Assignment Cover Sheet Course Name Biostatistics

PLAGIARISM

The University takes plagiarism seriously. All student work on the MPH is electronically screened for plagiarism and collusion. Any cases were this is suspected will be referred for investigation at Faculty level.

However this screening also allows the detection of students who are not referencing their work as carefully or as comprehensively as they should. The most common problem is the incorrect referencing of '*direct quotes*'.

Biostatistics - Marked Assignment 1

Notes:

1. When asked to do a statistical procedure or test you are expected to list and check all the assumptions of that procedure. **Marks will be allocated for getting the correct answer, interpreting the answer correctly (where applicable), using the correct procedure and for listing and testing the assumptions of that procedure.**
2. Computer output should be cut and pasted into the electronic document. Any statistic you are asked to calculated must be included in the text, i.e. you can get the statistical software package to calculate the number but I want to see that you know which number is the answer to the question. You can refer to graphs etc.
3. There are 5 questions
4. Remember if your work is clearly laid out and labeled it is easier for the marker to follow what you have done.

The data set 'growth study.xls' is a subset of data taken from a study of patterns of growth among South Asian and White European babies born in Manchester. There are 88 records in the database and it contains 11 variables on both the babies and their mothers.

ID	Participant ID number
Weight0	Weight at birth (Kg)
Length0	Length at birth (cm)
Weight3	Weight at 3 months (Kg)
Length3	Length at 3 months (cm)
Gender	Gender (0=female, 1=male)

Ethnicity	Ethnicity (0=South Asian, 1=White European)
Gestational Age	Gestational age (weeks)
Maternal weight	Mothers pre-pregnancy weight (Kg)
Employment	Mothers employment status (1=employed, 2=housewife, 3= unemployed, 4=student)
Smoke	Smoking status (0=no, 1=yes)

Answer the following questions:

Question 1 (16 marks)

a. **Use appropriate graphical and summary measures to illustrate the distribution of weight at birth (weight0). Marks will be allocated for using the most suitable type of graph, using the most appropriate summary statistics and justifying your choice. [8 marks]**

Solution:

The variable ‘weight at birth’ is a continuous variable and is measured using Interval scale. Therefore, we can compute descriptive statistics and histogram for the variable ‘weight at birth’. The descriptive statistics is calculated using SPSS (Analyze → Descriptive Statistics → Explore) and the output is given below:

Descriptives

			Statistic	Std. Error
Weight at birth (Kg)	Mean		3.2932	.05561
	95% Confidence Interval for Mean	Lower Bound	3.1827	
		Upper Bound	3.4037	
	5% Trimmed Mean		3.2836	
	Median		3.2000	
	Variance		.272	
	Std. Deviation		.52167	
	Minimum		2.20	
	Maximum		4.70	
	Range		2.50	
	Interquartile Range		.70	
	Skewness		.243	.257
	Kurtosis		-.330	.508

From the above table, we see that the mean weight at birth is 3.2932 kg with a standard deviation of 0.52167. The median weight at birth is 3.2 kg. This indicates that 50% of the sample data representing the weight at birth of babies falls below 3.2 kg and 50% of the sample data representing the weight at birth of babies falls above 3.2 kg. The minimum recorded birth weight of baby is 2.2kg and the maximum recorded birth weight is 4.7kg.

In order to determine whether the distribution of 'weight at birth' follows normal distribution, we carry out Kolmogrov-Smirnov test for normality. The null and alternate hypothesis is given below:

Null Hypothesis: H_0: The distribution follows approximately normal

Alternate Hypothesis: H_a: The distribution does not follow normal distribution:

The SPSS output of Kolmogrov-Smirnov test is given below

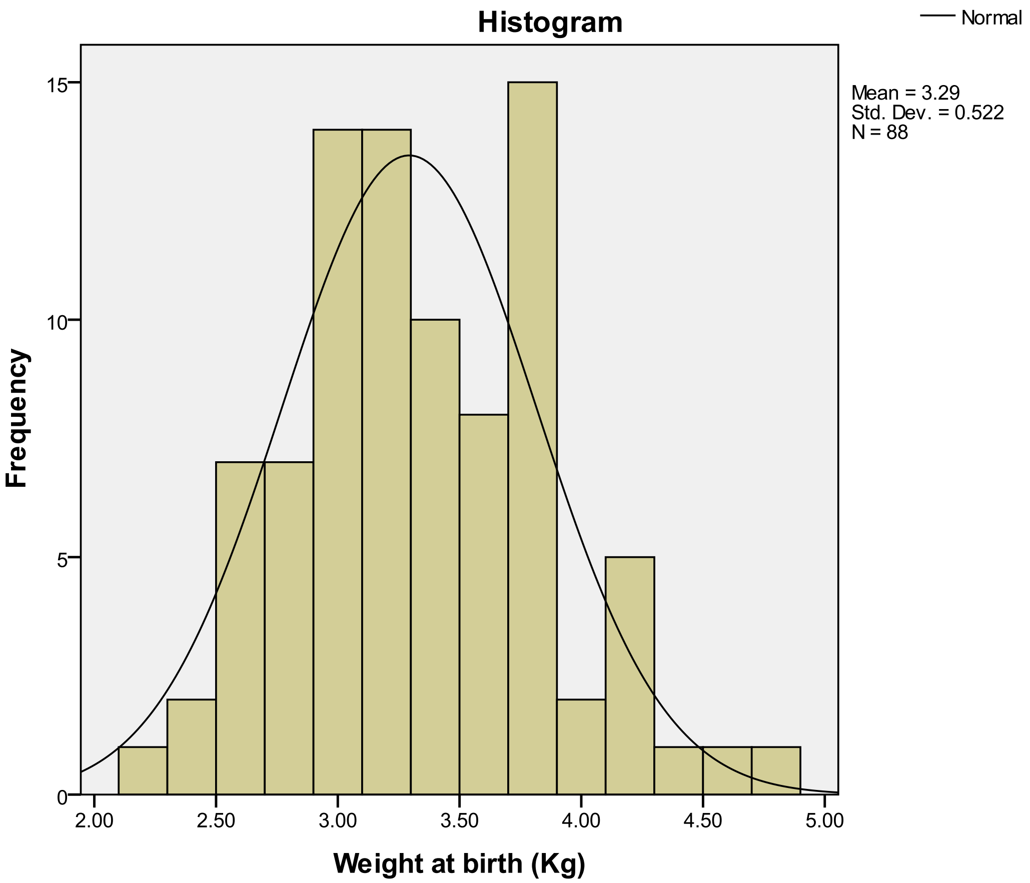
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Weight at birth (Kg)	.088	88	.092	.983	88	.296

a. Lilliefors Significance Correction

From the above table, we see that the value of Kolmogrov – Smirnov test statistic is 0.088 and its corresponding p – value is 0.092. Since the p – value of the test statistic is greater than 0.05, we do not have sufficient evidence to reject the null hypothesis. Therefore, we conclude that the distribution of 'Weight at Birth' follows normal distribution.

The histogram for the variable 'Weight at Birth' is given below:



From the above histogram, we see that the distribution of 'Weight at Birth' follows normal distribution.

b. A researcher has hypothesized that mean birth weight of these babies was 3Kg. Conduct a hypothesis test to test this hypothesis including testing all the assumptions. [5 marks]

Solution:

In order to determine whether the mean birth weight of the babies was 3Kg, we carry out single mean test. The null and alternate hypotheses are given below:

Null Hypothesis: H_0: $\mu = 3$

That is, the mean birth weight of these babies does not differ significantly from 3Kg.

Alternate Hypothesis: H_1: $\mu \neq 3$

That is, the mean birth weight of these babies differ significantly from 3Kg

The single mean test is carried out in SPSS (Analyze → Compare means → one mean test) and the output is given below:

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Weight at birth (Kg)	88	3.2932	.52167	.05561

One-Sample Test

	Test Value = 3					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Weight at birth (Kg)	5.272	87	.000	.29318	.1827	.4037

From the above table, we see that the value of t test statistic is 5.272 and its corresponding p – value is 0.000

Conclusion:

Since the p – value of the test statistic is less than 0.05; there is sufficient evidence to reject the null hypothesis. Therefore, we conclude that the mean birth weight of these babies differ significantly from 3Kg

c. Calculate and interpret the 95% confidence interval for the mean birth weight of the babies in this study. [3 marks]

The 95% confidence interval for mean birth weight of babies in this study is calculated by using the formula given below:

$$\left(\bar{x} - z_{\alpha/2} * \frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2} * \frac{s}{\sqrt{n}}\right)$$

From the Descriptive Statistics table (Refer Descriptive table given in (a)), we have

Sample mean weight at birth = $\bar{x}$ = 3.29

Sample standard deviation weight at birth = s = 0.52167

Sample size = n = 88

Using normal distribution table, we have

$Z_{0.05/2}$ = 1.96

$$\left(\bar{x} - z_{\alpha/2} * \frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2} * \frac{s}{\sqrt{n}}\right) = \left(3.2932 - 1.96 * \frac{0.52167}{\sqrt{88}}, 3.2932 + 1.96 * \frac{0.52167}{\sqrt{88}}\right)$$

= (3.2932 – 1.96 * 0.05561, 3.2932 + 1.96 * 0.05561)

= (3.2932 - 0.108996, 3.2932 + 0.108996)

= (3.1842, 3.4022)

Therefore, the 95% confidence interval for the mean birth weight of the babies in this study is **(3.1842, 3.4022)**

Question 2 (15 marks)

a. Use appropriate summary statistics and graphical techniques to describe the employment status of the mothers in this study. Marks will be allocated for using the most appropriate summary statistics and graphs, justifying your choice and interpreting the output correctly. [5 marks]

Solution:

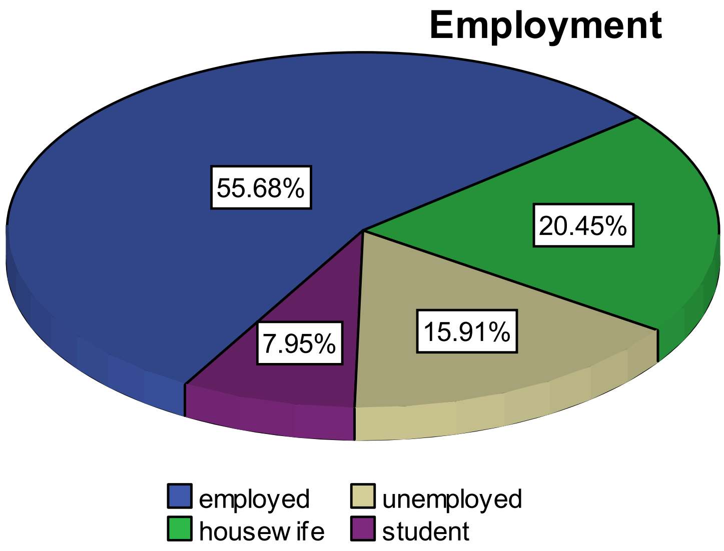
The variable 'Employment status of the mother' is a categorical variable and is measured under nominal scale. Totally, the employment status of mother is categorized into four categories, namely

- ✓ Employed
- ✓ Housewife
- ✓ Unemployed and
- ✓ Student

The table given below shows the frequency distribution of employment status of mother

Employment

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	employed	49	55.7	55.7	55.7
	housewife	18	20.5	20.5	76.1
	unemployed	14	15.9	15.9	92.0
	student	7	8.0	8.0	100.0
	Total	88	100.0	100.0	



From the above table, we see that nearly 55.7% of the mothers are employed, 20.5% of them are housewife, 16% of them are unemployed and 8% of them were students

b. **Calculate and interpret the point estimate and 95% confidence interval for the true proportion of mothers who were smokers. [4 marks]**

Solution:

Samples of 88 respondents opinion was taken for this study. The table given below shows the information about the proportion of mothers who were smokers

Smoke

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	79	89.8	89.8	89.8
	Yes	9	10.2	10.2	100.0
	Total	88	100.0	100.0	

From the above table, we see that 9 out of 88 mothers were smokers. That is 0.102 proportions of the mothers were smokers. The 95% confidence interval for the true proportion of mothers who were smokers is calculated by using the formula given below:

$$\left(p - z_{\alpha/2} * \sqrt{\frac{p*(1-p)}{n}}, p + z_{\alpha/2} * \sqrt{\frac{p*(1-p)}{n}}\right) =$$

$$\left(0.102 - 1.96 * \sqrt{\frac{0.102*(1-0.102)}{88}}, 0.102 + 1.96 * \sqrt{\frac{0.102*(1-0.102)}{88}}\right)$$

= (0.102 – 1.96 * 0.0323, 0.102 + 1.96 * 0.0323)

= (0.102 – 0.0632, 0.102 + 0.0632)

= (0.0388, 0.1652)

Therefore, the required 95% confidence interval for the true proportion of mothers who were smokers is **(0.0388, 0.1652)**

This indicates that when repeated samples are taken from the same population, then 95 out of 100 times, the true proportion of mothers who were smokers will fall within this interval

c. **Conduct a hypothesis test to test whether the proportion of South Asian mothers who smoke is the same as the proportion of White European mothers who smoke including testing all the assumptions. [6 marks]**

Solution:

The table given below shows the information about the association between ethnicity and smokers

Smoke * Ethnicity Cross tabulation

			Ethnicity		Total
			South Asian	White European	
Smoke	No	Count	56	23	79
	Yes	Count	7	2	9
Total		Count	63	25	88

From the above table, we have

Proportion of South Asian mothers who smoke = p_1 = 7/63 = 0.111

Proportion of White European mothers who smoke = p_2 = 2/25 = 0.08

In order to test whether the proportion of South Asian mothers who smoke is the same as the proportion of White European mothers who smoke, we carry out two proportion z test. The null and alternate hypotheses are given below:

Null Hypothesis: H_0: $P_1 = P_2$

That is, the proportion of South Asian mothers who smoke is the same as the proportion of White European mothers who smoke

Alternate Hypothesis: H_0: $P_1 \neq P_2$ (Two tailed test)

That is, the proportion of South Asian mothers who smoke is differ significantly from the proportion of White European mothers who smoke

Test Statistic:

The test statistic is:

$$z = \frac{p_1 - p_2}{\sqrt{P*(1-P)*\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Where

$$P = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{63*0.111+25*0.08}{63+25} = \frac{9}{88} = 0.102$$

Thus,

$$z = \frac{0.111-0.08}{\sqrt{0.102*(1-0.102)*\left(\frac{1}{63}+\frac{1}{25}\right)}}$$

$$= \frac{0.0311}{0.071623}$$

= 0.4344

Thus, the required value of z test statistic is 0.4344

Critical Region:

The critical value of z corresponding to 0.05 level of significance is:

$Z_{\alpha/2} = Z_{0.05/2} = \pm 1.96$ (Two tailed test)

The critical region is:

CR = {X: Z < - 1.96 or Z > 1.96}

Conclusion:

Since the value of the test statistic does not fall in the critical region, there is no sufficient evidence to reject the null hypothesis at 5% level of significance. Therefore, we conclude that the proportion of South Asian mothers who smoke is the same as the proportion of White European mothers who smoke

Question 3 (10 marks)

a. **Conduct a hypothesis test to test whether there is a difference in the length of babies at birth between boys and girls, including testing all the assumptions. [5 marks]**

Solution:

In order to test whether there is a difference in the length of babies at birth between boys and girls, we carry out independent sample t test. The null and alternate hypotheses are given below:

Null Hypothesis: H_0: $\mu_1 = \mu_2$

That is, the mean difference in the length of babies at birth between boys and girls does not differ significantly

Alternate Hypothesis: H_1: $\mu_1 \neq \mu_2$

That is, the mean difference in the length of babies at birth between boys and girls does not differ significantly

The independent sample t test is carried out in SPSS (Analyze → Compare means → independent sample t test) and the output is given below:

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Length at birth (cm)	Female	43	50.3105	2.05761	.31378
	Male	45	50.1111	2.92718	.43636

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Length at birth (cm)	Equal variances assumed	1.455	.231	.368	86	.714	.19935	.54166	-.87744	1.27615
	Equal variances not assumed			.371	79.109	.712	.19935	.53746	-.87042	1.26913

The Levene’s test is used to test for the equality of variances between the two groups. From the above table, we see that p – value of the levene’s test statistic is 0.231. Since the p – value is greater than 0.05, we conclude that the assumption of equality of variances holds true

Now, the value of the t test statistic is 0.368 and its corresponding p – value is 0.714. Since the p – value of t test statistic is greater than 0.05, there is no sufficient evidence to reject the null hypothesis. Therefore, we conclude that the mean difference in the length of babies at birth between boys and girls does not differ significantly

b. Conduct a hypothesis test to test whether there is a difference in the length of babies at 3 months between boys and girls, including testing all the assumptions. [5 marks]

Solution:

In order to test whether there is a difference in the length of babies at 3 months between boys and girls, we carry out independent sample t test. The null and alternate hypotheses are given below:

Null Hypothesis: H_0: $\mu_1 = \mu_2$

That is, the mean difference in the length of babies at 3 months between boys and girls does not differ significantly

Alternate Hypothesis: H_1: $\mu_1 \neq \mu_2$

That is, the mean difference in the length of babies at 3 months between boys and girls does not differ significantly

The independent sample t test is carried out in SPSS (Analyze → Compare means → independent sample t test) and the output is given below:

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Length at 3 months (cm)	Female	43	61.1895	2.37336	.36193
	Male	45	62.6858	1.94248	.28957

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Length at 3 months (cm)	Equal variances assumed	.155	.694	-3.243	86	.002	-1.49624	.46141	-2.41350	-.57898
	Equal variances not assumed			-3.228	81.21	.002	-1.49624	.46352	-2.41846	-.57403

The Levene’s test is used to test for the equality of variances between the two groups. From the above table, we see that p – value of the levene’s test statistic is 0.694. Since the p – value is greater than 0.05, we conclude that the assumption of equality of variances holds true

Now, the value of the t test statistic is – 3.243 and its corresponding p – value is 0.002. Since the p – value of t test statistic is less than 0.05, there is sufficient evidence to reject the null hypothesis. Therefore, we conclude that the mean difference in the length of babies at 3 months between boys and girls differ significantly. Going through the mean values, we see that the mean length of male babies at 3 months (62.6858) is significantly high than that of their female counterparts (61.1895)

Question 4. (8 marks)

Conduct a hypothesis test to answer the following question. Was the mean change in length between birth and 3 months different for boys and girls? Discuss your results with respect to the answers to questions 3a and b. [8 marks]

Solution:

In order to test whether there is a difference in the mean change in length between birth and 3 months different for boys and girls; we carry out independent sample t test. In order to test this claim, we first need to create a variable representing the difference length between birth and 3 months. The variable created in SPSS (Transform → Compute Variable → variable name as “mean_change” → formula (length0 – length3))

The null and alternate hypotheses are given below:

Null Hypothesis: H_0: $\mu_1 = \mu_2$

That is, the mean change in length between birth and 3 months does not differ significantly for boys and girls

Alternate Hypothesis: H_1: $\mu_1 \neq \mu_2$

That is, the mean change in length between birth and 3 months differ significantly for boys and girls

The independent sample t test is carried out in SPSS (Analyze → Compare means → independent sample t test) and the output is given below:

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
mean_change	Female	43	-10.8791	2.67139	.40738
	Male	45	-12.5747	3.34036	.49795

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
mean_change	Equal variances assumed	2.847	.095	2.622	86	.010	1.69560	.64662	.41015	2.98104
	Equal variances not assumed			2.636	83.448	.010	1.69560	.64336	.41607	2.97512

The Levene's test is used to test for the equality of variances between the two groups. From the above table, we see that p – value of the levene's test statistic is 0.095. Since the p – value is greater than 0.05, we conclude that the assumption of equality of variances holds true

Now, the value of the t test statistic is 2.622 and its corresponding p – value is 0.010. Since the p – value of t test statistic is less than 0.05, there is sufficient evidence to reject the null hypothesis. Therefore, we conclude that the mean change in length between birth and 3 months differ significantly for boys and girls. Going through the mean values, we see that the mean length of male babies at 3 months (-12.5747) is significantly less than that of their female counterparts (- 10.8791)

On comparing with the results of 3b, we see that the difference in the mean length of male babies at birth and at 3 months is significantly high when compared to that of the difference in the mean length of male babies at birth and at 3 months. This indicates that male babies grow faster than female babies

Question 5. (9 marks)

a. **Determine how many subjects you need in a study if you wished to have 80% power to detect a difference of 300g in birth weight between two independent groups. Assume a significance level of 5%. Any other information you need can be derived from the database. [3 marks]**

Solution:

The information in the given problem can be represented with the following notations:

Power of the test = β = 0.80

Level of significance = α = 0.05

Difference in birth weight = $\mu_0 - \mu_a$ = 300g

Standard deviation of birth weight = s = 0.52167

The required sample size is calculated by using the formula given below:

$$n = \left[\frac{(Z_\alpha - Z_\beta) * \sigma}{(\mu_0 - \mu_a)} \right]^2$$

Where

Z_α = z value providing an area of α in the tail of a standard normal distribution

Z_β = z value providing an area of β in the tail of a standard normal distribution

σ = Population standard deviation

μ_0 = the value of the population mean in the null hypothesis

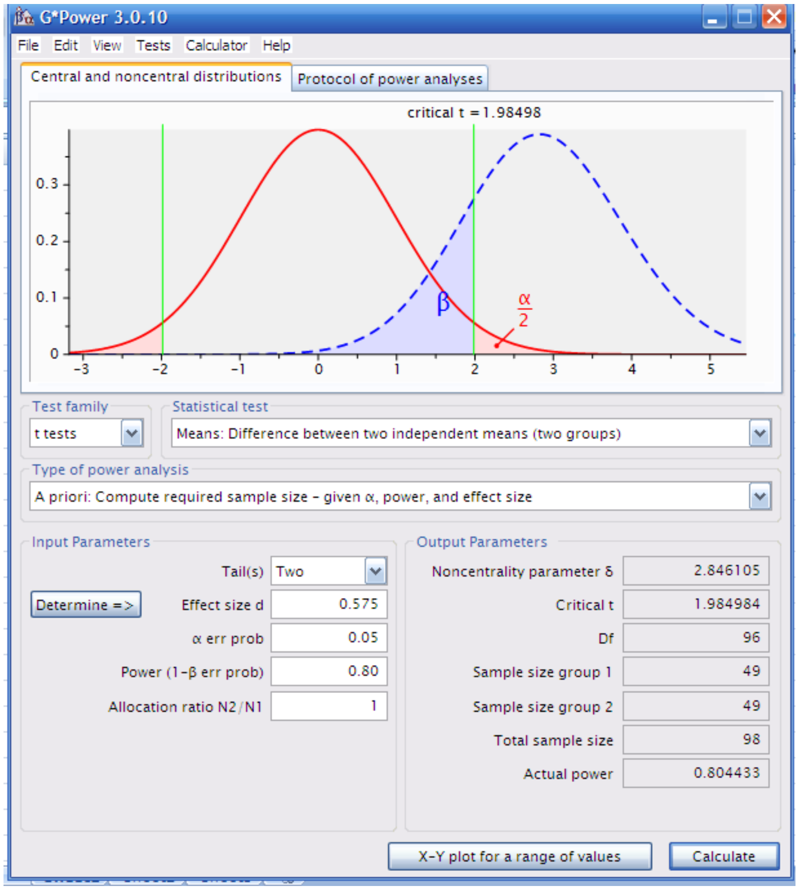
μ_a = the value of the population mean used for the type II error

Using normal distribution table, we have

$Z_{\alpha/2}$ = 1.96

Z_β = - 0.84162

The analysis is carried out in G * Power 3 calculator and the screenshot of the workings is given below:



Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Tail(s)	Two
Effect size d	0.575
α err prob	0.05
Power (1−β err prob)	0.80
Allocation ratio N2/N1	1

Output Parameters

Noncentrality parameter δ	2.846105
Critical t	1.984984
Df	96
Sample size group 1	49
Sample size group 2	49
Total sample size	98
Actual power	0.804433

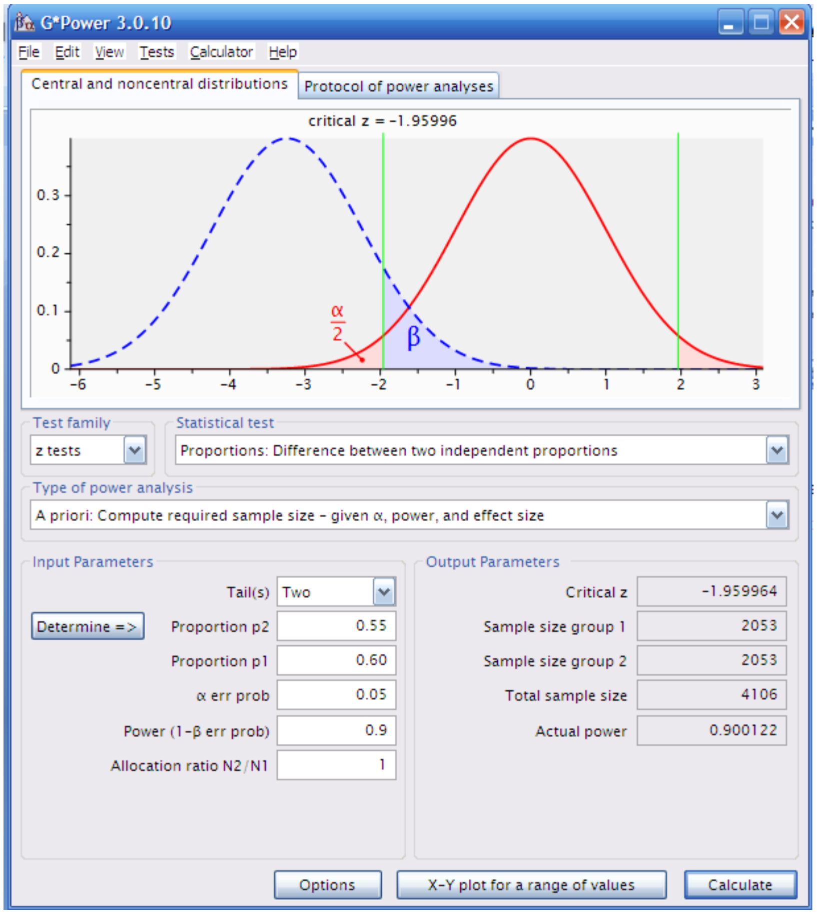
From the above output, we see that the required **sample size is 98**

b. **Determine how many subjects you need in a study to have 90% power to detect an absolute difference of 5% in the percentage of mothers who were smokers between two groups? Assume a significance level of 5%. Any other information you need can be derived from the database. [3 marks]**

Solution:

The sample size for the study having 90% power to detect an absolute difference of 5% in the percentage of mothers who were smokers between two groups with significance level of 5% is calculated using G * Power 3.

The analysis is carried out in G * Power 3 calculator and the screenshot of the workings is given below:



Test family: z tests

Statistical test: Proportions: Difference between two independent proportions

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Tail(s)	Two
Proportion p2	0.55
Proportion p1	0.60
α err prob	0.05
Power (1-β err prob)	0.9
Allocation ratio N2/N1	1

Output Parameters

Critical z	-1.959964
Sample size group 1	2053
Sample size group 2	2053
Total sample size	4106
Actual power	0.900122

From the above output, we see that the required **sample size is 4106**

c. **Consider the case where a drug (drug A) has been established as clinically effective and safe for use in reducing blood pressure. Another drug company has produced their own drug for reducing blood pressure (drug B) and have been asked to submit their drug to the Drug Standards Authority (DSA) for testing against the standard drug (A). However instead of submitting their own drug (B) they accidently submit a version of the standard drug (A). The DSA carries out the test (of drug A vs drug A) and they reject the null hypothesis and conclude that there is a difference between the standard drug (A) and the submitted drug (also A). In statistical terms, what type of error could have been made? Give reason for your answer. How often could this error be expected to occur if the significance level of 0.05 is used in the test? [3 marks]**

Solution:

Type I error occurs when one rejects the null hypothesis when it is true. In our case, the DSA carries out the test for the same type of drug (Drug A Vs Drug A) and they concluded that there is a difference between the standard drug (A) and the submitted drug (also A). But in real, there is no difference between the standard drug (A) and the submitted drug (also A). This is a typical example of Type I error.