STAT 221/231 Final Examimation: December 13, 2016, 4:00-6:30 p.m., PAC, Sections 4-8

Seating is predetermined so please check your seat assignment at https://odyssey.uwaterloo.ca/teaching/schedule.

Bring your Watcard and a ruler. Only Pink Tie or Blue Goggles Calculators may be used.

You may bring one (1) double-sided, letter sized (8.5 x 11 inches), handwritten page of notes to the exam (no photocopies).

Normal, Chi-squared and t tables will be provided

To avoid round off errors carry as many decimal places as possible while making your calculations. Your final answers should be given to 3 decimal places.

The Final Examination covers all the material in the Course Notes excluding Sections 4.8, 5.4 and 6.4

You should do the following problems in the Course Notes:

Chapter 1: 1-20
Chapter 2: 1-19
Chapter 3: 1-7
Chapter 4, Problems: 1-30
Chapter 5 Problems: 1-13
Chapter 6, Problems: 1-19
Chapter 7, Problems: 1-11
Chapter 8, Problems: 1-8

Sample Midterm Test 1
Sample Midterm Test 2
Sample Final Exam

There will be multiple choice and short answer questions based on the R code in Assignments 1-5.

To aid you in creating your sheet of notes here is a list of the ideas, definitions and theorems covered in the course:

Empirical Studies (Chapter 1)

units, populations and processes (page 1)
variate and types of variates (page 3)
response versus explanatory variates (page 24)
attributes (page 4)
types of studies (sample surveys, observational studies, experimental studies)

Numerical Summaries (Section 1.3)

measures of location: sample mean, median, mode (pages 7-8)
measures of variability: sample variance, sample standard deviation, range, IQR (page 8 and Definition 3, page 12)
measures of shape: skewness, kurtosis (page 8)
sample percentiles and quantiles (Definition 1, page 11)
lower or first quartile, upper or third quartile (Definition 2, page 11)
five number summary (Definition 4, page 12)
numerical summaries for bivariate data: sample correlation (Definition 5, page 14), relative risk (Definition 6, page 15)

Graphical Summaries (Section 1.3)

relative frequency histograms (page 16)
empirical cumulative distribution function (page 19)
boxplots (page 20-21)
scatterplots (page 21)

Data Analysis and Statistical Models

descriptive statistics, statistical inference (inductive versus deductive reasoning) (page 25)
discrete statistical models: Binomial(n, θ), NegativeBinomial(k, θ), Poisson(θ), Geometric(θ)
continuous statistical models: Exponential(θ), G(μ, σ), Multinomial$(n, \theta_1, \theta_2, \ldots, \theta_k)$ (pages 45-46)

Graphical Checks of the Univariate Models (Section 2.6)

You should know how to check the fit of a model by:

1) calculating expected frequencies using the assumed model (discrete and continuous distributions) and comparing them with observed relative frequencies

2) examining a graph of the probability density function of the assumed model superimposed on a relative frequency histogram for the observed data (continuous distributions)

3) examining a graph of the cumulative distribution function of the assumed model superimposed on a graph of the empirical cumulative distribution function for the observed data (continuous distributions)

4) examining a Normal or Gaussian qqplot

Planning and Conducting Empirical Studies (Chapter 3)

the steps of PPDAC (page 87)
the elements of each step of PPDAC (Section 3.2)
types of problems (decriptive, causative, predictive) (page 88)
units, target population and target process (Definition 14, page 89)
study units, study population and study process (Definition 17, page 90)
study error (Definition 18, page 90)
sampling protocol and sample size (Definition 19, page 91)
sample error (Definition 20, page 91)
measurement error (Defintion 21, page 91)
fishbone diagram (page 95)

Estimation (Chapter 4)

Point Estimation

definition of a point estimate of a parameter θ (Definition 7, page 47)
likelihood functions for discrete distributions (Definition 8, page 48)
maximum likelihood estimate (Definition 9, page 48)
relative likelihood function (Definition 10, page 51)
log likelihood function (Definition 11, page 52)
likelihood function for a random sample (page 53)
likelihood functions for continuous distributions (Definition 12, page 57)
likelihood function for multinomial distribution (page 59)
invariance property of maximum likelihood estimates (Theorem 13, page 61)
definition of a point estimator of a parameter θ (Definition 22, page 113)
sampling distribution of an estimator (Section 4.2)
$\chi^2(k)$ distribution and its properties and how to read Chi-squared tables (Section 4.5)
$t(k)$ distribution and its properties and how to read the t tables (Section 4.5)

Interval Estimation

relative likelihood function (Definition 23, page 113)

$100p\%$ likelihood interval (Definition 24, p. 114)

log relative likelihood function (Definition 25, page 114)

how to obtain likelihood intervals **from a graph** of the relative likelihood function or the log relative likelihood function

behaviour of likelihood functions, log relative likelihood functions, and likelihood intervals as the sample size increases (page 114)

interpretation of likelihood intervals (Table 4.2, page 116)

definition of a confidence interval (Definition 27, page 117)

interpretation of a $100p\%$ confidence interval (page 118)

behaviour of the width of a confidence interval as the sample size increases (page 118)

definition of a pivotal quantity (Definition 28, page 118)

how to find a confidence interval given a pivotal quantity or approximate pivotal quantity

$100p\%$ confidence interval for the mean of a Gaussian distribution with known standard deviation (Example 4.4.2, page 119)

approximate $100p\%$ confidence interval for the Binomial proportion (Example 4.4.3, page 120)

choosing a sample size for a Binomial experiment (Example 4.4.5,pages 121-122)

approximate $100p\%$ confidence interval for the mean of a Poisson distribution (Problem 4.24(c))

confidence interval for θ for a random sample from Exponential(θ) distribution (Problem 4.23)

likelihood ratio statistic and its approximate distribution for large n (Theorem 33, page 128)

likelihood based confidence intervals - a 15% likelihood interval is an approximate 95% confidence interval and a 10% likelihood interval is an approximate 97% confidence interval.

$100p\%$ confidence interval for the mean μ of a Gaussian distribution with unknown standard deviation σ (p. 133)

$100p\%$ confidence interval for the variance σ^2 and the standard deviation σ of a Gaussian distribution with unknown mean μ (page 136)

choosing a sample size for a Gaussian experiment (p. 135)

Tests of Hypothesis (Chapter 5)

null hypothesis, alternative hypothesis (pages 157-158)

test statistic or discrepancy measure (Definition 38, page 158)

$p-value$ (Definition 39, page 159) and its interpretation (page 160)

listen to the $p-value$ bears: http://www.youtube.com/watch?v=ax0tDcFkPic&feature=related

guidelines to be used in STAT 231 for interpreting $p-values$ (Table 5.1, page 160)

important points about tests of hypotheses including practical versus statistical significance (see pp. 162-163)

relationship between confidence intervals and tests of hypotheses (p. 167)

hypothesis test for the mean μ of a Gaussian, with unknown standard deviation σ (Section 5.2)

hypothesis test for the variance σ^2 of a Gaussian distribution with unknown mean μ (Section 5.2)

how to use the likelihood ratio test statistic to test $H_0 : \theta = \theta_0$ for each of the following:

Binomial(n, θ), Geometric(θ), Negative Binomial(k, θ), Poisson(θ), Exponential(θ), $G(\theta, \sigma)$ where σ is known

how to use the likelihood ratio test statistic to test $H_0 : \theta = \theta_0$ given a model and a set of data, and how to obtain the approximate $p-value$ using Normal tables (Section 5.3)

Gaussian Response Models (Chapter 6)

Definition of a Gaussian Model (Definition 40, page 189)

Simple Linear Regression (Section 6.2)

model assumptions for simple linear regression model:

$$Y_i \backsim G(\alpha + \beta x_i, \sigma) \qquad i = 1, 2, \ldots, n \text{ independently}$$

where α, β and σ are unknown parameters and the x_i's are known constants.

maximum likelihood estimates and least squares estimates of α and β (pages 194-195)

unbiased estimate of σ^2 (page 195)

derivation of the distribution of the maximum likelihood estimator of β (pages 195-196)

confidence interval for β (page 197)

how to test the hypothesis of no relationship ($H_0 : \beta = 0$) (page 197)

confidence interval for mean response at x: $\mu(x) = \alpha + \beta x$ (pages 198-199)

prediction interval for response Y at x (pages 201-202)

see summary of distributions for simple linear regression (page 205)

how to check simple linear regression model assumptions (pages 206-208) using:

(1) scatterplot of data and fitted line

(2) residual plots:

(i) $(x_i, \hat{r}_i)$, $i = 1, 2, \ldots, n$ where $\hat{r}_i = y_i - \hat{\mu}_i$ and $\hat{\mu}_i = \hat{\alpha} + \hat{\beta} x_i$.

(ii) $(x_i, \hat{r}_i^*)$, $i = 1, 2, \ldots, n$ where $\hat{r}_i^* = \hat{r}_i / s_e$ (same as the graph in (i) except y-axis is rescaled).

(iii) $(\hat{\mu}_i, \hat{r}_i^*)$, $i = 1, 2, \ldots, n$.

(3) qqplot of the residuals

Comparing the Means of Two Populations (Section 6.3)

point estimators of the means (page 210)

pooled estimator of variance (page 210)

confidence interval for the difference between the means assuming equal variances (page 211)

confidence interval for the difference between the means with unequal variances and large sample sizes (page 214)

confidence interval for the difference between the means in a paired experiment (page 216)

when and why a paired experiment is better than two independent random samples (pages 217-218)

Multinomial Models (Chapter 7)

testing the fit of a model using the Multinomial model and the likelihood ratio goodness of fit test statistic (Section 7.1-7.2)

testing for independence of variates in two way tables (Section 7.3)

Cause and Effect (Chapter 8)

definition of causation

association between two variates does not imply a causal relationship (Section 8.1)

explanations for a positive correlation between two variates (page 257)

the importance of experimental studies, controlling variates and randomization in proving a causal relationship (Section 8.3)

observational studies and Simpson's paradox (Section 8.3)