

## Math 472 (Statistics), Spring 08

### Final Exam (take home) given 5/8, due 5/12

The exam is due on Monday, 5/12 at 4:30 pm, to be submitted at the Math department front desk, 310 Malott Hall (Heather Peterson).

All citations and hints in the exam refer to the cumulative lecture notes (last version), as appearing on

<http://home.twcny.rr.com/minus/math472/lecnotes/stat-08.pdf>

Note that for problem 1, practice material and a quick guide to inference methods treated in the course appear on

<http://home.twcny.rr.com/minus/math472/final.html>

**Problem 1** [18%] (*Which method ?*) Read each scenario and indicate:

- (1) which inference procedure would be appropriate (one proportion, 2 proportions, one mean, two means, matched pairs, goodness of fit, homogeneity / independence, or regression slope);
- (2) the proper test statistic ( $z$ ,  $t$ , or  $\chi^2$ ); and
- (3) the number of degrees of freedom (if appropriate).

NOTE: Don't do the inference; just specify the method you would use!

Explanations are not required for your answers on this question. (*Independence / Homogeneity is treated as one type of inference* )

- a) Mean morning and evening commuting times on May 8, 2008 for a total of 200 randomly chosen drivers in 25 cities were collected. It is desired to find out whether morning or evening commutes tend to take longer.
- b) On May 7, 2008, a census of prisoners in each of the four US regions northeast, south, midwest, and west was conducted. We already knew the proportion of the total population (including only people living in these regions) that live in each region. We want to find out if the four regions have different tendencies to imprison people.
- c) According to the U.S. Department of Education, 51.9% of women and 46.7% of men receive financial aid in undergraduate school. An education researcher wishes to determine whether the difference in these percentages nationwide has increased. A random sample of 250 undergraduate female students revealed that 200 receive aid, while a sample of 300 male students showed that 180 receive aid.
- d) Nationwide, it is reported that graduates entering the actuarial field earn \$40,000 on average. A college placement officer feels that this number is too low. She surveys 36 graduates entering the actuarial field and finds the average salary to be \$41,000 with a standard deviation of \$3000.

- e) For each of the 50 states, the percentage of people with health insurance and the percentage of people who have quit smoking are collected. It is desired to study whether health insurance availability has an affect on the likelihood of people quitting smoking.
- f) A football coach looks at the records of his team in its last 250 games. He classifies every game as a big win, close game, or big loss. (A game decided by one touchdown or less is viewed as close.) He also keeps track of whether each game was home or away. He wants to know whether being at home affects the chance of these different kinds of game results.
- g) A survey found that in a random sample, 27% of the 254 men and 18% of the 177 women have purchased books on-line. Does this indicate a significant difference between the purchasing habits of males and females?
- h) Who is paid more - teachers or policemen? Assume all cities under consideration have their own fixed starting salary for teachers and another one for policemen. We select a random sample of 25 cities and record the starting salaries of teachers and policemen in each.
- i) A random sample of 122 college seniors who found jobs asked how many math courses they took and how many job offers they received. We wonder if there is an association between math background and employability.

**Problem 2** [10%] (*Testing variance*) Consider i.i.d. observations  $X_1, \dots, X_n$  with distribution  $N(\mu, \sigma^2)$  where both  $\mu \in \mathbb{R}$  and  $\sigma^2 > 0$  are unknown.

(a) For a certain  $\sigma_0^2 > 0$ , consider hypotheses  $H_0 : \sigma^2 = \sigma_0^2$  vs.  $H_a : \sigma^2 > \sigma_0^2$ . Find an  $\alpha$ -test with rejection region of form  $(c, \infty)$  (i.e. a one-sided test) where  $c$  is a quantile of a  $\chi^2$ -distribution.

(b) Check that the test also observes level  $\alpha$  on the extended null hypothesis  $H_a : \sigma^2 \leq \sigma_0^2$ .

**Problem 3** [16%] (*Two sample problem, equality of variances*) Let  $X_1, \dots, X_{n_1}$  be independent  $N(\mu_1, \sigma_1^2)$  and  $Y_1, \dots, Y_{n_2}$  be independent  $N(\mu_2, \sigma_2^2)$ , also independent of  $X_1, \dots, X_{n_1}$  ( $n_1, n_2 > 1$ ) where  $\mu_1, \sigma_1^2$  and  $\mu_2, \sigma_2^2$  are all unknown. Define the statistic

$$F = F(\mathbf{X}, \mathbf{Y}) = \frac{s_x^2}{s_y^2}, \text{ where}$$

$$s_x^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X}_n)^2, \quad s_y^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y}_n)^2$$

(here  $(\mathbf{X}, \mathbf{Y})$  symbolizes the total sample).

Define the **F-distribution with  $k, l$  degrees of freedom** (denoted  $F_{k,l}$ ) as the distribution of the random variable

$$\frac{k^{-1}H_1}{l^{-1}H_2}$$

where  $H_i$  are independent r.v.'s with  $H_1 \sim \chi_k^2$  and  $H_2 \sim \chi_l^2$ .

- (a) Show that  $F(\mathbf{X}, \mathbf{Y})$  has an  $F$ -distribution if  $\sigma_1^2 = \sigma_2^2$ , and find the degrees of freedom.
- (b) Consider hypotheses  $H_0 : \sigma_1^2 = \sigma_2^2$  vs.  $H_a : \sigma_1^2 > \sigma_2^2$ . Show that the following is an  $\alpha$ -test: reject if  $F(\mathbf{X}, \mathbf{Y}) > c$  for a certain  $c > 0$ , and find  $c$ .
- (c) Show that the level  $\alpha$  is also observed on the extended null hypothesis  $H_a : \sigma_1^2 \leq \sigma_2^2$ .
- (d) Assume  $n_1 = n_2$  and consider two sided hypotheses  $H_0 : \sigma_1^2 = \sigma_2^2$  vs.  $H_a : \sigma_1^2 \neq \sigma_2^2$ . Show that the following is an  $\alpha$ -test: reject if  $F(\mathbf{X}, \mathbf{Y}) \notin [c^{-1}, c]$  for a certain  $c > 0$ , and find  $c$ .

**Problem 4** [16%] (*McNemar's Test*) Let  $\mathbf{M} = (M_1, \dots, M_k)^\top$  have a multinomial law  $\mathfrak{M}_k(n, \mathbf{p})$  with unknown probability vector  $\mathbf{p} = (p_1, p_2, \dots, p_k)^\top$ , where  $k > 2$ . Consider hypotheses on the first two components

$$H_0 : p_1 = p_2$$

$$H_a : p_1 \neq p_2.$$

A test that is often used, called *McNemar's Test*, rejects  $H_0$  if

$$M_{cN} := \frac{(M_1 - M_2)^2}{M_1 + M_2} > \chi_{1,\alpha}^2$$

where  $\chi_{1,\alpha}^2$  is the upper  $\alpha$  quantile of the distribution  $\chi_1^2$ .

Show that this is an asymptotic  $\alpha$ -test, i.e. that McNemar's statistic  $M_{cN}$  has an asymptotic  $\chi_1^2$  distribution as  $n \rightarrow \infty$ .

**Hint:** it is possible to use the multivariate CLT as in Theorem 4.2, p. 111, but it is easier to apply the ordinary (univariate) CLT.

**Problem 5** [10%] (*Testing correlation*) Suppose we observe a bivariate sample  $(X_i, Y_i)$ ,  $i = 1, \dots, n$  where the pairs  $(X_i, Y_i)$  are i.i.d. and follow a joint normal distribution with mean vector  $\boldsymbol{\mu} = (\mu_x, \mu_y)^\top$  and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

where  $\rho$  is the correlation coefficient between  $X$  and  $Y$ . Consider the sample covariance matrix

$$\hat{\Sigma} = \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix}$$

such that

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, \quad s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2,$$

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

and the empirical correlation coefficient

$$R = \frac{s_{xy}}{s_x s_y}.$$

Consider hypotheses

$$H_0 : \rho = 0$$

$$H_a : \rho \neq 0.$$

Define the statistic

$$Q(\mathbf{X}, \mathbf{Y}) = \sqrt{n-1} \frac{R}{\sqrt{1-R^2}}$$

(here  $(\mathbf{X}, \mathbf{Y})$  symbolizes the total sample). Show that  $Q(\mathbf{X}, \mathbf{Y})$  can be used as a test statistic to construct an  $\alpha$ -test, and indicate the distribution and the quantile used to find the rejection region.

**Hint:** In Theorem 5.12 (p. 141) a test was found for the null hypothesis  $H_0 : \beta = 0$  and nonrandom  $x_i$ , but now the  $X_i$  are random.

**Problem 6** (*Analysis of variance*) Consider observations formed by  $m$  normal samples of size  $l_j$ ,  $j = 1, \dots, m$ , i.e. observations

$$Y_{jk} = \mu_j + \varepsilon_{jk}, \quad k = 1, \dots, l_j, \quad j = 1, \dots, m,$$

where  $\varepsilon_{jk} \sim N(0, \sigma^2)$  are i.i.d. normal noise variables and  $\mu_j, \sigma^2$  are unknown parameters. The total number of observations then is  $n = \sum_{j=1}^m l_j$ . Assume that  $l_j > 1$  for at least one  $j$ . Consider hypotheses

$$H_0 : \mu_1 = \dots = \mu_m$$

$$H_a : \mu_i \neq \mu_j \text{ for at least one pair } i \neq j.$$

Define the statistic

$$F(\mathbf{Y}) = \frac{(m-1)^{-1} SS_b}{(n-m)^{-1} SS_w} \text{ where}$$

$$SS_w : = \sum_{j=1}^m \sum_{k=1}^{l_j} (Y_{jk} - \bar{Y}_j)^2, \quad SS_b := \sum_{j=1}^m l_j (\bar{Y}_j - \bar{Y}_{..})^2 \text{ and}$$

$$\bar{Y}_j : = \frac{1}{l_j} \sum_{k=1}^{l_j} Y_{jk}, \quad \bar{Y}_{..} := \frac{1}{n} \sum_{j=1}^m \sum_{k=1}^{l_j} Y_{jk}$$

and  $\mathbf{Y}$  represents the total data.

(a) [15%] For the special case that all the group sample sizes are equal:  $l_1 = \dots = l_m$ ; find an  $\alpha$ -test for the above hypotheses based on  $F(\mathbf{Y})$ , and indicate the distribution and quantile to be used.

(b) [15%] Generalize (a) to the case of unequal  $l_j$ ,  $j = 1, \dots, m$ . **Hint:** To treat  $SS_b$ , obtain first  $\bar{Y}_j = \sigma l_j^{-1/2} Z_j$  for i.i.d.  $Z_j \sim N(0, 1)$  and then generalize the argument of Proposition 3.3, p 80.

**Comment.** The groups  $j = 1, \dots, m$  are also called factors or treatments. The question of interest is "do the factors have an influence upon the measured quantity  $Y_{jk}$ ?" The difference to regression is that in ANOVA no numerical value  $x_i$  is attached to the groups  $j = 1, \dots, m$ ; they represent just different categories (qualitative in nature). An example for  $m = 2$  is the drug testing problem, where one has two samples, one for old and new drug ( $m = 2$ ), with the same variance  $\sigma^2$ . For ANOVA with  $m$  groups, an example would be that  $j$  represents different social groups and  $Y_{jk}$  the cholesterol level, or  $j$  might represent different cities and  $Y_{jk}$  are measurements of air pollution etc.

$SS_w$  is called the "within groups sum of squares" and  $SS_b$  the "between groups sum of squares". It can be shown that the total sum of squares

$$SS_t := \sum_{j=1}^m \sum_{k=1}^{l_j} (Y_{jk} - \bar{Y}_{..})^2$$

can be decomposed as

$$SS_t = SS_w + SS_b$$

hence the name "analysis of variance" (comp. relation (142), p. 142 for a similar decomposition in bivariate regression).