December 11, 2007

Final Exam — Math 102 / Core 143 DX

Points are in parentheses. Write your answers on Answers sheet; examination (blue) books are for scratch work only. An unsimplified answer like $12\sqrt{3.51} + 6/7$ is usually worth more than 23.3, because it is easier to understand where it came from.

1. *(25 points)* An educational researcher wants to test whether students using a new curriculum for teaching fractions learn the material better than those using an older curriculum. She teaches 100 students, chosen at random, by the new method, and then administers a standardized exam. Long experience has shown that using the old method, students score an average of 50 on this exam, but her students' average score is 54, with an SD of 20.

   (a) What kind of test (z or t or 2-sample z or χ^2) should be used to decide whether the new curriculum is better than the old?

   (b) Find the value of the variable (z or t or χ^2) used in this test.

   (c) Find the P-value, i.e., the probability that, if the new curriculum were no better than the old, her students would average 54 or more just by chance.

   (d) Should we conclude that the new curriculum results in higher scores than the old?

   (e) Suppose another researcher reaches a conclusion that is the opposite of ours. Should we conclude that his data is "tainted" (i.e., obtained by faulty methods or even perhaps falsified)?

2. *(20 points)* The cathat plant has flowers that are either pink, blue or lavender. A genetic model holds that a single gene controls the color of the flowers, with pure forms (genotypes p/p or b/b) having pink or blue flowers respectively; but neither is dominant, so that hybrids (p/b or b/p) have lavender flowers. In an experiment, lavender-flowered cathats are crossed with each other, and, out of 40 offspring selected at random, 15 have pink flowers, 22 have lavender flowers and 3 have blue flowers.

   (a) What kind of test would be used to decide whether these results are significant evidence against the genetic model?

   (b) In computing the value of the variable (z or t or χ^2) used in this test, should we use the numbers of plants with flowers of each color, or the percentages of each color? Or doesn't it matter?

   (c) Compute the value of the variable.

   (d) Is this data significant evidence against the model?

3. *(15 points)* An ecologist projects the level N of a certain nutrient in a stream by using linear regression on the amount F of fertilizer used in a few fields near the stream. Suppose the averages of F and N are 3 (tons) and 3.5 (parts per thousand) respectively, with standard deviations of 2.5 and 2 respectively, and a correlation of 0.3.

   (a) What is his regression equation for projecting N from F?.

   (b) Roughly how far off should the ecologist expect the projections to be as he makes them using his equation in (a)?

(c) Suppose his regression equation was $N = .4F + 5$ (which it isn't). If the nutrient level turns out to be 7 when 8 tons of fertilizer were used, what is the corresponding residue (or residual) relative to his projection?

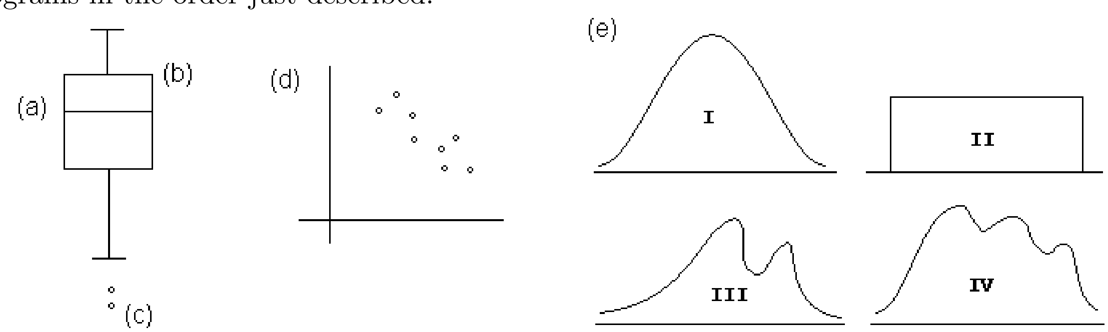
4. *(15 points)* A market survey of 400 randomly chosen households in a large city finds that 280 of them have an internet connection and the average amount of money spent per meal is \$8, with a standard deviation of \$4.

   (a) Find a 95% confidence interval for the percentage of households in the city that have internet connections.

   (b) How many households must be surveyed so that the confidence interval requested in (a) will turn out only a third as large as the one you found in (a)?

   (c) Find a 85% confidence interval for the average amount of money spent per meal in the city.

5. *(20 points)* On the modified boxplot, name the values (a), (b) and (c) (the dots) relative to the distribution. For the scatterplot (d), estimate the correlation. For (e), three of the histograms were generated from a very large survey of individuals' numbers of years of schooling; one is the actual data, another is the averages of samples of 50 taken from the data, a third is averages of samples of 500 taken from the data, and the fourth is not related. Arrange the histograms in the order just described.



6. *(20 points)* A pinochle deck has 48 cards, in the usual four suits but only six ranks (9-10-J-Q-K-A), two of each card. (Thus, for example, there are 8 aces and 12 spades.) In each case below, a card is selected only after the deck is shuffled.

   (a) If two cards are selected <u>without</u> replacement, what is the probability that both are kings?

   (b) If one card is selected, what is the probability that it is either a king or a club, or both?

   (c) If five cards are selected <u>with</u> replacement, what is the probability that at least one is a club?

   (d) If five cards are selected <u>with</u> replacement, what is the probability that exactly three are clubs?

7. *(10 points)* Relative to the article, "Food News Blues; Fat is bad, but good fat is good. What about fish? Wine? Nuts? A new appetite for answers has put science on a collision course with the media", by Barbara Kantrowitz and Claudia Kalb (with Anne Underwood and Pat Wingert): How do the sample sizes in the WHI and the chocolate study bring their results into question (in different ways)?

Formulas and Terms in Math 102 / Core 143

In a histogram,
# of scores (or total probability of outcomes) in an interval
= area under the histogram above that interval

For a list of n scores (x-values):

Average $\overline{x} = \mu = \frac{\sum x}{n}$, Standard deviation $\sigma = SD = \sqrt{\frac{\sum(x-\overline{x})^2}{n}}$

N-th percentile = x-value for which $N\%$ of scores are $\leq x$

Median = 50-th percentile, IQR = 75-th − 25-th percentile

Converting a score x to standard units (z-value or t-value):

$$z \text{ or } t = \frac{x-\overline{x}}{\sigma} \qquad \text{or} \qquad \frac{x-EV}{SE} \text{ (see below)}$$

Correlation: r = average of products $(z_x z_y)$, where z_q means q in standard units

Regression line of y on x (for predicting y from x, or for estimating average y within a vertical strip at x): Denote predicted y-value by $\hat{y}$. Then

$$\hat{y} - \overline{y} = r\left(\frac{\sigma_y}{\sigma_x}\right)(x-\overline{x})$$

Residual corresponding to data point (x, y): $y - \hat{y}$ or $(x, y - \hat{y})$.

RMS error for regression of y on x (= approximate standard deviation of data in any vertical strip, if scatter diagram is homoscedastic):

$$\sigma_y\sqrt{1-r^2}$$

Multiplication rule: Prob(A and B) = Prob(A)·Prob(B given A)

Independent events: Prob(B given A) = Prob(B)

Addition rule: Prob(A or B) = Prob(A) + Prob(B)− Prob(A and B)

Mutually exclusive events: Prob(A and B) = 0

Binomial probabilities: If an event has probability p on each trial, the probability of its occurring exactly k times in n independent trials:

$$C(n,k)\cdot p^k \cdot (1-p)^{n-k}\ , \qquad \text{where} \qquad C(n,k) = \frac{n!}{k!(n-k)!}$$

For a sample of size n from a population with average μ and standard deviation σ:

EV of sum of scores in sample $= n\mu$

SE of sum $= \sigma \cdot \sqrt{n}$

EV of average of sample $= \mu$

SE of average $= \sigma/\sqrt{n}$

For significance tests, approximate (bootstrap) population standard deviation σ with sample standard deviation $s = SD^+ = \sqrt{\dfrac{\sum(x-\overline{x})^2}{n-1}} = (SD \text{ of sample})\sqrt{\frac{n}{n-1}}$. (The null hypothesis will give a value to use for μ.) For large samples ($n \geq 30$), s is close to σ.

For confidence intervals, also approximate population average μ with sample average $\overline{x}$.

Special case: Population is 0's and 1's (or yeses and nos, or ins and outs, or ...), fraction of 1's is p, for a sample of size n:

EV of count $= np$

SE of count $= \sqrt{p(1-p)} \cdot \sqrt{n}$

EV of % (or proportion) $= p$

SE of % (or proportion) $= \sqrt{p(1-p)}/\sqrt{n}$

For CIs, approximate (bootstrap) population proportion p with sample proportion $\hat{p}$.

For use with confidence interval or t-test for significance on small ($n < 30$) samples: degrees of freedom $= n - 1$

$k\%$ confidence interval for the average of a population:

Let z_k denote the z-value for which k percent of the data is between $-z_k$ and z_k. Then the CI is

$$\overline{x} \pm z_k \cdot (SE \text{ for average})$$

(Similar for "proportion" in place of "average".)

For $n < 30$: Let $t_{(100-k)\%/2}$ denote the t-value for which $(100-k)\%/2$ of the probability in the t-distribution is to its right. Then the CI is

$$\overline{x} \pm t_{(100-k)\%/2} \cdot (SE \text{ for average})$$

For CI or significance test for difference of μ's in two populations:

SE for difference of averages of 2 samples $= \sqrt{(SE \text{ of first})^2 + (SE \text{ of second})^2}$

For significance test, EV of difference $= 0$ by H_0. (For more than 2 samples, use one-way ANOVA.)

[Technicality, mostly to be ignored: To get SE for difference of proportions in two populations:

For CI, use $\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$, where $\hat{p}_i$'s are proportions in samples, as above

For significance test: Pooled estimate for common population proportion is $\hat{p} = \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$,

and SE for difference $= \sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}$]

For deciding significance of differences in frequency distributions among categories:

$\chi^2 = \sum[(\text{observed} - \text{expected})^2/\text{expected}]$

degrees of freedom: in "list" distributions, # in list -1;

in "table" distributions, (# of rows -1) $\cdot$ (# of columns -1)

Normal table (Area between $-z$ and z)									
z	Area(%)	z	Area(%)	z	Area(%)	z	Area(%)	z	Area(%)
0.0	0.0	0.9	63.19	1.8	92.81	2.7	99.31	3.6	99.968
0.05	3.99	0.95	65.79	1.85	93.57	2.75	99.4	3.65	99.974
0.1	7.97	1	68.27	1.9	94.26	2.8	99.49	3.7	99.978
0.15	11.92	1.05	70.63	1.95	94.88	2.85	99.56	3.75	99.982
0.2	15.85	1.1	72.87	2	95.45	2.9	99.63	3.8	99.986
0.25	19.74	1.15	74.99	2.05	95.96	2.95	99.68	3.85	99.988
0.3	23.58	1.2	76.99	2.1	96.43	3	99.73	3.9	99.99
0.35	27.37	1.25	78.87	2.15	96.84	3.05	99.771	3.95	99.992
0.4	31.08	1.3	80.64	2.2	97.22	3.1	99.806	4	99.9937
0.45	34.73	1.35	82.3	2.25	97.56	3.15	99.837	4.05	99.9949
0.5	38.29	1.4	83.85	2.3	97.86	3.2	99.863	4.1	99.9959
0.55	41.77	1.45	85.29	2.35	98.12	3.25	99.885	4.15	99.9967
0.6	45.15	1.5	86.64	2.4	98.36	3.3	99.903	4.2	99.9973
0.65	48.43	1.55	87.89	2.45	98.57	3.35	99.919	4.25	99.9979
0.7	51.61	1.6	89.04	2.5	98.76	3.4	99.933	4.3	99.9983
0.75	54.67	1.65	90.11	2.55	98.92	3.45	99.944	4.35	99.9986
0.8	57.63	1.7	91.09	2.6	99.07	3.5	99.953	4.4	99.9989
0.85	60.47	1.75	91.99	2.65	99.2	3.55	99.961	4.45	99.9991

t-table: column head is $P(t \geq \text{entry})$						
df	25%	10%	5%	2.5%	1%	0.5%
1	1.00	3.08	6.31	12.71	31.82	63.66
2	0.82	1.89	2.92	4.30	6.96	9.92
3	0.76	1.64	2.35	3.18	4.54	5.84
4	0.74	1.53	2.13	2.78	3.75	4.60
5	0.73	1.48	2.02	2.57	3.36	4.03
6	0.72	1.44	1.94	2.45	3.14	3.71
7	0.71	1.41	1.89	2.36	3.00	3.50
8	0.71	1.40	1.86	2.31	2.90	3.36
9	0.70	1.38	1.83	2.26	2.82	3.25
10	0.70	1.37	1.81	2.23	2.76	3.17
11	0.70	1.36	1.80	2.20	2.72	3.11
12	0.70	1.36	1.78	2.18	2.68	3.05
13	0.69	1.35	1.77	2.16	2.65	3.01
14	0.69	1.35	1.76	2.14	2.62	2.98
15	0.69	1.34	1.75	2.13	2.60	2.95
16	0.69	1.34	1.75	2.12	2.58	2.92
17	0.69	1.33	1.74	2.11	2.57	2.90
18	0.69	1.33	1.73	2.10	2.55	2.88
19	0.69	1.33	1.73	2.09	2.54	2.86
20	0.69	1.33	1.72	2.09	2.53	2.85
21	0.69	1.32	1.72	2.08	2.52	2.83
22	0.69	1.32	1.72	2.07	2.51	2.82
23	0.69	1.32	1.71	2.07	2.50	2.80
24	0.68	1.32	1.71	2.06	2.49	2.80
25	0.68	1.32	1.71	2.06	2.49	2.79

χ^2-table: column head is $P(\chi^2 \geq \text{entry})$					
df	50%	30%	10%	5%	1%
1	0.46	1.07	2.71	3.84	6.64
2	1.39	2.41	4.60	5.99	9.21
3	2.37	3.67	6.25	7.82	11.34
4	3.36	4.88	7.78	9.49	13.28
5	4.35	6.06	9.24	11.07	15.09
6	5.35	7.23	10.65	12.59	16.81
7	6.35	8.38	12.02	14.07	18.48
8	7.34	9.52	13.36	15.51	20.09
9	8.34	10.66	14.68	16.92	21.67
10	9.34	11.78	15.99	18.31	23.21
11	10.34	12.90	17.28	19.68	24.73
12	11.34	14.01	18.55	21.03	26.22
13	12.34	15.12	19.81	22,36	27.69
14	13.34	16.22	21.06	23.69	29.14
15	14.34	17.32	22.31	25.00	30.58
16	15.34	18.42	23.54	26.30	32.00
17	16.34	19.51	24.77	27.59	33.41
18	17.34	20.60	25.99	28.87	34.81
19	18.34	21.69	27.20	30.14	36.19
20	19.34	22.78	28.41	31.41	37.57

Name: ______________________________

Your answers to Exam II

1. (a)
   (b)
   (c)
   (d)
   (e)
2. (a)
   (b)
   (c)
   (d)
3. (a)
   (b)
   (c)
4. (a)
   (b)
   (c)
5. (a)
   (b)
   (c)
   (d)
   (e)
6. (a)
   (b)
   (c)
   (d)
7. on back.

7.

Solutions to Final Exam

1. (a) A z-test.
   (b) $EV = 50$ by H_0, $SD = 20$ by bootstrapping, so $SE = 20/\sqrt{100} = 2$: $z = (54-50)/2 = 2$.
   (c) $P(\text{avg} \geq 54) = P(z \geq 2) = (100\% - 95\%)/2 = 2.5\%$
   (d) Yes, the probability of a score at least this high by chance is less than 5%, so we reject the null hypothesis that the average if all students used the new curriculum is only 50.
   (e) No, we know from our computer projects that significance tests can be wrong, so either he or we are making a legitimate mistake on the basis of validly obtained data.

2. (a) A χ^2-test.
   (b) The numbers must be used; percentages would give the wrong value.
   (c) Under the null hypothesis that the model is correct, we would expect 10 pink, 20 lavender and 10 blue. So $\chi^2 = (15-10)^2/10 + (22-20)^2/20 + (3-10)^2/10 = 7.6$.
   (d) From the χ^2-table with $df = 3 - 1 = 2$, we see that $\chi^2 = 7.6$ is significant, so we reject the null hypothesis.

3. (a) $N - 3.5 = .3(2/2.5)(F - 3)$
   (b) $2\sqrt{1 - (.3)^2}$
   (c) $7 - (.4(8) + 5) = -1.2$. Technically this is correct, but I won't take off points if the sign is positive.

4. (a) $280/400 \pm 2\sqrt{(280/400)(120/400)}/\sqrt{400} \approx 70\% \pm 4.6\%$
   (b) 9 times as many, or 3600
   (c) Using the normal table: $8 \pm 1.45(4/\sqrt{400}) = 8 \pm .29$ dollars

5. (a) Median.
   (b) Third quartile, or 75-th percentile.
   (c) Outliers.
   (d) Anything from $-.5$ to $-.9$ will be accepted.
   (e) III, IV, I, II (The distinction between III and IV is a bit thin, so I will give credit if they are reversed.)

6. (a) $(8/48)(7/47)$
   (b) $8/48 + 12/48 - 2/48$
   (c) $1 - (3/4)^5$
   (d) $(5!/(3!2!))(1/4)^3(3/4)^2$

7. The WHI study had a very large sample, but it was so big that the researchers could not monitor everyone and had to rely on the subjects to monitor and report accurately what they ate. The chocolate study was very small (only 20 or 30 subjects), so the results were unreliable (especially because it was run by the chocolate industry).