School of Mathematics and Statistics

Te Kura Mātai Tatauranga

STAT292 **Assignment 1 Solutions**

Note: Solutions must be either typed or written neatly, and questions must be answered in order. Where SAS is used to answer a question, relevant SAS output must be copied as an image file and included with your answer to the question and not at the end of the assignment.

1. **Which of the following variables are categorical?**

   a. **Water pressure (bars).**
   b. **Course grade (A, B, C, D, E, F).**
   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").**
   d. **Hospital admissions (patients per day).**
   e. **Yearly rainfall (centimeters).**
   f. **Phone number.**

   Answer: (b), (c), and (f). Course grade and level of approval of the Prime Minister's performance are both examples of ordinal categorical data (*i.e.*, there is an implicit hierarchy or ordering of the categories), whereas phone number is an example of nominal categorical data (*i.e.*, no implicit ordering of the categories). Even though both the level of approval of the Prime Minister's performance and phone numbers may be represented using numbers, those numbers do not have any intrinsic meaning and are simply placeholders for categories.

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.)**

Individuals' university degree (or equivalent) statuses must represent independent and identically distributed Bernoulli trials. This means that we are assuming that

- whether or not one individual has a university degree or equivalent in no way influences/changes the probability of another individual having a university degree or equivalent
- the probability of having a university degree or equivalent is the same for all individuals (specifically, 0.2).

An example of when this assumption would likely be violated is if we sampled multiple adults from the same family. Knowing whether or not one of the adults had a university degree or equivalent would impact the likelihood of other adults from the same family having a university degree or equivalent.

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?**

The mean number of these 40 people that would be expected to report having a university degree or equivalent at the time of the 2013 census (μ), along with corresponding variance (σ^2) and standard deviation (σ) are

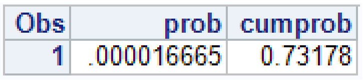
$$\begin{aligned}\mu &= \mathbb{E}(Y) = np = 40 \times 0.2 = 8 \text{ people} \\ \sigma^2 &= \mathbb{V}(Y) = np(1-p) = 40 \times 0.2 \times (1-0.2) = 6.4 \\ \sigma &= \sqrt{\sigma^2} = \sqrt{6.4} \approx 2.5298 \text{ people}\end{aligned}$$

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.**

Let Y denote the number of people with a university degree or equivalent. As shown in the SAS code and corresponding output below,

$$P(Y = 20) \approx 0.0000167$$

```
data;
/* P(Y = 20) */
prob = pdf("BINOMIAL", 20, 0.2, 40);
/* P(Y < 10) = P(Y <= 9) */
cumprob = cdf("BINOMIAL", 9, 0.2, 40);
proc print;
```



```
Obs        prob   cumprob
  1  .000016665   0.73178
```

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

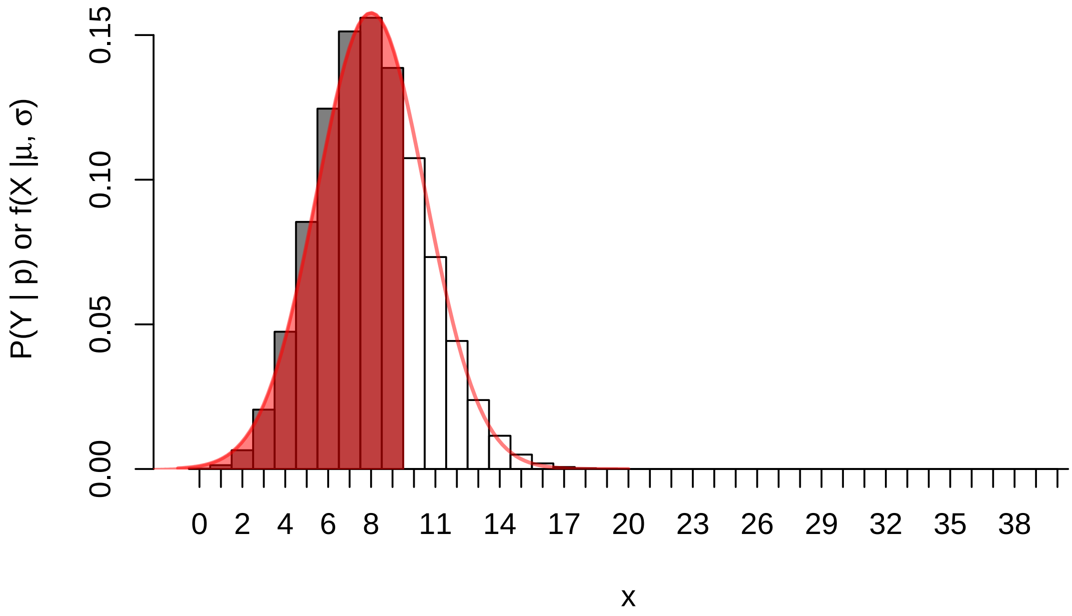
Here, we want to calculate $P(Y < 10)$, which is the same as calculating $P(Y \leq 9)$. Relevant SAS code for calculating $P(Y \leq 9)$ using the binomial distribution, along with output, is shown in part (c). This produces an exact probability of

$$P(Y < 10) \;=\; P(Y \leq 9) \;\approx\; 0.73178$$

To calculate the corresponding probability using a normal approximation, let X represent a normally distributed random variable with mean $\mu = np = 8$ and variance $\sigma^2 = np(1-p) = 6.4$ (*i.e.*, $X \sim \mathrm{N}(8, 6.4)$). Then

$$\begin{aligned} P(Y < 10) &\approx P(X < 9.5) \\ &\approx P\left(\frac{X-\mu}{\sigma} \leq \frac{9.5-8}{\sqrt{6.4}}\right) \\ &\approx P(Z \leq 0.5929) \\ &\approx 0.5 + P(0 \leq Z \leq 0.5929) \\ &\approx 0.5 + 0.2224 \\ &\approx 0.7224 \end{aligned}$$

The exact binomial probability is given by the area of the grey bars in the barplot shown below, and the normal approximation is given by the area shaded in red under the normal curve. The discrepancy between the exact probability and normal approximation corresponds to the difference in the area in grey falling outside of the normal curve and the area shaded in red that does not correspond to a grey bar.



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 diagnostic 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.

Have stomach or esophageal cancer?	Tested positive for 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 true proportion of false positives and the true 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.)**

The most conservative minimum sample sizes required would be obtained by using $p = 0.5$. Then, for a margin of error $\delta = 0.06$, the sample sizes must satisfy

$$\begin{aligned} n &\geq \left(\frac{z_{1-\frac{\alpha}{2}}}{\delta}\right)^2 p(1-p) \\ &\geq \left(\frac{z_{1-\frac{0.05}{2}}}{0.06}\right)^2 0.5(1-0.5) \\ &\geq \left(\frac{1.96}{0.06}\right)^2 0.5(1-0.5) \\ &\geq 266.768, \end{aligned}$$

so the most conservative minimum sample sizes required would be 267 for both those who have stomach or esophageal cancer and those who do not have stomach or esophageal cancer.

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

95% confidence intervals for the true proportion of false positives (+):

Standard:

$$\begin{aligned} \widehat{p}_+ &= \frac{y}{n} = \frac{32}{172} \approx 0.186 \\ S_{\widehat{p}_+} &= \sqrt{\frac{\widehat{p}_+\left(1-\widehat{p}_+\right)}{n}} = \sqrt{\frac{0.186 \times (1-0.186)}{172}} \approx 0.0297 \\ \widehat{p}_+ &\pm z_{1-\frac{\alpha}{2}} \times S_{\widehat{p}_+} \\ 0.186 &\pm 1.96 \times 0.0297 \\ & (0.1279, 0.2442). \end{aligned}$$

Agresti-Coull:

$$\begin{aligned}
\widehat{p}_+^* &= \frac{y+2}{n+4} = \frac{34}{176} \approx 0.1932 \\
S_{\widehat{p}_+^*} &= \sqrt{\frac{\widehat{p}_+^*\left(1-\widehat{p}_+^*\right)}{n+4}} = \sqrt{\frac{0.1932 \times (1-0.1932)}{176}} \approx 0.0298 \\
\widehat{p}_+^* &\pm z_{1-\frac{\alpha}{2}} \times S_{\widehat{p}_+^*} \\
0.1932 &\pm 1.96 \times 0.0298 \\
& (0.1349, 0.2515).
\end{aligned}$$

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?**

We want to test the hypotheses

$$\begin{aligned}
\mathcal{H}_0 : p_+ &= p_- \\
\mathcal{H}_1 : p_+ &\neq p_-
\end{aligned}$$

where p_+ denotes the true proportion of false positives and p_- denotes the true proportion of false negatives for the test. The sample proportions corresponding to p_+ and p_-, along with estimated common proportion and variance of the difference in proportions, are

$$\begin{aligned}
\widehat{p}_+ &\approx 0.186 \\
\widehat{p}_- &\approx 0.1963 \\
\widehat{p} &= \frac{32+32}{172+163} \approx 0.191 \\
S^2_{\widehat{p}_+-\widehat{p}_-} &= \widehat{p}\left(1-\widehat{p}\right)\left(\frac{1}{172}+\frac{1}{163}\right) \\
&\approx 0.191(1-0.191)\left(\frac{1}{172}+\frac{1}{163}\right) \\
&\approx 0.0018 \\
S_{\widehat{p}_+-\widehat{p}_-} &= \sqrt{S^2_{\widehat{p}_+-\widehat{p}_-}} \approx \sqrt{0.0018} \approx 0.043.
\end{aligned}$$

Then the test statistic is given by

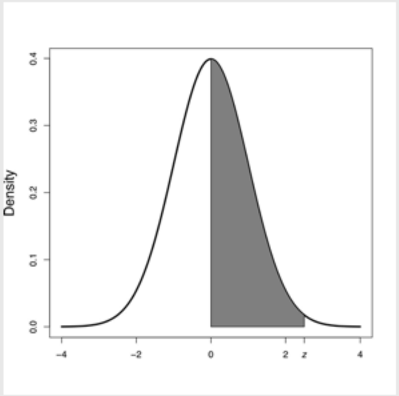
$$\begin{aligned}
z^* &= \frac{\widehat{p}_+ - \widehat{p}_-}{S_{\widehat{p}_+-\widehat{p}_-}} \\
&\approx \frac{0.186 - 0.1963}{0.043} \\
&\approx -0.239,
\end{aligned}$$

and the p-value is given by

$$\begin{aligned} p\text{-value} &= 2 \times P\left(Z > |z^{*}|\right) \\ &\approx 2 \times P\left(Z > |-0.239|\right) \\ &\approx 2 \times \left(0.5 - P\left(0 \leq Z \leq 0.239\right)\right) \\ &\approx 2 \times \left(0.5 - 0.0948\right) \ \approx \ 0.8104. \end{aligned}$$

As the p-value is much larger than any reasonable significance level (*e.g.*, $\alpha = 0.05$, $\alpha = 0.10$), we do not reject $\mathcal{H}_0$. Thus, there is insufficient evidence to suggest that the proportion of false positives and false negatives differ for this test.

Standard Normal Probabilities $P(0 \leq Z \leq z)$ for $Z \sim \mathrm{N}(0,1)$



z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000