Practice Final Exam

Statistical Methods and Models - Math 410, Fall 2011

December 4, 2011

You may use a calculator, and you may bring in one sheet (8.5" by 11" or A4) of notes. Otherwise closed book.

The real test will have about 10 questions.

Be sure when you do a hypothesis testing problem that you describe your parameters in terms of the problem when stating H_0 and H_a. Also be sure you rephrase you conclusion in terms of the original question being asked in the problem.

(1) You have a set of three dice.

(a) Calculate the probability of getting 0, 1, 2 and 3 sixes if the dice are fair.

(b) Suppose you roll the three dice 100 times. The results are

# of 6s	# of times
0	47
1	35
2	15
3	3

Test the null hypthesis that the dice are fair at $\alpha = .01$ significance level.

SOLUTION:

(a) *Note that if K is a binomial random variable with paramter $\pi = 1/6$ measuring the probability of a success (a 6). The probability of 0 sixes is $(5/6)^3 = .5787$. The probability of 1 six is $\binom{3}{1}(1/6)(5/6)^2 = .3472$. To get 2 sixes, $\binom{3}{2}(1/6)^2(5/6) = .0694$ and to get 3 sixes we have probability $(1/6)^3 = .0046$.*

(b) *We compute the χ^2 statistic. The table gives us observed values O_0, O_1, O_2 and O_3. We use the probabilities from the previous parts to calculated expected values by multiplying by 100. This gives*

$$\chi^2 = \frac{(57.87-47)^2}{57.87} + \frac{(34.72-35)^2}{34.72} + \frac{(6.94-15)^2}{6.94} + \frac{(.46-3)^2}{.46} = 25.43.$$

(i) *H_0 is that the distribution of the count of sixes be as given by the probabilities from the answer to part a.*

(ii) *H_a is that the distribution of sixes is otherwise.*

(iii) *The significance level is $.01$ as given.*

(iv) *The test statistic is $\chi^2 = 25.43$.*

(v) *The p-value will be the probability that (given H_0) χ^2 would be greater than 25.43. We have 4 classes, so 3* *degrees of freedom. The associated p-value is very small (well under α).*

(vi) *We reject H_0, concluding that the number of sixes on rolling our three dice is not represented by the probabilities described in the previous part, at the .01 significance level.*

(2) Males and females were asked about what they would do if they received a \$100 bill by mail, addressed to their neighbor, but wrongly delivered to them. Would they return it to their neighbor? 69 men were sampled, of whom 52 said yes. 131 females were sampled, of whom 120 yes. Does the data indicate that the proportions that said yes are different for male and female at a 5% level of significance?
SOLUTION:*Let π_1 be the proportion of men who would return the \$100 and π_2 be the proportion of women who would return the \$100. $p_1 = 52/69 = .7536$ is the proportion of men from our sample who would return the \$100. $p_2 = 120/131 = .916$ is the proportion of women from our sample who would return the \$100.*

(a) $H_0 : \pi_1 = \pi_2$.
(b) $H_a : \pi_1 \neq \pi_2$.
(c) $\alpha = .05$.
(d) *We make a test z statistic using our pooled population proportion $p = \frac{52+120}{69+131} = \frac{177}{200} = .885$. The standard error is $\sqrt{.885 \cdot .115(\frac{1}{131} + \frac{1}{69})} = .0475$.*

$$\overline{z} = \frac{p_1 - p_2}{SE} = \frac{-.1624}{.0475} = -3.4189.$$

(e) *We want $P(|z| \geq \overline{z}) = 2(1 - .99968) = 2 \cdot .00032 = .00064$. Since this is (much) less than α, we reject H_0 and conclude that the proportion of men who are likely to return the \$100 is smaller than the proportion of women.*

(3) Ten students have first and second exam scores as follows:

first	23	33	44	33	84	76	65	79	85	47
second	27	38	52	38	98	90	77	93	100	55.

Use this data to give a 95% confidence interval for the improvement in test scores from the first to the second exams.

SOLUTION: *We want to estimate the mean improvement for all hypothetical students who might take be taking this course. We use the data that is the improvements in scores for the 10 students actually in this course. This data is*

$$\{4, 5, 8, 5, 14, 14, 12, 14, 15, 8\}$$

The mean improvement is $\overline{x} = 9.1$. The sample standard deviation is 4.36. We use a t statistic $9 = 10 - 1$ df. $t_{.975} = 2.2622$. So we get

$$5.981 = 9.1 - 2.2622 \cdot 4.36/\sqrt{10} \leq \mu \leq 9.1 + 2.2622 \cdot 4.36/\sqrt{10} = 12.219.$$

with 95% confidence.

(4) Nineteen female rats are fed a specific diet between their 28th day and their 84th day. The groups were chosen randomly and 12 were fed a high protein diet while 7 were fed a low proteien diet.

The mean weight gain for the 12 rats on a high protein diet was 120 grams, and the sample standard deviation is 21.39. For the rats on the low protein diet it was 101 grams with a sample standard deviation of 20.62. Construct a 95% confidence interval for the difference between weight gain on the high protein diet and weight gain on the low protein diet.

SOLUTION: $120 - 101 = 19$. *The standard error for the different weight gains is*

$$SE = \sqrt{\frac{21.39^2}{12} + \frac{20.62^2}{7}} = 9.94.$$

We use 6 degrees of freedom since the smaller sample is size 7. Then $t_{.975} = 2.4469$. *So*

$$-5.32 = 19 - 2.4469 \cdot 9.94 \leq \mu_1 - \mu_2 \leq 19 + 2.4469 \cdot 9.94 = 43.32$$

with 95% certainty.

We could use Welch's formula to determine the degrees of freedom instead. This gives $df = 13.0845$, *and so* $t_{.975} = 2.1604$, *and so*

$$-2.4744 = 19 - 2.1604 \cdot 9.94 \leq \mu_1 - \mu_2 \leq 19 + 2.1604 \cdot 9.94 = 40.4744$$

which is of course a slightly narrower range.

(5) In a 1974 study, 862 Marine recruits were assigned randomly to either take daily 2 gram vitamin C supplements, or a placebos for eight weeks. During the study some recruits were removed from their platoons, and others were excluded from analysis because they didn't take their pills for the entire 8 weeks. At the end of the study there were 331 recruits in the vitamin C group, and 343 in the placebo group.

The vitamin C group spend a mean of 20.3 days with a cold, with standard deviation .879 days. The placebo group spent a mean of 20.7 days with a cold, with standard deviation .642.

Is there evidence at the 5% significance level that the vitamin C treatment reduces the length of cold symptoms from this study?

SOLUTION: *Let* μ_1 *be the mean number of days over an eight week period that marines who are taking the vitamin C supplement have cold symptoms. Let* μ_2 *be the mean number of days over an eight week period that marines who are not taking the vitamin C suppleents have colds.*

(a) $H_0 : \mu_1 = \mu_2$.
(b) $H_a : \mu_1 < \mu_2$.
(c) $\alpha = .05$.

(d) *To calculate our test statistic, we would use a t statistic, but since n is large, we can approximate by a z statistic. Our standard error is*

$$SE = \sqrt{\frac{.879^2}{331} + \frac{.642^2}{343}} = .059.$$

So

$$\overline{z} = \frac{20.3 - 20.7}{.059} = -6.7797.$$

(e) *The p-value is very small, so that $P(|z| \geq \overline{z}) < \alpha$. So we reject H_0, and conclude that vitamin C supplements change the number of days with cold symptoms.*

(6) Use a one-sided χ^2 contingency table at the $\alpha = .05$ level to test whether acupuncture is more successful than placebo using the following table of data:

	Success	Failure	Total
Acupuncture	7	6	13
Placebo	4	13	17
Total	11	19	30

SOLUTION: *Our table of expected counts is*

	Success	*Failure*	*Total*
Acupuncture	4.77	8.23	13
Placebo	6.23	10.77	17
Total	11	19	30

(a) H_0 : *We see the same percentage of failures and successes in people treated with acupuncture as in people treated with a placebo.*

(b) H_a : *We see a higher percentage of successes in the acupuncture treated group.*

(c) $\alpha = .05$.

(d) *Our test statistic is*

$$\chi^2 = \frac{(7-4.77)^2}{4.77} + \frac{(4-6.23)^2}{6.23} + \frac{(6-8.23)^2}{8.23} + \frac{(13-10.77)^2}{10.77} = 2.90673.$$

(e) *Our p-value is*

$$\frac{1}{2}P(\chi^2 \geq 2.907) < \frac{1}{2}P(\chi^2 \geq 2.7055) = \frac{1}{2}.1 = .05.$$

(f) *Since the p-value is less than α, we reject H_0 and conclude that acupuncture is more effective than placebo treatment.*

(7) Suppose independent random samples from two normal populations gave the following results:

$$n_X = 25, \overline{x} = 16, s_x = 4.7$$

$$n_Y = 45, \overline{y} = 20, s_y = 2.3.$$

Find a conservative .95 confidence interval for the difference $\mu_X - \mu_Y$.
SOLUTION: *$\overline{x} - \overline{y} = -4$. The standard error is*

$$SE = \sqrt{\frac{4.7^2}{25} + \frac{2.3^2}{45}} = 1.00058.$$

We use 24 df and $t_{.975} = 2.0639$. So we get

$$-6.0651 = -4 - 2.0639 \cdot 1.00058 \leq \mu_X - \mu_Y \leq -4 + 2.0639 \cdot 1.00058 = -1.9349/$$

(8) How large a sample is needed to obtain an estimate of the mean IQ of college students to within a 95% margin of error of 1 point? (Use $\sigma = 15$.)

SOLUTION: *Let $\overline{x}$ be our sample mean, and n be the sample size. Then using $z_{.975} = 1.96$, we get*

$$\overline{x} - 1.96 \cdot 15/\sqrt{n} \leq \mu \leq \overline{x} + 1.96 \cdot 15/\sqrt{n}.$$

We want $1.96 \cdot 15/\sqrt{n} \leq 1$. This is the same as $1.96 \cdot 15 \leq \sqrt{n}$, which is in turn the same as $1.96^2 \cdot 225 \leq n$. So we need n at least 865 since $1.96^2 \cdot 225 = 864.36$.

(9) A professional basketball player used to make 55% of his free throws. His agent claims he has improved, and offers as evidence that he has made 40 out of his last 60 free throws. Test the hypothesis at the .05 significance level that his free throw percentage has improved from 55%.

SOLUTION: *Our sample proportion is $p = 2/3$ and our sample size is 60. Let π be the percentage of free throws our basketball player is now making.*

(a) $H_0 : \pi = .55$.
(b) $H_a : \pi > .55$.
(c) $\alpha = .05$.
(d) *Our test statistic is 40 successes in our sample. The associated z-statistic is*

$$\overline{z} = \frac{40 - .55 \cdot 60}{\sqrt{60 \cdot .55 \cdot .45}} = 1.8165.$$

(e) *Our P-value is $P(z \geq 1.8165) = (1 - .965) = .035$.*
(f) *The P-value is less than α so we reject H_0 and conclude that the basketball player's freethrow rate is now above 55%. Note that this is contingent on treating his last 60 throws as a random sample of his current ability. This is open to question.*

(10) Assume an airline flies a plane whose safe passenger load is 16000 pounds. Assume the passengers weigh 190 pounds on average (overall) with a standard deviation of 40 pounds, and that on each flight we have a random sample of 80 passengers. How often will the plane be overloaded?

SOLUTION: *We assume that the weight of passengers is approximately normal. We are asking how likely 80 passengers are to weigh more than 16000 pounds. The meain weight of 80 passengers is* $80 \cdot 190 = 15200$, *and the standard deviation is* $\sqrt{8040} = 357.71$.

$$16000 = 15200 + 800 = 15200 + 2.236 \cdot 357.71.$$

So we have to ask for the probability that the total weight is more than 2.236 standard deviations about the mean. Since the distribution is normal, this probability is about .013.

(11) Compute μ, $E(X^2)$ and σ for the random variable X with pdf

$$f(x) = \begin{cases} |x| & -1 \le x \le 1 \\ 0 & \text{else.} \end{cases}$$

SOLUTION:

$$\mu = E(X) = \int_{-1}^{0} -x^2 \, dx + \int_{0}^{1} x^2 \, dx = 0.$$

$$E(X^2) = \int_{-1}^{0} -x^3 \, dx + \int_{0}^{1} x^3 \, dx = \frac{2}{4} = .5.$$

The variance is $E(X^2) - E(X)^2 = .5$, *so the standard deviation is* $1/\sqrt{2}$.

(12) Compute μ, $E(X^2)$ and σ for the random variable X with pdf

$$f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{else.} \end{cases}$$

We assume $a < b$ here.

SOLUTION:

$$\mu = E(X) = \int_{a}^{b} \frac{x}{b-a} \, dx = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}.$$

$$E(X^2) = \int_{a}^{b} \frac{x^2}{b-a} \, dx = \frac{b^3 - a^3}{3(b-a)} = \frac{b^2 + ab + a^2}{3}.$$

The variance is

$$E(X^2) - E(X)^2 = \frac{b^2 + ab + a^2}{3} - \frac{b^2 + 2ab + a^2}{4} = \frac{b^2}{12} - \frac{2ab}{12} + \frac{a^2}{12} = \frac{(a-b)^2}{12}.$$

Thus the standard deviation is $(b-a)/2\sqrt{3}$.

(13) The length of time necessary to perform a specific manufacturing operation is a random variable T with mean 12 and standard deviation 2 minutes. Use Chebyshev's Inequality to find a lower bound for $P(6 < T < 18)$.

SOLUTION: *We are looking at values within 3 standard deviations of the mean. So we are asking for* $P(-3 < Z < 3) > 1 - \frac{1}{9} = \frac{8}{9}$.

(14) An urn contains 50 black balls and 10 white balls. You pick 10 balls at random. Give a formula for the probability of picking all black balls.

SOLUTION: *This is*

$$\frac{50}{60}\frac{49}{59}\frac{48}{58}\cdots\frac{41}{51} = \frac{50!50!}{40!60!}$$

This is about 14%.

(15) Give the definition of E being independent of F. Prove that if E is independent of F then F is independent of E.

SOLUTION: *E is independent of F means $P(E) = P(E|F)$. The definition of $P(E|F)$ is $P(E \cap F)/P(F)$. So E independent of F means*

$$P(E) = P(E|F) = \frac{P(E \cap F)}{P(F)}.$$

To show F is independent of E, we want to show $P(F) = P(F|E)$. We take the equation above and multiply by $P(F)$ and then divide by $P(E)$. We get then

$$P(F) = \frac{P(E \cap F)}{P(E)} = P(F|E)$$

which is what we want to prove.

(16) State Bayes's Theorem. A test for a particular cancer has sensitivity 2/3, specificity .91. Prevalence among the relevant population is .015. Calculate the predictive value positive.

SOLUTION:

(a) *Bayes's Theorem:*

$$P(E|F) = \frac{P(E)P(F|E)}{P(E)P(F|E) + P(\overline{E})P(F|\overline{E})}$$

(b) *This is the probability of the disease if the test is positive.*

$$PV^{+} = P(\text{disease}|\text{positive}) = \frac{P(\text{disease})P(\text{positive}|\text{disease})}{P(\text{disease})P(\text{positive}|\text{disease}) + P(\text{no disease})P(\text{positive}|\text{no disease})}$$

$$= \frac{.015 \cdot (2/3)}{.015 \cdot (2/3) + .985 \cdot .09} = \frac{.01}{.01 + .08865} = .101368.$$

(17) Suppose $\pi = .55$ (a population proportion). But you don't know π and you are testing the hypothesis $\pi = .5$ against the hypothesis $\pi > .5$. Your procedure will be to pick a random sample of 20 individuals and do a z-test at the .02 significance level. What is the power of this test?

SOLUTION: *The power of the test tells us how likely we are to reject H_0 when H_0 is false. So o calculate the power here (as H_0 is indeed false) we need the probability of rejecting H_0.*

- *We reject H_0 if $P(z \geq \overline{z}) \leq .02$. That corresponds to $\overline{z} \geq 2.0537$.*

- *If K is the number of successes in our sample of 20, then under the hypothesis $H_0 : \pi = .5$, we'll have*
$$\overline{z} = \frac{K - 10}{\sqrt{20 \cdot .5 \cdot .5}}$$
- *So $\overline{z} \geq 2.0537$ corresponds to*
$$\frac{K-10}{2.23607} \geq 2.0537 \text{ that is } K - 10 \geq 4.5922.$$
- *So we'll reject H_0 when $K \geq 15$.*
- *Given $\pi = .55$, we now need to calculate the probability that $K \geq 15$. We could do this directly, or by using the normal approximation with the continuiuty correction. We'll do it directly:*
$$P(K \geq 15) = \sum_{i=15}^{20} \binom{20}{i} .55^i .45^{20-i} = .055.$$