Solutions to STAT 300 Sample Midterm Exam

Problem 1 (11 pts). Circle the answer which is the *closest* to the correct answer (choose one answer only, unless indicated otherwise).

1. (3 pts) In a hypothesis testing, the power of the test does <u>*not*</u> depend on

   (a) the specific alternative

   (b) the sample size

   (c) *** the p-value**

   (d) the test statistic being used.

2. (3 pts) The Chi-Squared test may perform poorly if

   (a) the model does not hold

   (b) *** the sample size is too small**

   (c) the conditions for the central limit theory do not hold

   (d) none of the above

3. (5 pts) A new medication designed to reduce fever is being tested for efficacy and side effects. For convenience, we call this new medication Drug X. The researcher wants to test whether Drug X is more effective than the currently existing Drug Y. The experiment enrolls 1200 patients (600 men and 600 women) with high fever. The primary outcome measure is the drop in body temperature 3 hour after taking the treatment. The 1200 patients were first divided in two groups based on gender, then subjects within each group are randomly assigned to the two treatment groups (Drug X and Drug Y). The study included which of the following (*select all that apply. You will lose 1 pt for each wrong answer*):

   (a) Blinding

   (b) ***Randomization***

   (c) ***Blocking***

   (d) A placebo

   (e) Missing values

Problem 2 (10 pts). A soft drink company has invented a new drink, and would like to find out if it will be as popular as the existing favorite drink. For this purpose, its research department arranges 7 participants for taste testing. Each participant tries the new drink and rates it on a 5-point scale (1= terrible, ..., 5 = excellent). The ratings are:

2 5 3 4 1 4 5

The company will sell the new drink if the median rating is at least 3. Fill the following blanks as appropriate:

(a) (4 pts) Under the null hypothesis, the test statistic follows a *Bin(n=7, p=0.5)* distribution (clearly specify the parameter values or degrees of freedom as appropriate).

(b) (3 pts) Is the hypothesis test one-sided or two-sided? Answer: *One-sided*

(c) (3 pts) Is a large sample required for the null distribution of the test statistic to be reasonably accurate? Answer: *No* (Yes or No).

Problem 3 (17 pts). Creed *et al.* describe a study in which patients with severe irritable bowel syndrome (IBS) were randomly allocated to one of three treatment groups. One group received eight hours of psychotherapy, one group was placed on a course of the antidepressant Paroxeline while the third received the standard treatment (routine care by a gastroenterologist or general practitioner). Subjects were assessed at the start of the trial and one year after the end of treatment. One outcome measured was the number of days with "restricted activity" in the year following the treatment. Suppose a sample of the data for this response were as below:

Psychotherapy	122	20	125	67	99	
Paroxeline	100	180	75	127	118	222
Standard	54	216	127	208	166	355

1. (3 pts) For this study, which of the following methods could be the **most** appropriate to test a suitable null hypothesis using the data?

   (a) Two–sample t-test

   (b) Wilcoxon Rank Sum test

   (c) ***Kruskal-Wallis test***

   (d) Signed Rank test

   (e) ANOVA

2. For the test you have selected:

   (a) (3 pts) State the null hypothesis clearly in words (in one sentence).

   The distributions of the number of days with "restricted activity" (in the year following the treatment) are the same for the three treatment groups.

   (b) (3 pts) State the alternative hypothesis clearly in words (in one sentence).

   At least one of the distributions of the number of days with "restricted activity" (in the year following the treatment) is different for the three treatment groups.

   (c) (4 pts) Under the null hypothesis, the test statistic follows a χ^2_2 distribution (clearly specify the parameter values or degrees of freedom as appropriate).

   (d) (4 pts) Suppose the test statistic takes the value a, say. With the aid of a rough sketch of the distribution from (c), illustrate how the p–value would be found for the test. (Make sure to shade the appropriate area and label the axes.)

   Figure adds here

Problem 4 (10 pts). Many statistical methods assume that the data follow Normal distributions. In practice, this Normality assumption should be checked before applying the methods.

1. (6 pts) Name **three** statistical methods/models which rely on Normality assumptions for the data. (If you name more than three, you lose two points for each wrong answer).

   Any three of these methods/models: one-sample t-test, two-sample t-test, ANOVA, test of proportions using normal approximation, linear regression.

2. (4 pts) Suggest a method to check whether data follow a Normal distribution, and use *one sentence* to briefly describe the method.

Any one of the following methods:
(a) Normal QQ plot: plot the quantiles of data against quantiles of N(0,1), and a near straighline is expected if normality holds.
(b) Normal probability plot: plots the ordered data against an ordered sample (with equal size) from the standard normal N(0,1), and a near straighline is expected if normality holds.
(c) Chi-squared test: divide data into intervals, count the number of observations in each interval, and then compare the counts with the expected counts if normality holds.
Note: a histogram cannot really be used, since many other distributions such as t-distribution are also symmetric.

Problem 5 (17 pts). To investigate a possible association between two types of chemotherapy drugs and the recurrence of a certain type of cancer, a total of 400 cancer patients were randomly selected for a study. Half of the sample received treatment with Drug A, the remaining half was treated with Drug B. After one year of follow-up for all individuals, the disease had recurred in 25 people. Based on the collected data, individuals treated with Drug A were four times as likely to experience recurrence compared to patients treated with Drug B.

1. (6 pts) Use the information to complete a contingency table, clearly labelling your rows and columns and providing marginal totals.

	Recurrence		
	Yes	No	Total
Drug A	20	180	200
Drug B	5	195	200
Total	25	375	400

or

	Drug A	Drug B	Total
Recurred	20	5	25
Not recurred	180	195	375
Total	200	200	400

2. (3 pts) State the null hypothesis clearly in words (in one sentence).

   There is no association between the two types of chemotherapy drugs and the recurrence of a the type of cancer.

3. (4 pts) Under the null hypothesis, the test statistic follows a χ^2_1 distribution approximately (clearly specify the parameter values or degrees of freedom as appropriate).

4. (4 pts) After computing expected counts for all four cells, a test statistic with value 1.82 is obtained. The 95-th percentile of the null distribution of this test statistic is 3.84. What conclusion do you draw?

   Do not reject the null hypothesis.

   Explain your conclusion in words, in the context of this problem.

   There is not enough evidence to conclude that there is an association between the two types of chemotherapy drugs and the recurrence of a the type of cancer.

Problem 6 (10 pts). Suppose that the cancer recurrence under drug C is 10% based on previous studies. A researcher wishes to check if the cancer recurrence rate has increased in recent years. He designed a new experiment, collected a sample of 20 patients, and decided to reject previous conclusion (i.e., 10% recurrence rate) if the observed recurrence percentage is more than 20%.

Note: In both problems below, you only need to write down the mathematical equations to calculate the desired probability or the sample size n without actually solving the equation. (i.e., you only need to show the methods, and you do not need to compute the final answers. Also, either an approximate method or an exact method would be fine.)

1. (5 pts) What is the type I error probability based on the decision rule the researcher uses? You only need to show the method without doing the calculations.

   Let p be the proportion of cancer recurrence under drug C. We are testing

   $$H_0 : p = 0.1 \qquad vs \qquad H_1 : p > 0.1.$$

   Let X be the number of patients who have experienced cancer recurrence in the sample of $n = 20$ patients. Then, under H_0, we have

   $$X \sim B(20, 0.1).$$

   Type I error probability is given by

   $$P(X/20 > 0.2|p = 0.1) = P(X > 4|p = 0.1) = \sum_{k=5}^{20} C_{20}^k 0.1^k 0.9^{20-k} = 1 - \sum_{k=0}^{4} C_{20}^k 0.1^k 0.9^{20-k}.$$

2. (5pts) The researcher wishes to have 80% power to detect a possible 15% recurrence rate. How many data should he collect (i.e., what should be the sample size n)? Again, you only need to show the method without doing calculations.

   Power $= P(X/n > 0.2|p = 0.15) = 0.8$. *So, solve the following equation for n*

   $$\sum_{k=[0.2n]}^{n} C_n^k 0.15^k 0.85^{n-k} = 0.8,$$

   where $[0.2n]$ is the smallest integer which is greater than $0.2n$. This equation can be solved numerically or by computer simulation. An approximate answer is fine.