Name:
SID:

Statistics 135 Fall 2008
Final Exam

Show your work. The number of points each question is worth is shown at the beginning of the question. There are 10 problems.

1. [2] The normal equations for the least squares estimates are $X^T(Y - X\hat{\beta}) = 0$. Show that if the design matrix X contains a column of 1's, then the average value of the residuals is 0.

   The residuals are $Y - X\hat{\beta}$. If a column of X is all 1's, then from that row in the normal equations, $\sum(Y_i - x_i\hat{\beta}) = 0$ so the average is 0.

2. Recall that the probability mass function of a binomial random variable with m trials and success probability θ is $p(k) = \binom{m}{k}\theta^k(1-\theta)^{m-k}$, $k = 0, 1, \ldots, m$. Suppose that $X_1, X_2, \ldots, X_n$ are independent binomial random variables, with common success probability θ, but with differing numbers of trials, $m_1, m_2, \ldots, m_n$.

   (a) [4] Find the maximum likelihood estimate (MLE) of θ.

   $\text{Lik}(\theta) = \prod_{i=1}^n \binom{m_i}{x_i}\theta^{x_i}(1-\theta)^{m_i - x_i} = C\theta^T(1-\theta)^{N-T}$, where C is the product of the binomial coefficients, $N = \sum m_i$ and $T = \sum x_i$. Maximizing gives $\hat{\theta} = T/N$

   (b) [2] Is the MLE unbiased?

   Since $E(X_i) = m_i\theta$, it follows that $E(T) = \sum m_i\theta = N\theta$, so the MLE is unbiased.

   (c) [2] What is the standard deviation of the MLE? Since $Var(X_i) = m_i\theta(1-\theta)$, $Var(T) = \theta(1-\theta)\sum m_i = N\theta(1-\theta)$, so $Var(\hat{\theta}) = \theta(1-\theta)/N$. The large sample approximation gives the same result.

   (d) [2] What is the probability distribution of the MLE?

   T is binomial(N, θ), so $P(\hat{\theta} = t/N) = P(T = t)$, for $t = 0, 1, \ldots, T$ and $P(T = t)$ is a binomial probability. A normal approximation could also be used.

   (e) [2]. If the prior distribution of θ is uniform, what is the posterior distribution?

   The posterior distribution is proportional to the likelihood times the prior, and the prior equals 1, so the posterior is proportional to the likelihood. That distribution is Beta$(T+1, N-T+1)$

3. This problem arises in particle physics. An experiment is run to estimate the background rate at which particles are produced. The particle count X_1 is measured and it is assumed that $X_1 \sim Poisson(\lambda_b)$, where λ_b, the background rate, is unknown. With the experimental apparatus turned on a count X_2 is measured and it is assumed that $X_2 \sim Poisson(\lambda_b + \lambda_s)$, where the experimental apparatus produces "source" counts which are Poisson(λ_s) independent of

the background particles and what is observed is background counts plus source counts. The counts X_1 and X_2 are thus observed. Explain carefully and clearly how to construct a generalized likelihood ratio test of the null hypothesis that $\lambda_s = 0$, versus the alternative that $\lambda_s > 0$, that is, under the null hypothesis, the experimental source does not produce any particles?

(a) [4] Derive the test statistic.

The likelihood is $\text{Lik}(\lambda_b, \lambda_s) = \frac{\lambda_b^{x_1} e^{-\lambda_b}}{x_1!} \frac{(\lambda_b+\lambda_s)^{x_2} e^{-\lambda_b-\lambda_s}}{x_2!}$. Under H_0, $\lambda_s = 0$.
This has to be maximized under $H_0 : \lambda_s = 0$ and under $H_A : \lambda_s > 0$. Doing so we find that under H_0, $\hat{\lambda_b} = \bar{x}$ and under H_A, $\hat{\lambda_b} = x_1$ and $\hat{\lambda_s} = \max(0, x_2 - x_1)$. Substituting these into to likelihood ratio, and assuming for simplicity that $x_2 \geq x_1$, we find

$$\log \Lambda = x_1 \log(\bar{x}/x_1) + x_2 \log(\bar{x}/x_2)$$

(b) [2]Identify the approximate null distribution.

Under H_0, $-2 \log \Lambda$ approximately follows a chi-square distribution with 1 df

(c) [2]Explain how to form the rejection region for a level α test.

Reject if $-2 \log \Lambda$ is larger than the upper α quantile of the chi-square distribution with 1 df.

4. In a test of sensory perception, each of n panelists is presented with three items to taste, in randomized order. Two of the items are identical (eg Coke) and one is different (eg Pepsi). Each panelist is asked to identify the one that is different. Explain carefully and clearly how you could test the null hypothesis that the items are indistinguishable.

(a) [4]Propose a test statistic.

Let T be the number of successful identifications. (Many students set this up as a chi-square test for a multinomial with either two or three cells, and I gave credit for such solutions.)

(b) [2] What is the null distribution of the test statistic?

Binomial$(n, 1/3)$

(c) [2] Identify the rejection region of the test with significance level α.

The upper tail of the binomial distribution

5. Consider the standard linear model, $Y = X\beta + \epsilon$.

(a) [2]Give expressions for the fitted values and the residuals.

$\hat{Y} = X\hat{\beta}$ where $\hat{\beta} = (X^T X)^{-1} X^T Y$.
$\hat{e} = Y - \hat{Y}$

(b) [2] Derive the covariance matrix of the residuals.

Let $P = X(X^T X)^{-1} X^T$. Then $P = P^2 = P^T$
$\Sigma_{\hat{e}\hat{e}} = \sigma^2(I - P)I(I - P) = \sigma^2(I - P)$

(c) [2] Derive the cross-covariance matrix of Y and the residuals. Are they uncorrelated?

$\Sigma_{Y\hat{e}} = \sigma^2 I(I - P) = \sigma^2(I - P)$, so they are correlated.

6. SAT scores (x) and freshman year GPAs (Y) are recorded for a group of male and female students. A model of the form $Y = \beta_1 I_M + \beta_2 I_F + \beta_3 x$ is to be fit to the data, where $I_M = 1$ if the student is male and 0 if female and $I_F = 1$ if the student is female and 0 if male.

(a) [2] Explain clearly how to write this a a linear model of the form $Y = X\beta + \epsilon$. What is X?

Y is the vector of observations – the i-th entry is the GPA of the i-th person. X is a matrix whose i-th entry first column is 1 if the i-th person is male and 0 if female, and similarly the 2nd column is 0 if male and 1 if female. The 3rd column is the vector of SAT scores.

(b) [2] Explain how to estimate the standard error of $\hat{\beta}_1$.

$s^2 = RSS/(n - 3)$. The estimated standard error is s times the square root of the (1,1) entry of $(X^T X)^{-1}$

(c) [2] Explain how to estimate the variance of $\hat{\beta}_1 - \hat{\beta}_2$.

$\hat{\beta}_1 - \hat{\beta}_2 = [1\ \ -1\ 0]\hat{\beta}$ so the variance of this quantity is $\sigma^2[[1\ \ -1\ 0](X^T X)^{-1}[1\ \ -1\ 0]^T$. This variance is estimate by using s^2 in place of σ^2.

(d) [2] Explain how to test $H_0 : \beta_1 - \beta_2 = 0$.

Refer the statistic $(\hat{\beta}_1 - \hat{\beta}_2)/s_{\hat{\beta}_1 - \hat{\beta}_2}$ to a t distribution with $n - 3$ df.

7. [4]A group of patients who smoked were seen by family practice physicians at a clinic, They were then classified by gender and by whether they had been advised to stop smoking, with results shown in the table below. Is there a relationship between gender and being advised to cease smoking?

	Advised	Not Advised
Male	48	47
Female	80	136

Using the chi-square test of independence, the expected counts are

```
39.09968  55.90032
 88.90032 127.09968
```

and the test statistic is 4.96 with 1 df. The p-value is $.05 < p < .01$

8. Suppose that under the null hypothesis a random variable X has a uniform distribution on $[0, 1]$, and under the alternative it has a density $f(x) = 2x$.

   (a) [2] What is the most powerful test at level $\alpha = 0.10$?

   By the Neyman-Pearson Lemma, the likelihood ratio test is most powerful. $\Lambda = 1/2x$, so the test rejects for large x. For $\alpha = .10$, the critical value is $x_c = 0.9$

   (b) [2] What is the power of the test?

   The cdf under the alternative is $F(x) = x^2$, so the power is $1 - F(.9) = 0.19$

   (c) [2] Is this test uniformly most powerful for testing against the alternative $H_A : f(x) = \lambda x^{\lambda-1}$, for $\lambda > 1$? Why or why not?

   The test is UMP since it is the same for all $\lambda > 1$

9. Suppose independent samples are taken under each of two conditions: $x_1 = 32, x_2 = 11, x_3 = 4$ and $y_1 = 44, y_2 = 15$. Model the X's as being independent realizations from a continuous distribution F_X and the Y's being independent realizations from a continuous distribution F_Y.

   (a) [2] Give an estimate for $P(X < Y)$.

   $\hat{P} = (1/6)\sum_i \sum_j 1(X_i < Y_j) = 5/6$

   (b) [2] Explain how to test $H_0 : P(X < Y) = 1/2$ versus $H_A : P(X < Y) > 1/2$. Identify the test statistic.

   This hypothesis can be tested by the Mann-Whitney test based on the sum of the ranks of the Y's, which is 5+3=8.

   (c) [2] What is the null distribution of the test statistic?

   The null distribution can be computed by considering all possible ways to assign ranks to the Y's. There are 10 possible ways. In this way, the null distribution is found to be: $P(T = 9) = P(T = 8) = 1/10$. $P(T = 7) = P(T = 6) = P(T = 5) = 2/10$. $P(T = 4) = P(T = 3) = 1/10$

   (d) [2] What is the p-value of the test for this data?

   From the probabilities above, $P(T \geq 8) = 2/10$, the p-value

10. A nonparametric test for the slope of a regression line. Suppose that one observes pairs (Y_i, x_i), $i = 1, 2, \ldots, n$ and that $x_1 < x_2 < \cdots < x_n$. For simplicity assume that the Y_i are distinct, that is there are no ties. To test for postive association between the x's and Y's consider the test statistic

$$T = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I(D_j - D_i)$$

where $I(z)$ is 1 if $z > 0$ and equals -1 if $z < 0$.

(a) [2] Consider the null hypothesis that the slope is 0. Explain why large values of T are evidence against the null hypothesis.

If the slope is postive, one would expect that $Y_j - Y_i$ would tend to be postive when $x_j - x_i > 0$. This should give rise to large values of T.

(b) [2] How could the null distribution of T be determined?

Under the null hypothesis, the Y's would not be expected to systematically increase, or decrease. The null distribution can thus be constructed by forming all possible permutations of the Y's, calculating the corresponding values of T, and using the resulting distribution. The test would reject if the observed value of T was in the right tail of this distribution, the critical value being determined by α.