Statstical Inference

Home Work no.3

Hadiseh mesbah

810102253

Problem 1:

Part one:

$H_0: M = 50$ $\quad$ $H_A: m > 50$

$d_1 = |43-50| = 7$ $\quad$ $d_2 = |47-50| = 3$ $\quad$ $d_3 = |52-50| = 2$ $\quad$ $d_4 = |68-50| = 18$

$d_5 = |72-50| = 12$ $\quad$ $d_6 = |55-50| = 5$ $\quad$ $d_7 = |61-50| = 11$ $\quad$ $d_8 = |44-50| = 6$

$d_9 = |58-50| = 8$ $\quad$ $d_{10} = |63-50| = 13$ $\quad$ $d_{11} = |54-50| = 4$ $\quad$ $d_{12} = |57-50| = 7$

$d_{13} = |77-50| = 17$ $\quad$ $d_{14} = |36-50| = 14$ $\quad$ $d_{15} = |80-50| = 30$ $\quad$ $d_{16} = |53-50| = 3$

$d_{17} = |60-50| = 10$

$R_1 = 7$ $\quad$ $R_2 = 2.5$ $\quad$ $R_3 = 1$ $\quad$ $R_4 = 14$ $\quad$ $R_5 = 15$ $\quad$ $R_6 = 5$ $\quad$ $R_7 = 11$

$R_8 = 6$ $\quad$ $R_9 = 8$ $\quad$ $R_{10} = 12$ $\quad$ $R_{11} = 4$ $\quad$ $R_{12} = 7$ $\quad$ $R_{13} = 16$ $\quad$ $R_{14} = 13$ $\quad$ $R_{15} = 17$

$R_{16} = 5.2$ $\quad$ $R_{17} = 10$

$$W = \sum_{i: X_i > M} R_i = R_3 + R_4 + R_5 + R_6 + R_7 + R_9 + R_{10} + R_{11} + R_{12} + R_{13} + R_{15} + R_{16} + R_{17} = 124.5$$

$$W \sim \text{normal}\left(\frac{17 \times 18}{4}, \sqrt{\frac{17 \times 18 \times 35}{24}}\right) = N(76.5, 21.12)$$

$$P\_value = P(W \geq 124.5) = P\left(Z > \frac{124.5 - 76.5}{21.12}\right) = P(Z > 2.27)$$

$$= 1 - 0.988$$

$$0.011 < 0.05$$

So reject H_0, we know that the median salary is above 50 \$

Part two:

For use the Mann-whitney-wilcoxon Rank sum Test, need another group or data to compare these together. there is not another dataset so, we can't.

Problem two:

Part one:

1 based on data, It looks like, on average, more words reported if they are presented to the right ear. How ever it's not a big difference, and not all participants show it.

Part Two:

Rank

Participant	Left Ear	Right Ear
1	25 7	32 19.5
2	29 11.5	30 14
3	10 3	7 2
4	31 16	36 24
5	27 9.5	20 4.5
6	24 6	32 19.5
7	27 9.5	26 8
8	29 11.5	33 23
9	30 14	32 19.5
10	32 19.5	32 19.5
11	20 4.5	30 14
12	5 1	32 19.5

$weight_{left} = 113$

$weight_{right} = 187$

$$U_{left} = mn + \frac{n(n+1)}{2} - W_{left} = 12\times12 + \frac{12\times13}{2} - 113 = 109$$

$$U_{right} = mn + \frac{m(m+1)}{2} - W_{right} = 12\times12 + \frac{12\times13}{2} - 187 = 35$$

$U_{right} = 35$

$U_{left} = 109$

$$\mu_{U_i} = \frac{mn}{2} = 72 \qquad \sigma_{U_i} = \sqrt{\frac{mn(m+n+1)}{12}} = 17.32$$

Because of number of m and n, we can use Normal Distirbuin

$n, m \geq 10$

$$z = \frac{35-72}{17.32} = -2.13 \qquad 1 - P(z < 2.13) = 0.98 = 0.016$$

P_value $= 1 - 0.016 > 0.05$ $\longrightarrow$ there is a significant defference between the group.

Problem Four :

Part 3:

$$e_{ij} = P_{ij} \times f = (P_i, P_j) f = \left(\frac{f_i}{f} \times \frac{f_j}{f}\right) f = \frac{f_i \times f_j}{f}$$

$$f = \sum_{i=1}^{3} \sum_{j=1}^{3} f_{ij} = 300$$

$e_{11} = \left(\frac{99 \times 156}{300}\right) = 51.4$ $\quad e_{12} = \left(\frac{99 \times 84}{300}\right) = 27.7$ $\quad e_{31} = \left(\frac{99 \times 60}{300}\right) = 19.8$

$e_{21} = \left(\frac{94 \times 156}{300}\right) = 48.8$ $\quad e_{22} = \left(\frac{94 \times 84}{300}\right) = 26.3$ $\quad e_{31} = \left(\frac{94 \times 60}{300}\right) = 18.8$

$e_{31} = \left(\frac{107 \times 156}{300}\right) = 55.6$ $\quad e_{32} = \left(\frac{107 \times 84}{300}\right)^2 = 29.9$ $\quad e_{33} = \left(\frac{107 \times 60}{300}\right) = 21.4$

$$\chi = \sum_{i=1}^{3} \sum_{j=1}^{3} \frac{(e_{ij} - f_{ij})^2}{e_{ij}}$$

As. $\chi^2 = 10.57$ more than $\chi^2_{0.05,4} = 9.49$, we can reject H_0

H_0 is about independent of 2 var

Problem Five:

Using χ^2 test

	No moustache	Wears a moustache	
Between 18 and 30	28	12	40
Over 30	52	8	60
	80	20	100

$$\chi^2 = \frac{\Sigma (o_{ij} - E_{ij})^2}{E_{ij}} \qquad E_{ij} = \frac{F_i F_j}{\text{total}}$$

$$E_{11} = \frac{40 \times 80}{100} = 32 \quad E_{12} = \frac{40 \times 20}{100} = 8 \quad E_{21} = \frac{80 \times 60}{100} = 48 \quad E_{22} = \frac{60 \times 20}{100} = 12$$

$$\chi^2 = \frac{(28-32)^2}{32} + \frac{(12-8)^2}{8} + \frac{(52-48)^2}{48} + \frac{(8-12)^2}{12} = 4{,}166$$

$df = (\text{number of rows} - 1)(\text{number of columns} - 1) = (2-1)(2-1) = 1$

$\rightarrow$ Critical value : 3,841 $\quad 4{,}166 > 3.841 \quad H_0$ be rejected

Problem Six:

	O	A	B	AB	
Rh positive	82	89	54	19	244
Rh negative	13	27	7	9	56
	95	116	61	28	300

$$E_{11} = \frac{244 \times 95}{300} = 77.27 \quad E_{12} = \frac{244 \times 116}{300} = 94.35 \quad E_{13} = \frac{244 \times 61}{300} = 49{,}61 \quad E_{14} = \frac{244 \times 28}{300} = 27{,}7$$

$$E_{21} = \frac{56 \times 95}{300} = 17.74 \quad E_{22} = \frac{56 \times 116}{300} = 21{,}26 \quad E_{23} = \frac{56 \times 61}{300} = 11{,}4 \quad E_{24} = \frac{56 \times 28}{300} = 5{,}2$$

$$df = (2-1)(4-1) = 3$$

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{4} \left(\frac{f_{ij} - e_{ij}}{e_{ij}}\right)^2 = 11{,}8 > 7{,}81 \longrightarrow \text{reject the } H_0$$

Problem 7:

Part one: Based on Empirical cdf $\longrightarrow F_n(y_i) = \frac{i}{n}$, $n \longrightarrow 5$

then:

$F_n(y_1) = \frac{1}{5} = 0{,}2 \quad F_n(y_2) = \frac{2}{5} = 0{,}4 \quad F_n(y_3) = \frac{3}{5} = 0{,}6 \quad F_n(y_4) = \frac{4}{5} = 0{,}8$

$F_n(y_5) = \frac{5}{5} = 1$

Thus $D_n = 0{,}1$

Part Two:

$F(y_1) \leq 0{,}2 \quad F_n(y_1) = 0{,}2 \quad D_n = |F(y_1) - F_n(y_1)| \quad D_n \leq 0{,}2$

$0{,}2 \leq F(y_2) \leq 0{,}4 \quad F_n(y_2) = 0{,}4 \quad D_n \leq 0{,}2$

$0{,}4 \leq F(y_3) \leq 0{,}6 \quad F_n(y_3) = 0{,}6 \quad D_n \leq 0{,}2$

$0{,}6 \leq F(y_4) \leq 0{,}8 \quad F_n(y_4) = 0{,}8 \quad D_n \leq 0{,}2$

max

$0{,}8 \leq F(y_5) \leq 1 \quad F_n(y_5) = 1 \quad D_n \leq 0{,}2$

$\longrightarrow D_n \leq 0{,}2$

Problem eight:

i	$y_i = F(y_i)$	$F_n(y_i)$
1	0,01	0,04
2	0,06	0,08
3	0,08	0,11
4	0,09	0,16
5	0,11	0,20
6	0,16	0,24
7	0,22	0,28
8	0,23	0,32
9	0,29	0,36
10	0,30	0,40
11	0,35	0,44
12	0,38	0,48
13	0,40	0,52
14	0,41	0,56
15	0,42	0,60
16	0,48	0,64
17	0,57	0,68
18	0,66	0,72
19	0,71	0,76
20	0,75	0,80
21	0,78	0,84
22	0,79	0,86
23	0,82	0,92
24	0,88	0,96
25	0,90	1

the max value of $|F_n(x) - F(x)|$ is at $x = y_{15}$ where it's value is $0,6 - 0,42 = 0,18$

, since $n = 25$, $n^{\frac{1}{2}} D_n^* = 0,9$

$\rightarrow H(0,9) = 0,6073$, Therefore tail area corresponding to the observed value of D_n^* is

$1 - 0,6073 = 0,3927$

Part 2:

i	y_i	$F(y_i)$	$F_n(y_i)$
1	0,01	0,015	0,04
2	0,06	0,09	0,08
3	0,08	0,12	0,12
4	0,09	0,135	0,16
5	0,11	0,165	0,2
6	0,16	0,24	0,24
7	0,22	0,33	0,28
8	0,23	0,345	0,32
9	0,29	0,435	0,36
10	0,3	0,45	0,4
11	0,35	0,525	0,44
12	0,38	0,57	0,48
13	0,4	0,6	0,52
14	0,41	0,615	0,56
15	0,42	0,63	0,60
16	0,48	0,72	0,64
17	0,57	0,785	0,68
18	0,66	0,83	0,72
19	0,71	0,855	0,76
20	0,75	0,875	0,80
21	0,78	0,89	0,84
22	0,79	0,895	0,88
23	0,82	0,91	0,92
24	0,88	0,94	0,96
25	0,9	0,95	1

Supremum of $|F_n(x) - F(x)|$ occurs as $x \rightarrow y_{18}$

$F(x) \rightarrow 0,83$ while F_n remains at $0,68$

$D_n^* = 0,83 - 0,68 = 0,15 \longrightarrow n^{\frac{1}{2}} D_n^*$ is $1 - 0,3728 = 0,6272$

Part 3:

Since PdF of uniform distribution is equal to 1, the value of the joint PdF of 25 obsrv. under uniform distribution has the value $L_1 = 1$, and 16 of obsrv. are less than $\frac{1}{2}$ and 9 are greater than $\frac{1}{2}$. therefore, the value of joint pdf of obsrv. under other distribution is $L_2 = \left(\frac{3}{2}\right)^{16} \left(\frac{1}{2}\right)^{9} = 1,2829$.

the probibility $\longrightarrow \dfrac{\frac{1}{2} L_1}{\frac{1}{2} L_2 + \frac{1}{2} L_1} = 0,438$

Problem nine:

x_i	y_i	$F_m(x)$	$G_n(x)$
-2.47		0,04	0
-1.73		0,08	0
-1.28		0,12	0
-0,82		0,16	0
-0,74		0,2	0
	-0,71	0,2	0,05
-0,56		0,24	0,05
-0,4		0,28	0,05
-0,39		0,32	0,05
	-0,37	0,32	0,1
-0,32		0,36	0,1
	-0,3	0,36	0,15
	-0,27	0,36	0,2
-0,06		0,4	0,2
	0	0,4	0,25
0,05		0,44	0,25
0,06		0,48	0,25
	0,26	0,48	0,3
0,29		0,52	0,3
0,31		0,56	0,3
	0,36	0,56	0,35
	0,38	0,56	0,4

x_i	y_i	$F_m(x)$	$G_n(x)$
	0,44	0,56	0,45
0,51		0,6	0,45
	0,52	0,6	0,5
0,57		0,64	0,5
0,61		0,68	0,5
0,64		0,72	0,5
	0,66	0,72	0,55
	0,7	0,72	0,6
	0,76	0,72	0,65
1,05		0,76	0,65
1,06		0,8	0,65
1,09		0,84	0,65
1,31		0,88	0,65
	1,38	0,88	0,7
	1,5	0,88	0,75
	1,56	0,88	0,8
1,64		0,92	0,8
	1,66	0,92	0,85
1,77		0,96	0,85
	2,2	0,96	0,90
	2,31	0,96	0,95
2,36		1	0,95
	3.29	1	1

Part 1:

shall donate 25 ordered observ in first sample by $x_1 < \dots < x_{25}$ and next sample by $y_1 < \dots < y_{20}$. in above table max value $|F_m(x) - G_n(x)|$

$x = -0,39 \longrightarrow 0,32 - 0,05 = 0,27$

$\longrightarrow D_{mn} = 0,27$ since $m = 25$ $n = 20$ $\left(\frac{mn}{m+n}\right)^{\frac{1}{2}} D_{mn} = 0,9$

$H(0,9) = 0,6073$ Hence, tail area corresponding to the observed value of D_{mn} is $1 - 0,6073 = 0,3927$

Part 2:

Shall add 2 units to each of the value in the first sample and then carry out same procedure as privous part.

maximum value of $|F_n(x) - G_n(x)|$ when $x = 1,56$

$$0,8 - 0,241 = 0,56$$

Thus. $D_{mn} = 0,56$ and $\left(\frac{mn}{(m+n)}\right)^{\frac{1}{2}} D_m = 1,8667$

$H(1.8667) = 0,998$

therefore, the tail area corresponding to the observed value of D_{mn} is $1 - 0,998 = 0,002$

x_i	y_i	$F_m(x)$	$G_n(x)$
	-0,71	0	0,05
-0,47		0,04	0,05
	-0,37	0,04	0,1
	0,3	0,04	0,15
	-0,27	0,04	0,2
	0	0,04	0,25
	0,26	0,04	0,3
0,27		0,08	0,3
	0,36	0,08	0,35
	0,38	0,08	0,4
	0,44	0,08	0,45
	0,52	0,08	0,5
	0,66	0,08	0,55
	0,7	0,08	0,6
0,72		0,12	0,6
	0,96	0,12	0,65
1,18		0,16	0,65
1,26		0,2	0,65
	1,38	0,2	0,7
1,44		0,24	0,7
	1,50	0,24	0,75
	1,56	0,24	0,8
1,6		0,28	0,8
1,61		,32	,80
	1,66	,32	,85
1,68		,36	,85
1,94		,4	,85
2,05		,44	,85
2,6		,48	,85
	2,2	,48	,9
2,29		,52	,9
2,31		,56	,95
2,81		,6	,95

x_i	y_i	$F_m(x)$	$G_n(x)$
2.59		,64	,95
2.61		,68	,95
2.64		,72	,95
3,05		,76	,95
3,06		,8	,95
3,09		,84	,95
	3,29	,84	1
3,31		,88	1
3,64		,92	1
3.77		,96	1
4.36		1	1

Part 3:

Shall multiply each of the obser. in second sample by 3 and then carry out same procedure in part 1.

the max value of $|F_m(m) - G_n(m)|$

$x = 1,06 \longrightarrow 0,8 - 0,3 = 0,5$ $D_{mn} = 0,56$ $\left(\frac{mn}{m+n}\right)^{\frac{1}{2}} D_{mn} = 1,8667$

$H(1,667) = 0,992$

therefore the tail area corresponding to the observed value of D_{mn} is

$1 - 0,992 = 0,008$

x_i	y_i	$F_m(m)$	$G_n(m)$
-2.97		0,04	0
	-2.13	,0.4	,0.5
-1,73		0,08	,05
-1,28		,12	0,05
	-1,11	,12	0,1
	-0,9	,12	0,15
-0,82		,16	0,15
	-0,81	,16	0,2
-,74		,2	0,2
-0,56		,24	0,2
-0,4		,28	0,2
-0,37		,32	0,2
-0,32		,36	0,2
-0,06		,4	0,2
	0	,4	0,25
0,03		,44	0,25
0,06		,48	0,25
0,26		,52	0,25
0,31		,56	0,25
0,51		,6	0,25
0,59		,64	0,25
0,61		,68	0,25
0,64		,72	0,25
	0,78	0,72	0,3
1,05		0,76	0,3
1,06		0,8	0,3
	1,08	0,8	,35
1,07		0,84	,35
	1,14	0,84	,4
1,31		0,86	,4
	1,32	0,88	,45
	1,56	0,88	,5

x_i	y_i	$F_m(m)$	$G_n(m)$
1,64		,92	,5
1,77		,96	,5
	1,98	,96	,5
	2,1	,96	,55
2,36		1	,6
	2,88	1	,6
	4,14	1	,65
	4.5	1	,7
	4,68	1	,75
	4,98	1	,8
	6,6	1	,85
			,90
	6,98	1	,95
	9,87	1	1

Problem eleven:

a) true

بله ممکن است هر آزمونی حتی با درصد بالا هم به دلیل شانس اشتباه باشه. چون این آزمون های آماری صد درصد درست نیستند مثلا اگر $\alpha=0.05$ باشه بازهم درصدی احتمال خطا داریم و همیشه احتمال خطای نوع یک هست.

b) False

یک عدد بزرگ به معنی significant بودن نیست. اهمیت اصلی مربوط به حجم نمونه، effect size، متغیر بودن داده‌ها است. مقداری از داده‌ها ممکن است به دلیل خطای نمونه یا تغییرات تصادفی ... بیشتر و یا ارتباطی ذاتی را منعکس کنند پس موارد دیگه هم هستند.

c) True

برای مقایسه‌ی دو مقدار P-value با مقدار α توجه می‌کنیم که آستانه‌ی رد فرض صفر هستند. اگر α برابر با 0.05 باشد، هر دو مقدار P-value بالاتر از آستانه هستند و ما نمی‌توانیم فرض صفر را رد کنیم، اگر α 0.01 باشه مقدار P-value 4.7% زیر آستانه است و فرض صفر رد می‌شود و می‌توان گفت P-value 5.2% بالاتر از آستانه است، نمی‌توان فرض صفر را رد کنیم. در حالت کلی می‌توان گفت هر چه مقدار P-value کمتر باشه با اطمینان بیشتری می‌توانیم فرض صفر را رد کنیم

Part 2:

"The answer is (A)

Since the null hypothesis is designed in such a way that weather is there any significance difference due to chance or not."

Part 3: given data : $H_0: \mu_0 = 50$ Two tailed test: $n_1 = 100$ $S_1 = 10$

$H_1: \mu \neq 50$ $n_2 = 900$ $S_2 = 10$

therefore formulla is $Z = \dfrac{\bar{X} - 50}{S/\sqrt{n}} = \bar{X} - 50$ $\therefore \dfrac{\bar{x} - 50}{10/\sqrt{900}} = \dfrac{\bar{x} - 50}{0.333}$

Ans. depending upon SE the arrage Function of sample Size is false

Part 3:

Suppose every firm in the city hires employees through a process equivalent to simple random sampling. In that case, it implies that the hiring process is impartial and not biased toward any specific race.

Simple random sampling ensures that each individual is equally likely to be selected for employment, regardless of race.

Given that the black population in the city is 10%, we would expect the representation of black employees in each firm to be approximately 10% by probability if the hiring process is unbiased. This is because the hiring process is similar to randomly selecting individuals from the city's population.

If the firms adhere to this unbiased hiring process, none of them would be found guilty of discrimination by the z-test or any statistical test. The z-test determines whether the observed difference between two proportions is statistically significant, indicating a deviation from what would be expected due to chance.

In this case, if the firms follow unbiased, simple random sampling and the percentage of black employees matches the percentage of black individuals in the surrounding region (10%), there would be no statistically significant difference between the observed proportion of black employees and the expected proportion based on the city's demographics.

Therefore, in this scenario, none of the firms would be found guilty of discrimination based on the z-test or any statistical test because they would be meeting the anticipated representation of black employees based on the population's demographics.

Problem 12:

Coin is fair $p = .5$

$H_0: \; p = 0.5$
$H_1: \; p \neq 0.5$

if number of heads in single sample of 100 tosses is between 40 and 60, we will not reject H_0.

X number of heads in 100 tosses of coin $\quad X \sim \text{Binomial}(n = 100, p)$

P[rejecting the hypothesis when it's actually correct]

when hypothesis correct i.e coin is fair $\quad X \sim \text{Binomial}(n = 100, p = 0.5)$

Hence, applying the CLT $\quad np > 10$, $\quad n(1-p) > 10$

reject when

$P(x < 40)$

$$P(X < 39.5) = P\left(z < \frac{39.5 - 50}{50}\right) = P(z < 2.1) = 0.01786$$

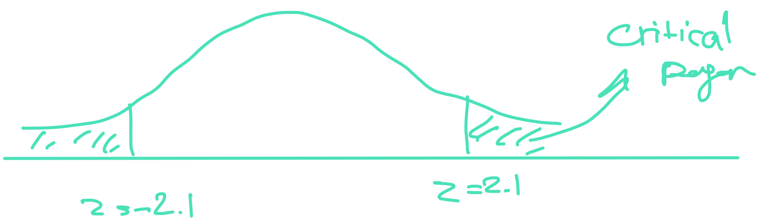
also.

$$P(x > 60.5) = P\left(z > \frac{60.5 - 50}{50}\right) = P(z > 2.1) = 0.01786$$

the probability of rejecting the hypothesis when it is actually true

$0.01786 + 0.01786 = 0.03572$

Part 2) type one error is when we reject the H_0 when we shouldn't have, and the probability of doing it is α (significant level)



Problem 13:

Part one: Since we have 2 samples from 2 diffrent population, and we need to determine if the mean score are diffrent, so we must choose the z-test.

Part two:

Public school: $\bar{X}_1 = 70$ $\sigma_1 = 4$ $n_1 = 50$

Catholic School: $\bar{X}_2 = 74$ $\sigma_2 = 6$ $n_2 = 65$

$$z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \longrightarrow \frac{\overset{-4}{(70-74)}}{\sqrt{\frac{16}{50} + \frac{36}{65}}} = -4.279$$

$\alpha = 0.025$

$Z_\alpha = 1.96$ $z \leq -1.96$ or $z \geq 1.96$ thus. H_0 will be rejected

Problem 14:

This scenario is an example of an observational study, not a controlled experiment. There are a few reasons why:

1. Lack of Controlled Variables: In a controlled experiment, the researcher manipulates one or more independent variables while controlling other potential confounding variables. However, in this case, the researcher is not manipulating how much the students play video games; they are simply observing two groups based on their playing habits.

2. Observation of Existing Groups: The two groups - students who play video games for less than an hour and those who play at least ten hours a week - are pre-existing groups. The researcher observes these groups as they are, without assigning participants based on any controlled condition.

3. No Random Assignment: In a controlled experiment, participants are randomly assigned to different groups to ensure that each group is similar except for the tested intervention. However, in this scenario, there is no indication of random assignment; the groups are based on their existing video game-playing behavior.

In summary, this study is categorized as observational because the researcher observes existing behaviors and conditions without manipulating or controlling any variables.

Part 2:

$$SE = \sqrt{\frac{Var_1}{n_1} + \frac{Var_2}{n_2}}$$

$Var_1 = 400$ $Var_2 = 2500$ $n_1 = 15$ $n_2 = 20$ dgree freedom $\Rightarrow 33$

$$t = \frac{mean_1 - mean_2}{SE}$$

P_value $= 2 \times (1 - CDF$ t-distribution $(t))$

$mean_1 = 120$ $mean_2 = 100$ $SE = \sqrt{\frac{400}{15} + \frac{2500}{20}}$ t-statistic $\approx 1{,}62$

P_value $\approx 0{,}114$

At 0.05 level, P-value greater than 0,05, difference is not significant.

At 0,01 level, P-value greater than 0,01, difference is not significant.

Part 3:

We cannot conclusively determine the accuracy of this indication for the current experiment as it is solely based on an observational study. Other factors, such as age, education, or innate ability, could influence both video games and spatial aptitude, which are called confounding variables. Additionally, selection bias may also be present. Therefore, it is necessary to conduct a controlled experiment to obtain accurate and reliable results.

Problem 15:

Part one:

$$d = \frac{n^{\frac{1}{2}} (\mu_A - \mu_0)}{\sigma} = \frac{7 \times 0{,}5}{7} = 0{,}5$$

$$Power = \Phi(\underset{1{,}645}{\cancel{Z_\alpha}} + d) = 0{,}984$$

P_value $= 1 - \Phi(d) = 0{,}3085$ So we have 30% to reject H_0, Part two: $\beta = 1 - \text{power} = 1 - 0{,}98402 = 0{,}01597$

$n \simeq 91$

Part 3: Power $= \Phi(z_{\text{stat}})$ Type your text $0{,}99 = \Phi\left(1{,}645 + \frac{n^{\frac{1}{2}} \times 0{,}5}{7}\right)$ $14 \times 0{,}681 = \sqrt{n}$

Problem 16:

Calculate α: $\alpha = P(X < 40) + P(X > 60)$

$= P(X \leq 39) + P(X \geq 61) = 0{,}0176 + 0{,}0176 = 0{,}352$

Part Two is in code part.

Problem 17:

given: $f(x|\theta) = \theta e^{-\theta x}$

if $\vec{x} = (x_1, \dots, x_n)$ i.i.d $\longrightarrow f(\bar{x}|\theta) = \theta^n e^{-\theta(x_1 \dots x_n)} = (\theta e^{-\theta \bar{x}})^n = f(\bar{x}|\theta)$

Log likhood $\ell(\theta) = n(\ln\theta - \bar{x}\theta)$ $\frac{dL}{d\theta} = n(1/\theta - \bar{x}) \longrightarrow$ mLE of $\theta \Rightarrow \hat{\theta} := 1/\bar{x}$

liklyhood ratio test for $H_0: \theta = \theta_0$, $H_1: \theta \neq \theta_0$

$H_0: \theta \in \Theta_0$ $H_1: \theta \in \Theta_0^c$

$\lambda(x) = \frac{\sup_{\Theta_0} L(\theta|x)}{\sup_{\Theta} L(\theta|x)} \xrightarrow[\text{above}]{\text{for}} \lambda(x) = \frac{L(\hat{\theta}_0|x)}{L(\hat{\theta}|x)}$

reject H_0 if: $\lambda(\bar{x}) = \frac{f(\bar{x}|\theta_0)}{f(\bar{x}|\theta)} \leq k$ So we have:

$$\frac{f(\bar{x}|\theta_0)}{f(\bar{x}|\hat{\theta})} = \frac{(\theta_0 e^{-\theta_0 \bar{x}})^n}{(\hat{\theta} e^{-\hat{\theta}\bar{x}})^n} = \left(\frac{\theta_0 e^{-\theta_0 \bar{x}}}{(1/\bar{x}) e^{-1}}\right)^n = (e \bar{x} \theta_0 e^{-\theta_0 \bar{x}})^n$$

thus test is of the form

reject H_0 if: $\bar{x} e^{-\theta_0 \bar{x}} \leq c$ where $c = \frac{k^{\frac{1}{n}}}{e\theta_0}$

Problem 18

in this case we can use T-distribution

$H_0: \mu \geq 28$

$$t = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}} \qquad \bar{x} = 25.9 \quad \mu_0 = 28 \quad S = 5.6 \quad n = 50$$

$$t = \frac{25.9 - 28}{\frac{5.6}{\sqrt{50}}} \simeq -2.65 \qquad df = n-1 = 50-1 = 49$$

$\alpha = 0.05$ if t-statistic is less than Critical value or if p-value is less than α reject H_0,

p-value $\simeq 0.005$ and critical value $\simeq -1.6$ then we reject H_0.

we conclude that there is sufficent eveidance based on the sample that the population mean is less than 28

Part 2:

Critical value of $\bar{y}$ to reject the null hypothesis for left tailed test

$$y - t \times SE = 28 - 1.68 \times 0.79 = 26.67$$

that is we reject the null hypothesis, when the sample mean is less than 26

Probibility of making a Type II error $= Pr(\text{Not reject } H_0 | H_0 \text{ is false})$

$$P(\bar{y} > 26.67 | \mu = 27) = P(t > (26.67 - 27)/0.7919) = P(t > 0.42)$$

$$= 0.6618$$

Problem 12:

Part one:

Mean Duration: The average duration of the surgical procedure is estimated to be 130 minutes, based on a sample of 40 surgeries.

Confidence Interval: The 95% confidence interval ranges from 128.5 to 131.5 minutes, suggesting that the true average duration is likely within this range.

Precision and Reliability: The narrow confidence interval indicates a precise estimate, and the use of the bootstrap method with 1000 replications provides a robust measure that's less sensitive to non-normal data or outliers.

Practical Use: This information is valuable for surgical scheduling and resource planning, ensuring efficient use of hospital resources.

Part 2:

To construct a confidence interval for the median using the bootstrap method:

Resample: Create many bootstrap samples from your original data by sampling with replacement.

Calculate Medians: Compute the median for each bootstrap sample.

Form Median Distribution: Create a distribution from these medians.

Determine Confidence Interval: Find the 2.5th and 97.5th percentiles of this distribution for a 95% confidence interval. These percentiles are the lower and upper bounds of your interval.

This method does not rely on the assumption of normality or any specific distribution of the data, making it versatile and robust, especially for statistics like the median, where the sampling distribution might be unknown or difficult to derive analytically. It's particularly useful when the sample size is not large enough to rely on asymptotic normality or when the data are skewed or contain outliers.

Part there:

The bootstrap method is a statistical technique that offers several advantages over parametric methods, particularly in the context of the study you described. These advantages include:

1. Fewer assumptions about data distribution: Parametric methods usually require data to follow a specific distribution, like the normal distribution. The bootstrap method, however, does not depend on such assumptions, making it more flexible and applicable to a broader range of data distributions, even those with outliers or skewness.

2. Robustness to small sample sizes: When the sample size is small, parametric methods may not hold well because their assumptions, like the central limit theorem, do not apply. Bootstrap methods, on the other hand, rely on the empirical distribution of the sample, allowing for more accurate estimates and confidence intervals.

3. Applicability to complex statistics: Parametric methods are often limited to estimating simple statistics like means and standard deviations. The bootstrap method can construct confidence intervals and test hypotheses for more complex statistics, such as medians, percentiles, or custom-defined statistics.

4. Intuitive interpretation: The bootstrap method is conceptually simple and easy to understand. It involves resampling from the observed data and simulating obtaining new samples from the underlying population.

5. Versatility in applications: The bootstrap method can be adapted to various statistical tasks, including hypothesis testing, regression, classification, and time-series analysis.

In this study, we aim to estimate the average duration of a specific type of surgical procedure. The bootstrap method provides a non-parametric way to construct a confidence interval for the mean or median duration, which is particularly useful if the distribution of durations is not normal or if the sample size is not large enough to justify the use of parametric methods.

Part four:
The bootstrap method is versatile and valuable, especially for complex or non-normal data, but it's not always the best choice due to:

Efficiency: Traditional parametric methods might be more accurate and less computationally intensive for large datasets or simple parameters.

Sample Size: Bootstrap may perform poorly with very small samples.
Outliers: It can be sensitive to outliers or highly skewed data.

Computational Demand: It can be computationally heavy, especially with many resamples.

While the bootstrap method is a flexible and powerful tool, especially for complex or non-standard problems, there are better choices than this one. The decision to use the bootstrap method should be based on the specifics of the data, the research question, the available computational resources, and the goals of the analysis. It's often beneficial to compare the results of the bootstrap method with those from parametric methods to assess the robustness and reliability of the estimates.

Part five:
Bootstrap statistics are estimates derived from the bootstrap method, a resampling technique used to assess the distribution of a statistic by repeatedly sampling with replacement from the data set. The process for calculating bootstrap statistics typically involves the following steps:

1. Original Sample: Start with an original sample from your data, representing a population.

2. Resampling:
   - Generate a large number of new samples (bootstrap samples), each of which is the same size as the original sample.
   - Create these samples by randomly selecting observations from the original sample with replacements (meaning the same observation can be chosen multiple times for a single sample).

3. Calculate Statistic for Each Sample:
   - For each bootstrap sample, calculate the statistic of interest (mean, median, variance, etc.).

4. Form the Bootstrap Distribution:
   - Use the statistics calculated from the bootstrap samples to form a statistic distribution known as the bootstrap distribution.

5. Summarize the Bootstrap Distribution:
   - Extract information from this distribution, such as:
     - The mean of the bootstrap statistics to estimate the statistics of the population.
     - Percentiles of the bootstrap distribution to create confidence intervals for the statistic.
     - The standard deviation or variance of the bootstrap statistics to estimate the standard error of the statistic.

Bootstrap statistics are powerful because they make minimal assumptions about the underlying data distribution and can be used to estimate the distribution of almost any statistic. This flexibility makes the bootstrap method particularly useful for complex statistics or when traditional parametric methods are not applicable.