STAT292
Assignment 1
Isabelle (Izzy) Southon
300597453

**1. Which of the following variables are categorical, and why?**

A, B, and F are categorical variables.

Inland Revenue Department (IRD) tax number is categorical because it is not quantitative data. Although an IRD tax number is a number, it has no numerical meaning. You also cannot measure or count different IRD tax numbers together as they have no numeric value. IRD tax number would be a categorical qualitative unordered nominal variable.

Level of agreement with the federal government's COVID response (1 = "Strongly disagree", 2 = "Disagree", 3 = "Neither agree nor disagree", 4 = "Agree", 5 = "Strongly agree") is a categorical variable. This is because the levels of agreement (given as numbers) are being given as options for categorical values. The numbers have no numeric value because they correspond to categories based on agreement level. Level of agreement is a qualitative, ordered ordinal variable.

Phone number is a categorical variable because the values cannot be counted/added together because they have no numeric value. If asked for your phone number, the response to that question would be a grouping of numbers, yet they do not mean anything important for data collection purposes.

**2.**

**a) Clearly explain what we are assuming about these 60 children in representing the number of these children living in poverty in the financial year ending in June 2021 by a binomial distribution. Provide an example of when this assumption would likely be violated. (Your answer must clearly refer to the situation described in the problem.)**

We are assuming that the 60 children are either not living in poverty 'success' or are living in poverty 'failure'. There is a set number of trials, being 60 children. We are assuming that the outcome of one child's living situation does not influence the outcome of another child's living situation in relation to poverty. Lastly, we are assuming that each randomly selected child has the same probability of 'success' of not living in poverty.

An example of when this assumption would likely be violated is that not all randomly selected children may have the same probability of 'success' or not living in poverty. Different regions of New Zealand have higher rates of poverty, so although the sample is randomly selected, children from different regions may have different probabilities of success. For example, if we randomly selected a child from the region in NZ with the lowest poverty rate and a child from the region with the highest poverty rate, their probabilities of 'success' would be different.

**b) What is the mean number of these 60 children that would be expected to have lived in poverty in the financial year ending in June 2021? What are the corresponding variance and standard deviation? (3 marks)**

STAT292
Assignment 1
Isabelle (Izzy) Southon
300597453

Let Y denote the number of children that would be expected to have lived in poverty in the financial year ending in June 2021.
Y ~ Bin(n,p)
Y ~ Bin(60, 0.16) = 9.60
μ = E(Y) = np = 60 x 0.16 = 9.60 children
$\sigma^2$ = V(Y) = np (1-p) = 60 x 0.16 x 0.84 = 8.064
σ = $\sqrt{V(Y)} = \sqrt{8.064} \approx$ 2.83972 children

We can expect a mean of 9.60 (10) children would have been expected to have lived in poverty in the financial year ending in June 2021. The corresponding variance is 8.064 children and the standard deviation is 2.84 children. Since children is a numerical integer data type, if we had to, we would round the variance to 8 children, and the standard deviation to 3 children.

**c) Using R, calculate the probability (to at least 3dp) that exactly 15 of these 60 children lived in poverty in the financial year ending in June 2021. (Be sure to show your R code and output.) (2 marks)**

P(X = 15), n = 60, p = 0.16

```
> # Calcuate P(X=15) for X~Bin(n=60, p =0.16)
> dbinom(x = 15, size=60, prob=0.16)
[1] 0.02400109
```

The probability that exactly 15 of these 60 children lived in poverty in the financial year ending in June 2021 is 0.024(3dp) children.

**d) What is the probability (to at least 4dp) that more than 10 of these 60 children lived in poverty in the financial year ending in June 2021? Calculate this probability:**

**i. exactly using R and the binomial distribution (be sure to include your R code and output) and**

P(X > 10) or P(X >=10) , n = 60, p = 0.16

```
> # Calculate P(X > 10) = P(X >= 10) for X ~ Bin(n=60,p= 0.16)
> pbinom(q = 10, size = 60, prob = 0.16)
[1] 0.6377971
```

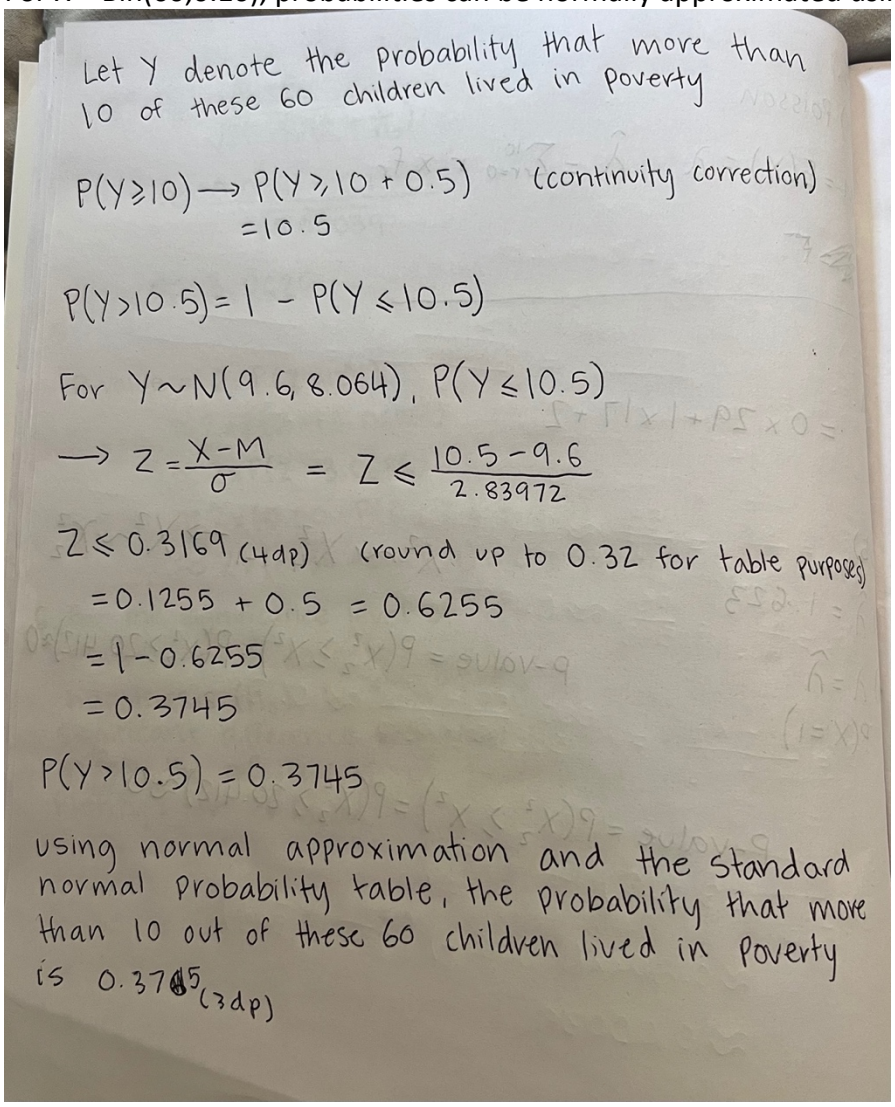
1 – 0.6377971 = 0.3622 (4dp)
The probability that more than 10 of these 60 children lived in poverty is 0.3622(4dp) children. This probability includes 10 children and more.

**ii. by hand using a normal approximation and the standard normal probability table. (7 marks)**

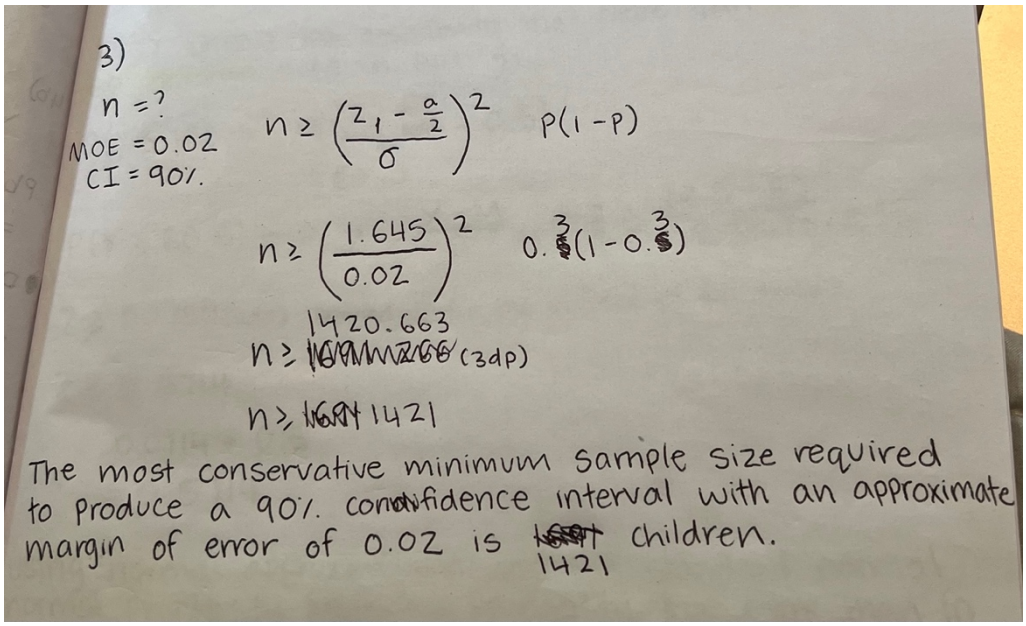
A binomial distribution with a random variable X, can be approximated in terms of a normal random variable Y if n is sufficiently large, n = 60. If np >= 5, n = 60 and p = 0.16, 60 x 0.16 = 9.6. This means np >= 5. 60(1 – 0.16) = 50.4, and therefore n(1-p) >= 5.

STAT292
Assignment 1
Isabelle (Izzy) Southon
300597453

For X ~ Bin(60,0.16), probabilities can be normally approximated using Y ~ N(9.6,8.064).




**3.**



**4.**

**a) Produce both a standard and an Agresti-Coull 95% confidence interval (to at least 3dp) for the proportion of children from Southland who lived in poverty in the financial year ending in June 2021. (5 marks)**

```
> library(PropCIs)
> # Proportion of children from Southland is 500 + 55 = 555
> # 95% Clopper-Pearson exact confidence interval
> exactci(x = 55, n = 555, conf.level = 0.95)



data:

95 percent confidence interval:
 0.07552973 0.12703494

> # 95% Agresti-Coull adjusted confidence interval
> add4ci(x = 55, n = 555, conf.level = 0.95)



data:

95 percent confidence interval:
 0.07688248 0.12705312
sample estimates:
[1] 0.1019678
```

The 95% standard confidence interval is 0.0755, 0.1270. The Agresti-Coull 95% confidence interval is 0.0769, 0.1271. All confidence intervals have been rounded to 4 decimal places.

**b) Test whether the proportions of children who lived in poverty in the financial year ending in June 2021 are different for the Southland and Bay of Plenty regions using a test for proportions. Be sure to report the hypotheses, test statistic, p-value, and your conclusion at the $\alpha = 0.05$ significance level. Your working should be done by hand, and the p-value should be calculated using the standard normal probability table. (8 marks)**

STAT292
Assignment 1
Isabelle (Izzy) Southon
300597453

4b) A test for proportions

$H_0: p_1 = p_2$

$H_1: p_1 \neq p_2$

1. $\hat{p}_1 = \frac{x_1}{n_1} = \frac{55}{555} \approx 0.0991$ (4dp)

$\hat{p}_2 = \frac{x_2}{n_1} = \frac{95}{595} \approx 0.1597$ (4dp)

2. $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{55 + 95}{555 + 595} \approx 0.1304$

3. $SE^2_{\hat{p}_1 - \hat{p}_2} = \hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

$\approx 0.1304 \times (1 - 0.1304)\left(\frac{1}{555} + \frac{1}{595}\right)$

$\approx 0.00039$

4. $Z^* = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

$\approx \frac{0.0991 - 0.1597}{\sqrt{0.00039}}$

$\approx 3.0686$

5. P-value $= 2 \times P(Z > |z^*|)$

$\approx 2 \times P(Z > |3.0686|)$

$\approx 2 \times P(Z > 3.0686)$

$\approx 2 \times (0.50 - P(0 \leq Z \leq 3.07))$

$\approx 2 \times (0.50 - 0.4989) = 0.0022$ or $2.2 \times 10^{-3}$

6. For the significance level of $\alpha = 0.05$, we have P-value $< \alpha$ $(0.0022 < 0.05)$ so we reject the null hypothesis $(H_0)$. We don't fail to conclude that there is a significant difference between in poverty for children between the regions of Southland and Bay of Plenty.

STAT292
Assignment 1
Isabelle (Izzy) Southon
300597453

**5.**

**a) Find the mean number of flight cancellations per day (to at least 3dp). (3 marks)**

$\mu = E(X) = \lambda$

$$\hat{\lambda} = \frac{\sum_{r=0}^{9} r \times f_r}{69}$$

$$= \frac{0 \text{ x } 29 + 1 \text{ x } 17 + 2 \text{ x } 9 + 3 \text{ x } 4 + 4 \text{ x } 3 + 5 \text{ x } 2 + 6 \text{ x } 0 + 7 \text{ x } 1 + 8 \text{ x } 0 + 9 \text{ x } 4}{69}$$

$$\frac{112}{69} = 1.623$$

The mean number of flight cancellations per day is 1.623(3dp) flights.

**b)**

X >=9 or 1-X < 8

```
> # 5b
> # Test whether the number of flight cancellations per day is consistent with
> # a Poisson Distribution
>
> # H0 : The population distribution is Poisson
> # H1 : The population distribution is not Poisson
>
> r <- 0:9
> observed <- c(29,17,9,4,3,2,0,1,0,4)
> n <- sum(observed)
> lambda.hat <- (sum(r * observed))/n
> prob <- c(dpois(0:8,lambda.hat),
+           1-ppois(8,lambda.hat))
> expected <- n * prob
>
> data.frame("r" = r,
+            "observed" = observed,
+            "probabilities"= prob,
+            "expected freq" = expected)
   r observed probabilities expected.freq
1  0       29  1.972687e-01  13.611541857
2  1       17  3.202043e-01  22.094096928
3  2        9  2.598760e-01  17.931440985
4  3        4  1.406092e-01   9.702035702
5  4        3  5.705881e-02   3.937057966
6  5        2  1.852344e-02   1.278117369
7  6        0  5.011172e-03   0.345770882
8  7        1  1.162011e-03   0.080178755
9  8        0  2.357703e-04   0.016268153
10 9        4  5.060004e-05   0.003491403
```

STAT292
Assignment 1
Isabelle (Izzy) Southon
300597453

```
> r.new <- 0:4
> observed.new <- c(29,17,9,4,(3+2+0+1+0+4))
> sum(observed.new)
[1] 69
> prob.new <- c(dpois(0:3,lambda.hat),
+               1-ppois(3,lambda.hat))
> expected.new <- n * prob.new
>
> sum(prob.new) #checks
[1] 1
> sum(expected.new)
[1] 69
>
> data.frame("r" = r.new,
+            "observed" = observed.new,
+            "probabilities"= prob.new,
+            "expected freq" = expected.new)
  r observed probabilities expected.freq
1 0       29     0.1972687     13.611542
2 1       17     0.3202043     22.094097
3 2        9     0.2598760     17.931441
4 3        4     0.1406092      9.702036
5 4       10     0.0820418      5.660885
>
> test.stat <- sum(((observed.new - expected.new)^2)/expected.new)
>
> df <- 5 - 1 - 1 #number of rows p - 1 - 1
>
> p.val <- pchisq(test.stat,df,lower.tail = FALSE)
> paste("p-value is:", p.val)
[1] "p-value is: 1.59767147400638e-06"
>
> # The p-value of 0.00000016 is < than the 0.05 significance level. This means
> # that we have evidence to reject the null hypothesis and say that the number
> # of flight cancellations per day is not consistent with a poisson distribution.
>
```

Grouping categories with expected frequencies > 5

The test statistic is 29.6976443388869.

Under $H_0$: $X^2 \sim X^2_{5-1-1} \longrightarrow X^2 \sim X^2_2$