Solutions to STAT 300 Midterm Exam

Problem 1 (40 pts, 4 pts each).

Version I: 1. C 2. B 3. C 4. B 5. A 6. C 7. C 8. C 9. D 10. C (or D).

Version II: 1. B 2. C 3. B 4. A 5. C 6. C 7. C 8. D 9. C (or D) 10. C.

Problem 2 (27 pts). A drug company tested two new pain relief drugs for headache sufferers. They randomly selected 18 patients who suffered from headaches and randomly assigned them into three groups of 6 each, with 12 patients taking one of the two drugs and 6 patients taking a placebo (one that does not have medicinal ingredients). After the experiment, each subject was asked to report his/her pain on a scale of 1 to 10, with 10 being most pain.

(a) (7 pts, 1 pt for the first 5 answers) Suggest **three** methods to compare the effectiveness of the two drugs (write the names of the methods below).

two-sample t-test, Wilcoxon rank sum test, permutation test

For these three methods, **the Wilcoxon rank sum test (or the permutation test)** may be most reliable/appropriate for this dataset. The **two-sample t-test** may be most powerful if all required assumptions hold, because **it uses most assumptions/information (independence, normality, equal variance** (2pts).

(b) (2 pts) To compare the two new drugs, if the ranks of the pain scores in one group are

$$1,\ 3,\ 4,\ 6,\ 8,\ 10$$

in the combined sample of the two drug groups, the test Wilcoxon rank sum test may be used, and the value of the test statistic is given by **32**.

(c) (10 pts) To compare all three groups, the company performed a one-way ANOVA and obtained the following table:

Source	df	SS	MS	Test Statistic	P-value
Between group (drug)	2	28.22	–	11.91	<0.0001
Within group (residual)	24	28.44	1.19		

(i) The value of the between group MS (mean square) is **14.11** (1 pt). Under the null hypothesis, the test statistic follows a **F(2, 24)** distribution (2 pts), because **it is the ratio of two χ^2 variates with degrees of freedom of 2 and 24 respectively** (2 pts).
(ii) Assuming equal variance, the value of the common variance is estimated to be **1.19** (2 pts).
(iii) To perform pair-wise comparisons at 5% level, the significance level for each two-sample comparison should be **0.05/3** (1 pt), because **there are 3 simultaneous tests so the significance level (or type I error probability) for each test must be adjusted.** (2 pts).

(d) (4 pts) Suppose that each patient in the third group (placebo group) took both drugs at different times (instead of taking a placebo). To compare effectiveness of the two drugs for this group, suggest two methods: **paired t-test, sign test** (2 pts). Reason: **the two measurements from each patient are *correlated (not independent)*** (2 pts).

(e) (4 pts) If the effectiveness of a drug depends on the gender of the patients, we say that there is an **interaction** (1 pt). Does it affect the significance of the drug? **Yes** (1 pt). In this case, what should we do in data analysis? **Evaluate the drug effects for male and female patients separately** (2 pts).

Problem 3 (21 pts). In an air pollution study, a random sample of 20 households were selected from each of two communities A and B. A respondent in each household was asked whether or not anyone in the household was bothered by air pollution. Here are the collected data:

		Community (Y)		
		A	B	Total
Bothered by air pollution (X)?	Yes	4	8	12
	No	16	12	28
	Total	20	20	40

A researcher wishes to know if people in the two communities are equally bothered by air pollution.

(a) (2 pts) Write the hypotheses in words

H_0**: the two communities are equally bothered by air pollution (or X and Y are independent).**

H_1**: the two communities are not equally bothered by air pollution (or X and Y are not independent).**

(b) (5 pts) Suggest **two** methods to perform the test for H_0 versus H_1:

Chi-square test, Fisher's exact test (2 pts)

For these two methods, **the Fisher's exact test** (1 pt) may be more reliable for this dataset, because **the cell count 4 is less than 5 (or the sample size is small)** (2 pts).

(c) (3 pts) If X and Y are independent, the expected cell count for the first cell (i.e., the cell with count 4) is (2 pts)

$$\frac{\mathbf{12}}{\mathbf{40}} \times \frac{\mathbf{20}}{\mathbf{40}} \times \mathbf{40} = \mathbf{6}.$$

and the difference between the expected cell count and the observed cell count for the first cell is **2** (1 pt).

(d) (5 pts) If the sum of the differences between all the observed cell counts and the corresponding expected cell counts is large, it suggests that the **alternative hypothesis** (1 pt) is more likely to hold. To determine if the sum of the differences is large or not, we can compare the test statistic (write a formula, 1pt)

$$T = \sum_{i,j} (o_{ij} - e_{ij})^2 / e_{ij}$$

to the 95% percentile of the χ^2_1 distribution approximately (2 pt), and we reject H_0 if $T > \chi^2_1(0.95)$ **(or T exceeds the 95% percentile)**.

(e) (2 pts) If there are three communities (instead of two), assuming a large sample, the null distribution is χ^2_2.

(f) (4pts) Suppose that we also wish to test if there are equal number of individuals who are bothered by and not bothered by air pollution, ignoring which community they are from. We decide to reject the null hypothesis of equal number if 25 or more individuals (out of the total 40) are not bothered by air pollution. The power of the test when in fact 60% individuals are not bothered by air pollution is given by (Show the key steps. No need to compute the final answer):

Let X and p be number and proportion of individuals who are not bothered by air pollution respectively. The desired power is

$$P(X \geq 25 | p = 0.60) = \sum_{k=25}^{40} C_{40}^k 0.60^k 0.40^{40-k},$$

which can be computed by software.

Problem 4 (12 points). Referring to Problem 3. Suppose that we also wish to test if there are equal numbers of individuals who are bothered by and not bothered by air pollution, ignoring which communities they are from. We decide to reject the null hypothesis of equal number if 25 or more individuals (out of the total 40) are not bothered by air pollution.
(a) (6 pts) Under the null hypothesis, the test statistics follows a **Binomial (40, 0.5) distribution** (2 pts). Is a large sample required for the null distribution of the test statistic to be reasonably accurate? Answer: **No** (1 pt). We can also approximate the null distribution of the test statistic by a **normal distribution** (1 pt) with parameter(s) given by $\mu = 40 \times 0.5 = 20, \sigma^2 = 40 \times 0.5^2 = 10$ (2 pts).

(b) (6 pts) If we wish a power of at least 80% if in fact 60% individuals are not bothered by air pollution, (i) use an exact method to compute required the sample size, and (ii) use an approximate method to compute the required sample size.

Let n be the required sample size. Note that $25/40 = 0.625$. An exact solution can be found by solving the following equation numerically (3 pts)

$$P(\frac{X}{n} \geq \frac{25}{40} \mid p = 0.60) = \sum_{k=[0.625n]}^{n} C_n^k 0.60^k 0.40^{n-k} = 0.80.$$

Note that $B(n, 0.6)$ can be approximated by $N(0.6n, 0.24n)$. An approximate answer can be obtained by using the following normal approximation (3 pts)

$$P(\frac{X}{n} \geq \frac{25}{40} \mid p = 0.60) \approx P(Z \geq \frac{0.625n - 0.6n}{\sqrt{0.24n}}) = 0.80,$$

where $Z \sim N(0, 1)$.