THE UNIVERSITY OF MINNESOTA

Statistics 5021 March 30, 2003

Statistics 5021 Sample Second Midterm Examination II with Solutions

1. A grocery store attempts to stock many food items, among them asparagus (A), rye bread (B), and cheddar cheese (C). On any randomly chosen day an item may be available or not available. The following table lists symbols for some events defined in terms of availability of these foods.

Food	Available	Not Available
Asparagus	A+	A–
Rye bread	B+	B–
Cheddar cheese	C+	C–

Here is a table of the probabilities of all combinations of these events.

	B+			B–			
	A+	A–	Totals	A+	A–	Totals	Totals
C+	.25	.10	.35	*.20*	.15	.35	.70
C–	.10	.05	.15	.10	.05	.15	.30
Totals	.35	.15	.50	.30	.20	.50	1.00

[underlining not in original]

For example, P(Cheddar cheese and asparagus are available but not rye bread) = P(**C+** ***and*** **A+** ***and*** **B–**) = .20 (value in *italics* in the table)

(a) Find the conditional probability P(**C–** | **A–**) = P(cheese not available given asparagus not available)

P(**C–** | **A–**) = P(**C–** *and* **A–**)/P(**A–**)
From the table P(**A–**) = .15 + .20 = .35 and P(**C–** *and* **A–**) = .05 + .05 = .10
Therefore P(**C–** | **A–**) = .10/.35 = .286

(b) Find P(**(C–** ***and*** **A–)** ***or*** **B–**)

P(**(C–** ***and*** **A–)** ***or*** **B–**) = P(**C–** ***and*** **A–**) + P(**B–**) – P(**C–** ***and*** **A–** ***and*** **B–**)
= .10 + .50 – .05 = .55
Or just add the underlined cells in the table

(c) Are events **C+** and **B+** independent? Justify your answer.

This requires computing probabilities. You have to check just one of the following:

P(**C+** ***and*** **B+**) = P(**C+**)P(**B+**) .35 = .70×.50 => independent
P(**C+** | **B+**) = P(**C+**) .35/.50 = .70 => independent
P(**B+** | **C+**) = P(**B+**) .35/.70 = .50 => independent
They are independent.

(d) Are outcomes **C–** and **A–** independent? Justify your answer.

Again you need to check just one of the following

P(**C–** ***and*** **A–**) = P(**C–**)P(**A–**) .10 = .30×.35 = .105 =>*not* independent
P(**C–** | **A–**) = P(**C–**) .10/.35 ≠ .30 => *not* independent
P(**A–** | **C–**) = P(**A–**) .10/30 ≠ .35 => *not* independent
They are *not* independent.

2. In a test for ESP (extrasensory perception), the experimenter uses cards that are hidden from a subject. Each card contains one of a star ☆, a circle ○, a wave ≈, a square □ or a cross +. The experimenter looks at many cards in random order, thinks of the shape on the card, and the subject guesses what the shape actually is. If a subject does not have ESP, the probability of a correct guess should be 1/5 = .2 and getting success on different cards should be independent.

Suppose the experiment consisted of guesses as to the contents of 800 cards.

(a) If the subject does not have ESP, what are the mean and standard deviation of the proportion of correct guesses out of the 800 guesses?

Under the assumptions we have n = 800 trials with constant probability of success p = 20. Thus the number of successes X is binomial. The proportion is $\hat{p}$ = X/n

The mean of $\hat{p}$ is p = .20 and standard deviation is $\sigma_{\hat{p}} = \sqrt{p(1-p)/n} = \sqrt{.2\times .8/800}$ = 0.014142.

(b) If the subject does not have ESP, what is the approximate probability that the percent of correct guesses is less than 17.5%.

What is wanted is P($\hat{p}$ < .175). Using the normal approximation

P($\hat{p}$ < .175) = P(Z < (.175 – .2)/ 0.014142) = P(Z < –1.77) = 0.0384 using Table A.

The exact value computed in MacAnova by `cumbin(800*.175,800,.2)` is 0.040728.

(c) If the subject does not have ESP, find the number y of successful guesses such that P(number of correct guesses ≥ y) = .01.

Solve the problem for Z and then unstandardize.

From Table D, the upper 1% value for Z is 2.326.

Since the count of successes is 800×$\hat{p}$ its mean is 800×.2 = 160 and its standard deviation is 800× 0.014142 = 11.31 (or use $\mu = np$ and $\sigma = \sqrt{n\times p(1-p)}$). Thus the required value is

160 + 11.31×2.326 = 186.3 ~ 186

3. The level of calcium in the blood of healthy young adults varies with mean about 9.5 milligrams per deciliter and standard deviation about $\sigma = .4$. Here are data on the blood calcium level of 40 healthy rural Guatemala pregnant women

Calcium level in blood of 40 women									
10.63	9.25	9.66	9.20	9.97	9.28	9.65	9.57	9.16	9.42
9.73	9.57	9.78	9.17	9.63	9.79	8.92	9.91	10.20	9.69
9.93	10.11	9.83	9.65	9.53	9.69	9.71	9.41	9.47	8.86
9.79	10.16	9.52	9.48	9.50	9.67	10.00	8.85	9.94	9.10

Here's a little MacAnova output

```
Cmd> x <- \
   vector(10.63,9.25,9.66,9.20,9.97,9.28,9.65,9.57,9.16,9.42,\
          9.73,9.57,9.78,9.17,9.63,9.79,8.92,9.91,10.20,9.69,\
          9.93,10.11,9.83,9.65,9.53,9.69,9.71,9.41,9.47,8.86,\
          9.79,10.16,9.52,9.48,9.50,9.67,10.00,8.85,9.94,9.10)

Cmd> sum(x)
(1)       384.38

Cmd> sum((x - sum(x)/40)^2)
(1)       5.5344
```

s = is not needed here since you know $\sigma = .4$, but if you didn't,

$$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} = \sqrt{\frac{5.5344}{39}} = .3767$$

(a) Is there statistical evidence that the mean blood calcium level in this population of women is higher than that in the population of healthy young adults. Use $\alpha = .05$. State hypotheses and show your work which should include an approximate P-value.

H_0: $\mu = 9.5$, H_a: $\mu > 9.5$, so you use a one (right) tail test.

The test statistic is

$$z = (\bar{x} - 9.5)/\sigma_{\bar{x}} = (\bar{x} - 9.5)/(.4/\sqrt{40})$$

$$\bar{x} = \sum x/n = 384.38/40 = 9.6095$$

$$z = (9.6095 - 9.5)/0.06325 = 1.7312$$

From Table D, the upper 5% critical value is 1.645. Since 1.731 > 1.645, you reject H_0; there *is* statistical evidence.

(b) Find a 95% confidence interval for the mean calcium level in this population,
Again from Table D, the critical value for a 95% confidence interval is 1.96 so the confidence interval is $9.6095 \pm 1.96 \times 0.06325$ = (9.486. 9.734).

4. In the following, decide whether each underlined statement is **true** or **false**. Then circle your choice and *explain why*.

(a) Based on a readings from 12 radiation detectors exposed to radon (a radioactive gas), a confidence interval for the mean radioactivity level μ was calculated to be (98.16 , 110.10).

The probability is .95 that μ is between 98.16 and 110.10 **True** **False**

The confidence level was accidentally omitted, but was irrelevant. The confidence level has nothing to do with the probability. Assuming the intended confidence level was .95, in a long history of computing 95% confidence intervals μ will be inside the interval about 95% of the time. But at this point, either μ is between 98.16 and 110.10 (in which case the probability is 1), or it's not (in which case the probability is 0).

(b) Using the same data, a test of H_0: μ = 105 vs H_a: μ ≠ 105 yielded a P-value = 0.755.

The probability that H_a is true is 1 – .755 = .245 **True** **False**

A P-value is the probability, computed under H_0, of observing a value of the test statistic which is at least as extreme as the observed one. It has no valid interpretation as the probability of a hypothesis.

(c) The 7 Twin Cities area counties are Anoka, Carver, Dakota, Hennepin, Ramsey, Scott, and Washington. Let Y be the number of these counties in which there will be a thunderstorm on July 1, 2001.

Y is binomial with n = 7 and unknown probability p **True** **False**

Because of the nature of thunderstorms, a day on which there is a thunderstorm in one county is likely to have a thunderstorm in a neighboring county. Hence, even though there are 7 events, with two outcomes, they are not independent events and hence the distribution is *not* binomial. Moreover, an argument might be made that that probabilities are not the same.