Name: _____

Qualifying Exam, April 2009
Real Analysis I

THIS IS A CLOSED BOOK, CLOSED NOTES EXAM
Solve 4 of the following 5 problems. You must clearly indicate which 4 are to be graded.

Problem 1 (25 points.)
If μ^* is an outer measure on X and $\{A_j\}_1^\infty$ is a sequence of disjoint μ^*-measurable sets, then for any $E \subset X$,

$$\mu^*\left(E \cap (\cup_{j=1}^\infty A_j)\right) = \sum_{j=1}^\infty \mu^*\left(E \cap A_j\right).$$

Problem 2 (25 points.)
Let $C \subset [0,1]$ be the Cantor set. Define $f : \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} x & \text{if } x \notin C; \\ 0 & \text{if } x \in C. \end{cases}$$

(a) Is f Lebesgue measurable on $\mathbb{R}$? Justify your answer.
(b) Is f Riemann integrable on $[0,1]$? Is f Lebesgue integrable on $[0,1]$? Justify your answer.

Problem 3 (25 points.)
Compute the following limit and justify the calculations. (Hint: Use the properties of the function $\frac{\sin y}{y}$.)

$$\lim_{n\to\infty} \int_0^\pi n \sin\left(\frac{x}{n}\right) dx$$

Problem 4 (25 points.)
Let $\{f_n\}$ be a sequence of real-valued functions on $\mathbb{R}$. Let m be the Lebesgue measure. Show that if $f_n \to f$ in $L^1(\mathbb{R}, m)$, then $f_n \to f$ in measure. Is the converse true? Justify your answer.

Problem 5 (25 points.)
Let f and g be real-valued absolutely continuous functions on $[a,b]$, $a < b$.
(a) Show that the product fg is also absolutely continuous on $[a,b]$. (Hint: first show that f and g are bounded.)
(b) Show that $\int_a^b [f'(x)g(x) + f(x)g'(x)]dx = f(b)g(b) - f(a)g(a)$.

Ph.D. Qualifying Examination in Probability

April 6, 2009

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. Let $Y_1, Y_2, \ldots$ be independent and identically distributed random variables with mean 0 and variance 1. Define

$$X_n = \frac{\sum_{i=1}^{n} Y_i}{(2n \log\log n)^{1/2}}.$$

(a) The following fact is true (do not provide proof): with probability 1

$$\limsup_{n\to\infty} X_n = 1. \tag{1}$$

Use this to conclude that X_n *does not* $\to 0$ with probability 1.

(b) Show that $X_n \to 0$ in probability.

(c) Show that X_n *does not* $\to N(0,1)$ (standard normal) in distribution.

(d) Show that $X_n \to 0$ in mean square.

(e) Show that the random variable $\sup_{n\geq 1} X_n$ is finite with probability 1.

---

2. Let P and Q be probability measures on $(\mathbb{N}^+, \mathcal{A})$, where $\mathcal{A}$ is the class of all subsets of the positive integers $\mathbb{N}^+$. Let $d(P,Q)$ denote the total variation metric

$$\sup_{A \in \mathcal{A}} |P(A) - Q(A)|.$$

(a) Verify that

$$d(P,Q) = \frac{1}{2} \sum_{m=0}^{\infty} |P(m) - Q(m)|.$$

(b) Let P be Bernoulli(p) and Q Poisson(p). Show that

$$d(P,Q) = p(1 - e^{-p}) \leq p^2.$$

(c) Let P be Bernoulli(p) and Q Poisson(λ), where $\lambda = -\ln(1-p)$. Show that

$$d(P,Q) = p - \lambda e^{-\lambda} = 1 - e^{-\lambda} - \lambda e^{-\lambda} \leq \frac{1}{2}\lambda^2.$$

(d) Let P be the distribution of $\sum_1^n X_i$, where $X_1, \ldots, X_n$ are independent and X_i is Bernoulli(p_i), $1 \leq i \leq n$, and let Q be Poisson $(\sum_1^n \lambda_i)$, where $\lambda_i = -\ln(1-p_i)$. Show that

$$d(P,Q) \leq \frac{1}{2} \sum_1^n \lambda_i^2.$$

3. Let F be a cumulative distribution function. Its *median* is $\nu = F^{-1}(1/2) = \inf\{x : F(x) \geq 1/2\}$. Assume that ν is the *unique* solution of $F(x-) \leq 1/2 \leq F(x)$. Consider the *sample median*, based on a sample of independent and identically distributed random variables $X_1, \ldots, X_n$ from F, defined as

$$\hat{\nu}_n = X_{\lfloor \frac{n+1}{2}:n \rfloor},$$

where $X_{1:n} \leq \cdots \leq X_{n:n}$ denote the ordered sample values and $\lfloor y \rfloor$ denotes the greatest integer $\leq y$.

(a) The following exponential probability inequality for $\hat{\nu} - \nu$ is true (do not provide proof): for every $\varepsilon > 0$,

$$P(|\hat{\nu}_n - \nu| > \varepsilon) \leq 2e^{-2n\Delta_\varepsilon^2}, \tag{2}$$

with

$$\Delta_\varepsilon = \min\{F(\nu + \varepsilon) - 1/2, 1/2 - F(\nu - \varepsilon)\}.$$

Use this to conclude that $\hat{\nu}_n \to \nu$ with probability 1.

(b) Suppose that F is differentiable at ν with $F'(\nu) > 0$. Show that with probability 1

$$|\hat{\nu}_n - \nu| = O\left(\sqrt{\frac{\log n}{n}}\right), \ n \to \infty.$$

(c) Use (2) to obtain an exponential probability inequality for

$$\sup_{m \geq n} |\hat{\nu}_m - \nu|.$$

Ph.D. Qualifying Examination in Statistical Inference

April 8, 2009

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. Let $\xrightarrow{d}$ denote convergence in distribution and $\xrightarrow{p}$ denote convergence in probability.
   (a) Suppose that a statistic T_n for estimation of θ satisfies

$$\frac{T_n - \mu_n}{\sigma_n} \xrightarrow{d} N(0,1),\ n \to \infty, \tag{1}$$

where $\sigma_n \to 0$, $n \to \infty$, and

$$\mu_n \to \theta,\ n \to \infty. \tag{2}$$

Show that T_n is *consistent* for esimation of θ:

$$T_n \xrightarrow{p} \theta. \tag{3}$$

   (b) In order also to have

$$\frac{T_n - \theta}{\sigma_n} \xrightarrow{d} N(0,1),\ n \to \infty, \tag{4}$$

what (if any) additional condition(s) are needed?

---

2. Let $X_1, \ldots, X_n$ be a sample of i.i.d. random variables with univariate distribution $F(x-\theta)$, where F is given and has density f, and the location parameter θ is unknown and takes values in the real line $\mathbb{R}$.

(a) Consider testing the null hypothesis $H : \theta = 0$ versus the alternative $K : \theta = 1$. Give the form of a test that is most powerful of its size.

(b) For $H : \theta \leq 0$ versus $K : \theta > 0$, and $F =$ standard normal, show that the test in (a) is *uniformly most powerful* (UMP) of its size.

(c) For $H : \theta \leq 0$ versus $K : \theta > 0$, and f the *Cauchy* density

$$f(x) = \frac{1}{\pi} \frac{1}{1+x^2},$$

show that there does not exist a *uniformly most powerful* (UMP) test.

(d) When a UMP test does not exist, one may consider a *locally most powerful* (LMP) test, say ϕ, which maximizes the slope of the power function at the boundary between H and K among all tests of the same size. That is, ϕ is LMP if for any other test ϕ^* satisfying

$$\int \phi^*(x) f(x) dx \leq \int \phi(x) f(x) dx,$$

we have

$$\frac{d}{d\theta} \left[\int \phi^*(x) f(x-\theta) dx \right]_{|\theta=0} \leq \frac{d}{d\theta} \left[\int \phi(x) f(x-\theta) dx \right]_{|\theta=0}.$$

Show that for f the above Cauchy density, the LMP test exists and rejects H for sufficiently large values of

$$\sum_{i=1}^{n} \frac{2X_i}{1+X_i^2}.$$

3. Let F denote the distribution Binomial(k, p), where k is a known positive integer and $0 < p < 1$ is unknown. We are interested in estimating the probability of exactly one success, i.e., the parameter $\theta = P_p(X = 1) = kp(1-p)^{k-1}$, based on a random sample of size 2, X_1 and X_2, from F.

(a) Find the distribution of $X_1 + X_2$. Show that this statistic is complete and sufficient.

(b) Explicitly find a UMVUE for θ. Carefully justify your answer.

(c) Is your estimator in (b) unique? Why?

Ph.D. Qualifying Exam: Spring 2009
Linear models

- Number of questions = 3. Answer all of them. Total points = 50.
- Simplify your answers as much as possible and carefully justify all steps to get full credit.

---

1. Consider the simple linear regression model,

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

where the ϵ_i follow independent $N(0, \sigma^2)$ distribution, with β_0, β_1 and σ^2 as the unknown parameters. Let $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$ denote the estimators of these parameters derived using the standard least squares theory. Our interest is in computing a $100(1-\alpha)\%$ confidence interval for $\phi = -\beta_0/\beta_1$ by first constructing an interval for $\delta = -\phi + \overline{x}$ and then applying a transformation.

   (a) Show that $\overline{Y} - \delta\hat{\beta}_1 \sim N(0, \sigma^2 w)$, where [7 points]

$$w = \frac{1}{n} + \frac{\delta^2}{\sum_i (x_i - \overline{x})^2}.$$

   (b) Show that $T = (\overline{Y} - \delta\hat{\beta}_1)/(\hat{\sigma}\sqrt{w})$ follows a t_{n-2} distribution. [5 points]

   (c) Use the result in (b) to find a $100(1-\alpha)\%$ confidence interval for δ. [4 points]

   (d) Use the interval in (c) to find a $100(1-\alpha)\%$ confidence interval for ϕ. [2 points]

   (e) What difficulties, if any, would you face if you try to directly compute a confidence interval for ϕ by starting with the distribution of $\hat{\phi} = -\hat{\beta}_0/\hat{\beta}_1$? [2 points]

2. Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

where $\overline{x} = 0$ and the errors are *correlated* $N(0, \sigma^2)$ random variables with $cov(\epsilon_i, \epsilon_j) = \sigma^2\rho$, $i \neq j$. Assume that the correlation $0 < \rho < 1$ is known. The parameters β_0, β_1 and σ^2 are unknown.

   (a) Derive the best linear unbiased estimator of the 2×1 vector $\beta = (\beta_0, \beta_1)$, say $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$. Specify the distribution of $\hat{\beta}$. [8 points]

(b) Develop a suitable test for $H_0 : \beta_1 = 0$ versus, $H_1 : \beta_1 \neq 0$. [7 points]

3. Consider the linear model

$$Y_{ij} = \beta_i + \epsilon_{ij}, \quad j = 1, 2, \ldots, m_i, \; i = 1, 2, \ldots, n,$$

where the ϵ_{ij} follow independent $N(0, \sigma^2)$ distributions. Derive an F-test for [15 points]

$$H_0 : \beta_i = i, \; i = 1, \ldots, n, \quad \text{versus,} \quad H_1 : \text{ not } H_0.$$

Ph.D. Qualifying Exam in Statistical Methods

April 11, 2009

Project "Study of Depression", the data set is available on www.utdallas.edu/~mbaron/Qual

A study of 3,189 high school students has been conducted in order to find socioeconomic and family factors that may be associated with stress and depression. Data Set "depression.txt" (or "depression.sas7bdat" for SAS users) contains some variables obtained from this study.

Column	Variable
1	Participant's identification number
	Possible factors affecting depression
2	Gender (female or male)
3	Guardian status: 0=does not live with both natural parents 1=lives with both natural parents
4	Community Cohesion Score (16-80) - it shows how strongly the participant is connected to the community
	Response variables
5	Total Depression Score (0–60)
6	Clinical Diagnosis of Major Depression: (for the clinical sample of patients only) 1=positive diagnosis; 0=negative diagnosis

Use these data to answer the following questions. Support all your conclusions with suitable tests and model selection methods. State the necessary assumptions being used and verify them. Check for outliers and influential observations. Apply remedial measures if necessary.

1. What factors appear to have significant effect on the level of depression measured by the depression score?
2. Are effects of gender and guardian status fixed or random? Explain.
3. Is there any significant interaction, and what does its presence (or absence) mean?
4. Construct a 90% prediction interval for the total depression score of Jane, a female student who lives with both parents and has a cohesion score of 30.
5. Ann is also a female student with a cohesion score of 30, however, she lives with her aunt, separately from her parents. Derive the form and construct a 90% prediction interval for the *difference* in depression scores of Jane and Ann.
6. The students fall into 4 groups according to their gender and guardian status. Find all the significant differences in mean depression scores for these groups, keeping the experimentwise error rate not exceeding 0.05.
7. Do the high school students living with natural parents have a lower chance to be diagnosed with major depression?
8. Estimate the *probabilities* of being diagnosed with major depression for both Jane and Ann. Which factors are significant in estimating these probabilities?

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

Test significance of variable XYZ. Use SAS, PROC ... with option ... The F test gives a p-value of ... . 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.

Email the report to mbaron@utdallas.edu.