IGP 304 (Spring 2006) Midterm Exam

Problems 1-4 (15 points each): Multiple choice questions. Choose a single best answer.

1. For a normal distribution with mean 2 and standard deviation 3, the area between –1 and 5 is about:
   (a) 5% (b) 50% (c) 68% (d) 95%

(c). [For a normal distribution with mean 2 and standard deviation 3, the area between –1 and 5 is from one SD below the mean to one SD above the mean. The area between them is about 68%.]

2. The distribution of diastolic blood pressures can be approximated by normal distributions. The expected value and standard deviation of the distribution of diastolic blood pressures in 35-44 years old men are 80 mm Hg and 12 mm Hg. If a 41-year-old male subject has diastolic blood pressure 92, what percent of men between 35 and 44 have higher diastolic blood pressure than him?
   (a) 5% (b) 50% (c) 32% (d) 16%

(d). [92 is one SD above 80. The question asks for $\Pr(x \geq 92) = \Pr[z \geq (92 - 80)/12] = \Pr[z \geq 1] = 16\%$.]

3. Suppose you go to a hospital to record the sex of the newborn babies. Which is more likely: (i) in the next 100 deliveries, you see 54 or more boys; (ii) in the next 1000 deliveries, you see 540 or more boys?
   (a) (i) (b) (ii) (c) (i) and (ii) have equal chance (d) Need more information to tell.

(i). [The larger the sample size is, the more centralized the results are. In fact, the probability of the former is 0.24 while that of the latter is 0.006.]

4. Consider all possible diastolic blood pressure (DBP) measurements from a mother and her first-born child. Let A = {mother's DBP ≥ 95} and B = {first-born child's DBP ≥ 80}. Suppose Pr(A) = 0.1, Pr(B) = 0.2, and Pr(A and B) = 0.05. Then A and B are
   (a) independent (b) dependent (c) Need more information to tell

(b). [This is because $\Pr(A \text{ and } B) \neq \Pr(A) \times \Pr(B)$.]

Problems 5-12 (15 points each): True/false questions: Select true or false. If false, state the correct result/statement/conclusion that is beyond just grammatically negating a false statement.

5. Confidence interval is a way of estimating an effect. It often gives more information than a point estimate.
   (a) True (b) False

True. [This is why a confidence interval is also called an interval estimate.]

6. In a study on systolic blood pressure of infants, the researchers measured birthweight and the age (in days) when the blood pressure was measured. They fit a linear regression with systolic blood pressure as the response variable and both birthweight and age as input variables. In the result table, the t-test for age is not significant at level 0.05. This suggests age is not an important factor on infant's systolic blood pressure.
   (a) True (b) False

False. The result suggests that age is not significant towards explaining SBP *after* birthweight has been taken into account.

7. Correlation coefficient can be used to quantify the level of relatedness between two variables. A correlation coefficient near 1 indicates strong correlation and a correlation coefficient near –1 indicates weak correlation.
   (a) True (b) False

False. A correlation coefficient near 1 indicates strong (positive) correlation; one near –1 indicates strong (negative) correlation; one near 0 indicates weak correlation.

8. In 1998 the UK government started a campaign to reduce smoking among teenagers from the then national average of 13% to a lower percentage. In 2001, 1000 teenagers were sampled in a survey and 123 were found to smoke. They carried out a z-test to see if the smoking rate in 2001 was different from 13%. The p-value was 0.51. This means the probability of seeing 123 smokers among 1000 teenagers was 0.51.
   (a) True (b) False

False. A p-value is the probability of observing data as extreme as or more extreme than the real data, not the probability of observing exactly the data, which often is small even when p-value is large.

9. Blood pressure may change as people grow older. Suppose we have measured systolic blood pressures (SBP) on 400 adult men and 400 adult women, and have recorded their ages. We may analyze the relationship between SBP and age in men and women separately. We also may combine the two groups and carry out a single analysis on the relationship between SBP and age. When we combine the groups and do a simple linear regression of SBP on age, we find the p-value for the regression fit is 0.003. This suggests that combining the groups was not appropriate and the gender variable has a significant effect on SBP and should not have been ignored.
   (a) True (b) False

False. The p-value in the simple linear regression tells that when considered alone, age is a significant factor to explain SBP. The analysis doesn't have anything to reflect the effect of combining groups compared to not combining groups.

10. In a study on socioeconomic status and education, the researchers recorded socioeconomic status as four levels and education as four levels. They carried out a Pearson's chi-squared test. The degrees of freedom of the test should be eight.
    (a) True (b) False

False. The DF is $(4 – 1)\times(4 – 1) = 9$.

11. In a study on the effect of a vaccine to prevent influenza infection, the subjects were randomly assigned to two groups: vaccine group and placebo group. After the flu season they collected data on influenza incidences of these subjects. Then risk difference, risk ratio, and odds ratio are different ways of measuring the effect of the vaccine compared to that of the placebo.
    (a) True (b) False

True.

12. In a survey to study the relationship between gender and the risk of having asthma, we can test for the relationship by calculating (i) the odds ratio of having asthma in women and in men, and (ii) the standard error of the odds ratio, and then divide (i) by (ii) to obtain a z-statistic.
    (a) True (b) False

False. The z-statistic is logOR/s.e.(logOR), not OR/s.e.(OR).

Problems 13-14: The cut-off values at significance level 0.05 for various distributions

1.96	standard normal
1.982	t distribution with degrees of freedom between 100 and 110
1.967	t distribution with degrees of freedom between 350 and 360
1.965	t distribution with degrees of freedom between 460 and 470
3.84	chi-squared distribution with 1 degree of freedom
5.99	chi-squared distribution with 2 degrees of freedom
7.81	chi-squared distribution with 3 degrees of freedom
9.49	chi-squared distribution with 4 degrees of freedom

13. (40 points) A study was performed concerning risk factors for carotidartery stenosis (narrowing) among 464 men born in 1914 and residing in the city of Malmö, Sweden. The data reported for blood glucose level are in the following table

	No stenosis (n = 356)		Stenosis (n = 108)	
	Mean	SD	Mean	SD
Blood glucose (mmol/L)	5.3	1.4	5.1	1.3

(a) What test can be performed to see if average blood glucose levels differ between babies with and without stenosis? If there are multiple tests, answer the simplest (but still valid) one.

A two-sample t-test is appropriate.

(b) Is the information enough to calculate the test statistic? If no, what extra information is needed? If yes, carry out the test and state your conclusion on the basis of the test result.

Yes. The information is enough.

The null hypothesis is: The mean blood glucose levels are the same for the two groups.

The statistic is $t = (\bar{x}_1 - \bar{x}_0)/s.e.$, where $\bar{x}_1 - \bar{x}_0 = 5.1 - 5.3 = -0.2$ and the standard error is (EMS p.66)

$$s.e. = \sqrt{\left[\frac{(n_1-1)s_1^2 + (n_0-1)s_0^2}{n_1+n_0-2}\right]\left[\frac{1}{n_1}+\frac{1}{n_0}\right]} = \sqrt{\left[\frac{107\times 1.3^2 + 355\times 1.4^2}{108+356-2}\right]\left[\frac{1}{108}+\frac{1}{356}\right]} = 0.1513\,.$$

Then t = – 0.2/0.1513 = –1.32 with DF = $n_1 + n_0 - 2 = 462$. Since |–1.32| < 1.965, the cut-off value for t_{462} at level 0.05, we don't reject the null hypothesis that the two groups have different means at significance level 0.05.

[Note: If you used the formula on p.62, then $s.e. = \sqrt{s_1^2/n_1 + s_0^2/n_0} = \sqrt{1.3^2/108 + 1.4^2/356} = 0.1454$ and t = – 0.2/0.1454 = –1.38. Since |–1.38| < 1.96, the cut-off value of standard normal distribution at level 0.05, we don't reject the null hypothesis that the two groups have different means at significance level 0.05.]

14. (30 points) In a case-control study, researchers want to investigate if carrying a known genetic variant on a candidate gene has effect on the risk of a disease. They find that 50 of 100 cases and 40 of 200 controls carry the genetic variant. Carry out a Pearson's chi-squared test on the data and state your conclusion.

The null hypothesis is that carrying genetic variant has no effect on disease risk (or, carrying the variant and not carrying the variant have the same effect on disease risk).

The observed data are

	Variant	No variant	Total
Case	50	50	100
Control	40	160	200
Total	90	210	300

The expected counts are

	Variant	No variant	Total
Case	30	70	100
Control	60	140	200
Total	90	210	300

Then the test statistic is $(50 - 30)^2/30 + (50 - 70)^2/70 + (40 - 60)^2/60 + (160 - 140)^2/140 = 28.57 >$ 3.84, the cut-off value of the chi-squared distribution with DF = 1 at level 0.05. Thus, we reject the null hypothesis at significance level 0.05.