Name: ______________________________

Qualifying Exam, April 2008
Real Analysis I

THIS IS A CLOSED BOOK, CLOSED NOTES EXAM

Problem 1 (20 points.) Prove or disprove (by a counterexample) the following statements:
a. A countable subset of $\mathbb{R}$ has Lebesgue measure zero.
b. If a subset of $\mathbb{R}$ has Lebesgue measure zero then it is countable.

Problem 2 (20 points.)
Let $(X, \mathcal{M}, \mu)$ be a measure space.
a. Prove that μ is continuous from below, that is, if $\{E_j\}_{j=1}^{\infty} \subset \mathcal{M}$ and $E_1 \subset E_2 \subset \cdots$, then $\mu(\cup_{j=1}^{\infty} E_j) = \lim_{j\to\infty} \mu(E_j)$.
b. Let $\{E_j\}_{j=1}^{\infty}$ be a sequence of measurable sets in X and let $E = \cup_{k=1}^{\infty} \cap_{j=k}^{\infty} E_j$. Prove that $\mu(E) \leq \liminf \mu(E_j)$.

Problem 3 (20 points.) Let f be a real-valued function on $\mathbb{R}$. Which of the following statements are true? Justify your answers.
(i) If f is measurable, then $|f|$ is measurable.
(ii) If $|f|$ is measurable, then f is measurable.

Problem 4 (20 points.)
Let (f_n) be a sequence of integrable functions on $[0,1]$ such that $0 \leq f_{n+1} \leq f_n$ for all n and $f = \lim_{n\to\infty} f_n$. Show that $f = 0$ a.e. iff $\lim_{n\to\infty} \int f_n = 0$.

Problem 5 (20 points.) Compute the following limit and justify the calculations. (Hint: Use the dominated convergence theorem.)

$$\lim_{n\to\infty} \int_0^{\infty} \frac{2n^2 + \sin(n^2x^2+1)}{n^2+x^2} e^{-x} dx$$

Name: ______________________________

Ph.D. Qualifying Examination in Probability Theory

April 7, 2008

1. Consider the measure space $(R, \mathcal{B}, \lambda)$ with the Borel σ-field $\mathcal{B}$ and Lebesgue measure λ. Which of the following classes of sets is a π-system, a λ-system, a monotone class, a field, or a σ-field:

   - class A of all sets of measure 0
   - class B of all nonmeasurable sets
   - class C of all finite and countable sets
   - class D of all finite, countable, cofinite, and cocountable sets?

   In the table, circle all that apply. Explain your answers for the class B.

	π-system	λ-system	monotone class	field	σ-field
A	yes no	yes no	yes no	yes no	yes no
B	yes no	yes no	yes no	yes no	yes no
C	yes no	yes no	yes no	yes no	yes no
D	yes no	yes no	yes no	yes no	yes no

2. Let $\{X_n\}$ be a sequence of random variables with a common mean μ and a common variance $\sigma^2 \in (0, \infty)$. Show that for any $\varepsilon > 0$, with probability 1,

   (a) there exists a number N such that
   $$\text{for all } n \geq N, \ |X_n| \leq \varepsilon n.$$

   (b) if X_n are independent and identically distributed, then $|X_n| > \varepsilon/n$ for infinitely many n.

3. The *density* of a set of positive integers A is defined as the limit
   $$\lim_{n\to\infty} \frac{\text{number of elements in } A \cap [1, n]}{n}.$$
   Use inclusion-exclusion formulas to show that the set of integers not divisible by a perfect cube ~~equals~~ has density
   $$\prod_p \left(1 - \frac{1}{p^3}\right),$$
   where the product is taken over all prime numbers (assume without a proof that this product converges).

Ph.D. Qualifying Examination in Statistical Inference

April 9, 2008

Instructions:

(a) *There are three problems, each of equal weight. You may submit work on all three.*

(b) *Extra credit will be given for a problem with all parts solved well.*

(c) *Look over all three problems before beginning work.*

(d) *Start each problem on a new page, and number the pages.*

(e) *On each page, indicate problem number and part, and write your name.*

(f) *Indicate your lines of reasoning and what background results are being applied.*

---

1. Consider a statistical model defined as a family of densities on $\mathbb{R}^d$: $\mathcal{P} = \{f(\boldsymbol{x}, \theta), \theta \in \Theta\}$, with Θ a parameter space. Assume that all densities in $\mathcal{P}$ have the same support.

Let $\boldsymbol{X}$ denote an observation on some distribution in $\mathcal{P}$. Consider the usual likelihood function,

$$L(\theta, \boldsymbol{X}) = f(\boldsymbol{X}, \theta), \ \theta \in \Theta.$$

(a) State the *factorization theorem*, which gives a necessary and sufficient criterion for a statistic $S(\boldsymbol{X})$ to be *sufficient* for the family $\mathcal{P}$.

(b) Apply the factorization theorem to show that, considered as a statistic, the likelihood function $L(\theta, \boldsymbol{X})$, $\theta \in \Theta$, is sufficient for $\mathcal{P}$.

(c) For a fixed value $\theta_0 \in \Theta$, define the *likelihood ratio*

$$\Lambda(\theta, \boldsymbol{X}) = \frac{L(\theta, \boldsymbol{X})}{L(\theta_0, \boldsymbol{X})}, \ \theta \in \Theta.$$

Apply the factorization theorem to show that, considered as a statistic, the function $\Lambda(\theta, \boldsymbol{X})$, $\theta \in \Theta$, is sufficient for $\mathcal{P}$.

(d) A sufficient statistic $T(\boldsymbol{X})$ is *minimal sufficient* if, for every other sufficient statistic $S(\boldsymbol{X})$, $T(\boldsymbol{X})$ depends on $\boldsymbol{X}$ only through $S(\boldsymbol{X})$, i.e., $T(\boldsymbol{X})$ may be expressed as some function g of $S(\boldsymbol{X})$: $T(\boldsymbol{X}) = g(S(\boldsymbol{X}))$. Apply the factorization theorem to show that, considered as a statistic, the function $\Lambda(\theta, \boldsymbol{X})$, $\theta \in \Theta$, is *minimal sufficient* for $\mathcal{P}$.

(e) (i) Is the maximum likelihood estimator (MLE) of θ a function of every sufficient statistic $S(\boldsymbol{X})$?

(ii) Is the MLE of θ itself always sufficient?

---

2. Let $X_1, \ldots, X_n$ be i.i.d. random variables with univariate distribution F. Let ξ_p denote $\inf\{x : F(x) \geq p\}$, the pth quantile of F, for $0 < p < 1$. A popular measure of spread of a distribution F is the parameter $\theta = \xi_{0.75} - \xi_{0.25}$, the so-called *interquartile range*.

(a) In terms of the data $X_1, \ldots, X_n$, give a consistent estimator of θ, in the sense of convergence in probability.

(b) Let F be Normal(μ, σ^2), with (μ, σ^2) unknown.

  (i) Express θ in terms of (μ, σ^2).

  (ii) Derive the maximum likelihood estimator, $\widehat{\theta}_{\text{MLE}}$, of θ.

  (iii) Justify that $\widehat{\theta}_{\text{MLE}}$ is consistent for estimation of θ.

  (iv) Show that $\widehat{\theta}_{\text{MLE}}$ is *biased* for estimation of θ, with bias $E(\widehat{\theta}_{\text{MLE}}) - \theta < 0$.

(c) Let F be Normal$(0, \sigma^2)$. For the null hypothesis $H_0 : \theta = \theta_0$, derive a test statistic that is most powerful among tests of equal or lesser significance level.

3. Let $\boldsymbol{Z} = (Z_1, \ldots, Z_d)'$ be a random (column) vector in $\mathbb{R}^d$ distributed as Normal$(\boldsymbol{0}, \boldsymbol{I}_d)$, where $\boldsymbol{I}_d$ denotes the $d \times d$ identity matrix. For $\boldsymbol{\mu}$ a given vector in $\mathbb{R}^d$ and Σ a given $d \times d$ nonsingular covariance matrix, put

$$\boldsymbol{Y} = \Sigma^{1/2}\boldsymbol{Z} + \boldsymbol{\mu},$$

where $\Sigma^{1/2} \times \Sigma^{1/2} = \Sigma$.

(a) Show that $\boldsymbol{Y}$ is Normal$(\boldsymbol{\mu}, \Sigma)$.

(b) The *Euclidean distance* of $\boldsymbol{Z}$ from the origin in $\mathbb{R}^d$ is

$$\mathrm{ED}(\boldsymbol{Z}) = \|\boldsymbol{Z}\| = \sqrt{\sum_1^d Z_i^2}.$$

Show that $(\mathrm{ED}(\boldsymbol{Z}))^2 = \|\boldsymbol{Z}\|^2$ is distributed as χ_d^2, chi-square with d degrees of freedom. (Useful fact: The 4th moment of the univariate standard normal distribution is 3.)

(c) The *Mahalanobis distance* of $\boldsymbol{Y}$ from $\boldsymbol{\mu}$ is

$$\mathrm{MD}(\boldsymbol{Y}) = \sqrt{(\boldsymbol{Y} - \boldsymbol{\mu})' \Sigma^{-1} (\boldsymbol{Y} - \boldsymbol{\mu})}.$$

Show that $(\mathrm{MD}(\boldsymbol{Y}))^2$ is distributed as χ_d^2.

(d) Show that for large dimension d, $\mathrm{MD}(\boldsymbol{Y})$ is approximately $\sqrt{d}$. That is, show that

$$\frac{(\mathrm{MD}(\boldsymbol{Y}))^2}{d} \xrightarrow{p} 1, \; d \to \infty.$$

(Hint: use Chebyshev's inequality.)

(e) Discuss extension of (d) to the case of arbitrary distribution F for $\boldsymbol{Z}$.

Ph.D. Qualifying Exam: Spring 2008
Linear models

- Number of questions = 2. Answer both of them. Total points = 60.
- Simplify your answers as much as possible and carefully justify all steps to get full credit.

---

1. Suppose we have two categorical explanatory variables A and B, each with 4 categories. A response variable Y is observed at some — but not all — combinations of A and B. We would like to study the effect of A on $E(Y)$ and B serves as a nuisance factor. Let Y_{ij} be the value of Y when A is in its ith category and B is in its jth category. The following table describes the combinations of A and B at which Y is observed and the totals of the observed responses.

		B 1	2	3	4	Total
A	1	Y_{11}	Y_{12}	-	Y_{14}	$Y_{1.}$
	2	-	Y_{22}	Y_{23}	Y_{24}	$Y_{2.}$
	3	Y_{31}	Y_{32}	Y_{33}	-	$Y_{3.}$
	4	Y_{41}	-	Y_{43}	Y_{44}	$Y_{4.}$
Total		$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	$Y_{.4}$	$Y_{..}$

The combinations of A and B that are not observed as marked as "–" in this table. Consider the following linear model for these data.

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad i = 1, \ldots, 4; \quad j = 1, \ldots, 4; \quad (i, j) \text{ is in the design;}$$

where μ is a common intercept, α_i is the effect of the ith category of A, β_j is the effect of the jth category of B, and ϵ_{ij} is the random error term. The phrase "(i, j) is in the design" means that the model is applicable to only those combinations of A and B at which Y is observed. We assume that the errors follow mutually independent $N(0, \sigma^2)$ distributions. Further, since the design matrix corresponding to the above model is two less than the full rank, we assume two constraints on the parameters, namely, $\sum_i \tau_i = 0 = \sum_j \beta_j$.

   (a) Derive the normal equations for estimating $\mu, \alpha_i, \beta_j,\ i, j = 1, \ldots, 4$, using the method of least squares. [Hint: Write the error sum of squares as $\sum_i \sum_j n_{ij}\epsilon_{ij}^2$, where $n_{ij} = 1$ if (i, j) is the design, otherwise $n_{ij} = 0$.] [10 points]

(b) Use the answer in (a) to show that the least squares estimators of μ, α_i and β_j are [10 points]

$$\hat{\mu} = \frac{Y_{..}}{12},$$
$$\hat{\alpha}_i = \frac{3}{8}\left(Y_{i.} - \frac{1}{3}\sum_j n_{ij} Y_{.j}\right),$$
$$\hat{\beta}_j = \frac{Y_{.j}}{3} - \hat{\mu} - \frac{1}{3}\sum_i n_{ij}\hat{\alpha}_i.$$

(c) Use the answer in (b) to show that $E(\hat{\alpha}_1 - \hat{\alpha}_2) = \alpha_1 - \alpha_2$. [10 points]

(d) Use the answer in (b) to show that $var(\hat{\alpha}_1 - \hat{\alpha}_2) = \frac{3}{4}\sigma^2$. [10 points]

2. Consider data collected before and after some type of intervention has occurred in the process being modelled. For example, economic data may be collected pre- and post-war. Or, physiological data may be collected pre- and post-treatment. Of interest is whether the same linear regression model can be used to fit the pre- and post-intervention data.

Specifically, suppose the model for the n_1 pre-intervention observations is $Y_1 = X_1\beta_1 + \epsilon_1$ and the model for the n_2 post-intervention data is $Y_2 = X_2\beta_2 + \epsilon_2$, where X_1 and X_2 are design matrices of the same k explanatory variables, both with rank k.

(a) Show how the two sets of data can be combined into a single regression model of the form $Y = X\beta + \epsilon$ so that the hypothesis $H_0 : \beta_1 = \beta_2$ can be expressed in terms of this single regression model as $H_0 : L'\beta = 0$. Carefully define all notation you use, including Y, X, β, ϵ and L. [10 points]

(b) Develop an F-test for testing this hypothesis. State assumptions that you make about the model and the data for the test to be valid. [10 points]

2008 Statistics Qualifying Exam: Methods

Instructions: attach answers including graphic files to an email and send to `ammann@utdallas.edu`.

1. The data for this problem can be found at
   http://www.utdallas.edu/~ammann/ToothGrowth.dat
   This experiment was conducted to determine what effect Vitamin C has on the growth of teeth in guinea pigs. Three dosages were tested, 0.5, 1, 2 mg, and two different delivery methods were used, orange juice or ascorbic acid. Ten guinea pigs were randomly assigned to each of the resulting 6 groups.

   (a) Treat *Dose* as a categorical variable and fit an appropriate linear model to predict *Growth* based on *Method* and *Dose*.

   (b) Verify the assumptions used with this model.

   (c) Use the model to construct a 95% confidence interval for the *mean* tooth growth of guinea pigs who receive 1 mg of Vitamin C using orange juice as the delivery method.

2. The data for this problem can be obtained from
http://www.utdallas.edu/~ammann/InsectSprays.dat
This experiment represents the count of insects observed after applying one of six different insecticide sprays.

   (a) Fit a model to predict *Count* based on *Spray*.

   (b) Verify the assumptions required by the model and justify any transformations that are required.

   (c) Construct an informative plot of the data after any transformations that shows how *Count* varies with *Spray*.

3. The data for this problem can be obtained from
http://www.utdallas.edu/~ammann/Wasp.dat
This is a data frame with 100 observations on a species of wasp. *caste* indicates whether the observation came from a Queen or Worker. The other variables are physical measurements of the wasp bodies.

   (a) Fit a model to predict G1L based on TL, TW, HH, G1H.

   (b) Test for significance of the terms in the model.

   (c) Verify the assumptions used.

   (d) Determine if any variables can be removed from this model.

   (e) Fit a model to predict *caste* based on TL, TW, HH, G1H.

   (f) Compare actual caste with the caste predicted by this model.