Real Analysis I. Qualifying Exam 2007

1) Show that every open set of real numbers is measurable.

2) Show that if f is a measurable function and $f = g$ almost everywhere, then g is measurable.

3) Let f be a nonnegative integrable function. Show that

$$F(x) = \int_{-\infty}^{x} f$$

is continuous (by using the Monotone Convergence Theorem).

4) Argue that if f is absolutely continuous, then f has a derivative almost everywhere.

5) Let g be an integrable function on $[0, 1]$. Show that there is a bounded measurable function f such that $||f|| \neq 0$ and

$$\int fg = ||g||_1 ||f||_\infty.$$

Ph.D. Qualifying Examination in Probability

April 9, 2007

Instructions:

(a) There are three problems. Submit all three for full credit.
(b) Start each problem on a new page, indicate the problem number, and number the pages.
(c) Indicate any concept or theorem used, and show how it is being applied.

1. Let $v_1 > 0$, $v_2 > 0$, and $0 \le \alpha \le 1$.

(a) Prove that, for any real u_1 and u_2,

$$\frac{[\alpha u_1 + (1-\alpha)u_2]^2}{\alpha v_1 + (1-\alpha)v_2} \le \alpha \frac{u_1^2}{v_1} + (1-\alpha)\frac{u_2^2}{v_2}.$$

(b) Use (a) to show that the Fisher information for location defined on distribution functions F with absolutely continuous density f,

$$I(F) = \int \left(\frac{f'(x)}{f(x)} \right)^2 dF(x),$$

is a convex function of the argument F.

(c) Define $F_\varepsilon = \varepsilon F_1 + (1-\varepsilon)F_0$ for two given distributions F_0 and F_1 and $0 \le \varepsilon \le 1$. Show that

$$J(\varepsilon) = I(F_\varepsilon)$$

is a convex function of the argument ε.

2. Let $X_n, n \ge 1$ be i.i.d. random variables with $P(X_n = 1) = P(X_n = -1) = 1/2$, and put $S_n = \sum_{i=1}^n X_i, n \ge 1$.

(a) Show that

$$\frac{S_n}{\sqrt{n} \log \log n} \xrightarrow{P} 0.$$

(b) Show that

$$\frac{S_n}{\sqrt{n}} \overset{P}{\not\to} 0.$$

3. In a probability space $(\Omega, \mathcal{F}, P)$, let $\{A_n, n \ge 1\}$ be independent events satisfying

$$\sum_{n=1}^{\infty} P(A_n) = \infty.$$

(a) Show that, for any $m \ge 1$,

$$P(\cap_{n=m}^{\infty} A_n) = 0.$$

(b) Use (a) to show that

$$P(A_n \text{ infinitely often}) = 1.$$

Ph.D. Qualifying Exam: Spring 2007
Statistical Inference

- Number of questions = 3. Answer all of them. Total points = 75.
- Simplify your answers as much as possible and carefully justify all steps to get full credit.

1. Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean $\mu = \theta$ and standard deviation $\sigma = \nu\theta$, where the population *coefficient of variation* $\sigma/\mu = \nu$ is known. Consider a class $C = \{\delta_b = b\overline{X}, b \in \mathcal{R}\}$ of estimators of θ, where $\overline{X}$ is the sample mean.

   (a) Find the estimator of θ in C that minimizes the mean squared error (MSE). Find the MSE of the optimal estimator. [10 points]

   (b) Show that the optimal estimator in (a), which uses the additional information that ν is known, is better than $\overline{X}$, which does not use this information. [5 points]

2. For given $-\infty < \mu < \infty$ and $\sigma^2 > 0$, let $\mathrm{X} = (X_1, \ldots, X_n)$ represent a random sample from a normal(μ, σ^2) distribution. Also let $\overline{X} = n^{-1}\sum_i X_i$ and $S^2 = (n-1)^{-1}\sum_i(X_i - \overline{X})^2$ denote the sample mean and variance. Assume that the prior distributions for μ and $\log\sigma$ are independent uniforms on $(-\infty, \infty)$, or equivalently, the joint prior probability density of (μ, σ^2) is,

$$\pi(\mu, \sigma^2) \propto (\sigma^2)^{-1}.$$

   This prior distribution is *improper* — its joint density does not integrate to one. But this will not cause any trouble in answering the following questions.

   (a) Show that the joint posterior density of (μ, σ^2) is, [7 points]

$$\pi(\mu, \sigma^2|\mathrm{x}) \propto (\sigma^2)^{-(n+2)/2} \exp\left[-\frac{1}{2\sigma^2}\{(n-1)s^2 + n(\overline{x} - \mu)^2\}\right].$$

   (b) Derive $\pi(\mu|\mathrm{x})$, i.e., the marginal posterior distribution of μ. [8 points]

   (c) Show that the marginal posterior distribution of $(s/\sqrt{n})^{-1}(\mu - \overline{x})$ is a t-distribution with $(n-1)$ degrees of freedom. [5 points]

   (d) Derive an HPD credible interval for μ that has $(1-\alpha)$ posterior probability. [5 points]

(e) Consider testing the hypotheses: $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$. Find posterior probability that the null hypothesis is true, $P(\mu \leq 0|\mathbf{x})$. [5 points]

(f) Compare your answer in (d) with the usual classical two-sided $(1-\alpha)$ level *confidence interval* for μ based on t-distribution. What do you conclude? [5 points]

(g) Compare your answer in (e) with the *p-value* for the usual classical t-test of H_0 versus H_1. What do you conclude? [5 points]

3. Let X_{ij}, $i = 1, \dots, n_j$, denote a random sample of size n_j from a normal(μ_j, σ_j^2) distribution, $j = 1, 2$. The two samples are mutually independent. We assume that σ_1^2 is *known*, but σ_2^2 is *unknown*. Our interest lies in testing the null hypothesis $H_0 : \mu_1 - \mu_2 = \Delta_0$, where Δ_0 is a specified constant. This is the setting of the usual two-sample t-test with the exception that one variance is known and the other is unknown.

Let $\overline{X}_j = n_j^{-1} \sum_{i=1}^{n_j} X_{ij}$ and $S_j^2 = (n_j - 1)^{-1} \sum_{i=1}^{n_j} (X_{ij} - \overline{X}_j)^2$ be the sample mean and variance for the j-th sample, $j = 1, 2$. A reasonable statistic for testing H_0 is

$$T = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

(a) Argue that, under H_0,

$$Y = \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^{-1} \left(\frac{\sigma_1^2}{n_1} + \frac{S_2^2}{n_2}\right)$$

does not have an exact χ^2-distribution, and hence T does not have an exact t-distribution. [5 points]

(b) We would like to approximate the distribution of νY as a χ_ν^2- distribution, where ν is estimated using a Satterthwaite-type "moment matching" technique. In particular,

i. Show that $E(\nu Y) = E(\chi_\nu^2)$. [3 points]

ii. Find ν such that $Var(\nu Y) = Var(\chi_\nu^2)$. [5 points]

iii. Find a natural estimator $\hat{\nu}$ of ν. [2 points]

(c) Based on the approximation derived in (b), what will you use as the approximate distribution of T under H_0 to derive a critical region for the test? [5 points]

Ph.D. Qualifying Examination: Spring 2007
Linear Models

April 11, 2007

Instructions

- Number of questions = 2, total points = 30.
- Show your work and justify all steps to get full credit.

Problems

1. (20 points) Consider n independent replications of the three–dimensional normal vector Y that has the mean

$$\mu = (\mu_1, \mu_2, \mu_3)$$

and the variance–covariance matrix C with the elements

$$c_{11} = c_{12} = c_{13} = 1,$$

$$c_{21} = 1, \; c_{22} = c_{23} = 2,$$

$$c_{31} = 1, \; c_{32} = 2, c_{33} = 3.$$

   (a) (5 points) Derive the least squares estimate for the two following functions of μ. $L_1(\mu)) = \mu_2 - \mu_1$ and $L_2(\mu) = \mu_3 - \mu_2$.

   (b) (5 points) Find the covariance matrix of the (*unconstrained*) estimate $\hat{\mu}$ for μ.

   (c) (5 points) Consider the null hypothesis $H : \mu_1 = \mu_2 = \mu_3$ and describe the distribution of the test statistic.

   (d) (5 points) Find the eigenvectors and eigenvalues of the matrix C.

2. (10 points) Linear model with idependent zero–mean normal errors has the form

$$Y_j = j\beta + \epsilon_j \quad ((1 \leq j \leq n)$$

where the variance (σ_j^2) of $\epsilon_j = j^2\sigma^2$.

   (a) (5 points) Apply the weighted least squares technique to find the minimum variance unbiased estimate for β.

   (b) (5 points) Find the value of this variance.

Ph.D. Qualifying Examination in Statistical Methods

April 13, 2007

Exam consists of two projects (70 points + 30 points).
Both data sets are available on http://www.utdallas.edu/~mbaron/Qual07.

PROJECT I "Coronary heart disease", Data Set 1 (70 points)

A health insurance company collected information on 788 of its subscribers who had made claims resulting from coronary heart disease. Data were obtained on total costs and nature of services provided to these 788 subscribers during a two-year period. Each line in the data set has an identification number and provides information on nine other variables for each subscriber:

Column	Variable
1	Subscriber dentification number
2	Total cost of claims ($)
3	Age (years)
4	Gender (FEMALE OR MALE)
5	Number of interventions of procedures carried out
6	Number of prescribed drugs
7	Number of emergency room visits
8	Number of other complications
9	Number of comorbidities (subscriber's other diseases during the same period)
10	Duration of treatment (days)

The company needs a good model for predicting the total cost of insurance claims. Is a linear model adequate, or do we need a nonlinear model? Which variables are significant? Do we need separate models for females and males? Support your conclusions with suitable tests and model selection methods.

As part of your analysis, verify assumptions being used, check for outliers and influential observations. Try to apply remedial measures if necessary.

Finally, predict the total cost of claims submitted by a 55-year old woman during a two-year period provided that she never visits the emergency room, has no other diseases, and has 4 prescribed drugs. Compute a 90% prediction interval.

Instructions

– Load the data sets from http://www.utdallas.edu/~mbaron/Qual07

– Conduct the necessary data analysis using *software of your choice.*

– Submit a report, written or typed, hard copy or e-mail. If you choose to e-mail the report, send it to both ammann@utdallas.edu and mbaron@utdallas.edu.

– In the report, describe every step of your analysis: method, reasons, and results. For example:

> *Test significance of variable XXX. Use SAS, PROC ... with option ... The F test gives a p-value of 0.0003. Therefore, ...*
>
> *Verify assumptions of the test. Use ... ... Variable ... violates assumption ... because ... Therefore, ...*

– Attach your computer programs and only relevant parts of the output. Do not attach the parts of output that were not used to answer questions.

PROJECT II “Restaurants”, Data Set 2 (30 points)

In a study of restaurant behavior, twenty couples were observed. Ten of them made orders around 6 pm, the other ten made orders around 9 pm (these times were chosen randomly). Experimenters recorded the price of the entreé ordered by one person in each pair, and categorized subjects on the basis of whether or not they pay the bill (host and guests). The data follow, the analysis was carried out using SAS (see the attached printout).

6 pm		9 pm	
Host	Guest	Host	Guest
8.00	8.25	9.75	8.75
7.00	8.75	10.25	9.00
8.25	9.75	9.50	9.25
9.00	8.00	9.00	8.50
8.25	9.25	10.50	8.75

1. The printout contains 7 different F-tests (F-values are 6.64, 10.43, 0.21, 9.27, 1.1249, 0.0230, 9.2681). Explain clearly which hypotheses were tested using these F-statistics.

2. Is there a sufficient evidence to conclude that hosts order cheaper entreés than guests? Assuming normal distribution of errors, give a 95% confidence interval for the mean difference of entreé prices ordered by guests and hosts.

3. Explain how method of moments estimators of variance components were obtained.

4. Do the two factors interact? Use suitable plots in addition to the F-test to support your answer.

```
data;
infile Qual07;
input time $ host $ entree;

PROC GLM;
CLASS host time;
MODEL entree = host time host*time;
RANDOM time host*time / TEST;

PROC VARCOMP method = type1;
CLASS host time;
MODEL entree = host time host*time / fixed=1;

run;
```

```
==============================================================================

       General Linear Models Procedure
           Class Level Information

           Class    Levels    Values

           HOST          2    G H
           TIME          2    F M


       Number of observations in data set = 20


Dependent Variable: ENTREE

Source                DF       Sum of Squares       Mean Square    F Value    Pr > F

Model                  3       7.30937500           2.43645833     6.64       0.0040
Error                 16       5.87500000           0.36718750
Corrected Total       19      13.18437500

R-Square                          C.V.          Root MSE          ENTREE Mean

0.554397                       6.818115         0.60595998        8.88750000



Source                DF       Type III SS          Mean Square     F Value   Pr > F

HOST                   1       3.82812500           3.82812500      10.43     0.0052
TIME                   1       0.07812500           0.07812500       0.21     0.6508
HOST*TIME              1       3.40312500           3.40312500       9.27     0.0077


       Tests of Hypotheses for Mixed Model Analysis of Variance

Source: HOST
Error: MS(HOST*TIME)
                           Denominator      Denominator
DF      Type III MS                 DF               MS       F Value     Pr > F
1         3.828125                   1         3.403125       1.1249     0.4813

Source: TIME
Error: MS(HOST*TIME)
                           Denominator      Denominator
DF      Type III MS                 DF               MS       F Value     Pr > F
1         0.078125                   1         3.403125       0.0230     0.9043

Source: HOST*TIME
Error: MS(Error)
                           Denominator      Denominator
DF      Type III MS                 DF               MS       F Value     Pr > F
1         3.403125                  16        0.3671875       9.2681     0.0077


Variance Component                          Estimate

Var(TIME)                                -0.33250000
Var(HOST*TIME)                            0.60718750
Var(Error)                                0.36718750
```