Statistics

Exercise 1

1) Stem-and-leaf Display

62, 65, 68, 70, 73, 75, 75, 78, 81, 83, 84, 85, 87, 89, 92, 95, 96, 98, 100

stem	leaf
6	2 5 8
7	0 3 5 5 8
8	1 3 4 5 7 9
9	2 5 6 8
10	0

2) Box plot

| 55 60 62 63 65 66 68 70 72 75 77 78 80 85 88 |

a) 5-number summary

lower (inner) fence → 64 − 20.25 = 43.75

upper (inner) fence → 77.5 + 20.25 = 97.75

minimum → 55

25th quartile → 64

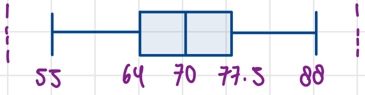
50th quartile → 70

75th quartile → 77.5

maximum → 88

b) IQR → 77.5 − 64 = 13.5

13.5 x 1.5 = 20.25



no outliers

Exercise 2

1) Trimean

10, 12, 15, 18, 21, 24, 27, 30, 33, 36, 39, 42, 45, 48, 50

$Q_1 = 18$ Trimean $= (Q_1 + 2Q_2 + Q_3)/4$

$Q_2 = 30$ $= (18 + 2\cdot30 + 42)/4$

$Q_3 = 42$ $= 128/4 = 32$

2) Geometric mean

+5%, +10%, −3%, +6%

$$[(1+0.05)(1+0.1)(1-0.03)(1+0.06)]^{1/4} - 1$$

$= 1.0004 - 1 = 0.00045$

$100 \times 0.0004 = 0.045 = 4.5\%$

3) Trimmed Mean

10% trim [65, 70, 72, 75, 80, 85, 90, 92, 95, 100]

$n = 10$ $10\% \times 10 = 1$

$$\frac{70+72+75+80+85+90+92+95}{8} = 82\,375$$

Exercise 3

1) 8 beople, 4 in a row

order matters, abcd, diff from bacd

$_8P_4 = 1680$

2) 7 books, 4 taken

order ≠ matter

$_7C_4 = 35$

3) 10 red, 15 blue, select 5 random (no replace) 3 balls red

$10 + 15 = 25$

$_{25}C_5 = 53130$

$_{10}C_3 = 120$

$_{15}C_2 = 105$

$120 \times 105 = 12600$

$P = \frac{12600}{53130} \approx 0.237$

$\approx 23.7\%$

Exercise 4

1) percentage returns

10%, 15%, −5%, 8%, 12%

geometric mean → $[(1+0.1)(1+0.15)(1-0.05)(1+0.08)(1+0.12)]^{1/5} - 1$

$= 0.00079976$

$0.00079976 \times 100 = 0.079976$ → 7.998%

2) Box Plots

A: 7, 9, 12, 13, 14, 15, 16

B: 5, 7, 8, 10, 12, 15, 18

a) A: q_1 → 9 min → 7

q_2 → 13 max → 16

q_3 → 15

IQR → $15 - 9 = 6$

$6 \times 1.5 = 9$

lower inner → $9 - 9 = 0$

upper inner → $15 + 9 = 24$

B: q_1 → 7 min → 5

q_2 → 10 max → 18

q_3 → 15

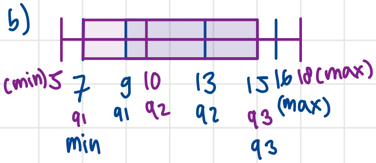
IQR → $15 - 7 = 8$

$8 \times 1.5 = 12$

lower inner → $7 - 12 = -5$

upper inner → 27

b)



c) group A has a higher median

there are no outliers

3) probability

card is drawn from 52 cards, coin is flipped

probability for 'king' and 'tail'

king → $\frac{4}{52}$

tail → $\frac{1}{2}$

'king' and 'tail' = $\frac{\cancel{4}^{2}}{52} \times \frac{1}{\cancel{2}} = \frac{2}{52} = \frac{1}{26}$

4) Stem and leaf

leaf X	stem	leaf Y
2 4 7 9	1	3 6 8
1 4 6 8	2	0 3 5 7 9
0 2	3	1 3

5) probability

3 heads exactly when flip coin 5 times

$2^5 = 32$ diffrent outcomes

$5C3 = 10$

$10/32 = 5/16$

6) probability (binomial distribution)

success rate of 80%, 15 free throws, at least12 successful?

$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

k number of successes in n trials where p is probability

atleast 12 → $X \geq 12$

$\left(\frac{8}{10}\right)^{12} \cdot \left(\frac{2}{10}\right)^{3} = 0.25014$

$\left(\frac{8}{10}\right)^{13} \cdot \left(\frac{2}{10}\right)^{2} = 0.2309$

$\left(\frac{8}{10}\right)^{14} \cdot \left(\frac{2}{10}\right) = 0.13194$

$\left(\frac{8}{10}\right)^{15} = 0.03518$ +

0.64016 → 64.0%

7) Pearson correlation

	x	y	x^2	y^2	xy
	2	10	4	100	20
	4	15	16	225	60
	6	20	36	400	120
	8	25	64	625	200
	10	30	100	900	300
sum	30	100	220	2250	700

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

$$= \frac{5(700) - (30)(100)}{\sqrt{[5\cdot 220 - 30^2][5\cdot 2250 - (100)^2]}}$$

$= 1$ ← perfect correlation

p-value $< .00001$ ← statistically significant

therefore reject H_0.

H_0 = no correlation between hours of sunlight and plant height ($r = 0$)

H_1 = there is a correlation between hours of sunlight and plant height ($r \neq 0$)

There is a statistically significant correlation between hours of sunlight and plant height

df (pearson) $= n - 2$

$= 5$ (data points) $- 2 = 3$

$r(3) = 1, < .001$

Exercise 5

1) Standard Deviation [70, 85, 78, 90, 80]

$N : 5$

$\sum x = 411$

$M = 82.2$

$\sigma^2 = 53.76$

$\sigma = 7.332$

2) probability

30% prefer coffee over tea, select 100 people, fewer than 25 people prefer coffee

$q = 1 - p = 0.7$

$\mu = n \cdot p = 100 \times 0.3 = 30$

$\sigma = \sqrt{n \cdot p \cdot q}$

$= \sqrt{30 \cdot 0.7} = \sqrt{21} \approx 4.583$

fewer than 25 → $X < 25$

$\leq a$ (at most a) $a + 0.5$

$< a$ (less than a) $a - 0.5$

$\geq a$ (at least a) $a - 0.5$

$> a$ (more than a) $a + 0.5$

use $P(X < 24.5)$ for continuity correction

$$z = \frac{X - \mu}{\sigma} = \frac{24.5 - 30}{4.583} \approx -1.2 \rightarrow 0.11507 \rightarrow 11.5\%$$

3) Probability

$n = 100 \quad p = 0.4$ 45 sucesssres

$q = 1 - p = 0.6$

$\mu = n.p = 100 \times 0.4 = 40$

$\sigma = \sqrt{n.p.q}$

$= \sqrt{40 . 0.6} = \sqrt{24} \approx 4.9$

at least 25 → $X \geq 45$

use $P(X \geq 44.5)$

$$Z = \frac{X - \mu}{\sigma} = \frac{44.5 - 40}{4.9} = 0.918 \rightarrow 0.821$$

$P(X \geq 44.5) = 1 - 0.821 = 0.179$

$= 17.9\%$

P value → 0.000009378

reject H_0

training program significanty reduced weight.

Exercise 6

1) T-test

$\mu = 1000$ → two-tailed

950, 960, 970, 980, 1020, 1030, 990, 1010, 1000, 995

H_0 = the mean lifespan of the bulbs is 1000 hours

H_1 = the man lifespan differs significantly from 1000 hours

mean : 990.5

SD . 24.54

N : 10

t-score = −1,22 (t-statistic)

df = N − 1 = 9

± 2 262 (t-critical)

−1,22 falls in range of −2.262 and 2.262

cannot reject H_0, tuerefore the mean lifespan of the bulbs is 1000 hours.

2)

client	before	after	difference
1	85	82	−3
2	78	75	−3
3	90	85	−5
4	76	74	−2
5	88	85	−3
6	81	78	−3
7	79	76	−3
8	92	89	−3

H_0 = training program doesn't significantly reduce weight

H_1 = training program significantly reduces weight

paired t-test

mean = −3.125

SD = 0.78006

N = 8

t-statistic = 11,323

df = 7

t-critical → ± 2,365

11,323 doesn't fall in range, reject H_0, accept H_1

3) test if new diet plan (A) significantly improves weight loss compared to standard diet plan (B) - Independent t-test

group	sample size (n)	mean weight loss (x)	standard deviation (s)
A	25	8 kg	2
B	25	6 kg	2.5

H_0 = A doesn't significantly improve weight loss compared to B

H_1 = A significantly improves weight loss compared to B

t-statistic → 3,1235

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}$$

$df = n_1 + n_2 - 2$ (equal variances)

$= 47$

t-critical ± 2,012

reject H_0, accept H_1

Exercise 7

1) one-way ANOVA

fertilizer A	fertilizer B	fertilizer C
15	20	23
16	22	27
14	19	26
15	21	28
17	20	24

H_0 = the type of fertilizer doesn't significantly affect plant growth

H_1 = the type of fertilizer significantly affects plant growth

Groups	N	Σx	mean	Σx^2	sd
A	5	77	15.4	1191	1.14
B	5	102	20.4	2086	1.14
C	5	130	26	3390	1.58
total	15	309	20.6	6667	

source	df	SS	MS	F statistic
between groups	2	281.2	140.6	82.706
within groups	12	20.4	1.7	
total	14	301.6		

for degrees of freedom 2 & 12 with $\alpha = 0.05$ critical F-value is 3.885

F test > critical F-value (82.706 > 3.885) reject H_0, accept H_1

2) Chi-Square Test

Fertilizer	plant A	plant B	plant C	total
X	10 (13.33) [0.83]	20 (15.56) [1.27]	10 (11.11) [0.11]	40
Y	15 (10.00) [2.50]	10 (11.67) [0.24]	5 (8.33) [1.33]	30
Z	5 (6.67) [0.42]	5 (7.78) [0.99]	10 (5.56) [3.56]	20
total	30	35	25	90

H_0 = the two groups are independent

H_1 = the two groups are not independent

p-value is 0.023 > 0.05, reject H_0, accept H_1

3) Two-way ANOVA

Language	self-study	Instructor-led
python	78, 82, 85	90, 88, 92
Java	72, 75, 74	85, 80, 84
C++	65, 68, 70	78, 75, 80

H_01 = the mean test scores across all programing languages are the same

H_02 = the mean test scores across all study methods are the same

H_03 = there is no interaction between programming language and study method on test scores

Source	SS	df	MS	F	P
rows	523.44	2	261.72	40.96	< .0001 → reject
columns	382.72	1	382.72	59.9	< .0001 → reject
r x c	2.11	2	1.06	0.17	0.8457 → accept
error	76.67	12	6.39		
total	984.94	17			

$F(2, 12) = 3.89$

$F(1, 12) = 4.74$

$40.96 > 3.89$, reject H_01

$59.9 > 4.74$, reject H_02

$0.17 < 3.89$, accept H_03