

Prob/Stat Exam
May 20, 2005

Do as many problems as you can, including some from each section. Be sure to tackle some problems from each section.

MATH STAT QUESTIONS:

1. Suppose the random variable X has density $f_X(x) = \frac{3}{8}x^2$, $0 < x < 2$.
 - (a) Find $P(X > 1)$.
 - (b) Find the mean $E(X)$ and standard deviation $SD(X)$.
 - (c) Find the density of \sqrt{X} .

2. Suppose X, Y are random variables with means $E(X) = 1, E(Y) = 2$, standard deviations $SD(X) = 3, SD(Y) = 4$, and correlation $\rho(X, Y) = .3$. Find the mean and standard deviation of $2X + Y$.

3. A population has density $f(x) = \exp(-x)$, $x > 0$. We take a random sample of 30 observations from this population. We want to know the probability that the sum of the 30 observations exceeds 25.
 - (a) Show how to approximate the desired probability with the aid of a normal table. (Not having a normal table, you will not be able actually to do this.)
 - (b) Write down a one dimensional integral which is exactly equal to the desired probability.
 - (c) Suppose we are given a computer file of 100,000 randomly generated numbers from a random number generator which spits out deviates uniformly distributed on the unit interval.. How could we use these along with the computer to approximate the desired probability?

4. Suppose the random variables (X, Y) have the joint density
$$f_{X,Y}(x, y) = 2(x + y), \quad 0 < x, 0 < y, 0 < y < x < 1.$$
 - (a) Find $P(X + Y < 1)$.
 - (b) Find the marginal density $f_X(x)$.
 - (c) Find the conditional probability $P(Y < .5 | X > .25)$.
 - (d) Find the conditional probability $P(Y < .5 | X = x)$. (Your answer should be a function of x .)

5. Suppose X has the gamma distribution with $\alpha = 3, \beta = 1.5$. Thus

$$f_X(x) = \frac{x^2 \exp(-x/1.5)}{(1.5)^3 2!}, \quad x > 0. \text{ Find the expected value of } X^{-2}.$$

6. Suppose the output of a random mechanism is a random variable X with mass function $f(x|\lambda) = \exp(-\lambda) \frac{\lambda^x}{x!}$, $x = 0, 1, 2, \dots$; i.e., suppose X is Poisson with unknown mean $\lambda > 0$. Taking a Bayesian tack, give λ a Gamma prior with $\alpha = 3, \beta = 2$. Thus $\pi(\lambda) = \frac{\lambda^2 \exp(-\lambda/2)}{2^3 2!}$, $\lambda > 0$. Given 25 independent observations x_1, \dots, x_{25} of X , what is the posterior distribution $\pi(\lambda | x_1, \dots, x_{25})$?

7. Consider a normal population with density

$$f(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \text{ where } \mu \in \mathbb{R}, \sigma^2 > 0 \text{ are unknown.}$$

Suppose we plan a random sample x_1, \dots, x_n from the population.

- Define and explain intuitively: $T(x_1, \dots, x_n)$ is a sufficient statistic for (μ, σ^2) . You should define both "sufficient" and "statistic".
- Find a minimal sufficient statistic for (μ, σ^2) . You should show that you have done this, and also define "minimal sufficient statistic".

8. Consider a normal population with density $f(x|\tau) = (2\pi\tau)^{-1/2} \exp\left(\frac{-x^2}{2\tau}\right)$, $x \in \mathbb{R}$,

where $\tau > 0$ is unknown. Thus the population mean is known to be 0 and the population variance $\tau > 0$ is unknown. Suppose we plan a random sample x_1, \dots, x_n from this population.

- Find the maximum likelihood estimator for τ .
- Show that the MLE is unbiased for τ .
- Find the Cramer-Rao lower bound for unbiased estimators of τ .
- Does the MLE achieve the CRLB? If so, in what sense is it optimal? Explain.

9. Consider a population with density $f(x|\lambda) = \lambda \exp(-\lambda x)$, $x > 0$, where $\lambda > 0$ is unknown.

- What is the uniformly most powerful test of $H_0: \lambda = 1$ versus $H_1: \lambda > 1$, based on a single observation x_1 , of significance level $\alpha = .05$? Justify your answer.
- What is the power of your test corresponding to $\lambda = 2$?

LINEAR MODELS QUESTIONS:

1. Consider the model $D_i = r t_i + \varepsilon_i$, $i = 1, \dots, N$, where the errors ε_i are independent, and $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$ for each i . Find the least squares estimator of the rate r and derive its variance.

2. The dataset "European OECD," taken from the Data and Story Library (<http://lib.stat.cmu.edu/DASL>) includes the per capita income (PCINC) of 20 European OECD countries for 1960 along with the percentages of the labor force employed in agriculture (AGR), industry (IND), and services (SER) for each country.

- (a) The sample correlation matrix for the four variables is shown below, followed by the summary output for the full multiple regression model. Notice the p-value for the F statistic for the full model. What does this say about the explanatory merit of the full model? Should we simply discard the model? Now notice the p-values for the individual covariates. What do these seem to say about the explanatory values of the individual covariates? What is going on here? Can you do anything to fix this?

	PCINC	AGR	IND	SER
PCINC	1.000	-0.795	0.745	0.689
AGR	-0.795	1.000	-0.934	-0.868
IND	0.745	-0.934	1.000	0.635
SER	0.689	-0.868	0.635	1.000

Call:

```
lm(formula = PCINC ~ AGR + IND + SER, data = econ)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-4660.13	16298.65	-0.286	0.779
AGR	40.93	163.03	0.251	0.805
IND	59.89	162.89	0.368	0.718
SER	59.24	162.49	0.365	0.720

Residual standard error: 288.5 on 16 degrees of freedom

Multiple R-Squared: 0.6346, Adjusted R-squared:
0.5661

F-statistic: 9.262 on 3 and 16 DF, p-value: 0.0008716

- (b) An AIC-based stepwise procedure selected the model $PCINC = 19.032 \cdot IND + 18.521 \cdot SER$. An ANOVA test comparing this to a model with only IND gave the following output:

Analysis of Variance Table

Model 1: PCINC ~ IND + SER

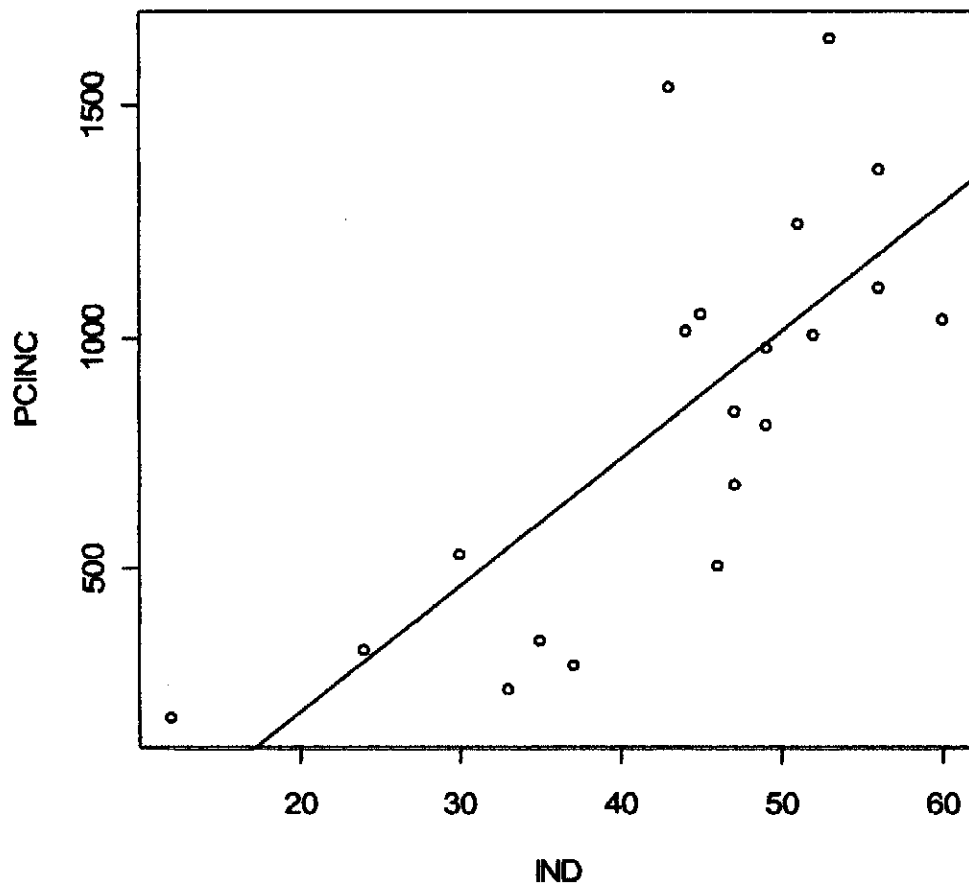
Model 2: PCINC ~ IND

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	17	1336707				
	18	1620762	-1	-284055	3.6126	0.07444

Thus, the model PCINC ~ IND has residual sum of squares 1620762 with residual df 18 while the model PCINC ~ IND + SER has residual sum of squares 1336707 with residual df 17.

Indicate how the F statistic 3.6126 comparing these models is found from the other information above. Based on these results, would you prefer the smaller or the larger model?

(c) The plot on the next page shows the fitted line for the simple linear regression model $PCINC = -359.31 + 27.49 \cdot IND$. How would you interpret the fit to the data? Which assumptions of the linear regression model appear to be violated, and what could you do to improve the model?



MARKOV CHAIN QUESTIONS:

1. Consider the following urn model: Two urns contain a total of two balls. Thus at each time $n = 0, 1, 2, \dots$ urn 1 contains $X_1(n)$ balls (possible values for $X_1(n)$ are 0, 1, 2) and urn 2 contains $X_2(n) = 2 - X_1(n)$ balls. After each time n , one of the two available balls is chosen at random and moved from its current urn to the other urn.

- Define a state space for this Markov chain.
- Write down the transition matrix.
- Show the chain is irreducible and aperiodic.
- Give the long run stable distribution. That is, for very large n , what is the approximate probability that $X_1(n) = 0$? $= 1$? $= 2$?
- after a very long time period, approximately what fraction of the time have there been 0 balls in urn 1? One ball in urn 1? Two balls in urn 1? Explain any theoretical connection between your answers in (d) and (e).

2. Now consider a more complicated urn model. Two urns contain a total of three balls. After each time n the following transition rule is in effect: if urn 2 contains two or more balls, one ball is moved from urn 2 to urn 1; if urn 1 contains two or more balls, one ball is chosen at random from the three available balls and moved from its current urn to the other urn.

Carry out an analysis of the long run behavior of this Markov chain. Set up a state space and transition matrix; identify recurrent and transient states. Discuss irreducibility. Find, or if pressed for time indicate clearly how to find, the long run stable distribution of the chain. Over the long haul, what fractions of the time will urn 1 have 0, 1, 2, 3 balls?