Assignment 1

STAT 292 Applied Statistics 2A — 300296259

1. Which of the following variables are categorical?

a. Water pressure (bars).

Not categorical, water pressure is continuous.

b. Course grade (A, B, C, D, E, F).

Categorical.

c. Level of approval of the Prime Minister's performance (1 = "Strongly disapprove", 2 = "Disapprove", 3 = "Neither approve nor disapprove", 4 = "Approve",5 = "Stongly approve").

Categorical.

d. Hospital admissions (patients per day).

Not categorical, patients per day is numeric.

e. Yearly rainfall (centimeters).

Not categorical, rainfall in centimetres is continuous.

f. Phone number.

Categorical. Phone numbers are numeric but they aren't counted or measured so it wouldn't make sense to perform numeric operations on them.

2. Results from the 2013 New Zealand Census suggest that 20% of adults in New Zealand had a university degree or equivalent at the time of the census. Consider a random sample of 40 New Zealanders who participated in the 2013 census, and suppose that the number of these people who reported having a university degree or equivalent at the time of the 2013 census can be represented by a random variable following a binomial distribution.

a. Cleary explain what we are assuming about these 40 people in representing the number of them who reported having a university degree or equivalent at the time of the 2013 census 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're assuming that the probability of someone having a degree or not having a degree isn't influenced by someone else having a degree/not having a degree thus making them independent. We're also assuming that the probability of having a degree/not having a degree is the same for every person thus making the probbailities identically distributed.

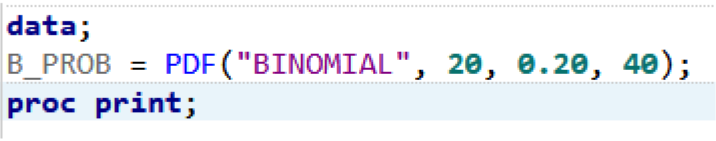
b. What is the mean number of these 40 people that would be expected to have reported having a university degree or equivalent at the time of the 2013 census? What are the corresponding variance and standard deviation?

$\mu = E(Y) = np = 40 \times 0.20 = 8$

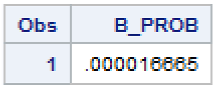
$\sigma^2 = Var(Y) = np(1-p) = 40 \times (0.20 \times 0.80) = 40 \times 0.16 = 6.4$

$\sigma = \sqrt{Var(Y)} = \sqrt{6.4} = 2.52982212813$

c. Using SAS, calculate the probability that exactly half of these 40 people reported having a university degree or equivalent at the time of the 2013 census.



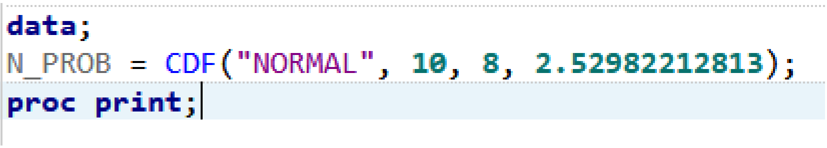
```
data;
B_PROB = PDF("BINOMIAL", 20, 0.20, 40);
proc print;
```



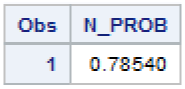
```
Obs  B_PROB
  1  .000016665
```

d. What is the probability that fewer than 10 of these 40 people reported having a university degree or equivalent at the time of the 2013 census? Calculate this probability

- **exactly using SAS and**



```
data;
N_PROB = CDF("NORMAL", 10, 8, 2.52982212813);
proc print;
```



```
Obs  N_PROB
  1  0.78540
```

- **by hand using a normal approximation and the normal probability table.**

$P(X < 10) \approx P(\frac{X-\mu}{\sigma} < \frac{10-\mu}{\sigma})$

$P(X < 10) \approx 0.5 + P(Z < \frac{10-8}{2.52982212813})$

$P(X < 10) \approx 0.5 + P(Z < 0.79)$

$P(X < 10) \approx 0.5 + 0.2852$

$P(X < 10) \approx 0.7852$

3. Medical diagnostic tests are subject to one of two types of errors:

- **false positive:** A person who does not have the disease or condition returns a test result that suggests that they do have the disease or condition.

- **false negative:** A person who has the disease or condition returns a test result that suggests that they do not have the disease or condition.

These two errors typically have quite different probabilities of occurring with the probability of a false positive most commonly being higher than the probability of a false negative because it is nearly always more catastrophic to miss those who have the disease or condition.

A recent report by the European CanCer Organisation (2017) into a non-invasive diagnostic test for stomach and esophageal cancers reported results on test results for 335 people across three different hospitals. The diagnosis test was administered to roughly equal numbers of people with and without stomach or esophageal cancer to assess the efficacy of the test. Results for the test are as shown in the table below.

	Tested positive for stomach or esophageal cancer?		
Have stomach or esophageal cancer?	No	Yes	n
No	140	32	172
Yes	32	131	163

Source: The European CanCer Organisation (29 January 2017). "Breath test could help detect stomach and esophageal cancers." *ScienceDaily.*

a. Suppose we wish to separately estimate the proportion of false positives and the proportion of false negatives and produce 95% confidence intervals for these proportions. Find the most conservative minimal sample sizes required for those who have stomach or esophageal cancer and those who do not have stomach or esophageal cancer to produce confidence intervals with an approximate margin of error of 0.06. (Note that you need only carry out one sample size calculation.The most conservative minimal sample size required will be the same sample size required to estimate each of the proportion of false positives and the proportion of false negatives to within the specified margin of error.)

$n \geq (\frac{z_1 - \frac{\alpha}{2}}{\delta})^2 p(1-p)$

$p = \frac{172}{335} = 0.51343283582$

$n \geq (\frac{1.96}{0.06})^2 \times 0.51343283582 \times 0.48656716418$

$n \geq 1067.11111111 \times 0.51343283582 \times 0.48656716418$

$n \geq 266.585227098$

$n \geq 267$

b. Using these data, produce both a standard and an Agresti-Coull 95% confidence interval for the proportion of false positives. Be sure to show all working.

Standard 95% Confidence Interval for the proportion of false positives:

$\hat{p} \pm Z_1 - \frac{\alpha}{2} \times S_{\hat{p}}$

$\hat{p} = \frac{32}{335} = 0.09552238805$

$1.96 \times S_{\hat{p}}$

$S_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$S_{\hat{p}} = \sqrt{\frac{0.09552238805 * 0.90447761195}{335}}$

$S_{\hat{p}} = \sqrt{\frac{0.08639786143}{335}}$

$S_{\hat{p}} = \sqrt{0.00025790406} = 0.01605939164$

$\hat{p} \pm (Z_1 - \frac{\alpha}{2} \times S_{\hat{p}})$

$Z_1 - \frac{\alpha}{2} \times S_{\hat{p}} = 1.96 \times 0.01605939164 = 0.03147640761$

$0.09552238805 + 0.03147640761 = 0.12699879566$

$0.09552238805 - 0.03147640761 = 0.06404598044$

I am 95% confident that the true proportion of false positives is somewhere between 6.00% - 12.70%

Agresti-Coull 95% Confidence Interval for the proportion of false positives:

$\hat{p}^* \pm z_1 - \frac{\alpha}{2} \times S_{\hat{p}^*}$

$\hat{p}^* = \frac{y+2}{n^*}$

$n^* = n + 4$

$\hat{p}^* = \frac{32+2}{335+4} = \frac{34}{339} = 0.10029498525$

$1.96 \times S_{\hat{p}}^*$

$S^*_{\hat{p}} = \sqrt{\frac{\hat{p}^*(1-\hat{p}^*)}{n^*}}$

$S^*_{\hat{p}} = \sqrt{\frac{0.10029498525 \times 0.89970501475}{339}} = \sqrt{0.00026618259} = 0.01631510312$

$z_1 - \frac{\alpha}{2} \times S^*_{\hat{p}} = 1.96 \times 0.01631510312 = 0.03197760211$

$0.10029498525 + 0.03197760211 = 0.13227258736$

$0.10029498525 - 0.03197760211 = 0.06831738314$

I am 95% confident that the true proportion of false positives is somewhere between 6.83% - 13.23%

c. Test whether the proportion of false positives is significantly different from the proportion of false negatives. Carry out the test at the $\alpha = 0.05$ significance level, showing all working. Be sure to report the test statistic, p-value, and your conclusion based on the p-value. What does this result suggest about this particular diagnostic test?

p_1: false positive

p_2: false negative

$H_0 : p_1 = p_2$

$H_1 : p_1 \neq p_2$

$\hat{p_1} = \frac{y1}{n1} = \frac{32}{163} = 0.1963190184$

$\hat{p_2} = \frac{y2}{n2} = \frac{32}{172} = 0.18604651162$

H_0 assumes $p1 = p2$ thus:

$\hat{p} = \frac{y1+y2}{n1+n2} = \frac{32+32}{163+172} = \frac{64}{335} = 0.19104477611$

$S^2_{\hat{p_1}-\hat{p_2}} = \hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})$

$S^2_{\hat{p_1}-\hat{p_2}} = 0.19104477611 \times 0.80895522389 \times (\frac{1}{163} + \frac{1}{172})$

$S^2_{\hat{p_1}-\hat{p_2}} = 0.19104477611 \times 0.80895522389 \times 0.01194892281$

$S^2_{\hat{p_1}-\hat{p_2}} = 0.00184666622$

$z^* = \frac{\hat{p_1}-\hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}$

$z^* = \frac{0.1963190184-0.18604651162}{\sqrt{0.19104477611 \times 0.80895522389 \times 0.01194892281}}$

$z^* = \frac{0.01027250678}{\sqrt{0.00184666622}}$

$z^* = \frac{0.01027250678}{0.04297285445}$

$z^* = 0.23904641456$

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

$p - value = 2 \times P(Z > 0.24)$

$p - value = 2 \times 0.0948 = 0.1896$

$0.1896 > 0.05$

Since our p-value is less than our threshold (our significance level, 0.05), we reject H_1 in favour of H_0, thus there is not a significant difference between the proportion of false positives ve the proportion of false negatives.