Math 115 Practice Final

Also look at the practice and real midterms. At least one question on the final will be off of your homework.

1. The following is a list of tests and procedures that we've discussed this semester: 1-sample t test, 2-sample t test, ANOVA F test, matched-pairs t test, chi-square test, linear regression, 1-sample z test, 2-sample z test. Define parameters and state hypotheses to answer each of the following questions, and identify the appropriate statistical technique. (Some questions may have more than one correct answer.)

(a) Do Macs and PC's take the same time to restart? You record the time to restart 10 Macs and 10 PC's.

(b) Do computers restart faster or slower when they are connected to the internet? You record restart times with and without a connection for each of 20 Macs.

(c) You and a friend play 25 games of ping-pong to test whether or not you are both equally likely to win in a particular game.

(d) Are the mean calcium levels the same for women in different U.S. states, or do they differ? You record the calcium level for samples of women in each of 3 randomly chosen states.

(e) Does a student's score on a midterm help to predict his or her score on the final? You have midterm and final scores for 50 students.

(f) Are seniors more or less likely than juniors to participate in a varsity sport? You consider a random sample of ASC juniors and seniors.

(g) Does birth order (first born, second born, etc.) affect a student's choice of major? You consider a random sample of college seniors.

(a) 2-sample t test. Let μ_1 **= mean restart time for Macs,** μ_2 **= mean restart time for PC's.** $H_0 : \mu_1 = \mu_2$. $H_a : \mu_1 \neq \mu_2$. **Use test statistic** $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$. **Could use df=10-1=9.**

(b) Matched-pairs t test. Let μ_1 **= mean restart time for Macs connected to the internet,** μ_2 **= mean restart time for Macs not connected to the internet.** $H_0 : \mu_1 = \mu_2$. $H_a : \mu_1 \neq \mu_2$. **For each Mac, record (restart time with connection) - (restart time without connection). Find the average** $\bar{x}$. **Use test statistic** $t = \dfrac{\bar{x} - 0}{\frac{s}{\sqrt{n}}}$, **df = 20 - 1 = 19.**

(c) Let p = probability you win. $H_0 : p = 0.5$. $H_a : p \neq 0.5$. **Assuming independence, the number of times you win in 25 games is a binomial random variable,** $B(25, .5)$. **Since we expect** $25 \cdot .5 = 12.5$ **wins and 12.5 losses, both of which are** ≥ 10, **the binomial is approximately normal, so we perform a 1-sample z test. The test statistic is** $z = \dfrac{\hat{p} - .5}{\sqrt{\frac{(.5)(.5)}{25}}}$.

(d) ANOVA F test. Let μ_1 **= mean calcium level of women in state 1, and similarly for** μ_2, μ_3. $H_0 : \mu_1 = \mu_2 = \mu_3$. H_a**: at least one mean is different from the others. df for the numerator is 3-1=2. For the denominator df is n-3, where n is the total number of women sampled.**

(e) Let X = midterm score, Y = final score. Do linear regression to get model $\hat{y} = a + bx$. $H_0 : b = 0$ **(knowing midterm score doesn't help predict final score).**

$H_a : b \neq 0$ (knowing midterm score does help predict final score); or $H_a : b > 0$ (higher midterm score predicts higher final score).

(f) 2-sample z test. Let p_1 = proportion of seniors who play varsity sports, p_2 = proportion of juniors who play varsity sports. $H_0 : p_1 = p_2$. $H_a : p_1 \neq p_2$. Test statistic $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$. (Could also use chi-square.)

(g) Chi-square test. H_0: proportions in each major are the same for all birth order groups. H_a: they're not all the same. Test statistic $X^2 = \sum \dfrac{(\text{exp. - obs.})^2}{\text{exp.}}$.

2. Pfizer, the company that manufactures the impotence drug Viagra, conducted a clinical trial involving 100 male subjects complaining of impotence. 50 subjects were assigned at random to receive Viagra, with the remaining 50 subjects receiving a placebo. 30 of the 50 treated men reported "success," compared with only 12 of the untreated men.

(a) Construct a 99% confidence interval for the probability of "success" among those men taking the placebo.

(b) Pfizer would like to report that taking Viagra improves the success probability to something greater than 0.5. Is this claim supported by the data at level $\alpha = 0.05$? State null and alternative hypotheses and report a P-value for the test.

(c) The CEO of Pfizer says that "We saw a 60% success rate, so of course we can say its over 0.5!" Explain why this isn't true in a way that a statistically challenged CEO might understand.

(d) Find a 95% confidence interval for the difference in success probabilities for the two treatment groups.

(a) Let p_1 be the success rate for men on the placebo. Use the Wilson estimate: $\tilde{p}_1 = \frac{X_1+2}{n_1+4} = \frac{12+2}{50+4} = .26$. So the C.I. is $\tilde{p}_1 \pm z^* \sqrt{\frac{(.26)(1-.26)}{54}} = .26 \pm (2.576)(.05963) = .26 \pm (.15) = (.11, .41)$.

(b) Let p_2 be the success rate for men taking Viagra. $H_0 : p_2 = .5$. $H_a : p_2 > .5$ (one-sided). The sample proportion $\hat{p}_2 = 30/50 = .6$, so the test statistic (1-sample z) is $z = \dfrac{.6 - .5}{\sqrt{\frac{(.5)(.5)}{50}}} = \dfrac{.1}{.0707} = 1.414$. P-value: $P(z > 1.414) = 1 - P(z \leq 1.414) = 1 - .9207 = .0793$. This is not less than .05, so we cannot conclude with 95% confidence that the success probability for men on Viagra is over 50%.

(c) Well, that could have been just luck. For example, if you flip a coin three times and get all heads, that doesn't necessarily mean that the coin's weighted. If the real success rate is 50%, then just by luck we'd have about an 8% chance of seeing as many successes as we did.

(d) We want to do a 2-sample z test. Now our Wilson estimates are $\tilde{p}_1 = \frac{12+1}{50+2} = .25$ and $\tilde{p}_2 = \frac{30+1}{50+2} = .60$. The standard error is $\sqrt{\dfrac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \dfrac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}} = \sqrt{\dfrac{(.25)(.75)}{52} + \dfrac{(.60)(.40)}{52}} = .09067$. The CI is $(\tilde{p}_2 - \tilde{p}_1) \pm z^*(SE) = .35 \pm (1.96)(.09067) = .35 \pm .1777 = (.17, .53)$.

3. A baker recorded the number of delicious blueberry pies that she made each day over an 11-day period. Here are the data:

$$33, 38, 43, 30, 29, 40, 51, 27, 42, 23, 31$$

(a) Construct a stemplot.
(b) Give the five-number summary.

(a)

stem	**leaf**
2	**379**
3	**0138**
4	**023**
5	**1**

(b) The five-number summary is (min, Q1, median, Q3, max), so here it's (23, 29, 33, 42, 51).

4. In a class survey at Angus Scott College in Scotland, 11 of the 67 females (16.42%) were left-handed, compared to 7 of the 31 males (22.58%). Is this strong evidence that there is an association between gender and handedness? State hypotheses, carry out a test, and report your conclusion.

Let p_1 be the proportion of females who are left-handed, and p_2 the proportion of males who are left-handed. So our sample proportions are $\hat{p}_1 = 11/67 = .1642$, and $\hat{p}_2 = 7/31 = .2258$. $H_0 : p_1 = p_2$. $H_a : p_1 \neq p_2$ (two-sided). We'll do a 2-sample z test (we could also do a chi-square test). Our test statistic is $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$. The overall proportion $\hat{p} = \frac{11+7}{67+31} = .184$, so our z is $\frac{-.0616}{.165} = -.373$. The P-value is $2P(z < -.373) = 2(.3557) = .7114$. The P-value is large, so we can't reject the null; this isn't enough evidence to conclude that gender affects handedness. The difference in the sample proportions could easily be due to chance.

5. You plan to do a temperature study to see if normal human body temperature really is different from 98.6 degrees Fahrenheit. Assume that the standard deviation of body temperatures is approximately $\sigma = 0.8$ degrees Fahrenheit.

(a) How large a sample would you need to get a margin of error of 0.1 for a 95% CI?
(b) With this sample size, what values of the sample average would lead you to reject the null hypothesis of 98.6 at the 0.01 significance level?

(a) For a 95% CI, we would have $\bar{x} \pm (1.96)\dfrac{\sigma}{\sqrt{n}}$. So we want the margin of error $(1.96)\dfrac{\sigma}{\sqrt{n}} = (1.96)\dfrac{.8}{\sqrt{n}}$ to be ≤ 0.1, or $(1.96)\dfrac{.8}{.1} \leq \sqrt{n}$, or $15.68 \leq \sqrt{n}$, or $245.86 \leq n$. Thus we choose $n = 246$ or larger.

(b) Let μ be the mean body temperature. $H_0 : \mu = 98.6$. $H_a : \mu \neq 98.6$ (two-sided). The test statistic would be $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{\bar{x} - 98.6}{.8/\sqrt{246}} = \dfrac{\bar{x} - 98.6}{.051}$. We would reject the null if $2P(Z < z) \leq .01$, or $P(Z < z) \leq .005$. From Table D, we see that we will reject if $z < -.2576$ or $z > .2576$. If $z < -.2576$, then $\frac{\bar{x}-98.6}{.051} < -.2576$, or $\bar{x} - 98.6 < -.13125$, or $\bar{x} < 98.47$. Similarly, $z > .2576$ when $\bar{x} > 98.73$. Thus we would reject the null if the sample mean temperature is less than 98.47 or greater than 98.73.

6. You've decided to spend next year wandering through Nepal in search of the abominable snowman of the Himalayas, also known as the elusive yeti. You plan to sell your used Chrysler

LeBaron so that you can purchase a yak when you get there. LeBarons of the year and mileage of yours are selling for a mean of \$6940 with a standard deviation of \$250. Your research shows that yaks in Nepal are going for about 65,000 Nepalese rupees with a standard deviation of 500 rupees. You have to survive on your profit, so you want to estimate what you can expect in your pocket (in rupees) after the sale and subsequent purchase. One U.S. dollar is worth about 43 Nepalese rupees.

Let the random variable A be the sale price of your LeBaron (in dollars), B the price of a yak (in rupees), and D your profit (in rupees). Then D = 43A - B. Means add, so $\mu_D = 43\mu_A - \mu_B = 43(6940) - 65000 = 233,420$. **It's reasonable to assume that the prices A and B are independent, so the variances add.** $\sigma_D^2 = \sigma_{43A}^2 + \sigma_B^2 = (43)^2\sigma_A^2 + \sigma_B^2 = (43)^2(250)^2 + (500)^2 = 115,812,500$, **so** $\sigma_D = \sqrt{115,812,500} = 10,762$ **rupees.**

7. Everyone knows that crunchy peanut butter is better than creamy, so if you're smart, you should prefer crunchy. To see if there's statistical evidence for this, we surveyed 200 smart people and 200 dumb people, with the following results:

	Prefer crunchy	**Prefer creamy**	**No preference**	**Total**
Smart	100	80	20	200
Dumb	50	120	30	200
Total	**150**	**200**	**50**	**400**

Is there an association between intelligence level and peanut butter preference? State hypotheses and perform an appropriate statistical test. What do you conclude?

H_0: **there is no association.** H_a: **there is an association. We'll do a chi-square test. Overall, .375 is the proportion that prefers crunchy, .5 creamy, and .125 no preference. So, if there's no association, we'd expect to see**

	Prefer crunchy	**Prefer creamy**	**No preference**	**Total**
Smart	**75**	**100**	**25**	**200**
Dumb	**75**	**100**	**25**	**200**
Total	**150**	**200**	**50**	**400**

So the chi-square statistic is $X^2 = (100-75)^2/75 + (80-100)^2/100 + (20-25)^2/25 + (50-75)^2/75 + (120-100)^2/100 + (30-25)^2/25 = 26.67$. **The degrees of freedom** $= (2-1)(3-1) = 2$, **so our P-value (from Table F) is** $< .0005$. **Thus there is very strong evidence to reject the null hypothesis, and we conclude that there is an association between intelligence level and peanut butter preference.**

8. A researcher wishes to try three different techniques to lower the blood pressure of individuals diagnosed with high blood pressure. The subjects are randomly assigned to three groups; the first group takes medication, the second group exercises, and the third group follows a special diet. After four weeks, the reduction in each person's blood pressure is recorded. We want to run an ANOVA to find out whether there is evidence at the level $\alpha = 0.05$ that there is any difference among the means. The data are as follows:

Medication	**Exercise**	**Diet**
10	6	5
12	8	9
9	3	12
15	0	8
13	2	4

(a) State null and alternative hypotheses.

(b) I calculated in my head that for these data, MSG=80.07, and MSE=8.73. Explain, in words, the meanings of MSG and MSE.
(c) Find the F statistic and the corresponding P-value.
(d) What do you conclude?

(a) H_0: All three means are equal. H_a: At least one mean is different from the others.

(b) MSG (mean square groups) is the variation among the means of the three groups. MSE (mean square error) is the variation among the samples in the same group.

(c) F=MSG/MSE, so F=80.07/8.73=9.17. We have df for the numerator = 3-1 = 2, and df for the denominator = 15 - 3 = 12. Consulting Table E, we see that our P-value is between .010 and .001.

(d) Our P-value is less than $\alpha = 0.05$, so we conclude that there is significant evidence that at least one mean is different from the others.

9. The ages and systolic blood pressures of six randomly selected subjects are in the table below.

Subject	**Age x**	**Pressure y**
A	43	128
B	48	120
C	56	135
D	61	143
E	67	141
F	70	152

(a) Compute the equation of the least-squares regression line giving blood pressure (y) as a function of age (x).
(b) What fraction of the variation in pressures is explained by the variation in ages?
(c) Use your regression line to predict the blood pressure of a 50-year-old. How confident are you of your result?
(d) Use your regression line to predict the blood pressure of a 20-year-old. How confident are you of your result?

(a) First, find the correlation r. We know that $r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$. **Here,** $n = 6$, $\bar{x} = 57.5$, $s_x = 10.60$, $\bar{y} = 136.5$, **and** $s_y = 11.40$, **so** $r = .897$. **The slope of the regression line is** $b = r\frac{s_y}{s_x} = (.897)(11.40)/(10.60) = 0.965$. **The intercept is** $a = \bar{y} - b\bar{x} = 81.01$, **so the equation of the regression line is** $\hat{y} = a + bx$, **or** $\hat{y} = 81.01 + .965x$.

(b) $r^2 = .80$, **so 80% of the pressure variation is due to the age variation.**

(c) $\hat{y} = 81.01 + .965(50) = 129$. **This might give an okay approximation for the blood pressure of the average 50-year-old, but we shouldn't trust it for any particular 50-year-old. Also, our sample size is very small, which makes our estimate unreliable.**

(d) $\hat{y} = 81.01 + .965(20) = 100.3$. **We shouldn't trust this at all. In addition to the problems mentioned in part (c), 20 is far outside the range of our data, so our extrapolation is likely to be worthless.**

10. A group of adults who swim regularly for exercise were evaluated for depression. It turned out that these swimmers were less likely to be depressed than the general population. The researchers said the difference was statistically significant.
(a) What does "statistically significant" mean in this context?
(b) Is this an experiment or an observational study? Explain.

(c) News reports claimed that this study proved that swimming can prevent depression. Explain why this conclusion is not justified by the study. Include an example of a possible lurking or confounding variable.

(a) That we would be unlikely to see such a low depression rate among the group of swimmers just by chance if the depression rate among all swimmers is the same as the overall population's.

(b) An observational study. The researchers aren't assigning subjects to different groups; they're just observing the results.

(c) Association isn't causation! Maybe healthy people are more likely to swim, and healthy people are less depressed (so health is a confounding variable). Or maybe people are more likely to swim when they're happy. Or maybe only wealthy people can afford to swim regularly, and wealthy people are less depressed (so wealth is a lurking variable). Or. . . .

11. According to the M&M's website, 16% of plain chocolate M&M's are green, 20% orange, 24% blue, 13% red, and 14% yellow, while the rest are brown. Assume that you have an enormous bag of over one million M&M's.

(a) If you pick an M&M at random, what is the probability that
  (i) it's brown?
  (ii) it's yellow or green?
  (iii) it's not red?

(b) If you pick four M$M's in a row, what is the probability that
  (i) they are all orange?
  (ii) the first is yellow and the second is green?
  (iii) the third one is the first one that's red?
  (iv) none are yellow?
  (v) under half (i.e., zero or one) are orange?

(c) If you pick four thousand M&M's in a row, what is the probability that over half are orange?

(d) Let Y be the number of letters in the color of an M&M picked at random (so if you pick red, $Y = 3$). Find the mean μ_Y and standard deviation σ_Y.

Since our bag is huge, we can assume that our picks are roughly independent. (If we had a small bag, this wouldn't be true: if the first one picked is red, then the second is less likely to be red, for example.)

(a)(i) .13 (ii) .16+.14=.30 (iii)1-P(red)=1-.13=.87

(b)(i) (P(1st is orange))(P(2nd is orange))(P(3rd is orange))(P(4th is orange)) = (.2)(.2)(.2)(.2)=.0016.

(ii) (P(1st yellow))(P(2nd green))=(.14)(.16)=.0224.

(iii) (P(1st isn't red))(P(2nd isn't red))(P(3rd is red)) = (.87)(.87)(.13) = .098.

(iv) (P(1st isn't yellow))(P(2nd isn't yellow))(P(3rd isn't yellow))(P(4th isn't yellow)) $= (1 - .14)^4 =$ **.547.**

(v) The number X **of orange M&M's in four draws has the binomial (4, .20) distribution. P(X<2) = P(X=0) + P(X=1) = (from Table C) .4096 + .4096 = .8192.**

(c) We'll use the normal approximation to the binomial: B(4000, .20)≈ N($4000(.2)$, $\sqrt{4000(.2)(1-.2)}$) **= N(800,25.3). So** $P(X > 2000) \approx P(z > \frac{2000-800}{25.3}) = P(z > 47.4) \approx 0$**: it's practically impossible that over half will be orange.**

(d) The distribution for Y is

Y	**3**	**4**	**5**	**6**
Prob	**.13**	**.24**	**.29**	**.34**

. So $\mu_Y = 3(.13) + 4(.24) + 5(.29) + 6(.34) = 4.84$**, and** $\sigma_Y = \sqrt{(3-4.84)^2(.13) + (4-4.84)^2(.24) + (5-4.84)^2(.29) + (6-4.84)^2(.34)} = 1.04$**.**

12. A spokesperson for the commonwealth of Pennsylvania claims that the average size of state parks in western Pennsylvania is at least 2000 acres. You suspect that the average size is actually smaller. A random sample of five parks is selected, and the number of acres is shown. At $\alpha = 0.01$, is there enough evidence to reject the spokeperson's claim?

959 1187 493 6249 541

Let μ be the mean park size. $H_0 : \mu = 2000$**.** $H_a : \mu < 2000$ **(one-sided). Use a 1-sample t test. The t statistic is** $t = \dfrac{\bar{x} - 2000}{s/\sqrt{n}} = \dfrac{1885.8 - 2000}{2456/\sqrt{5}} = -0.104$**. From Table D, using df = 5 -1 = 4, we get a P-value of greater than .25 (use the upper tail probability for $t > +0.104$) (actually, P=0.461). Thus, there is not much evidence to reject the null, and we can't conclude that the average park size is less than 2000 acres.**