School of Mathematics and Statistics

Te Kura Mātai Tatauranga

STAT 292 **Test 1: Due by Saturday, 9 May 2020 at 8:00 AM**

Instructions: There are 10 questions given on pages 2–9 worth a total of 100 marks.

Answer **ALL** questions.

Solutions must be either typed or written neatly, and questions must be answered in order.

Some SAS output and a standard normal probability table are provided on pages 10–11.

Be sure to submit your assignment as a PDF and follow the instructions specified on the submission system.

Section A: Multi-Choice (40 Marks)

For Section A questions, only record the letter corresponding to your answer. Do not present any working to support your choice of answer.

Use the following information to answer Questions 1 to 6.

When MetService reports that there is a 30% chance of rain in Wellington for a given day, it means that they estimate the probability of any rain in Wellington for that day to be 0.3. Consider 8 randomly selected days where MetService reported that there was a 30% chance of rain in Wellington. Suppose that rain was recorded in Wellington for 5 of those days.

1. **Assuming that MetService's reported probability of rain in Wellington for each of those days is correct, what is the probability (to 4dp) that exactly 5 of the 8 days had rain? (5 marks)**
   a. **0.0008**
   b. **0.0013**
   c. **Solution: 0.0467**
   d. **0.625**
   e. **0.9887**
2. **For a random sample of 8 days where MetService reports that there is a 30% chance of rain, what are the mean and variance (to 4dp) for the number of days Y that have rain? (5 marks)**
   a. $\mathbb{E}(Y) = 0.3$, $\mathbb{V}(Y) = 0.0262$.
   b. $\mathbb{E}(Y) = 0.3$, $\mathbb{V}(Y) = 0.162$.
   c. $\mathbb{E}(Y) = 2.4$, $\mathbb{V}(Y) = 1.2961$.
   d. **Solution:** $\mathbb{E}(Y) = 2.4$, $\mathbb{V}(Y) = 1.68$.
   e. $\mathbb{E}(Y) = 2.4$, $\mathbb{V}(Y) = 5.6$.
3. **Again assuming that MetService's reported probability of rain in Wellington for each of the 8 days is correct, use a normal approximation to find the probability (to 4dp) that rain was recorded on fewer than 5 of the 8 days. (For your reference, a standard normal probability table is presented on page 11.) (5 marks)**
   a. **0.7422**
   b. **0.8023**
   c. **0.8907**
   d. **0.8944**
   e. **Solution: 0.9474**
4. **Was it appropriate to use the normal approximation in Question 3? (5 marks)**
   a. **Solution: No. One of np and $n(1-p)$ is less than 5.**
   b. **No. Both np and $n(1-p)$ are less than 5.**
   c. **Yes. One of np and $n(1-p)$ is at least 5.**
   d. **Yes. Both np and $n(1-p)$ are at least 5.**
   e. **Yes. Both np and $np(1-p)$ are at least 5.**

5. **Now, suppose that the true probability of rain in Wellington for the 8 randomly sampled days is unknown. Using the observed number of days with rain (5), produce an Agresti-Coull adjusted 95% confidence interval (to 4dp) for the true probability of rain p. (5 marks)**
   a. **(0.2895, 0.9605)**
   b. **Solution: (0.3044, 0.8623)**
   c. **(0.441, 0.7257)**
   d. **(0.4538, 0.7962)**
   e. **None of the above.**
6. **Finally, suppose that you want to estimate the true proportion of days with rain in Wellington, and you plan to present your results using a 90% confidence interval. Find the most conservative minimum sample size required if the interval is to have an approximate margin of error of 0.03. (5 marks)**
   a. **632**
   b. **Solution: 752**
   c. **897**
   d. **1068**
   e. **None of the above.**

Use the following information to answer Questions 7 to 8.

For students who graduated from a particular university 5 years ago, the following data show the numbers of students who have not changed their jobs since they graduated for a random selection of 110 bachelor's degree students and 70 master's degree students:

Degree	Sample size	Number of students in the same job
Bachelor's	110	58
Master's	70	32

7. **Let p_B denote the proportions of bachelor's degree students who have not changed their jobs since they graduated and p_M denote the proportions of master's degree students who have not changed their jobs since they graduated. Calculate the test statistic z^* (to 4dp) for a test of**

$$\begin{aligned} \mathcal{H}_0 : p_B &= p_M \\ \mathcal{H}_1 : p_B &\neq p_M. \end{aligned}$$

   (5 marks)

   a. **Solution:** $z^* = 0.9174$
   b. $z^* = 0.5326$
   c. $z^* = -4.9566$
   d. $z^* = -5.4526$
   e. None of the above.

8. **For the hypothesis test in Question 7, calculate the p-value (to 4dp). Use the standard normal probability table on page 11 to calculate the p-value. (5 marks)**
   a. p-value = 0.0000
   b. p-value = 0.1788
   c. p-value = 0.2981
   d. **Solution:** p-value = 0.3576
   e. p-value = 0.5962

Section B: Written Answers (60 Marks)

For Section B questions, you must write your response to the question. Page 10 includes SAS output which may prove useful to answering parts of Question 9.

9. **(40 marks)**

Consider data published in the 1950s on a case-control study investigating the relationship between smoking and lung cancer. A breakdown of lung cancer by smoker status (where smokers are classified as those smoking at least 1 cigarette per day for a year) and reported sex of the individual is presented in the partial contingency tables below.

		Smoker status	
Sex	Has lung cancer?	Smoker	Non-smoker
Male	Yes	647	2
	No	622	27
Female	Yes	41	19
	No	28	32

a. **Estimate the conditional associations between incidence of lung cancer and smoker status, conditional on reported sex of the individual, using conditional odds ratios (to 4dp).**

Let X denote whether or not the person is a smoker (0 = Non-smoker, 1 = Smoker), Y denote whether or not the person has lung cancer (0 = No, 1 = Yes), and Z denote whether the person is a male (0 = Female, 1 = Male). Then

$$\widehat{\theta}_{XY(1)} = \frac{647 \times 27}{2 \times 622} \approx 14.0426$$
$$\widehat{\theta}_{XY(0)} = \frac{41 \times 32}{19 \times 28} \approx 2.4662$$

b. **Assuming that the conditional associations estimated in part (a) are indicative of the true conditional odds ratios, do reported sex of the individual and smoker status interact in their effect on incidence of lung cancer? Explain why or why not.**

Yes. If we assume that $\theta_{XY(1)} = \widehat{\theta}_{XY(1)}$ and $\theta_{XY(0)} = \widehat{\theta}_{XY(0)}$, then from part (a) we would have $\theta_{XY(1)} \neq \theta_{XY(0)}$, so there is an interaction between sex of the individual and smoker status.

Now consider the marginal table representing the relationship between lung cancer and smoker status, as shown below.

	Smoker status	
Has lung cancer?	Smoker	Non-smoker
Yes	688	21
No	650	59

c. **Using the marginal table, estimate the association between smoker status and lung cancer using the odds ratio (to 4dp). Interpret the estimated odds ratio, and present a corresponding 95% confidence interval (to 4dp).**

$$\widehat{\theta}_{XY} \;=\; \frac{688 \times 59}{21 \times 650} \;=\; 2.9738$$

and the corresponding 95% confidence interval for θ_{XY} is (1.7867, 4.9495):

$$\left(\exp\left(\log\widehat{\theta} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}}\right),\right.$$
$$\left.\exp\left(\log\widehat{\theta} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}}\right)\right)$$
$$\Longrightarrow \left(\exp\left(\log 2.9738 - 1.96 \times \sqrt{\frac{1}{688}+\frac{1}{21}+\frac{1}{650}+\frac{1}{59}}\right),\right.$$
$$\left.\exp\left(\log 2.9738 + 1.96 \times \sqrt{\frac{1}{688}+\frac{1}{21}+\frac{1}{650}+\frac{1}{59}}\right)\right)$$
$$\Longrightarrow (1.7867, 4.9494)$$

Interpreting $\widehat{\theta}_{XY}$, we would estimate that the odds of being a smoker for those who have lung cancer is approximately 2.9738 times that for those who do not have lung cancer.

d. **Using the marginal table, carry out a chi-square test of independence for smoker status and incidence of lung cancer. Relevant SAS output can be found on page 10 and may be used to answer this question (*i.e.*, hand calculations are not required). Be sure to answer the following questions:**

i. **Is a chi-square test of independence appropriate for the data presented in the marginal table? Why or why not?**

Yes. As shown in the SAS output, the smallest expected frequency is 40, so all expected frequencies are in excess of 5.

ii. **What are the hypotheses to be tested?**

$\mathcal{H}_0$: Smoker status and incidence of lung cancer are independent
$\mathcal{H}_1$: Smoker status and incidence of lung cancer are not independent

iii. **What are the Pearson and likelihood ratio chi-square test statistics? What are their distribution under the null hypothesis?**

From the SAS output,

$$\begin{aligned} X^2 &\approx 19.1292 \quad \text{(Pearson)} \\ G^2 &\approx 19.8780 \quad \text{(Likelihood ratio)} \end{aligned}$$

Under the null hypothesis, both X^2 and G^2 are asymptotically distributed χ^2_1.

iv. **What are the p-values corresponding to the Pearson and likelihod ratio chi-square test statistics?**

From the SAS output, the p-value corresponding to both test statistics is

$$p\text{-value} < 0.0001$$

v. **What is your conclusion at the $\alpha = 0.05$ significance level?**

As the p-value is less than the significance level, we reject $\mathcal{H}_0$ and conclude that there is strong evidence to suggest that smoker status and incidence of lung cancer are dependent, as the p-value is much less than the significance level of $\alpha = 0.05$.

e. **Again using the marginal table, now carry out Fisher's exact test to determine if smokers are more likely to have lung cancer than non-smokers. Relevant SAS output can be found on page 10 and may be used to answer this question (*i.e.*, hand calculations are not required). Be sure to answer the following questions:**

i. **What are the hypotheses to be tested?**

$$\begin{aligned} \mathcal{H}_0 : \theta &= 1 \\ \mathcal{H}_1 : \theta &> 1 \end{aligned}$$

where θ denotes the odds ratio representing the odds of a smoker having lung cancer relative to the odds of a non-smoker having lung cancer.

ii. **What is the p-value for the test? Clearly explain what row in the SAS output provides relevant information for this p-value.**

From the "Right-sided Pr >= F" row of the "Fisher's Exact Test" table, we see that the p-value is given by

$$p\text{-value} \quad < \quad 0.0001$$

iii. **Although it would be possible to calculate a mid-p-value in theory, explain why a mid-p-value would be unnecessary in this case. (Note: You are being asked a conceptual question, not to try to calculate a mid-p-value.)**

The mid-p-value is used to address issues with discreteness in p-values. As the sample size increases, the probability associated with individual tables tends toward 0, so the p-value and mid-p-value converge, and the p-value becomes closer to continuous. In the case we are considering, the samples size ($n = 1418$) is so large that issues with the discreteness in p-values are essentially nullified, meaning that there is no real benefit provided by the mid-p-value.

iv. **What conclusion would you make at the $\alpha = 0.05$ significance level?**

As the p-value is considerably less than the significance level, we reject $\mathcal{H}_0$ and conclude that there is strong evidence to suggest that $\theta > 1$, meaning that smokers are more likely to have lung cancer than non-smokers.

10. **(20 marks)**

The Otago region has a rabbit problem, and farmers are interested to know whether or not the distribution of rabbit holes in the region is random. A researcher took a random sample of 30 areas (each 100 m^2) and calculated an average of 1.6 rabbit holes per area. The frequency distribution for the number of rabbit holes per 100 m^2 is given below.

Number of rabbit holes (r)	Frequency (f_r)	$P(Y = r)$	$\widehat{f}_r$
0	6	0.2019	$\widehat{f}_0$
1	6	0.32303	9.6909
2	12	0.25843	7.7529
≥ 3	6	$P(Y \geq 3)$	6.4992

It is of interest to test the hypotheses

$\mathcal{H}_0$: The population distribution is Poisson.
$\mathcal{H}_1$: The population distribution is not Poisson.

using a chi-square goodness-of-fit test.

a. **What are the missing values for the probability (to 5dp) and expected frequency (to 4dp) corresponding to the black cells in the above table?**

$$\begin{aligned} \widehat{f}_0 &= 6.057 \\ P(Y \geq 3) &= 0.21664 \end{aligned}$$

b. **Calculate the test statistic (to 4dp).**

$$\begin{aligned} X^2 &= \frac{(6-6.057)^2}{6.057} + \frac{(6-9.6909)^2}{9.6909} + \frac{(12-7.7529)^2}{7.7529} + \frac{(6-6.4992)^2}{6.4992} \\ &= 3.7712 \end{aligned}$$

c. **Name the probability distribution that the test statistic follows under the null hypothesis. Explain why this distribution has two degrees of freedom.**

X^2 follows a chi-square distribution with 2 degrees of freedom. There are four frequencies that are estimated corresponding to outcomes of 0, 1, 2, and ≥ 3 rabbit holes. We lose one degree of freedom from estimating the rate parameter λ for the Poisson distribution (the researcher estimated $\widehat{\lambda} = 1.6$ rabbits per area) and a second degree of freedom due to the constraint that the frequencies must sum to the sample size (*i.e.*, one frequency is determined once the other three are known).

d. **The p-value for the test is 0.1517 (to 4dp). State what you would conclude at the $\alpha = 0.05$ significance level.**

As the p-value exceeds the significance level, we have insufficient evidence to reject the null hypothesis. We therefore conclude that the data are consistent with a Poisson population distribution.

Table Analysis

Results

The FREQ Procedure

Table of HAS_LUNG_CANCER by SMOKER_STATUS				
		SMOKER_STATUS		
		Smoker	Non_smoker	Total
HAS_LUNG_CANCER				
Yes	Frequency	688	21	709
	Expected	669	40	
	Col Pct	51.42	26.25	
No	Frequency	650	59	709
	Expected	669	40	
	Col Pct	48.58	73.75	
Total	Frequency	1338	80	1418

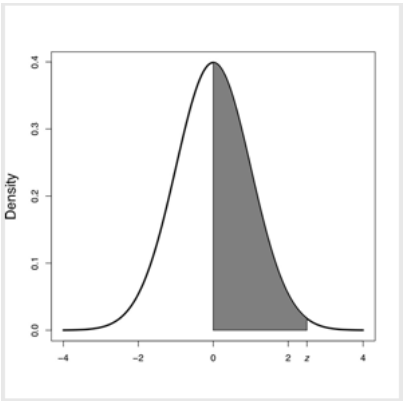
Statistics for Table of HAS_LUNG_CANCER by SMOKER_STATUS

Statistic	DF	Value	Prob
Chi-Square	1	19.1292	<.0001
Likelihood Ratio Chi-Square	1	19.8780	<.0001
Continuity Adj. Chi-Square	1	18.1357	<.0001
Mantel-Haenszel Chi-Square	1	19.1157	<.0001
Phi Coefficient		0.1161	
Contingency Coefficient		0.1154	
Cramer's V		0.1161	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	688
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	<.0001
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Sample Size = 1418

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