8C_0 / {}^{12}C_3 = \frac{4}{220} = \frac{1}{55}$	$14/55$
1	${}^4C_1\,{}^8C_2 / {}^{12}C_3 = \frac{112}{220} = \frac{28}{55}$	${}^4C_2\,{}^8C_1 / {}^{12}C_3 = \frac{48}{220} = \frac{12}{55}$	${}^4C_3\,{}^8C_0 / {}^{12}C_3 = \frac{4}{220} = \frac{1}{55}$	—	$28/55$
2	${}^4C_2\,{}^8C_1 / {}^{12}C_3 = \frac{48}{220} = \frac{12}{55}$	${}^4C_3\,{}^8C_0 / {}^{12}C_3 = \frac{4}{220} = \frac{1}{55}$	—	—	$12/55$
3	${}^4C_3\,{}^8C_0 / {}^{12}C_3 = \frac{4}{220} = \frac{1}{55}$	—	—	—	$1/55$
Total h(y)	$14/55$	$28/55$	$12/55$	$1/55$	1

marginal dist of y

b) $P(X, Y \in A) = P(x + y \geq 2)$

$P = 1 - P(x+y \leq 1)$ (Total Prob)

$= 1 - [P(0,0) + P(0,1) + P(1,0)]$

$= 1 - \left[\frac{1}{55} + \frac{6}{55} + \frac{6}{55}\right]$

$P(x+y \geq 2) = \frac{42}{55}$

Statistical independence

$f(x,y) = g(x)\,h(y)$

$f(x,y) = f(x|y)\,h(y)$

$$g(x) = f(x|y) = \int_{-\infty}^{\infty} h(y)\,dy = 1$$

.1: Joint Probability Distribution for Example 3.14

$f(x,y)$		x = 0	1	2	Row Totals
y	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
	1	$\frac{3}{14}$	$\frac{3}{14}$	0	$\frac{3}{7}$
	2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
Column Totals		$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

show not statistically independent

taking $f(0,1) = 3/14$

$g(0) = 5/14$

$h(1) = 15/28$

$f(0,1) = g(0)\,h(1)$

$3/14 = 5/14 \times 15/28$

$3/14 \neq \frac{75}{392}$

Slide 20

2.52 A coin is tossed twice. Let Z denote the number of heads on the first toss and W the total number of heads on the 2 tosses. If the coin is unbalanced and a head has a 40% chance of occurring, find

(a) the joint probability distribution of W and Z.

(b) the marginal distribution of W.

(c) the marginal distribution of Z.

(d) the probability that at least 1 head occurs.

Marginal Distribution → Discrete

$$g(x) = \sum_y f(x,y)\,dy$$

$$h(y) = \sum_x f(x,y)\,dx$$

H = 0.6
T = 0.4

a)

y \ x	z0 (no. of heads on first toss)	z1	Marginal dist g(x)
w0	(0.6)(0.6) = 0.36 TT	(0.4)(0.4) = 0 TT	0.36
w1	(0.4)(0.6) = 0.24 HT	(0.4)(0.6) = 0.24 TH	0.48
w2	(0.6)(0.6) = 0 HH	(0.4)(0.4) = 0.16 HH	0.16
Total h(w)	0.6	0.4	1

(Total no. of heads on 2nd tosses)

Marginal dist of y

b) Marginal distribution of w

$w=0 \quad = 0.36$

$w=1 \quad = 0.48$

$w=2 \quad = 0.16$

c) Marginal Distribution of z

$z=0 \quad = 0.6$

$z=1 \quad = 0.4$

d) P(of atleast 1 H) = P(w=1 or w=2)

= P(w=1) + P(w=2)

= 0.48 + 0.16

= 0.64

Slide 22

Let the number of phone calls received by a switchboard during a 5-minute interval be a random variable X with probability function

$$f(x) = \frac{e^{-3}3^x}{x!}, \quad \text{for } x = 0, 1, 2, \ldots$$

(a) Determine the probability that X equals 0, 1, 2, 3, 4, 5, and 6.

(b) Graph the probability mass function for these values of x.

(c) Determine the cumulative distribution function for these values of X.

x	f(x)	F(x) → CDF
0	0.1353	0.1353
1	0.2707	0.4060
2	0.2707	0.6767
3	0.1804	0.8571
4	0.0902	0.9473
5	0.0361	0.9834
6	0.1202	0.9954

Probability Density Function (PDF)

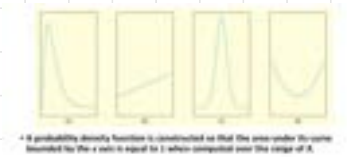
1. $f(u) \geq 0$, for all $u \in R$
2. $\int_{-\infty}^{\infty} f(u)\,du = 1$ → verify PDF
3. $P(a < X < b) = \int_a^b f(u)\,du$

Continous Probability Distribution

↳ F(u) → probability density function (PDF)

↳ Areas used to represent probabilities

↳ can not be given in tabular form



Example 1

Suppose that the error in the reaction temperature, in °C, for a controlled laboratory experiment is a continuous random variable X having the probability density function

$$f(x) = \begin{cases} \frac{x^2}{3}, & -1 < x < 2, \\ 0, & \text{elsewhere.} \end{cases}$$

(a) Verify that $f(x)$ is a density function.
(b) Find $P(0 < X \leq 1)$.

a) $\int_{-1}^{2} \frac{u^2}{3}\,du = \left| \frac{u^3}{9} \right|_{-1}^{2} = \frac{8}{9} + \frac{1}{9} = 1$ → as equal to 1 HENCE VERIFIED

b) $\int_{0}^{1} \frac{u^2}{3}\,du = \frac{1}{9}$

Example 3

- The Department of Energy (DOE) puts projects out on bid and generally estimates what a reasonable bid should be. Call the estimate b. The DOE has determined that the density function of the winning (low) bid is:

$$f(y) = \begin{cases} \frac{5}{8b}, & \frac{2}{5}b \leq y \leq 2b, \\ 0, & \text{elsewhere.} \end{cases}$$

- Find $F(y)$ and use it to determine the probability that the winning bid is less than the DOE's preliminary estimate b.

$F(y) = \int_{\frac{2}{5}b}^{y} \frac{5}{8b}\,dy = \left| \frac{5y}{8b} \right|_{\frac{2}{5}b}^{y} = \frac{5y}{8b} - \frac{1}{4} =$

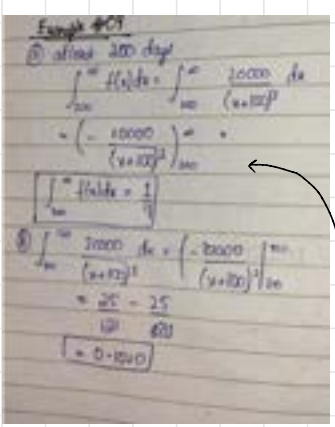
$P(y \leq b) = F(b) = \frac{5b}{8b} - \frac{1}{4} = \frac{3}{8}$

Example 4

The shelf life, in days, for bottles of a certain prescribed medicine is a random variable having the density function

$$f(x) = \begin{cases} \frac{20{,}000}{(x+100)^3}, & x > 0, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the probability that a bottle of this medicine will have a shelf life of
(a) at least 200 days;
(b) anywhere from 80 to 120 days.



$20000 \int_{200}^{\infty} \frac{1}{(u+100)^3}\,du$

$u = u + 100$

$\frac{du}{du} = 1 \Rightarrow du = \frac{du}{1}$

$20000 \int_{100}^{\infty} u^{-3}\,du$

$20000 \left| \frac{u^{-2}}{-2} \right|$

$20000 \left| \frac{(u+100)^{-2}}{-2} \right|$

$\left| \frac{-10000}{(u+100)^2} \right|_{200}^{\infty}$

Example 5

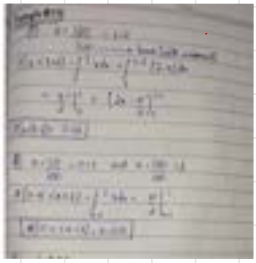
The total number of hours, measured in units of 100 hours, that a family runs a vacuum cleaner over a period of one year is a continuous random variable X that has the density function

$$f(x) = \begin{cases} x, & 0 < x < 1, \\ 2 - x, & 1 \le x < 2, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the probability that over a period of one year, a family runs their vacuum cleaner

(a) less than 120 hours;

(b) between 50 and 100 hours.



Q2) $f(u) = \begin{cases} 3u^{-4} & ; \ u > 1 \\ 0 & ; \ \text{elsewhere} \end{cases}$

a) Verify that $f(u)$ is a PDF

$$\int_1^\infty 3u^{-4}$$

$$\left[-\frac{3}{3} u^{-3} \right]_1^\infty$$

$$= \left[-u^{-3} \right]_1^\infty$$

$$\left(-\frac{1}{\infty^3} \right) - \left(-\frac{1}{1^3} \right)$$

$$= 0 - (-1) = 1$$

b) $F(u) = \text{CDF}$

use u as upper limit

$$\int_1^u \frac{3}{u^4} \, du$$

$$\left[-u^{-3} \right]^u$$

$$= -u^{-3} + 1$$

$$= 1 - u^{-3}$$

c) $P(u > 4)$

$$\int_4^\infty 3u^{-4} \, du$$

$$\left| -u^{-3} \right|_4^\infty$$

$$-1/64 - \frac{1}{\infty^3}$$

$$-\frac{1}{64}$$

Example 7

Measurements of scientific systems are always subject to variation, some more than others. There are many structures for measurement error, and statisticians spend a great deal of time modeling these errors. Suppose the measurement error X of a certain physical quantity is decided by the density function

$$f(x) = \begin{cases} k(3 - x^2), & -1 \le x \le 1, \\ 0, & \text{elsewhere.} \end{cases}$$

(a) Determine k that renders $f(x)$ a valid density function.

(b) Find the probability that a random error in measurement is less than $1/2$.

(c) For this particular measurement, it is undesirable if the magnitude of the error (i.e., $|x|$) exceeds 0.8. What is the probability that this occurs?

a) $\int_{-1}^{1} k(3 - u^2) \, du = 1$ → equate to 1 as valid PDF equates to 1

$$k(3u - u^3/3) = 1$$

$$k\left[3 - \frac{1}{3} - \left(3(-1) - \frac{(-1)^3}{3} \right) \right] = 1$$

$$k = \frac{3}{16}$$

b) $P(u < 1/2) = \int_{-1}^{1/2} \frac{3}{16} (3 - u^2) \, du \Rightarrow \frac{99}{128}$

c) $P[|u| < 0.8] = P[u < -0.8] + P[u \ge 0.8]$

↓ WHAT

$$= F(-0.8) + [1 - F(0.8)]$$

$$= 0.164$$

$$-0.8 < u < 0.8$$

Joint Density Function (JDF)

↳ for continuous variable

1. $f(x,y) \geq 0$
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$
3. $P[(X,Y) \in A] = \int_A \int f(x,y)\,dx\,dy$

Marginal Distribution → for continuous variable

$g(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$ → w.r.t. y

$h(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$ → w.r.t. x

Q3) $f(x,y) = \begin{cases} \frac{2}{5}(2x+3y) & 0<x<1,\ 0<y<1 \\ \text{elsewhere} & \end{cases}$

a) Verify that $f(x,y)$ is a valid JPDF

$$\int_0^1\int_0^1 \tfrac{2}{5}(2x+3y)\,dx\,dy$$

$$\tfrac{2}{5}\int_0^1 (x^2+3xy)\Big|_0^1\,dy$$

$$\tfrac{2}{5}\int_0^1 (1+3y)\,dy$$

$$\tfrac{2}{5}\left[y+\tfrac{3y^2}{2}\right]_0^1$$

$$\tfrac{2}{5}\left[1+\tfrac{3}{2}\right]$$

$$\tfrac{2}{5}\left[\tfrac{5}{2}\right]$$

$$1$$

b) $P[0<x<\frac{1}{2},\ \frac{1}{4}<y<\frac{1}{2}]$

$$= \tfrac{2}{5}\int_{1/4}^{1/2}\int_0^{1/2} (2x+3y)\,dx\,dy$$

$$= \tfrac{2}{5}\int_{1/4}^{1/2} (x^2+3xy)\Big|_0^{1/2}\,dy$$

$$= \tfrac{2}{5}\int_{1/4}^{1/2} \left(\tfrac{1}{4}+\tfrac{3y}{2}\right)dy$$

$$= \tfrac{2}{5}\left(\tfrac{y}{4}+\tfrac{3y^2}{4}\right)\Big|_{1/4}^{1/2}$$

$$= \tfrac{2}{5}\left[\tfrac{1}{8}+\tfrac{3}{16}-\left(\tfrac{1}{16}+\tfrac{3}{64}\right)\right]$$

$$= 0.08$$

c) find $g(x)$ & $h(y)$

Marginal $g(x) = \frac{2}{5}\int_0^1 (2x+3y)\,dy$

$$= \tfrac{2}{5}\left(2xy+\tfrac{3y^2}{2}\right)\Big|_0^1$$

$$= \tfrac{2}{5}\left(2x+\tfrac{3}{2}\right)$$

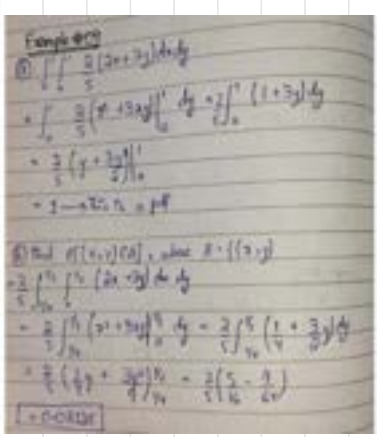
Marginal $h(y) = \frac{2}{5}\int_0^1 (2x+3y)\,dx$

Example 9

- A privately owned business operates both a drive-in facility and a walk-in facility. On a randomly selected day, let X and Y, respectively, be the proportions of the time that the drive-in and the walk-in facilities are in use, and suppose that the joint density function of these random variables is

$$f(x,y) = \begin{cases} \frac{2}{5}(2x+3y), & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0, & \text{elsewhere} \end{cases}$$

- (a) Verify for PDF?
  (b) Find $P[(X,Y) \in A]$, where $A = \{(x,y) \mid 0 < x < 1/2,\ 1/4 < y < 1/2\}$



mean of a random variable

when it is clear to which random variable we refer.

Let X be a random variable with probability distribution $f(x)$. The expected value of the random variable $g(X)$ is

$$\mu_{g(X)} = E[g(X)] = \sum_x g(x)f(x)$$

if X is discrete, and

$$\mu_{g(X)} = E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx$$

if X is continuous.

Let X be a random variable with probability distribution $f(x)$ and mean μ. The variance of X is

$$\sigma^2 = E[(X-\mu)^2] = \sum_x (x-\mu)^2 f(x), \qquad \text{if } X \text{ is discrete, and}$$

$$\sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx, \qquad \text{if } X \text{ is continuous.}$$

The positive square root of the variance, σ, is called the **standard deviation** of X.

Let X be a random variable with probability distribution $f(x)$. The **mean**, or **expected value**, of X is

$$\mu = E(X) = \sum_x x f(x)$$

if X is discrete, and

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\,dx$$

if X is continuous.

The reader should note that the way to calculate the expected value, or mean

Example 4.4: Suppose that the number of cars X that pass through a car wash between 4:00 P.M. and 5:00 P.M. on any sunny Friday has the following probability distribution:

x	4	5	6	7	8	9
$P(X=x)$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{1}{6}$

Let $g(X) = 2X - 1$ represent the amount of money, in dollars, paid to the attendant by the manager. Find the attendant's expected earnings for this particular time period.

Solution: By Theorem 4.1, the attendant can expect to receive

$$E[g(X)] = E(2X-1) = \sum_{x=4}^{9} (2x-1)f(x)$$
$$= (7)\left(\frac{1}{12}\right) + (9)\left(\frac{1}{12}\right) + (11)\left(\frac{1}{4}\right) + (13)\left(\frac{1}{4}\right) + (15)\left(\frac{1}{6}\right) + (17)\left(\frac{1}{6}\right) = \$12.67.$$

Let X and Y be random variables with joint probability distribution $f(x,y)$. The mean, or expected value, of the random variable $g(X,Y)$ is

$$\mu_{g(X,Y)} = E[g(X,Y)] = \sum_x \sum_y g(x,y)f(x,y)$$

if X and Y are discrete, and

$$\mu_{g(X,Y)} = E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)\,dx\,dy$$

if X and Y are continuous.

Example 4.6: Let X and Y be the random variables with joint probability distribution indicated in Table 3.1 on page 90. Find the expected value of $g(X,Y) = XY$. The table is reprinted here for convenience.

$f(x,y)$		y = 0	1	2	Row Totals
x	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
	1	$\frac{3}{7}$	$\frac{3}{7}$	0	$\frac{6}{7}$
	2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
Column Totals		$\frac{5}{7}$	$\frac{15}{28}$	$\frac{3}{28}$	1

Solution: By Definition 4.2, we write

$$E(XY) = \sum_{x=0}^{2}\sum_{y=0}^{2} xyf(x,y)$$
$$= (0)(0)f(0,0) + (0)(1)f(0,1) + (0)(2)f(0,2) + (1)(0)f(1,0) + (1)(1)f(1,1) + (1)(2)f(1,2) + (2)(0)f(2,0) + (2)(1)f(2,1) + (2)(2)f(2,2)$$
$$= f(1,1) = \frac{3}{14}.$$

variance & covariance of Random variables

$\mu = E(x) = \Sigma x f(x)$ → mean

$\sigma^2 = (x - \mu)^2$ → variance

$\sigma^2 = E(x^2) - [E(x)]^2$

($[E(x)]^2 = \mu^2$)

Let X be a random variable with probability distribution $f(x)$ and mean μ. The variance of X is

$$\sigma^2 = E[(X-\mu)^2] = \sum_x (x-\mu)^2 f(x), \quad \text{if } X \text{ is discrete, and}$$

$$\sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx, \quad \text{if } X \text{ is continuous.}$$

The positive square root of the variance, σ, is called the **standard deviation** of X.

MATHAMETICAL EXPECTATION

σ^2 → variance

$V(x) = E(x^2) - [E(x)]^2$

mean → expected

$\mu = E(x) = \sum_x x f(x)$

$V(Y) = E(Y^2) - [E(Y)]^2$

$E(xy) = \Sigma\Sigma\, xy\, f(x,y)$

$\text{Covariance}(x,y) = E(xy) - E(x).E(y)$

$\text{Corelation}(x,y) = \dfrac{\text{Covariance}(x,y)}{\sqrt{V(x)}\ \sqrt{V(Y)}}$

± 1 → strong

± 0.5 → moderate

0 → negligible

...mple 4.14: The fraction X of male runners and the fraction Y of female runners who compete in marathon races are described by the joint density function

$$f(x,y) = \begin{cases} 8xy, & 0 \le y \le x \le 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the covariance of X and Y.

Solution: We first compute the marginal density functions. They are

$$g(x) = \begin{cases} 4x^3, & 0 \le x \le 1, \\ 0, & \text{elsewhere,} \end{cases}$$

and

$$h(y) = \begin{cases} 4y(1-y^2), & 0 \le y \le 1, \\ 0, & \text{elsewhere.} \end{cases}$$

From these marginal density functions, we compute

$$\mu_X = E(X) = \int_0^1 4x^4\, dx = \frac{4}{5} \text{ and } \mu_Y = \int_0^1 4y^2(1-y^2)\, dy = \frac{8}{15}.$$

From the joint density function given above, we have

$$E(XY) = \int_0^1 \int_y^1 8x^2y^2\, dx\, dy = \frac{4}{9}.$$

Then

$$\sigma_{XY} = E(XY) - \mu_X\mu_Y = \frac{4}{9} - \left(\frac{4}{5}\right)\left(\frac{8}{15}\right) = \frac{4}{225}.$$

DISCRETE PROBABILITY DISTRIBUTION

1. Binomial Distribution

- The number X of successes in n Bernoulli trials is called a **binomial random variable**.
- The probability distribution of this discrete random variable is called the **binomial distribution**.
- The probability of a success in a binomial experiment can be computed with this formula:

$$b(x; n, p) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, \ldots, n$$

2. The Bernoulli Process

↳ repeated trials { FAIL / SUCESS } TWO POSSIBLE OUTCOMES

↳ the P(sucess) remains constant

↳ repeated trials are independent

↳ $P + q = 1$

Mean = np (p: Probability sucess, n: no of trials)

$q = 1 - P$ (Probability fail)

Variance = $npq \rightarrow \sigma^2$

Standard Deviation = $\sqrt{\text{variance}}$

PMF = $P(x) = {}^nC_x p^x q^{n-x}$

Q) A coin is tossed 4 times. find mean, variance, sd of heads that will be obtained

$n=4$, $P=\frac{1}{2}$ $q=\frac{1}{2}$

BD

1. mean = np
   $= 4(\frac{1}{2}) = 2$
2. Variance = npq
   $4(\frac{1}{2})(1-\frac{1}{2}) = 1$
3. SD = $\sqrt{\text{Variance}}$
   $= \sqrt{1} = 1$

F(x)

$2^4 = 16$

X	P(X)	XP(X)	X^2 P(X)
0	$\frac{1}{16}$		
1			
2			
3			
4			

$V(X) = E(X^2) - [E(X)]^2$
$= 5 - 4$
$= 1$

Q) A dice is rolled 480 times. Find the mean, variance of the no of 3s that will be rolled

$n=480$ $P=\frac{1}{6}$ $q=\frac{5}{6}$

Mean = np
$480(\frac{1}{6}) = 80$

SD = $\sqrt{npq}$
$\sqrt{480(\frac{1}{6})(\frac{5}{6})} = 8.16$

3. Hypergeometric Experiment

↳ Trial outcomes ↷ FAIL ↷ SUCESS

↳ P(sucess) changes each trial

↳ successive trials are dependent

↳ experiment repeated a fix no. of times

$$P(x) = \frac{{}^{a}C_{x}\ {}^{b}C_{n-x}}{{}^{a+b}C_{n}}$$

(a: one kind of items; b: another kind of items; n: sample size; selecting with out replacement)

a+b = total population

Q) • **Assistant Manager Applicants:** Ten people apply for a job as assistant manager of a restaurant. Five have completed college and five have not. If the manager selects 3 applicants at random, find the probability that all 3 are college graduates.

C: a → ${}^{5}C_{x}$

NC: b → ${}^{5}C_{n-x}$

$a+b = 10$ → ${}^{10}C_{3}$

P(all 3 college graduates) $= \frac{{}^{5}C_{3}\ {}^{5}C_{0}}{{}^{10}C_{3}}$

Q) • **House Insurance:** A recent study found that 2 out of every 10 houses in a neighborhood have no insurance. If 5 houses are selected from 10 houses, find the probability that exactly 1 will be uninsured.

Select 5

2 → not insured

8 → insured

P(1 will be insured) $= \frac{{}^{8}C_{4}\ {}^{2}C_{1}}{{}^{10}C_{5}} = \frac{5}{9}$ (insured, not insured)

I: a → ${}^{8}C_{x}$

NI: b → ${}^{2}C_{n-x}$

$a+b = 10$ → ${}^{10}C_{5}$

4. Geometric Experiment

↳ Trial outcomes ↷ FAIL ↷ SUCESS

↳ P(sucess) remain constant for each trial

↳ each trial is independent

↳ experiment repeated a variable no. of times, until first sucess is obtained

→ $g(x) = pq^{x-1}$

mean → $\mu = \frac{1}{p}$

variance → $\sigma^2 = \frac{1-p}{p^2}$

Poisson Distribution

→ $P(X) = e^{-\lambda}\left(\frac{\lambda^x}{x!}\right)$ (x: no of times event occured; λ: mean of x)

↳ avg no of sucess

↳ time interval

↳ mean

Q) • **Radioactive Particles:** During a laboratory experiment, the average number of radioactive particles passing through a counter in 1 millisecond is 4. What is the probability that 6 particles enter the counter in a given millisecond?

$\lambda = 1 \text{ ms} \times 4 = 4$

$P(x=6)$

$P(x=6, \lambda=4)$

$e^{-4}\left(\frac{4^6}{6!}\right) = 0.1042$

if 6 particles in 2ms

$\lambda = 2 \text{ ms} (4) = 8$

Discrete Random variable

- ↳ PMF
- ↳ CDF
- ↳ mean
- ↳ variance

Continuous Random variable

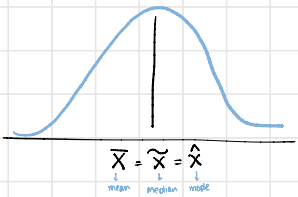
- ↳ PDF
- ↳ CDF
- ↳ mean
- ↳ variance

Normal Distribution

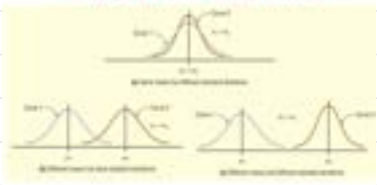
↳ Continous type Distribution

$$PDf: f(x) = \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sigma(\sqrt{2\pi})} \quad ; -\infty < x < \infty$$

(μ: population mean; σ: population sd)



Shapes of Normal Distribution



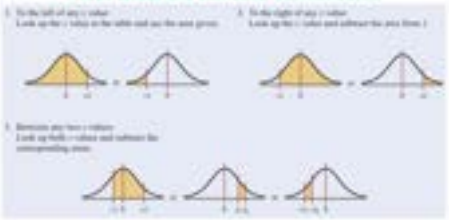
Standard Normal Distribution

↳ $\mu = 0$ and $\sigma = 1$

$$Z\text{-score} = \frac{x-\mu}{\sigma}, \quad y = \frac{e^{\frac{-z^2}{2}}}{\sqrt{2\pi}}$$

$$Z \sim SN(\mu=0, \sigma=1)$$

Areas under the standard Normal Curve



Example 01

a. Find the area to the left of $z = 2.06$

b. Find the area to the right of $z = -1.19$

c. Find the area between $z = 1.68$ and $z = -1.37$

$P(z < 2.06)$

$P(z > -1.19)$

$P(-1.37 < z < 1.68)$

Example # 12

•Gauges are used to reject all components for which a certain dimension is not within the specification 1.50 ± d. It is known that this measurement is normally distributed with mean 1.50 and standard deviation 0.2. Determine the value d such that the specifications "cover" 95% of the measurements.

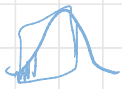
$$P(-1.96 < Z < 1.96) = 0.95$$

$$\pm 1.96 = \frac{(1.50 + d) - 1.50}{0.2} = Z$$

$\mu = 1.5 \quad \sigma = 0.2 \quad \alpha = 0.05$

AL: 0.95 + 0.025 . 0.975

CHP 6



9.177

6.5 Given a standard normal distribution, find the area under the curve that lies

(a) to the left of $z = -1.39$; 1-0.9177=0.0823

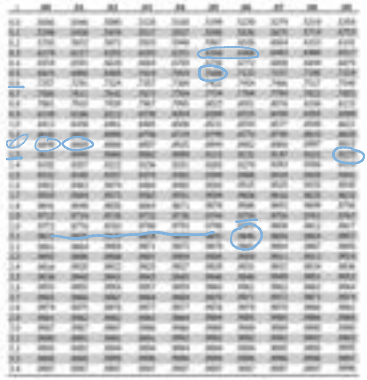
(b) to the right of $z = 1.96$; 1-0.9750=0.0250

(c) between $z = -2.16$ and $z = -0.65$; 1-0.9846 - 1-0.7422

(d) to the left of $z = 1.43$; 0.0154 - 0.2578 = 0.2424

(e) to the right of $z = -0.89$;

(f) between $z = -0.48$ and $z = 1.74$.



6.6 Find the value of z if the area under a standard normal curve

(a) to the right of z is 0.3622; 1-0.3622=0.6378 → 0.35

(b) to the left of z is 0.1131; 1-0.1131=0.8869 → 1.21

(c) between 0 and z, with $z > 0$, is 0.4838; 0.5 + 0.4838 = 0.9838 → 2.14

(d) between $-z$ and z, with $z > 0$, is 0.9500. 0.25+0.95=0.975 → 1.96

Z-score = $\frac{X-\mu}{\sigma}$

6.8 Given a normal distribution with $\mu = 30$ and $\sigma = 6$, find

(a) the normal curve area to the right of $x = 17$; $\frac{17-30}{6}$

(b) the normal curve area to the left of $x = 22$; $\frac{22-30}{6}$

(c) the normal curve area between $x = 32$ and $x = 41$; $\frac{32-30}{6}$ - $\frac{41-30}{6}$

(d) the value of x that has 80% of the normal curve area to the left; $x = z\times\sigma + \mu = 35.04$, $0.8\times 6+30$

(e) the two values of x that contain the middle 75% of the normal curve area.

$\frac{X-30}{6}$

6.9 Given the normally distributed variable X with mean 18 and standard deviation 2.5, find

(a) $P(X < 15)$; $\frac{15-18}{2.5} = -1.2$

(b) the value of k such that $P(X < k) = 0.2236$;

(c) the value of k such that $P(X > k) = 0.1814$;

(d) $P(17 < X < 21)$.

$\mu = 18$ 8.25 $\frac{X-18}{7.5}$

6.9 (a) $z = (15 - 18)/2.5 = -1.2$; $P(X < 15) = P(Z < -1.2) = 0.1151$.

(b) $z = -0.76$, $k = (2.5)(-0.76) + 18 = 16.1$.

(c) $z = 0.91$, $k = (2.5)(0.91) + 18 = 20.275$.

(d) $z_1 = (17 - 18)/2.5 = -0.4$, $z_2 = (21 - 18)/2.5 = 1.2$;
$P(17 < X < 21) = P(-0.4 < Z < 1.2) = 0.8849 - 0.3446 = 0.5403$.

Centeral limit theorem

z test

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

$\sigma/\sqrt{n}$ → standard error of mean

Formula	Use
1. $z = \frac{X - \mu}{\sigma}$	Used to gain information about an individual data value when the variable is normally distributed.
2. $z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$	Used to gain information when applying the central limit theorem about a sample mean when the variable is normally distributed or when the sample size is 30 or more.

Example 3

An upgrade of a certain software package requires the installation of 68 new files. Each file is installed sequentially. On average, it takes 15 seconds to install one file, with a variance of 11 sec^2.

(a) How likely is it that the whole package will be updated in less than 16 minutes?

(b) There is a new version of the package. It requires only N new files to be installed, and it is promised that 95% of the time upgrading takes less than 10 minutes. Based on this information, calculate N.

Solution:

(a) $n = 68$, $\mu =$ [illegible] $\text{Var}(X) =$ [illegible], $\sigma =$ [illegible] $= 960$sec

$$P(S_n \leq 960) = P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq \frac{960 - 1020}{\sqrt{748}}\right) = P(Z \leq -2.19) = 0.0143$$

(b) $N = ?$, $\mu =$ [illegible] $\text{Var}(X) = 11$sec^2, $\sigma =$ [illegible] $= 600$sec

$$P(S_n \leq 600) = P\left(\frac{S_n - N\mu}{\sigma\sqrt{N}} \leq \frac{600 - 15N}{\sqrt{11N}}\right) = P(Z \leq c) = 0.95$$

So for $N = 37$ [illegible] the correct probability more than 95%.

Finite Population correction Factor

$$\sqrt{\frac{N-n}{N-1}}$$

N → population size

n → sample size

- Estimation → point estimate (0.150 → 0.16) : $\bar{x}, s^2, p$ (represent their point as) / interval estimate ($16 < \mu < 25$) (with Probability) → CONFIDENCE INTERVAL
- Hypothesis testing
- ANOVA
- Regression and Corruption → Gallup, Pulse

Properties of good estimators

1. Unbiased estimator
   - $\bar{n}$ = parameter being estimated
2. Consistent estimator
3. Relatively efficient estimator
   - has the smallest variance

unbiased estimator of population mean (μ)

$\bar{X} \equiv \mu$

$\bar{X}$: statistic; μ: parameter

sampling error: $\bar{X} - \mu$

Q) 90% → $z^{\frac{(1-0.9)}{2}} = 1.65$

95% → $z^{\frac{(1-0.95)}{2}} = 1.96$

99% → $z^{\frac{(1-0.99)}{2}} = 2.58$

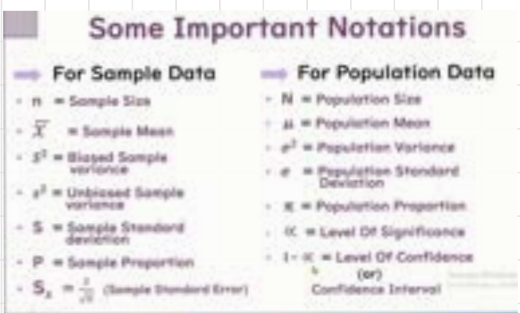
eg $n = 50, \bar{x} = 54, \sigma = 6$

95% of CI mean?

$54 \pm (1.96)\left(\frac{6}{\sqrt{50}}\right)$

$52.3 < \mu < 55.7$

95% chance that the above interval contains true population mean



PROB ASSIGNMENT K214688

Q1)

a) $f(x,y,z) = \begin{cases} ky^2z, & 0<x<4 \\ & y<1 \\ & 0<z<2 \end{cases}$

$$\int_0^2\int_0^1\int_0^1 ky^2z\,dx\,dy\,dz = 1$$

$$\int_0^2\int_0^1 \left| \frac{kx^2y^2z}{2} \right|_0^1 = 1$$

$$\int_0^2\int_0^1 \frac{ky^2z}{2} = 1$$

$$\int_0^2 \left| \frac{ky^3z}{6} \right|_0^1 = 1$$

$$\int_0^2 \frac{kz}{6} = 1$$

$$\left| \frac{kz^2}{12} \right|_0^2 = 1$$

$$\frac{k}{3} = 1$$

$$\boxed{k = 3}$$

b) $P(x<1/4,\ y>1/2,\ 1<z<2)$

$$3\int_1^2\int_{1/2}^1\int_0^{1/4} xy^2z\,dx\,dy\,dz$$

$$3\int_1^2\int_{1/2}^1 \left| \frac{x^2y^2z}{2} \right|_0^{1/4}$$

$$3\int_1^2 \left| \frac{y^2z}{32} \right|_{1/2}^1$$

$$3\int_1^2 \left| \frac{y^3z}{96} \right|_{1/2}^1$$

$$3\int_1^2 \frac{7z}{768}$$

$$3\left| \frac{7z^2}{1536} \right|_1^2$$

$$\frac{21}{512}$$

c) Marginal

$f(y)$: $3\int_0^2\int_0^1 xy^2z\,dx\,dz$

$$3\int_0^2 \left| \frac{x^2y^2z}{2} \right|_0^1$$

$$3\int_0^2 \frac{y^2z}{2}$$

$$3\left| \frac{y^2z^2}{4} \right|_0^2$$

$$= 3y^2$$

$f(z)$: $3\int_0^1\int_0^1 xy^2z\,dx\,dy$

$$3\int_0^1 \frac{y^2z}{2}$$

$$3\left| \frac{y^3z}{6} \right|_0^1$$

$$= \frac{z}{2}$$

$E(y) = \int_0^1 3y^3$

$= 3/4$

$E(z) = \int_0^2 \frac{z^2}{2}$

$= 4/3$

$E(yz) = 3/2\int_0^2\int_0^1 yz(y^2z)$

$= 3/2\int_0^2\int_0^1 y^3z^2$

$= 3/2\int_0^2 \left| \frac{y^4z^2}{4} \right|_0^1$

$= 3/2\int_0^2 \frac{z^2}{4}$

$= 3/2\left| \frac{z^3}{12} \right|_0^2$

$= 1/2$

$E(x^2) = \int_0^1 3y^4$

$\left| \frac{3y^5}{5} \right|_0^1$

$= 3/5$

$E(z^2) = \int_0^2 \frac{z^3}{2}$

$= \left| \frac{z^4}{8} \right|_0^2$

$= 2$

$\text{var}(y) = 3/5 - (3/4)^2$

$= 3/80$

$\text{var}(z) = 2 - (4/3)^2$

$= 2/9$

$\text{covariance} = \frac{1}{2} - \frac{3}{80}\left(\frac{2}{4}\right)$

$= \frac{59}{120}$

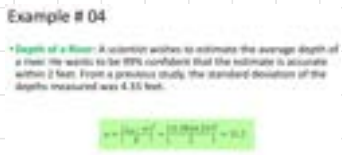
$\text{correlation} = \frac{59/120}{\sqrt{3/80}\sqrt{2/9}}$

$= 11.738$

ESTIMATION

Point estimate — Interval estimate

1. $\bar{X} = \frac{\Sigma x}{n}$ (sample mean; n → sample size)
2. $S = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$ (S.D)
3. $S_{\bar{x}} = \frac{s}{\sqrt{n}}$ (standard error mean; S.D)
4. $E = Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$ (margin of error; $Z_{\frac{\alpha}{2}}$ → critical region of z-distribution; $\frac{\sigma}{\sqrt{n}}$ → standard error of mean)
5. $n = \left(\frac{Z_{\frac{\alpha}{2}}\,\sigma}{E}\right)^2$


Example # 04

• Depth of a River: A scientist wishes to estimate the average depth of a river. He wants to be 95% confident that the estimate is accurate within 2 feet. From a previous study, the standard deviation of the depths measured was 4.35 feet.

Confidence interval for μ (1 distribution)

1. $\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ (population S.D)
   - ↳ σ known
2. $\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$ (sample S.D)
   - ↳ σ unknown
   - ↳ $n \geq 30$
3. $\bar{X} \pm t_{\frac{\alpha}{2(v)}} \frac{s}{\sqrt{n}}$ (sample S.D; critical value of t distribution; $n-1$)
   - ↳ σ unknown
   - ↳ $n < 30$

Confidence interval for $\mu_1 - \mu_2$ (2 distributions)

1. $(\bar{X}_1 - \bar{X}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}}$
   - ↳ σ_1, σ_2 known
2. $(\bar{X}_1 - \bar{X}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}}$
   - ↳ σ_1, σ_2 unknown
   - ↳ $n \geq 30$
3. $(\bar{X}_1 - \bar{X}_2) \pm t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}}$
   - ↳ σ_1, σ_2 unknown ?
   - ↳ $n < 30$

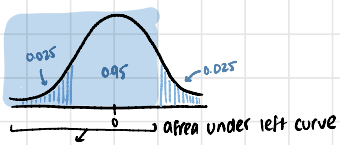
Z-distribution

Question # 1

An electrical firm manufactures light bulbs that have a length of life with mean , and a standard deviation of 40 hours. If a sample of 100 bulbs has an average life of 780 hours, Find a 95% Confidence Interval for the population mean of all bulbs produced by this firm.

$\sigma = 40$ hours $\quad n = 100 \quad \bar{x} = 780$ hours $\quad CI = 95\% \quad \mu = ?$

(1) as $n > 30$ we use z test



A_L = confidence level + α

$A_L = 0.95 + 0.025 = 0.975 \rightarrow$ find on Table

$Z_{\frac{\alpha}{2}} = 1.96$

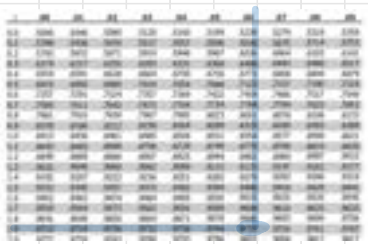
(2) $\bar{x} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

$= 780 \pm 1.96 \left(\frac{40}{\sqrt{100}}\right)$

$= 780 \pm 7.84$

$= 772.16 , 787.84$

$772.16 < \mu \leq 787.84$

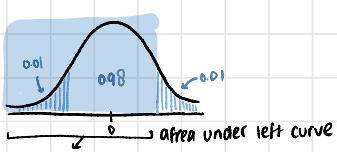


Question # 2

The height of a random sample of 50 college students showed a mean of 174.5 cm and a standard deviation of 6.9cm. Construct a 98% Confidence interval for the mean height of all college students.

$n = 50 , \bar{x} = 174.5 cm , s = 6.9 cm , CI = 98\% , \mu = ?$

(1) $n > 50$ so z test



A_L = confidence level + α

$A_L = 0.98 + 0.01 = 0.99 \rightarrow$ find on Table

$Z_{\frac{\alpha}{2}} = 2.3263$

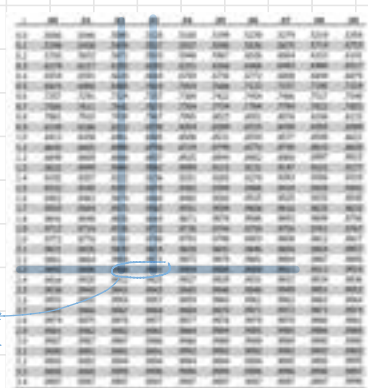
(2) $\bar{x} \pm Z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$

$= 174.5 \pm 2.3263 \left(6.9/\sqrt{50}\right)$

$= 174.5 \pm 2.27$

$= 172.23 , 176.77$

$172.33 < \mu \leq 176.77$



→ closest to 0.99
→ take avg
→ $\frac{0.01 + 0.01}{2} = 0.025$
→ 2.325

Question # 3

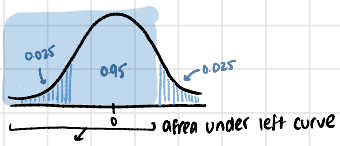
A random sample of 100 farms in a certain year gives an average yield of barley of 2100 lbs. Per acre. A random sample of 100 farms in the following years given an average yield of 2000 lbs. per acre. The S.D for two populations are 224 and 192 respectively. Compute a 95% C-I for the difference between two population means. Assuming data follows normal distribution

$n_1 = 100$ $\bar{X}_1 = 2100$ $CI = 95\%$

$n_2 = 100$ $\bar{X}_2 = 2000$ $\mu_1 - \mu_2 = ?$

$\sigma_1 = 224$ $\sigma_2 = 192$

(1) as $n \geq 30$ so z test



A_L = confidence level + α

$A_L = 0.95 + 0.025 = 0.975 \rightarrow$ find on Table

$Z_{\frac{\alpha}{2}} = 1.96$

(2) $(\bar{X}_1 - \bar{X}_1) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}}$

$= (2100 - 2000) \pm 1.96 \sqrt{\frac{224^2}{100} + \frac{192^2}{100}}$

$= 100 \pm 57.82$

$= 42.18 , 157.82$

$= 42.18 < \mu \leq 157.82$

Confidence Interval for Proportions

1. $\hat{P} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{P}\hat{q}}{n}}$ — $\hat{P}$: sample proportion; $\frac{x}{n}$: $x \rightarrow$ no of sample units; $\hat{q} = 1 - \hat{P}$

Example # 09

• **Covering College Costs:** A survey conducted by Sallie Mae and Gallup of 1404 respondents found that 323 students paid for their education by student loans. Find the 90% confidence of the true proportion of students who paid for their education by student loans.

$n = 1404$, $X = 323$, $CI = 90\%$

$P = \frac{323}{1404} = 0.23$ $q = 1 - 0.23 = 0.77$

$0.23 - 1.65 \sqrt{\frac{(0.23)(0.77)}{1404}}$

$= 21.1\% < p < 24.9\%$

Example #10

• **Religious Books:** A survey of 1721 people found that 15.9% of individuals purchase religious books at a Christian bookstore. Find the 95% confidence interval of the true proportion of people who purchase their religious books at a Christian bookstore.

$n = 1721$, $\hat{P} = 15.9\%$ (?), $CI = 95\%$

$\hat{q} = 1 - 0.159 = 0.841$

$= 0.159 \pm 1.96 \sqrt{\frac{(0.159)(0.841)}{1721}}$

$= 0.142 < p < 0.176$

Minimum 'n' needed Interval Estimate of Population Porpotion

$$n = \hat{p}\hat{q}\left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2$$ → rearrange confidence interval for porportions

Example # 11

- **Home Computers:** A researcher wishes to estimate, with 95% confidence, the proportion of people who own a home computer. A previous study shows that 40% of those interviewed had a computer at home. The researcher wishes to be accurate within 2% of the true proportion. Find the minimum sample size necessary.

CI = 95%, $\hat{p}$ = 40%, E = 2%.

$\hat{q}$ = 1 − 40 = 60%.

$n = (0.4)(0.6)\left(\frac{1.96}{0.02}\right)^2$

$= 2304.96$

Example # 12

- **M&M Colors:** A researcher wishes to estimate the percentage of M&M's that are brown. He wants to be 95% confident and be accurate within 3% of the true proportion. How large a sample size would be necessary?

CI = 95%, E = 3%.

$\hat{p}$ = 5% $\hat{q}$ = 5%.

$n = (0.5)(0.5)\left(\frac{1.96}{0.03}\right)$

$= 1067.1$

Chi Square Distribution

↳ random sample

↳ normal distribution

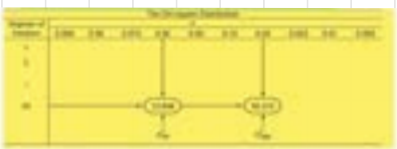
1. $\frac{(n-1)s^2}{\chi^2_{right}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{left}}$

2. $df = n-1$

Example # 14

Find the values for χ^2_{right} and χ^2_{left} for a 90% confidence interval when $n = 25$.

χ_{right}: $\frac{1-0.90}{2} = 0.05$, χ_{left}: $1-0.05 = 0.95$, df: $25-1 = 24$

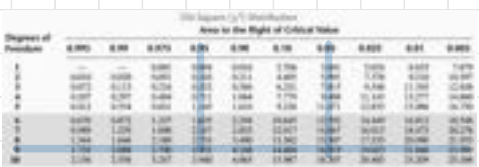


$\chi^2_{left} = 13.848$ $\chi^2_{right} = 36.415$

Example # 16

•Cost of Ski Lift Tickets Find the 90% confidence interval for the variance and standard deviation for the price in dollars of an adult single-day ski lift ticket. The data represent a selected sample of nationwide ski resorts. Assume the variable is normally distributed.

59	54	53	52	51
39	49	46	49	48



$n=10$, CI: 90

χ_{right}: $\frac{1-0.90}{2} = 0.05$, χ_{left}: $1-0.05 = 0.95$, $df = 9$

$\chi^2_{left} = 16.919$ $\chi^2_{right} = 3.325$

$\frac{(10-1)(28.2)^2}{16.919} < \sigma^2 < \frac{(10-1)(28.2)^2}{3.325}$?

$15.0 < \sigma^2 < 76.3$

S.D

$s = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$

$= \sqrt{\frac{25254}{10} - \left(\frac{500}{10}\right)^2}$

$= 5.039$

Example # 15

•Nicotine Content: Find the 95% confidence interval for the variance and standard deviation of the nicotine content of cigarettes manufactured if a sample of 20 cigarettes has a standard deviation of 1.6 milligrams.

CI: 95%, $n=20$, s.d: 1.6

$X_{right}: \frac{1-0.95}{2} = 0.025$, $X_{left}: 1-0.025 = 0.975$, $df: 19$

$X^2_{right} = 32.852$, $X^2_{left} = 8.907$

$$\frac{(20-1)(1.6)^2}{32.852} < \sigma^2 < \frac{(20-1)(1.6)^2}{8.907}$$

$$1.5 < \sigma^2 < 5.5$$

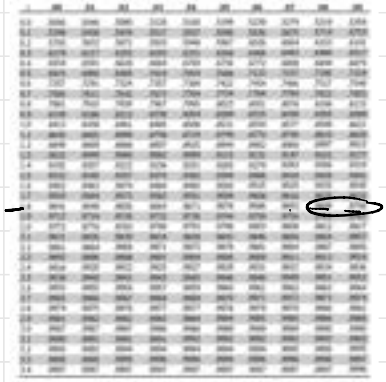
Chi-Square (χ^2) Distribution

Area to the Right of Critical Value

Degrees of freedom	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

CHP 9

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



9.3 Many cardiac patients wear an implanted pacemaker to control their heartbeat. A plastic connector module mounts on the top of the pacemaker. Assuming a standard deviation of 0.0015 inch and an approximately normal distribution, find a 95% confidence interval for the mean of the depths of all connector modules made by a certain manufacturing company. A random sample of 75 modules has an average depth of 0.310 inch.

$\sigma = 0.0015$ $CI = 95\%$

$n = 75$ $\bar{x} = 0.310$

$0.310 \pm 1.96\left(\frac{0.0015}{\sqrt{75}}\right)$

$0.3094 < \mu < 0.3106$

$\frac{1-0.95}{2} = 0.025$

$A_L = 0.025 + 0.95 = 0.975$

$P \sim 1.96$

9.6 How large a sample is needed in Exercise 9.2 if we wish to be 96% confident that our sample mean will be within 10 hours of the true mean?

$n = ?$ $CI = 96$ $\bar{x} = 10$ $\sigma = 40$ $E = 10$

$E = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

$n = \left(\frac{z_{\frac{\alpha}{2}} \sigma}{E}\right)^2$

$n = \left(\frac{2.055 \times 40}{10}\right)^2 = 68$ round up

$\frac{1-0.96}{2} = 0.02$

$A_L = 0.98$

$P = 2.055$

9.35 A random sample of size $n_1 = 25$, taken from a normal population with a standard deviation $\sigma_1 = 5$, has a mean $\bar{x}_1 = 80$. A second random sample of size $n_2 = 36$, taken from a different normal population with a standard deviation $\sigma_2 = 3$, has a mean $\bar{x}_2 = 75$. Find a 94% confidence interval for $\mu_1 - \mu_2$.

$n_1 = 25$ $\bar{x}_1 = 80$ $\sigma_1 = 5$ $CI = 94$

$n_2 = 36$ $\bar{x}_2 = 75$ $\sigma_2 = 3$

$80 - 75 \pm 1.885\sqrt{\frac{5^2}{25} + \frac{3^2}{75}}$

$2.9 < \mu_1 - \mu_2 < 7.1$

$\frac{1-0.94}{2} = 0.03$

$A_L = 0.97$

$P = 1.88$

HYPOTHESIS TESTING

Null hypothesis : H_0 : $\mu = 20$

Alternate Hypothesis : H_1 : $\mu \neq 20$ or $\mu > 20$ or $\mu < 20$

level of significance = α : 0.05 or 0.01 or 0.10

α : P [rejecting H_0 when H_0 is true] : P [Type - I Error]

β : Accepting false H_0 : Type II error

Statistical Tests

↳ Z-test

↳ t- test

↳ f- test

Two Tailed Test

$H_0 : \mu = K$

$H_1 : \mu \neq K$

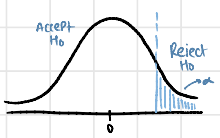
$\alpha = \frac{1 - \text{Confidence level}}{2}$

one Tailed Test

Right tailed test

$H_0 : \mu = K$

$H_1 : \mu > K$



Left-tailed test

$H_0 : \mu = K$

$H_1 : \mu < K$

$\alpha = 1 - \text{confidence level}$

Z-Test for mean

↳ $n \geq 30$

↳ σ is known

↳ independent sample

↳ population is normally distributed

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

($\bar{x}$: sample mean, μ: H_0 mean, σ: standard deviation, n: sample size)

$n \geq 30$

1. State H_0, H_1
2. calculate z formula
3. find rejection region
4. if calculated z inside rejection region
   ↳ reject H_0

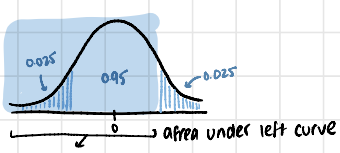
1. A factory has a machine that dispenses 80 mL of fluid in a bottle. An employee believes the average amount of fluid is not 80 mL. Using 40 samples, he measures the average amount dispensed by the machine to be 78 mL with a standard deviation of 2.5. (a) State the null and alternative hypotheses. (b) At a 95% confidence level, is there enough evidence to support the claim that the machine is not working properly?

$\bar{x} = 78$, $s = 25$, $n = 40$ → as > 30 so z-test

① $H_0 = \mu = 80$ (null hypotheses) $H_a = \mu \neq 80$ (alternate hypotheses)

② $z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{78 - 80}{25/\sqrt{40}} = -5.06$

③ $\alpha = \frac{1 - 0.95}{2} = 0.025$ 0.005 + 0.99



A_L = confidence level + α

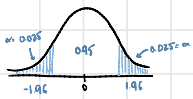
$A_L = 0.95 + 0.025 = 0.975$ → find on table

$Z_{\frac{\alpha}{2}} = 1.96$

④ z falls in rejection region

Reject H_0

so with a% confidence we believe that machine is not working properly



A random sample of size 36 is taken from a normal population with a known variance $\sigma^2 = 25$. If the mean of the sample is $\bar{x} = 42.6$, test the null hypothesis $\mu = 45$ against the alternative hypothesis $\mu < 45$ with $\alpha = 0.05$.

$n > 30$ so z-test

$n = 36$, $\sigma^2 = 25$, $\bar{x} = 42.6$, $\alpha = 0.05$

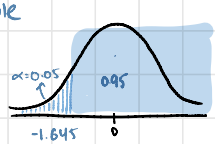
① $H_0 = 45$, $H_1 < 45$, $\sigma = 5$

② $z = \frac{42.6 - 45}{5/\sqrt{36}} = -2.88$

③ area under right curve

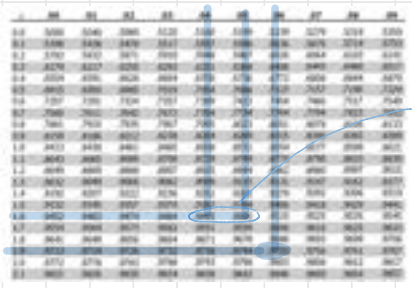
$A_R = 0.95$ → check table

$Z_{\alpha} = -1.645$



④ z falls in rejection region

Reject H_0



Z-test for two means

- $n > 30$
- σ known
- independent sample

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

sample mean; H_0 mean; standard deviation; sample size

$$\delta = \sqrt{\frac{n\Sigma x^2 - (\Sigma x)^2}{n(n-1)}}$$

An admission test was administered to incoming freshmen in the College of Medical Laboratory and Sciences and College of Radiologic Technology with 100 students each college randomly selected. The mean scores of the given samples were $\bar{x}_1 = 90$ and $\bar{x}_2 = 85$ and the variances of the test scores were 40 and 35 respectively. Is there a significant difference between the two groups? Use .01 level of significance.

$\bar{X}_1 = 90$ $\bar{X}_2 = 85$ $n = 100$

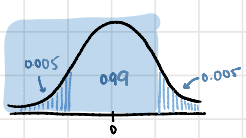
$s^2 = 40$ $s^2 = 35$ $\alpha = 0.01$

① $H_0: \mu = 0$

$H_1: \mu \neq 0$

② $$z = \frac{(90 - 85) - (0 - 0)}{\sqrt{\frac{40^2}{100} + \frac{35^2}{100}}} = 5.714$$

③

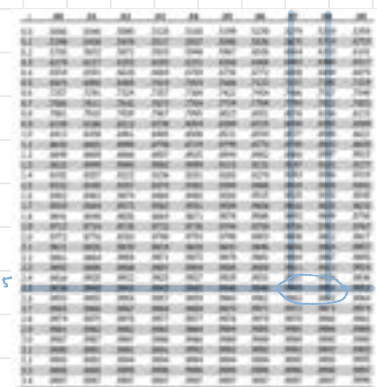


$AL = 0.005 + 0.99 = 0.995$

$z_{\frac{\alpha}{2}} = 2.575$

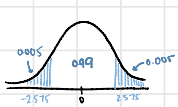
$= \frac{0.01 + 0.08}{2} = 0.075$

$= 2.575$



④ z falls in rejected region

H_0 rejected



T-Test for mean

↳ variance > 1

↳ $n < 30$

↳ σ is known

↳ independent sample

↳ population is normally distributed

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

$\bar{x}$: sample mean; μ: H_0 mean; σ: standard deviation; n: sample size

$df = n - 1$ (degree of freedom)

1. State H_0, H_1 — $n < 30$
2. calculate t formula
3. find rejection region
4. if calculated t inside rejection region
   ↳ reject H_0

2. A company manufactures car batteries with an average life span of 2 or more years. An engineer believes this value to be less. Using 10 samples, he measures the average life span to be 1.8 years with a standard deviation of 0.15. (a) State the null and alternative hypotheses. (b) At a 99% confidence level, is there enough evidence to discard the null hypothesis?

① $H_0 = \mu \geq 2$

$H_a = \mu < 2$

②

$\mu_0 = 2$, $\bar{x} = 1.8$, $s = 0.15$, $n = 10$ → as $n < 30$ so we use t test

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1.8 - 2}{0.15/\sqrt{10}} = -4.22$$

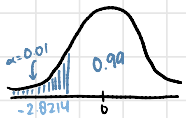
③ $df = n - 1$

$= 10 - 1 = 9$

$\alpha = 1 - 0.99 = 0.01$

find on table

$t_{\alpha} = -2.821$



④ t falls in rejection region

H_0 rejected

so with 99% confidence we belive that H_a

df \ α	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	318.3
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	22.33
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	10.21
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	7.173
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	5.893
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	5.208
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.785
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	4.501
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	4.297
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	4.144
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	4.025
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.930
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.852
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.787

T-test for two means

↳ $n < 30$

↳ σ known

↳ independent samples

Kaito grows tomatoes in two separate fields. When the tomatoes are ready to be picked, he is curious as to whether the sizes of his tomato plants differ between the two fields. He takes a random sample of plants from each field and measures the heights of the plants. Here is a summary of the results:

	Field A	Field B
Mean	1.3 m	1.6 m
Standard deviation	0.5 m	0.3 m
Number of plants	22	24

↳ $\sigma_1 \neq \sigma_2$ variance

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

($\bar{x}_1 - \bar{x}_2$: sample mean; $\mu_1 - \mu_2$: H_0 mean; σ: standard deviation; n: sample size)

$$df \approx \frac{\left(\frac{\delta_1^2}{n_1} + \frac{\delta_2^2}{n_2}\right)^2}{\frac{\delta_1^2}{n_1 - 1} + \frac{\delta_2^2}{n_2 - 1}}$$

① $H_0: \mu = 0 \rightarrow \mu_1 = \mu_2$

$H_1: \mu \neq 0 \rightarrow \mu_1 \neq \mu_2$

② $n_1 = 22$ $\sigma_1 = 0.5$ $\bar{x}_1 = 1.3$ $\alpha = 0.05$ (assume as not mentioned)

$n_2 = 24$ $\sigma_2 = 0.3$ $\bar{x}_2 = 1.6$

$$t = \frac{(1.3 - 1.6)(0 - 0)}{\sqrt{\frac{0.5^2}{22} + \frac{0.3^2}{24}}} = -2.4402$$

③ $df_1 = 21$ $df_2 = 23$ $\alpha = 0.05/2 = 0.025$

take $t_{\frac{\alpha}{2}}$ for the smallest df

$t_{0.025} = 2.080$

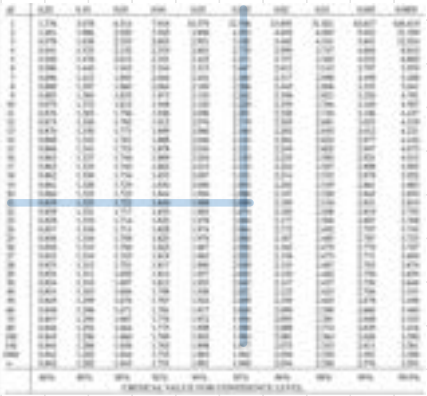
t lies in rejection region

H_0 rejected

↳ $\sigma_1 = \sigma_2$ variance

$$\delta = \frac{(n_1 - 1)(\delta_1)^2 + (n_2 - 1)(\delta_2)^2}{(n_1 + n_2) - 2}$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\delta \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$



T-test for two means

↳ $n < 30$

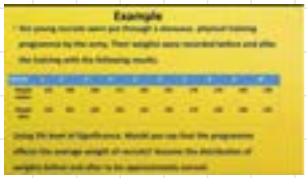
↳ σ unknown

↳ dependent samples

1. $D = X_1 - X_2$ 2. $D^2 = (X_1 - X_2)^2$

3. $\bar{X} = \frac{\Sigma D}{n}$ 4. $S.d = \sqrt{\frac{n\Sigma D^2 - (\Sigma D)^2}{n(n-1)}}$

5. $t = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ 6. $df = n-1$


Example

Six young recruits were put through a strenuous physical training programme by the army. Their weights were recorded before and after the training with the following results:

Weight before	125	195	160	171	140	201	170	176	195	139
Weight after	136	201	158	184	145	195	175	190	190	145

Using 5% level of significance, Would you say that the programme affects the average weight of recruits? Assume the distribution of weights before and after to be approximately normal.

① $H_0: \mu = 0$ $H_1: \mu \neq 0$

② $n = 10$, $\alpha = 0.05$

$\Sigma D = (125-136)+(195-201)+(160-158)+(171-184)+(140-145)$
$+(201-195)+(170-175)+(176+190)+(195-190)+(139-145) = -47$

$\bar{X} = \frac{47}{10} = 4.7$

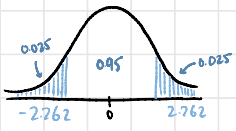
$\Sigma D^2 = (125-136)^2+(195-201)^2+(160-158)^2+(171-184)^2+(140-145)^2$
$+(201-195)^2+(170-175)^2+(176+190)^2+(195-190)^2+(139-145)^2 = 673$

$S.d = \sqrt{\frac{10(673)-(47)^2}{10(10-1)}} = 7.09$

③ $t = \frac{4.7-0}{7.09/\sqrt{10}} = 2.09$

④ $df = 9$, $\alpha = \frac{0.05}{2} = 0.025$

$t_{\frac{\alpha}{2}} = 2.262$



⑤ t lies in acceptance region

Ho accept

↳ $\sigma_1 = \sigma_2$ ← variance

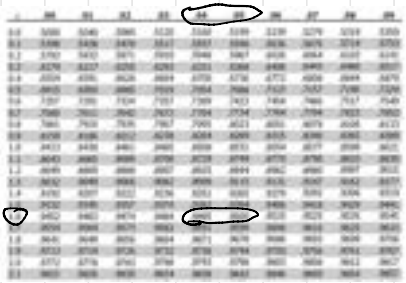
$\delta = \sqrt{\frac{(n-1)(\delta_1)^2+(n-1)(\delta_2)^2}{(n+n)-2}}$

$t = \frac{(\bar{X_1}-\bar{X_2})-(\mu_1-\mu_2)}{\delta\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

$\delta = \sqrt{\frac{n\Sigma x^2-(\Sigma x)^2}{n(n-1)}}$

df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	318.3
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	22.33
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	10.21
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	7.173
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	5.893
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	5.208
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.785
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	4.501
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	4.297
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	4.144
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	4.025
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.930
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.852
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.787

CHP 10



10.20 A random sample of 64 bags of white cheddar popcorn weighed, on average, 5.23 ounces with a standard deviation of 0.24 ounce. Test the hypothesis that $\mu = 5.5$ ounces against the alternative hypothesis, $\mu < 5.5$ ounces, at the 0.05 level of significance.

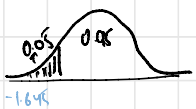
→ Z test

$n = 64 \quad \bar{x} = 5.23 \quad \delta = 0.24 \quad \alpha = 0.05$

$H_0: \mu = 5.5$

$H_1: \mu < 5.5$

$$z = \frac{5.23 - 5.5}{0.24/\sqrt{64}} = -9$$



$A_R = 0.95$

from table $= 1.645$

$P = 1.645$

H_0 reject $\mu < 5.5$

10.23 Test the hypothesis that the average content of containers of a particular lubricant is 10 liters if the contents of a random sample of 10 containers are 10.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.3, and 9.8 liters. Use a 0.01 level of significance and assume that the distribution of contents is normal.

$n = 10$ → T Test $\quad \alpha = 0.01$

10.3 10.4 10.2 10.1 10.1 9.9 10.3 9.8 9.7 9.8

$H_0: \mu = 10$

$H_1: \mu \neq 10$

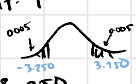
$\bar{x} = \frac{100.6}{10} = 10.06$

$\Sigma x^2 = 1012.64$

$$\delta = \sqrt{\frac{1012.64}{10} - \left(\frac{100.6}{10}\right)^2} = 0.246$$

$$t = \frac{10.06 - 10}{0.246/\sqrt{10}} = 0.77$$

$df = 9$

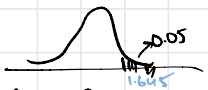


$P = 3.250$

df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	318.3
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	22.33
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	10.21
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	7.173
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	5.893
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	5.208
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.785
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	4.501
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	4.297
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	4.144
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	4.025
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.930
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.852
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.787

H_0 accept

10.25 It is claimed (H_1) that automobiles are driven on average more than 20,000 kilometers per year. To test this claim, 100 randomly selected automobile owners are asked to keep a record of the kilometers they travel. Would you agree with this claim if the random sample showed an average of 23,500 kilometers and a standard deviation of 3900 kilometers? Use a P-value in your conclusion.



$A_L = 0.95$

$P = 1.645$

z test

$H_0 \rightarrow \mu = 20{,}000$ $\quad n = 100$ $\quad \bar{x} = 23500$ $\quad \delta = 3900$

$H_1 \rightarrow \mu > 20{,}000$ $\quad \alpha = 0.05$ as not mentioned

$$z = \frac{23500 - 20000}{3900/\sqrt{100}} = 8.974$$

H_0 rejected

$\mu > 20{,}000$

10.27 A study at the University of Colorado at Boulder shows that running increases the percent resting metabolic rate (RMR) in older women. The average RMR of 30 elderly women runners was 34.0% higher than the average RMR of 30 sedentary elderly women, and the standard deviations were reported to be 10.5 and 10.2%, respectively. Was there a significant increase in RMR of the women runners over the sedentary women? Assume the populations to be approximately normally distributed with equal variances. Use a P-value in your conclusions.

Decision: Reject H_0 and claim $\mu > 220$ milligrams

10.27 The hypotheses are

$$H_0 : \mu_1 = \mu_2,$$
$$H_1 : \mu_1 > \mu_2.$$

Since $s_p = \sqrt{\frac{(29)(10.5)^2 + (29)(10.2)^2}{58}} = 10.35$, then

$$P\left[T > \frac{34.0}{10.35\sqrt{1/30 + 1/30}}\right] = P(Z > 12.72) \approx 0.$$

Hence, the conclusion is that running increases the mean RMR in older women.

z test δ known

$H_0 \rightarrow \mu_1 = \mu_2$ $\quad n = 30$ $\quad \bar{X}_1 - \bar{X}_2 = 34$ $\quad \delta_1 = 10.5$

$H_1 \rightarrow \mu_1 \neq \mu_2$ $\quad \delta = 10.2$

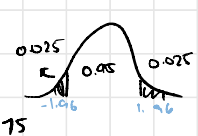
$$\delta = \frac{(29)(10.5)^2 - (29)(10.2)^2}{30 + 30 - 2} = 10.35$$

$$z = \frac{34}{10.35/\sqrt{\frac{1}{30} + \frac{1}{30}}} = 0.8481$$

H_0 accept

z table

$\alpha = 005$



$A_L = 0.975$

$P = 1.96$

10.35 To find out whether a new serum will arrest leukemia, 9 mice, all with an advanced stage of the disease, are selected. Five mice receive the treatment and 4 do not. Survival times, in years, from the time the experiment commenced are as follows:

Treatment	2.1	5.3	1.4	4.6	0.9
No Treatment	1.9	0.5	2.8	3.1	

At the 0.05 level of significance, can the serum be said to be effective? Assume the two populations to be normally distributed with equal variances.

T test σ unknown

$n_1 = 5$ $\quad \bar{x}_1 = \frac{14.3}{5} = 2.86$ $\quad s_1 = \frac{5(56.43) - (14.3)^2}{5(4)} = 3.883$

$n_2 = 4$ $\quad \bar{x}_2 = \frac{8.3}{4} = 2.075$ $\quad s_2 = \frac{4(21.31) - (8.3)^2}{4(3)} = 1.3625$

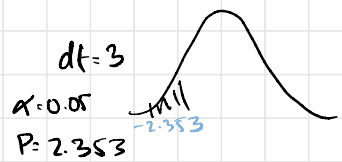
$\bar{x}_1 - \bar{x}_2 = 0.785$

$$s = \frac{3(1.3625)^2 + 4(3.883)^2}{5+4-2} = 1.674$$

$df = 3$

$\alpha = 0.05$

$P = 2.353$



$$t = \frac{0.785}{1.674\sqrt{\frac{1}{5} + \frac{1}{4}}} = -0.70$$

$H_0 \rightarrow \mu_1 - \mu_2 = 0$

$H_1 \rightarrow \mu_1 - \mu_2 < 0$

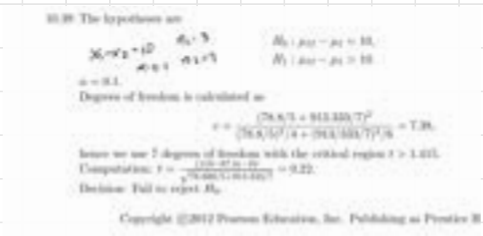
H_0 accept

normal with equal variances.

10.39 The following data represent the running times of films produced by two motion-picture companies:

Company	Time (minutes)						
1	102	86	98	109	92		
2	81	165	97	134	92	87	114

Test the hypothesis that the average running time of films produced by company 2 exceeds the average running time of films produced by company 1 by 10 minutes against the one-sided alternative that the difference is less than 10 minutes. Use a 0.1 level of significance and assume the distributions of times to be approximately normal with unequal variances.



ANALYSIS OF VARIANCE (ANOVA)

F test

- ↳ independent samples
- ↳ normal distribution
- ↳ σ are same

not t-test as the more means, the more t-tests needed

1. State H_0, H_a, α
2. Degree's of freedom
   - ↳ df between groups = $K-1$
   - ↳ df within groups = $N-K$ → no of groups
   - ↳ diff total = $N-1$ → total no of observations
   - ↳ F critical → using table (k, N)
3. Sum of square deviations from mean
   - ↳ each groups mean → $\bar{x}_1, \bar{x}_2, \bar{x}_3$
   - ↳ Grand mean → $\bar{\bar{x}}$
   - ↳ $SS_{total} = \Sigma (x - \bar{\bar{x}})^2$
   - ↳ $SS_{within} = \Sigma (x - \bar{x})^2$
   - ↳ $SS_{between} = SS_{total} - SS_{within}$
4. Mean Square
   - ↳ $MS_{between} = \frac{SS_{between}}{df_{between}}$
   - ↳ $MS_{within} = \frac{SS_{within}}{df_{within}}$
5. Calculate F statistic
   - ↳ $F = \frac{MS_{between}}{MS_{within}}$
   - ↳ conclusion

reject H_0	accept H_0
$F_{ratio} > F_{critical}$	$F_{ratio} < F_{critical}$
$F_{ratio} > 1$	$F_{ratio} \leq 1$

ANALYSIS OF VARIANCE (ANOVA)

E.G

	A	B	C
	1	2	2
	2	4	3
	5	2	4
mean	2.67	2.67	3

(1) $H_0 \rightarrow \mu_1 = \mu_2 = \mu_3$ → no diff in means

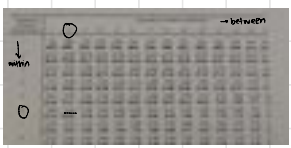
$H_a \rightarrow$ atleast 1 diff in the means

$\alpha = 0.5$ → generally

(2) $df_{between} = K - 1 = 3 - 1 = 2$

$df_{within} = N - K = 9 - 3 = 6$

diff total $= 6 + 2 = 8$



$F_{critical} = 5.14$

(3) $\bar{X}_1 = 2.67$, $\bar{X}_2 = 2.67$, $\bar{X}_3 = 3$, $\bar{\bar{X}} = \frac{25}{9} = 2.78$

$SS_{total} = \Sigma (x - \bar{\bar{X}})^2$

$= (1-2.78)^2 + (2-2.78)^2 + (5-2.78)^2 +$
$(2-2.78)^2 + (4-2.78)^2 + (2-2.78)^2 +$
$(2-2.78)^2 + (3-2.78)^2 + (4-2.78)^2 = 13.6$

$SS_{within} = \Sigma (x_n - \bar{X}_n)^2$

$= (1-2.67)^2 + (2-2.67)^2 + (5-2.67)^2 +$
$(2-2.67)^2 + (4-2.67)^2 + (2-2.67)^2 +$
$(2-3)^2 + (3-3)^2 + (4-3)^2 = 13.34$

$SS_{between} = SS_{total} - SS_{within}$

$= 13.6 - 13.34 = 0.23$

(4) $MS_{between} = \frac{SS_{between}}{df_{between}} = \frac{0.23}{2} = 0.12$

$MS_{within} = \frac{SS_{within}}{df_{within}} = \frac{13.34}{6} = 2.22$

1. State H_0, H_a, α
2. Degree's of freedom
   - ↳ df between groups $= K - 1$
   - ↳ df within groups $= N - K$ (N → total no of observations; K → no of groups)
   - ↳ diff total $= N - 1$
   - ↳ F critical → using table
3. Sum of square deviations from mean
   - ↳ each groups mean → $\bar{X}_1, \bar{X}_2, \bar{X}_3$
   - ↳ Grand mean → $\bar{\bar{X}}$
   - ↳ $SS_{total} = \Sigma (x - \bar{\bar{X}})^2$
   - ↳ $SS_{within} = \Sigma (x - \bar{X})^2$
   - ↳ $SS_{between} = SS_{total} - SS_{within}$
4. Mean Square
   - ↳ $MS_{between} = \frac{SS_{between}}{df_{between}}$
   - ↳ $MS_{within} = \frac{SS_{within}}{df_{within}}$
5. Calculate F statistic
   - ↳ $F = \frac{MS_{between}}{MS_{within}}$
   - ↳ conclusion

reject H_0	accept H_0
$F_{ratio} > F_{critical}$	$F_{ratio} < F_{critical}$
$F_{ratio} > 1$	$F_{ratio} \leq 1$

(5) $F = \frac{MS_{between}}{MS_{within}} = \frac{0.12}{2.22} = 0.05 \rightarrow < 1$

$F_{critical} = 5.14$

$0.05 < 5.14$ and $F < 1$

hence accept H_0

CHP 13

13.1 Six different machines are being considered for use in manufacturing rubber seals. The machines are being compared with respect to tensile strength of the product. A random sample of four seals from each machine is used to determine whether the mean tensile strength varies from machine to machine. The following are the tensile-strength measurements in kilograms per square centimeter $\times 10^{-2}$:

Machine					
1	2	3	4	5	6
17.5	16.4	20.3	14.6	17.5	18.3
16.9	19.2	15.7	16.7	19.2	16.2
15.8	17.7	17.8	20.8	16.5	17.5
18.6	15.4	18.9	18.9	20.5	20.1

Perform the analysis of variance at the 0.05 level of significance and indicate whether or not the mean tensile strengths differ significantly for the six machines.

(1) $H_0 \rightarrow \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$

$H_1 \rightarrow$ atleast 1 μ is diffent

$\alpha = 0.05$ $df_b = 5$ $F = 2.77$

$df_w = 18$

(2) $df_b = K-1 = 5$

$df_w = N-K = 24-6 = 18$

$df_t = 23$

(3)

μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_T
17.2	17.175	18.175	17.75	18.425	18.025	17.79

(4) $SS_T = 5.5003 + 9.6404 + 11.9004 + 21.6564 + 11.0804 + 8.2084$

$= 67.9863$

$SS_W = 4.1 + 8.1275 + 11.3075 + 21.65 + 9.4675 + 7.9875$

$= 62.64$

$S_b = 67.9863 - 62.64$

$= 5.3463$

$f = \frac{1.06926}{3.48} = 0.3072 < 1$ and < 2.77

accept H_0

(5) $MS_b = \frac{5.3463}{5} = 1.06926$

$MS_W = \frac{62.64}{18} = 3.48$

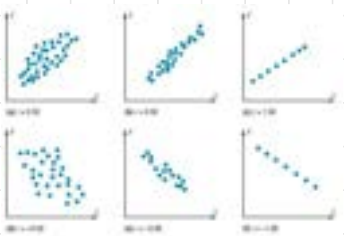
CORRELATION & REGRESSION

does a rs exist between variables

Scatter Plot

x - xais → independent variable → can be controlled

y - xais → dependent variable → cant be controlled



correlation coefficient

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

n → no of data pairs

hypothesis testing for correlation

$$t = r\sqrt{\frac{n-2}{1-r^2}} \qquad H_0: \rho = 0$$

$$df = n - 2$$

Correlation Coefficient (PPMC)

↳ if $r = 0$ then no correlation

↳ if $r \neq 0$ then correlation ($r > 0$, $n \uparrow \rightarrow y \uparrow$)

1. find r
2. state H_0, H_1
3. find t
4. find rejection region
5. If t lies in rejection region
   ↳ H_0 rejected
6. Type of rs

Type of rs

1. direct cause and effect rs (x causes y)
2. reverse cause and effect rs (y causes x)
3. caused by a third variable rs
4. curvilinear rs

Example # 05

- Test the significance of the correlation coefficient found in Example # 01.

Subject	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
A	43	126	5,418	1,849	15,876
B	49	125	6,125	2,401	15,625
C	56	156	8,736	3,136	24,336
D	59	141	8,319	3,481	19,881
E	67	141	9,447	4,489	19,881
F	70	190	13,300	4,900	36,100
$\Sigma x_i = 345$		$\Sigma y_i = 819$	$\Sigma x_i y_i = 51,345$	$\Sigma x_i^2 = 20,256$	$\Sigma y_i^2 = 132,699$

$H_0: \rho = 0 \quad H_1: \rho \neq 0$

$$r = \frac{6(47,634) - (345)(819)}{\sqrt{[6(20399) - (345)^2][6(112,443) - (819)^2]}} = 0.897$$

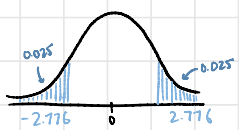
$$t = 0.897\sqrt{\frac{6-2}{1-0.897^2}} = 4.059$$

$df = 6 - 2 = 4 \quad \alpha = 0.025$ → why

t lies inside rejection region

H_0 rejected

So there is correlation



df \ α	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	318.3
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	22.33
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	10.21
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	7.173
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	5.893
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	5.208
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.785
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	4.501
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	4.297
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	4.144
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	4.025
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.930
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.852
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.787

Stopping Distances

In a study on speed control, it was found that the main reasons for regulations were to make traffic flow more efficient and to minimize the risk of danger. An area that was focused on in the study was the distance required to completely stop a vehicle at various speeds. Use the following table to answer the questions.

MPH	Breaking distance (feet)
20	39
30	83
40	91
50	133
60	269
80	411

Assume MPH is going to be used to predict stopping distance.

1. Which of the two variables is the independent variable?
2. Which is the dependent variable?
3. What type of variable is the independent variable?
4. What type of variable is the dependent variable?
5. Construct a scatter plot for the data.
6. Is there a linear relationship between the two variables?
7. Is the relationship positive or negative?
8. Can braking distance be accurately predicted from MPH?
9. List some other variables that affect braking distance.
10. Compute the value of r.
11. Is r significant at $\alpha = 0.05$?

Solve HOW

Simple Linear Regression

$y = mx + c$

↓

regression

$y = mx + c$

Example # 06

- Find the equation of the regression line for the data in Example # 01 (Slide 15), and graph the line on the scatter plot of the data.

Subject	Age, x	Pressure, y	xy	x^2	y^2
A	43	128	5,504	1,849	16,384
B	48	120	5,760	2,304	14,400
C	56	135	7,560	3,136	18,225
D	61	143	8,723	3,721	20,449
E	67	141	9,447	4,489	19,881
F	70	152	10,640	4,900	23,104
	$\Sigma x = 345$	$\Sigma y = 819$	$\Sigma xy = 47,634$	$\Sigma x^2 = 20,399$	$\Sigma y^2 = 112,443$

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n\sum x^2 - (\sum x)^2} = \frac{(819)(20,399) - (345)(47,634)}{(6)(20,399) - (345)^2} = 81.048$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{(6)(47,634) - (345)(819)}{(6)(20,399) - (345)^2} = 0.964$$

$$y' = 81.048 + 0.964x$$

sign of the correlation coefficient and the sign of the slope of the regression line

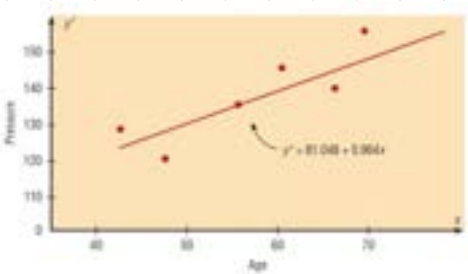
Prediction Using Regression Eq.

- Using the equation of the regression line found in Example # 06, predict the blood pressure for a person who is 50 years old.

Substituting 50 for x in the regression line $y' = 81.048 + 0.964x$ gives

$y' = 81.048 + (0.964)(50) = 129.248$ (rounded to 129)

In other words, the predicted systolic blood pressure for a 50-year-old person is 129.



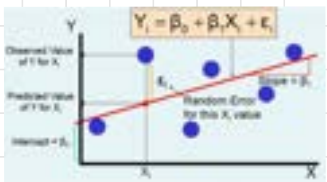
Method of Least Sq

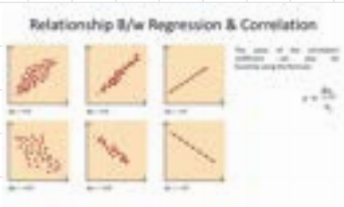
$$b_0 = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$\beta_1 = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$Y = \beta_0 + \beta_1 x + \epsilon$$ (random error)

$r^2 =$ explained variation / total variation





CHP 11

11.3 The amounts of a chemical compound y that dissolved in 100 grams of water at various temperatures x were recorded as follows:

x (°C)	y (grams)		
0	8	6	8
15	12	10	14
30	25	23	24
45	31	33	28
60	44	39	42
75	48	51	44

(a) Find the equation of the regression line.
(b) Graph the line on a scatter diagram.
(c) Estimate the amount of chemical that will dissolve in 100 grams of water at 50°C.

J
HOW

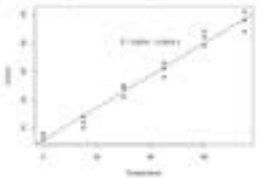
11.3 (a) $\sum x_i = 675$, $\sum y_i = 488$, $\sum x_i^2 = 37,125$, $\sum x_i y_i = 25,995$, $n = 18$. Therefore,

$$b = \frac{(18)(25,995) - (675)(488)}{(18)(37,125) - (675)^2} = 0.5676,$$

$$a = \frac{488 - (0.5676)(675)}{18} = 5.8254.$$

Hence $\hat{y} = 5.8254 + 0.5676x$

(b) The scatter plot and the regression line are shown below.



(c) For $x = 50$, $\hat{y} = 5.8254 + (0.5676)(50) = 34.205$ grams

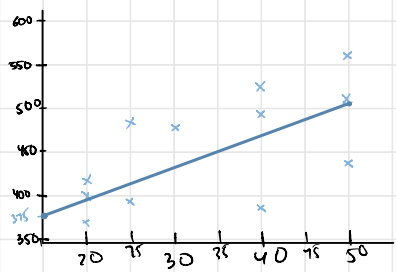
$\Sigma x = 225$ $\Sigma x^2 = 12375$ $\Sigma xy =$

$\Sigma y = 488$ Σy^2

11.9 A study was made by a retail merchant to determine the relation between weekly advertising expenditures and sales.

Advertising Costs ($)	Sales ($)
40	385
20	400
25	395
20	365
30	475
50	440
40	490
20	420
50	560
40	525
25	480
50	510

(a) Plot a scatter diagram.
(b) Find the equation of the regression line to predict weekly sales from advertising expenditures.
(c) Estimate the weekly sales when advertising costs are $35.
(d) Plot the residuals versus advertising costs. Comment.



b) $\Sigma x = 410$ $\Sigma x^2 = 15,650$ $\Sigma xy = 191325$

$\Sigma y = 5445$ $n = 12$

$$B_0 = \frac{(5445)(15650) - (410)(191325)}{12(15650) - 410^2} = 343.7056$$

$$B_1 = \frac{12(191325) - (410)(5445)}{12(15650) - 410^2} = 3.22081$$

$y = 343.7056 + 3.22081x$

c) $y = 343.7056 + 3.22081(35)$

$y = 456.434$

CHP 11

11.43 Compute and interpret the correlation coefficient for the following grades of 6 students selected at random:

Mathematics grade	70	92	80	74	65	83
English grade	74	84	63	87	78	90

x	x^2	y	y^2	xy
464	36354	476	38254	36926

$$r = \frac{6(36926) - (464)(476)}{\sqrt{[6(36354) - (464)^2][6(38254) - (476)^2]}}$$

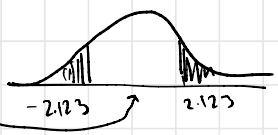
$r = 0.2396$

11.46 Test the hypothesis that $\rho = 0$ in Exercise 11.43 against the alternative that $\rho \neq 0$. Use a 0.05 level of significance.

$H_0: P = 0$ $\quad \alpha = 0.05$ $\quad n = 6$

$H_1: P \neq 0$ $\quad r = 0.2396$

$$t_c = 0.2396\sqrt{\frac{6-2}{1-0.2396^2}} = 0.4935$$



$df = 6 - 2 = 4$, t from table $= 2.123 \rightarrow$ rejection region

DONT REJECT H_0

11.45 With reference to Exercise 11.13 on page 400, and assume a bivariate normal distribution for x and y.

(a) Calculate r.

(b) Test the null hypothesis that $\rho = -0.5$ against the alternative that $\rho < -0.5$ at the 0.025 level of significance.

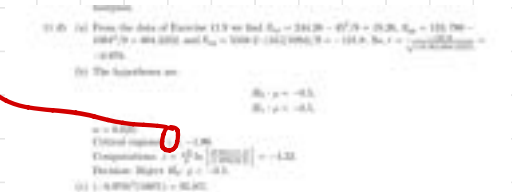
(c) Determine the percentage of the variation in the amount of particulate removed that is due to changes in the daily amount of rainfall.

Daily Rainfall, x (0.01 cm)	Particulate Removed, y ($\mu g/m^3$)
4.3	126
4.5	121
5.9	116
5.6	118
6.1	114
5.2	118
3.8	132
2.1	141
7.5	108

a) ... $r = -0.919$

$$t = -0.919\sqrt{\frac{9-2}{1-(0.919)^2}} = -12.71$$

WHY Z test



ishma hafeez notes

repshtree

BAYES RULE

$$P(A|B) = \frac{P(B|A_i)\,P(A_i)}{\sum_{i=1}^{k} P(B|A_i)P(A_i)}$$

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad P(B|A) = \frac{P(A \cap B)}{P(A)}$$

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

$$IQR = Q_3 - Q_1$$

$$L = Q_1 - 1.5(IQR)$$

$$U = Q_3 + 1.5(IQR)$$

Probability mass function (PMF)

1. $f(n) \geq 0$, for all $n \in R$
   - $\sum f(n)\,dn = 1$ → verify PMF
   - $P(X=x) = F(x) - F(x-1)$
   - $P(a < x < b) = F(b) - F(a)$

$P(n) = (n+1)^n$

Joint Probability Distribution

1. $f(n,y) \geq 0$
2. $\sum_{x} \sum_{y} f(n,y) = 1$
3. $P(X=n, Y=y) = f(n,y)$
   - $P[(X,Y) \in A] = \sum\sum_{A} f(n,y)$

σ^2 → variance

$$V(X) = E(X^2) - [E(X)]^2$$

mean

$$\mu = E(X) = \sum x f(x)$$

$$V(Y) = E(Y^2) - [E(Y)]^2$$

$$E(XY) = \sum\sum xy\, f(x,y)$$

$$\text{Covariance}(X,Y) = E(XY) - E(X)\,E(Y)$$

$$\text{Correlation}(X,Y) = \frac{\text{Covariance}(X,Y)}{\sqrt{V(X)}\,\sqrt{V(Y)}}$$

± 1 ⟶ strong
± 0.5 ⟶ moderate
0 ⟶ negligible

Emperical Rule

68% $\bar{x} \pm s$

95% $\bar{x} \pm 2s$

99.7% $\bar{x} \pm 3s$

Standard Normal Distribution

$$z\text{-score} = \frac{x-\mu}{\sigma}$$

Finite Population Correction Factor

$$\sqrt{\frac{N-n}{N-1}}$$

(N: population size; n: sample size)

Centeral limit theorem

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

(z test; $\sigma\sqrt{n}$: standard error of mean)

1. $\bar{x} = \frac{\sum x}{n}$ ($\bar{x}$: sample mean; n: sample size)
2. $s = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}$ (s: S.D)
3. $s_{\bar{x}} = \frac{s}{\sqrt{n}}$ ($s_{\bar{x}}$: standard error mean; s: S.D)
4. $E = z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$ (E: margin of error; $z_{\frac{\alpha}{2}}$: critical region of z-distribution; $\frac{\sigma}{\sqrt{n}}$: standard error of mean)
5. $n = \left(\frac{z_{\frac{\alpha}{2}}\,\sigma}{E}\right)^2$

Confidence Interval for Proportions

1. $\hat{P} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{P}\hat{q}}{n}}$ ($\hat{P}$: sample proportion; $\frac{x}{n}$: no of sample units; $\hat{q} = 1-\hat{P}$)

Chi Square Distribution

1. $\frac{(n-1)s^2}{\chi^2_{right}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{left}}$
2. $df = n-1$

Confidence Interval for μ (1 distribution)

2. $\bar{x} \pm z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$ (s: sample S.D)
   - ↳ σ unknown
   - ↳ $n \geq 30$

3. $\bar{x} \pm t_{\frac{\alpha}{2(v)}}\frac{s}{\sqrt{n}}$ ($t_{\frac{\alpha}{2(v)}}$: critical value of t distribution, $n-1$; s: sample S.D)
   - ↳ σ unknown
   - ↳ $n < 30$

Confidence Interval for $\mu_1 - \mu_2$ (2 distributions)

2. $(\bar{x}_1 - \bar{x}_2) \pm z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}}$
   - ↳ σ_1, σ_2 unknown
   - ↳ $n \geq 30$

3. $(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}}$
   - ↳ σ_1, σ_2 unknown
   - ↳ $n < 30$

Z-Test

↳ $n \geq 30$

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

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

(sample mean; H_0 mean; standard deviation; sample size)

$$\delta = \sqrt{\frac{n \Sigma x^2 - (\Sigma x)^2}{n(n-1)}}$$

T-Test for mean

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

$df = n - 1$

↳ $\sigma_1 \neq \sigma_2$ (variance)

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

(sample mean; H_0 mean; standard deviation; sample size)

$$df = \frac{(\delta_1 + \delta_2)^2}{\frac{\delta_1^2}{n-1} + \frac{\delta_2^2}{n-1}}$$

↳ $\sigma_1 = \sigma_2$ (variance)

$$\delta = \frac{(n-1)(\delta_1)^2 + (n-1)(\delta_2)^2}{(n+n) - 2}$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\delta \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\delta = \sqrt{\frac{n \Sigma x^2 - (\Sigma x)^2}{n(n-1)}}$$

1. State H_0, H_a, α
2. Degree's of freedom
   - ↳ df between groups = $K - 1$
   - ↳ df within groups = $N - K$ (no. of groups; total no. of observations)
   - ↳ diff total = $N - 1$
   - ↳ F critical → using table (a; b)
3. Sum of square deviations from mean
   - ↳ each groups mean → $\bar{x}_1, \bar{x}_2, \bar{x}_3$
   - ↳ Grand mean → $\bar{\bar{x}}$
   - ↳ $SS_{total} = \Sigma (x - \bar{\bar{x}})^2$
   - ↳ $SS_{within} = \Sigma (x - \bar{x})^2$
   - ↳ $SS_{between} = SS_{total} - SS_{within}$
4. Mean Square
   - ↳ $MS_{between} = \frac{SS_{between}}{df_{between}}$
   - ↳ $MS_{within} = \frac{SS_{within}}{df_{within}}$
5. Calculate F statistic
   - ↳ $F = \frac{MS_{between}}{MS_{within}}$
   - ↳ conclusion

reject H_0	accept H_0
$F_{ratio} > F_{critical}$	$F_{ratio} < F_{critical}$
$F_{ratio} > 1$	$F_{ratio} \leq 1$

Correlation Coefficient (PPMC)

(correlation coefficient; no. of data pairs)

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

(Hypothesis testing for correlation)

$$t = r\sqrt{\frac{n-2}{1-r^2}} \qquad H_0 : \rho = 0$$

$df = n - 2$

Simple Linear Regression

Method of least sq

$$B_0 = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$B_1 = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$Y = \beta_0 + \beta_1 x + \epsilon$$

(ϵ: random error)