

GROUND RULES:

- This exam contains 5 questions; each question is worth 20 points. Subpart point totals are given in []. The maximum number of points on this exam is 100.
- **All answers** (even Question 5: True/False) **are to be written in the accompanying blue Examination Book**. No accompanying work will be graded.
- You may use a calculator if you wish, but show all of your work and explain all of your reasoning!!!
- Tabled values for the standard normal, t , χ^2 , and F distributions are provided.
- Summary formulae for the continuous and discrete probability models and the STAT 512 “Distributional Results” handout are also provided.
- Any discussion or otherwise inappropriate communication between examinees, as well as the appearance of any unnecessary material, will be dealt with severely.
- Print your name at the top of this page in the upper right hand corner. Also, fill out the blue Examination Book front page information completely.
- You have 3 hours to complete this exam. GOOD LUCK!

1. Suppose that Y_1, Y_2, \dots, Y_n is an iid sample of Bernoulli observations with mean p , $0 < p < 1$, p fixed. In case you have forgotten, the Bernoulli(p) probability mass function is given by

$$f_Y(y; p) = \begin{cases} p^y(1-p)^{1-y}, & y = 0, 1 \\ 0, & \text{otherwise.} \end{cases}$$

(a) [10] Suppose that the goal is to test,

$$\begin{aligned} H_0 : p &= p_0 \\ &\text{versus} \\ H_a : p &> p_0. \end{aligned}$$

Show that the **uniformly most powerful (UMP)** level α rejection region has the form

$$\text{RR} = \left\{ \mathbf{y} : \sum_{i=1}^n y_i \geq k^* \right\},$$

where k^* is chosen so that the test is of level α . Write out an equation that describes how k^* would be chosen. Make sure to explain why is this rejection region is UMP. **Do not use any type of normal approximation on this part!**

(b) [7] We have shown in class that the test in part (a) can be performed by using a normal approximation to the sampling distribution of

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n Y_i.$$

In particular, we showed that an approximate level α rejection region has the form

$$\text{RR} = \{z : z > z_\alpha\},$$

where

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

and z_α is the upper α quantile of the standard normal distribution. Using this approximate level α rejection region, derive $K(p)$, the **power function** for the test. You should be able to express the power function in terms of the $\mathcal{N}(0, 1)$ cumulative distribution function.

(c) [3] The rejection region in part (a) is based on the **exact** distribution of the sufficient statistic

$$U = \sum_{i=1}^n Y_i.$$

The rejection in part (b) is based on the **asymptotic** distribution of $\hat{p} = U/n$. In a few sentences, explain (to an investigator) what the difference is between exact statistical procedures and procedures based on asymptotic theory.

2. Conditional on the value of θ , suppose that Y_1, Y_2, \dots, Y_n is an iid sample from

$$f_Y(y; \theta) = \begin{cases} \theta^2 y e^{-\theta y}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

In turn, θ is best regarded as a random variable with prior distribution

$$g(\theta) = \begin{cases} e^{-\theta}, & \theta > 0 \\ 0, & \text{otherwise.} \end{cases}$$

(a) [10] Turn the (proverbial) Bayesian crank to show that the posterior distribution of θ is $\text{gamma}(\alpha^*, \beta^*)$, where $\alpha^* = 2n + 1$ and $\beta^* = (u + 1)^{-1}$, where $u = \sum_{i=1}^n y_i$.

(b) [4] Let (θ_L, θ_U) denote an equal tail 95 percent credible interval for θ . Write out integral equations that show how θ_L and θ_U are computed. You don't have to evaluate the integrals.

(c) [6] I generated an iid sample from $f_Y(y; \theta)$; here are the data

2.10 0.45 4.12 3.84 1.69

With these data, report the posterior mean and posterior mode of θ . Recall that posterior distribution is given in part (a).

3. Consider our simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where $\epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2)$, for $i = 1, 2, \dots, n$. Recall that in this model, the x_i 's are regarded as fixed constants measured without error.

(a) [5] Show that $E(\bar{Y}) = \beta_0 + \beta_1 \bar{x}$ and $V(\bar{Y}) = \sigma^2/n$. The statistic \bar{Y} is the sample mean of Y_1, Y_2, \dots, Y_n .

(b) [8] Recall that the least squares estimator of β_1 is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n a_i Y_i,$$

where

$$a_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Use this fact to argue that $\hat{\beta}_1 \sim \mathcal{N}(\beta_1, c_{11}\sigma^2)$, where

$$c_{11} = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Note: On this part, I want you to do three things: (1) prove that $E(\hat{\beta}_1) = \beta_1$, (2) prove that $V(\hat{\beta}_1) = c_{11}\sigma^2$, and (3) argue that $\hat{\beta}_1$ has a normal distribution.

(c) [7] In class, I stated, without proof, that

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2),$$

where

$$\hat{\sigma}^2 = \text{MSE} = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2,$$

and also that MSE and $\hat{\beta}_1$ are independent statistics. Using this information, and the result from part (b), **derive completely** the $100(1 - \alpha)$ percent confidence interval for β_1 . Make sure that you define all of your notation.

4. A psychologist wanted to investigate the relationship between the physical characteristics of preadolescent boys and their maximal oxygen uptake (y , measured in milliliters of oxygen per kilogram of body weight). Four covariates were used: age (x_1 , in years), height (x_2 , in centimeters), weight (x_3 , in kilograms), and chest depth (x_4 , in centimeters). Data observed for $n = 10$ boys are shown below:

y	x_1	x_2	x_3	x_4
1.54	8.4	132.0	29.1	14.4
1.74	8.7	135.5	29.7	14.5
1.32	8.9	127.7	28.4	14.0
1.50	9.9	131.1	28.8	14.2
1.46	9.0	130.0	25.9	13.5
1.35	7.7	127.6	27.6	13.9
1.53	7.3	129.9	29.0	14.0
1.71	9.9	138.1	33.6	14.6
1.27	9.3	126.6	27.7	13.9
1.50	8.1	131.8	30.8	14.5

As a first step, the researcher wanted to consider the **full model**

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i,$$

for $i = 1, 2, \dots, 10$, where $\epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2)$, or, in matrix notation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where \mathbf{Y} is the 10×1 response vector, \mathbf{X} is the 10×5 matrix of covariates, $\boldsymbol{\beta}$ represents the 5×1 vector of regression parameters, and $\boldsymbol{\epsilon}$ represents a 10×1 multivariate normal random vector with mean $\mathbf{0}$ and variance-covariance matrix $\sigma^2 \mathbf{I}$. Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)'$ denote the least squares estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)'$. Let $\mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ denote the hat matrix. The researcher used SAS to fit the full model. She also asked SAS for the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 510.861 & -0.877 & -1.437 & 6.653 & -35.911 \\ -0.877 & 0.171 & -0.018 & -0.014 & 0.157 \\ -1.437 & -0.018 & 0.025 & -0.021 & -0.082 \\ 6.653 & -0.014 & -0.021 & 0.138 & -0.546 \\ -35.911 & 0.157 & -0.082 & -0.546 & 4.329 \end{bmatrix}$$

Here is the ANOVA table from the full model, obtained from SAS.

Source	DF	SS	MS	F	Pr > F
Model	4	0.20621	0.05155	38.17	0.0006
Error	5	0.00675	0.00135		
Corrected Total	9	0.21296			

Here are the parameter estimates, standard errors, and the t statistics used to test $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$, for $j = 0, 1, 2, 3, 4$. This output was obtained from SAS.

Variable	DF	Parm.Est	Std.Err.	t value	Pr > t
Intercept	1	-4.847	0.830	-5.84	0.0021
AGE	1	-0.034	0.015	-2.29	0.0709
HEIGHT	1	0.051	0.005	8.75	0.0003
WEIGHT	1	-0.024	0.013	-1.82	0.1277
CHEST	1	0.041	0.076	0.54	0.6113

Questions for you to answer:

- [2] Give the dimensions of \mathbf{M} and $V(\hat{\boldsymbol{\beta}})$.
- [2] Compute an estimate of $V(\hat{\beta}_1 - \hat{\beta}_2)$. Your answer should be a number.
- [2] Write out the set of hypotheses that can be tested with the F statistic in the ANOVA table on the previous page. Which hypothesis is more supported by the data?
- [3] What is the value of $\mathbf{y}'(\mathbf{I} - \mathbf{M})\mathbf{y}$? the value of $\mathbf{y}'\mathbf{y}$? the value of $(\mathbf{X}'\mathbf{X})_{1,1}$?
- [2] Does CHEST add to a model that includes the other three covariates? Explain.

A colleague participating in the study believes that a smaller model may be adequate for these data. Specifically, she believes the variables x_3 and x_4 are not important and, thus, the **reduced model** $Y_i = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \epsilon_i$ is adequate. Here is the ANOVA table from the reduced model, obtained using SAS.

Source	DF	SS	MS	F	Pr > F
Model	2	0.20020	0.10010	54.92	0.0005
Error	7	0.01276	0.00182		
Corrected Total	9	0.21296			

More questions for you to answer:

- [2] Consider writing the reduced model in the form $\mathbf{Y} = \mathbf{X}_0\boldsymbol{\gamma} + \boldsymbol{\epsilon}$. Give the form of \mathbf{X}_0 and $\boldsymbol{\gamma}$; that is, just write out what these are.
- [4] Perform a level $\alpha = 0.05$ test to assess whether or not the reduced model is adequate for the data. State your hypotheses, show how your test statistic is computed, state the rejection region, and write your conclusion.
- [2] The psychologist has a son with covariate information

$$(x_1, x_2, x_3, x_4)' = (8.2, 133.0, 29.0, 14.5)'$$

Predict the psychologist's son's oxygen uptake level, using the **full model** fit (the least squares estimates are at the top of the page). If she wanted to make an inferential statement for her son, would a confidence interval or prediction interval be appropriate? Explain. Note that I'm not asking you to construct an interval.

- [1] Explain why SST is the same for both the full and reduced models.

5. [20] True/False. A true statement is one that is always true. A false statement may be true some of the time, but not always. Each question is worth 1 point. No partial credit will be given, so no explanation is necessary.

- (a) The likelihood ratio statistic λ satisfies $\lambda \leq 1$.
- (b) The beta distribution is a conjugate prior for the binomial($1, p$) family.
- (c) In the simple linear regression model, normality is needed (on the errors ϵ_i) to show that the least-squares estimators are unbiased.
- (d) In the Bayesian paradigm, model parameters are treated as random variables with their own probability distributions.
- (e) A Jeffreys' prior distribution is an example of an informative prior.
- (f) Confidence intervals and credible intervals are interpreted in the same way.
- (g) In survival analysis, the Kaplan-Meier estimator can be viewed as a limiting life table estimator.
- (h) Under certain regularity conditions, $-2 \ln \lambda$, where λ denotes the likelihood ratio statistic, follows an approximate t distribution.
- (i) Suppose that the posterior distribution for θ is beta($2, 2$). The posterior mean and posterior median are equal.
- (j) The model $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$ is an example of a linear statistical model.
- (k) In a linear regression model, residuals obtained from a least-squares fit are neither independent nor homoscedastic.
- (l) In a non-Bayesian hypothesis test, small probability values are evidence that H_0 is true.
- (m) In bivariate situations where correlation methods are appropriate, the independent variable is best regarded as random.
- (n) In a hypothesis test, the power function is an increasing function of the parameter of interest.
- (o) Bayesian credible intervals can never extend outside the parameter space for the parameter under investigation.
- (p) In multiple linear regression, the hat matrix is symmetric and idempotent.
- (q) In survival analysis, the survivor function equals 1 minus the cumulative distribution function.
- (r) An increasing hazard rate function is associated with individuals who experience "wear-out" or aging.
- (s) The Kaplan-Meier estimator is an estimator of the survivor function.
- (t) In the analysis of data from clinical trials, parametric models are preferred over nonparametric models.