SPECIAL TOPICS 2002

Part I - STA 503 (Total of 50 points)

Please do all of the following problems from STA 503. If you need to go to CB 307 to perform a computer calculation that room is open and available for you to use.

1. **(10 points)** Suppose we have run the following SAS program

```
proc logistic data=test;
  model htattk = bp ;
run;
```

and obtained the output attached to the back of this exam. Assume this estimated model is correct.

   (a) Based on the estimated model, if a subject has a bp value = 155, what is the subject's probability of having heart desease?

   (b) Based on the estimated model, what is the bp value that corresponds to a 0.4 probability of having heart desease?

2. **(5 points)** A survey company plans to do two independent SRS's of size 100 from the same population. (Assume it is a normal population with a known $\sigma = 5$) From the first SRS they can compute a 90% confidence interval for the mean. This confidence interval may or may not contain the $\overline{X}$ from the second sample. What is the probability that it will?

3. **(15 points)** A t-test is often used when the population is normal. When the non-central t-table (or the non-central F table) is not readily available, people sometimes just use the normal distribution in place of the t-distribution. (i.e. they pretend the computed s^2 is the true population σ^2). I call it a *poor-man's* power/sample size calculation. If the sample size involved is not too small, (say at least 40), then the result should not be too far off.

   (a) Use the *poor-man's* method to for the following problem: If a one-sample t-test is to be used to test $H_0 : \mu \leq 1$ vs. $H_A : \mu > 1$ ($\alpha = 5\%$), and a power of at least 83% to detect a mean of at least 0.2 standard deviations above 1 is desired, what is the minimum sample size needed?

   (b) The *poor-man's* method only gives an approximate solution; will the exact solution (use non central t/F table or software) result in a larger n or smaller n? Why?

$RR: \left(\frac{\bar{x}-1}{s/\sqrt{n}} > 1.645\right)$

$P(RR) = P\left(\frac{\bar{x}-1}{s/\sqrt{n}} > 1.645\right) = P\left(\frac{\bar{x}-(1+0.25)}{s/\sqrt{n}} + 0.25 > 1.645\right) = P\left(N(0,1) + 0.3\sqrt{n} > 1.64\right) = 0.83$

(c) Now do the exact calculation (do not use *poor-man's* approximation) to compute the power of the following setup. Based on 12 observations use a t-test to test the hypothesis in (a) with $\alpha = 0.02$, and find the power of the test if the true population mean is 0.7σ over 1 i.e. $\mu = 1 + 0.7\sigma$.

4. **(7 points)** If a one sample Z-test (similar to t-test except the population variance is known) with a known variance of 4 was used to test $H_0 : \mu = -1$ vs. $H_A : \mu \neq -1$ based on a sample of size 49, what are the μ values that this test has at least 85% power to detect? ($\alpha = 0.05$). What if we use $\alpha = 0.02$?

5. **(5 points)** In testing a hypothesis $H_0 : \theta = 2$; $H_A : \theta \neq 2$, based on the same data, two test procedures gave two different P-values.

   (a) Suppose both test are valid tests in this situation and suppose we know that the alternative hypothesis is in fact true. Can we conclude that the test that gave a smaller P-value is a more powerful test? Why or why not?

   (b) Suppose we knew that the null hypothesis is true. Can we conclude that the test that gave a larger P-value is more powerful? Why or why not?

6. **(8 points)** In a two-way anova model of treatment and block (with no interaction), the means (μ_{ij}), in the model for some of the combinations are given in the following table.

   (a) Please fill the rest of the $\mu'_{ij}s$ in a way that make it satisfy a two-way anova model with no interaction.

		Trt 1	Trt 2	Trt3
	Block 1		97	95
Model Means	Block 2	93		
	Block 3	106		

   (b) Repeat (a) but now suppose the overall mean (grand mean) is 100.

		Trt 1	Trt 2	Trt3
	Block 1		97	95
Model Means	Block 2	93		
	Block 3	106		

Part II - STA 603 (Total of 50 points)

Please choose **two** of the following problems to work. Use only one side of your paper. If you answer more than two problems, the first two encountered on your answer sheets will be graded. It is important that you *do as much of each chosen problem as you can* in the time allotted.

1. (**25 points**) Assume $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\mathcal{E}(\boldsymbol{\epsilon}) = \mathbf{0}$, $\mathcal{D}(\mathbf{Y}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 1 & -1 \end{pmatrix}$, and $\mathbf{Y} = \begin{pmatrix} -1 \\ 3 \\ 2 \end{pmatrix}$. Please do each of the following:

   (a) (**10 points**) Find the *BLUE* estimate for $\boldsymbol{\beta}$.

   (b) (**7 points**) We know that any unbiased estimator, including the *BLUE* in part (a) can be derived from a complement to the range of $\mathbf{X}$ (say $\mathbf{W}$) by way of the two-step process we constructed in class. Recall our logic:

   Step 1 : given $\mathbf{v} \in \mathbf{R}^n$, uniquely decompose $\mathbf{v}$ as $\mathbf{u} + \mathbf{w}$, where $\mathbf{u}$ in $Rg(\mathbf{X})$ and $\mathbf{w} \in \mathbf{W}$.
   Step 2 : take any left inverse of $\mathbf{X}$, say $\mathbf{B}$, and *define* $\mathbf{Av} = \mathbf{Bu}$.

   Find a basis for the complement of $Rg(\mathbf{X})$ that would produce the *BLUE* estimator in (a) by way of this process. Make sure you show it is a basis.

   (c) (**8 points**) Let $\widehat{\boldsymbol{\beta}}^* = \mathbf{Ay}$, where $\mathbf{A} = \begin{pmatrix} 1 & -1 & 1 \\ 0 & 3 & 0 \end{pmatrix}$. Exhibit an estimator that is comparable to $\widehat{\boldsymbol{\beta}}^*$ under " $\leq$ ", and in fact better, but which is not the *BLUE* estimate from part (a).

2. (**25 points**) Assume a balanced two-way ANOVA model $\mathbf{y} = \mathbf{X}\boldsymbol{\mu} + \boldsymbol{\epsilon}$ with $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Suppose factor A has "a" levels and factor B has "b" levels, and suppose the common cell size is denoted by n. Consider the following hypothesis about contrasts in $\boldsymbol{\mu}_{abx1}$:

$$\mathbf{H}_0 : \sum_{i=1}^{a}\sum_{j=1}^{b} \alpha_{ij}\mu_{ij} = 0 \text{ and } \sum_{i=1}^{a}\sum_{j=1}^{b} \beta_{ij}\mu_{ij} = 0$$

where $\sum_{i=1}^{a}\sum_{j=1}^{b} \alpha_{ij}\beta_{ij} = 0$ and $\sum_{i=1}^{a}\sum_{j=1}^{b} \alpha_{ij}^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} \beta_{ij}^2 = 1$.

   (a) (**10 points**) Show that $\left\{ \begin{pmatrix} \alpha_{11}\mathbf{1}_{nx1} \\ \alpha_{12}\mathbf{1}_{nx1} \\ \vdots \\ \alpha_{1b}\mathbf{1}_{nx1} \\ \alpha_{21}\mathbf{1}_{nx1} \\ \vdots \\ \alpha_{ab}\mathbf{1}_{nx1} \end{pmatrix}, \begin{pmatrix} \beta_{11}\mathbf{1}_{nx1} \\ \beta_{12}\mathbf{1}_{nx1} \\ \vdots \\ \beta_{1b}\mathbf{1}_{nx1} \\ \beta_{21}\mathbf{1}_{nx1} \\ \vdots \\ \beta_{ab}\mathbf{1}_{nx1} \end{pmatrix} \right\}$ is an orthogonal basis for the G-space corresponding to this hypothesis, where $\mathbf{1}_{nx1}$ is the $nx1$ vector of all ones.

(b) **(8 points)** Show that the coordinates of the projection of $\mathbf{y}$ onto this G-space with respect to the basis in part (a) are given by $\begin{pmatrix} \sum_{i=1}^{a}\sum_{j=1}^{b} \alpha_{ij}\overline{y}_{ij} \\ \sum_{i=1}^{a}\sum_{j=1}^{b} \beta_{ij}\overline{y}_{ij} \end{pmatrix}$, where $\overline{y}_{ij}$ is the sample mean of the ij^{th} cell.

(c) **(7 points)** Briefly argue that $\frac{\|\mathbf{P}_G\mathbf{y}\|^2}{\sigma^2}$ has a non-central chi-squared distribution. You *do not* have to reprove any results we have from class.

3. **(25 points)** The Fisher-Cochran Theorem (mentioned briefly in class) is important in experimental design and many of you will see it in STA 643 this fall. This theorem says the following: Let $\mathbf{y}_{nx1}$ be destributed $N(\boldsymbol{\mu}, \mathbf{I})$ and suppose $\mathbf{y}^t\mathbf{y} = Q_1 + Q_2 + ...Q_k$, where $Q_i = \mathbf{y}^t\mathbf{A}_i\mathbf{y}$ with the rank of $\mathbf{A}_i$ denoted by r_i. Then a necessary and sufficient condition for quadratic forms Q_i, $i = 1, .., k$ to be independent and each be distributed as $\chi^2(n_i, \lambda_i)$ is simply that $\sum_{i=1}^{k} n_i = n$. In this problem you are asked to proved three relatively easy results that consititute most (not all) of a proof to Fisher-Cochran. You *do not* have to reprove any results from class that you would like to use. You *do not* have to prove the Fisher-Cochran Theorem.

   (a) **(5 points)** Suppose $\mathbf{W}$ is symmetric and $\mathbf{y}^t\mathbf{y} = \mathbf{y}^t\mathbf{W}\mathbf{y}$, for all $\mathbf{y}$ in $\mathbf{R}^n$. Show that $\mathbf{W} = \mathbf{I}_{nxn}$.

   (b) **(8 points)** Suppose $\mathbf{A}_1$and $\mathbf{A}_2$ are symmetric and idempotent with $\mathbf{y}^t\mathbf{y} = \mathbf{y}^t\mathbf{A}_1\mathbf{y} + \mathbf{y}^t\mathbf{A}_2\mathbf{y}$, for all $\mathbf{y}$ in $\mathbf{R}^n$. Show that the $rank(\mathbf{A}_1) + rank(\mathbf{A}_2) = n$.

   (c) **(12 points)** Now suppose $\mathbf{y}_{nx1}$ is destributed $N(\mathbf{0}, \mathbf{I})$ and assume that for all $\mathbf{y}$, $\mathbf{y}^t\mathbf{y} = \mathbf{y}^t\mathbf{A}_1\mathbf{y} + \mathbf{y}^t\mathbf{A}_2\mathbf{y}$, with $\mathbf{A}_1$ and $\mathbf{A}_2$ as in part b). Prove that $\mathbf{y}^t\mathbf{A}_1\mathbf{y}$ is independent of $\mathbf{y}^t\mathbf{A}_2\mathbf{y}$ and that $\mathbf{y}^t\mathbf{A}_i\mathbf{y}$ has a central χ^2 distribution on n_i degrees of freedom, where n_i =rank($\mathbf{A}_i$).

Code and Output for STA 503 Prob. 1

```
data test;
input bp htattk $;
cards;
112  YES
122  NO
132  NO
142  NO
152  YES
162  NO
177  YES
197  YES
[redacted]
124  NO
120  YES
155  YES
119  NO
135  YES
[redacted]
132  YES
[redacted]
;
run;
proc logistic data = test;
class htattk ;
model htattk = bp;
run;
```

The LOGISTIC Procedure

Model Information

Data Set	WORK.TEST
Response Variable	htattk
Number of Response Levels	2
Number of Observations	17
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	htattk	Total Frequency
1	NO	7
2	YES	10

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	25.035	24.981
SC	25.868	26.648
-2 Log L	23.035	20.981

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2.0535	1	0.1519
Score	1.8696	1	0.1715
Wald	1.6521	1	0.1987

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	4.4706	3.7240	1.4412	0.2299
bp	1	-0.0343	0.0267	1.6521	0.1987

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
bp	0.966	0.917	1.018

Association of Predicted Probabilities and Observed Responses

Percent Concordant	64.3	Somers' D	0.329
Percent Discordant	31.4	Gamma	0.343
Percent Tied	4.3	Tau-a	0.169
Pairs	70	c	0.664