STAT 35000 Sample Final Exam Spring 2016

Your name: _____

- **Turn off your cell phone before the exam begins!**
- Show your work on all questions. Unsupported work may not receive full credit.
- Report decimal answers to three decimal places.
- You are responsible for upholding IUPUI's standard for academic integrity. This includes protecting your work from the eyes of other students.
- You are allowed the following aids: a one-page (front and back) 8.5×11 handwritten formula sheet, a scientific/graphing calculator, and pens/pencils.
- When you are done, turn in both your exam and your formula sheet.

Problem	Points Possible	Points Earned
1	50	
2	35	
3	30	
4	55	
5	25	
6	55	
Total	250	

Problem 1. $(5 + 15 + 10 + 10 + 10 = 50)$ Let X = number of flaws on an electroplated automobile grill. Its distribution is modeled by the following PMF:

x	0	1	2	3
$p(x)$	0.8	0.1	0.05	0.05

1. What is the probability that there are 2 or more flaws on the grill, i.e. $P(X \geq 2)$?

$Pr(X \geq 2) = p(2) + p(3) = 0.05 + 0.05 = 0.1$

2. Let Y be total number of grills with 2 or more flaws in a random sample of 100 grills. What is the **exact** distribution of Y? Please specify the name and parameter value(s) for distribution of Y.

$Y \sim Bin(n = 100, p = 0.1)$

3. Find the probability that no grills from the sample of 100 has 2 or more flaws, using the **exact** distribution.

$Pr(Y = 0) = binompdf(100, 0.1, 0) = 2.656 \times 10^{-5}$

4. What is the **approximate** distribution of Y? Please specify the name and parameter value(s) for distribution of Y.

$Y \dot{\sim} N(np = 10, np(1-p) = 9)$

5. Find the probability that more than 15 grills have 2 or more flaws, using the **approximate** distribution of Y.

$Pr(Y > 15) = normalcdf(15, 9^9, 10, 3)$

$= 0.0478$

Problem 2. (5+5+5+10+10 = 35) Urban storm water can be contaminated by many sources, including discarded batteries. When ruptured, these battries release metals of environmental significance. The following are summary data on zinc mass (g) for two different brands of size D batteries found in urban areas around Cleveland, which are assumed to be normally distributed with common variance. We would like to decide whether on average the zinc mass is different for the 2 brands of batteries.

Brand	Sample Size	Mean	S.D.
A	15	138.52	7.76
B	20	149.07	5.52

1. For the two populations, two brands of batteries, what are the parameters of interest? Denote them by μ_A and μ_B respectively.

around cleveland

μ_A = the average zinc mass for brand A battery found in urban areas

μ_B: ~~~~~ B ~~~~~

2. State the null and alternative hypotheses in terms of μ_A and μ_B defined above.

$H_0: \mu_A = \mu_B$, $H_1: \mu_A \neq \mu_B$

3. Which method would you use?
(a) 1-sample T (b) 1-sample Z (c) 2-sample T (d) 2-sample Z

4. Calculate the 95% **confidence interval** for the mean difference $\mu_d = \mu_A - \mu_B$.

$$(\bar{X}_A - \bar{X}_B) \pm t^*_{\alpha/2}(n_1+n_2-2) \text{ Spooled } \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$= (-15.11, -5.988)$$

Spooled = 6.564

$t(133)$

95%

2.0345

5. Use the calculated confidence interval above to make decision for the hypotheses given in part 2. Is zinc mass different for the 2 brands of batteries on average?

We're 95% sure that $\mu_d = \mu_A - \mu_B$ is between -15.11 and -5.988, which are both negative.

$\Rightarrow$ We're at least 95% sure that the zinc mass is different for the 2 brands of batteries on average.

Problem 3. (5+5+5+10+5 = 30) Researchers are interested in determining if cooking of certain vegetables can cause a loss of vitamin C. Vitamin C amounts are measured before and after the vegetables are cooked, which are assumed to be normally distributed. Data can be found in the table below.

	VC					Mean	S.D.
Before	73	79	86	88	78	80.8	6.14
After	20	27	29	36	17	25.8	7.53
Change = After - Before	-53	-52	-57	-52	-61	-55.0	3.94

1. What is the parameter of interest in this problem?

μ_d = the average change of VC before and after the vegetables are cooked, $\mu_d = \mu_{after} - \mu_{before}$

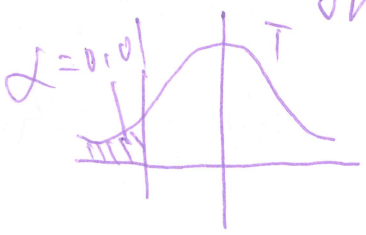
2. State the null and alternative hypotheses in terms of the parameter of your interest.

$H_0: \mu_d = 0$, $H_1: \mu_d < 0$

3. Which method would you use?
(a) 1-sample T (b) 1-sample Z (c) 2-sample T (d) 2-sample Z

4. Calculate the test statistic and find the **rejection region**, using $\alpha = 0.01$.

$$t_{obs} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{-55}{\frac{3.94}{\sqrt{5}}} = -31.14$$



RR: $T < invT(0.01, 4) = -3.74$

5. State your conclusion and interpret the result in the context for the researchers.

Since $t_{obs} = -31.14 < -3.74$, we reject H_0.

At 1% level, the data show that we have strong evidence to conclude that there's significant loss of VC after the vegetables are cooked.

Problem 4. $(5 + 10 + 5 + 10 + 15 + 10 = 55)$ A random sample of 100 data points was collected to test if the population mean μ is different from 24000 at $\alpha = 0.05$ level. That is,

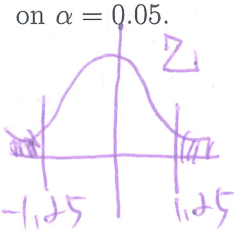
$$H_0 : \mu = 24000 \quad \text{vs.} \quad H_1 : \mu \neq 24000$$

The sample resulted in a mean of 23500. Assume the sample is from a Normal population with known standard deviation $\sigma = 4000$.

1. What does α represent? How do you interpret it?

α is the probability to make type I error.
If $\mu = 24000$, we only have 5% chance to conclude that $\mu \neq 24000$.

2. Calculate the p **value** of the your testing procedure and make your decision based on $\alpha = 0.05$.



$$z_{obs} = \frac{23500 - 24000}{\frac{4000}{10}} = -1.25$$

p-value = 2 normalcdf(1.25, 99, 0, 1) $\approx 0.211 > 0.05$

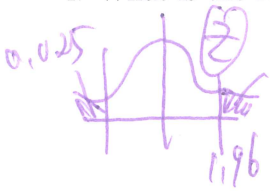
$\Rightarrow$ Accept H_0.

3. Interpret the above p **value** in the context.

If $\mu = 24000$, we have 21.1% chance to observe ~~$\bar{x} = 23500$~~ or more extreme that $|z_{obs}| > 1.25$.

$z_{obs} = -1.25$

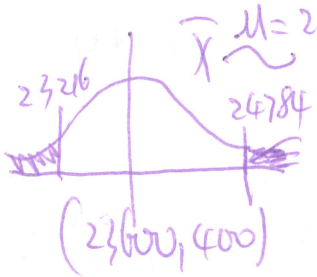
4. What is the rejection region in terms of the sample average $\bar{X}$?



RR(z): $|z| > 1.96$

i.e. $\left|\frac{\bar{X} - 24000}{400}\right| > 1.96$ i.e. $\bar{X} > 24784$ or $\bar{X} < 23216$

5. Suppose that the true population mean is actually 23600. For the test above, what is the power to successfully detect the alternative.



$N(23600, 400^2)$

power = normalcdf(-99, 23216, 23600, 400) + normalcdf(24784, 99, 23600, 400)

6. If we want to increase the above power to exceed 99%, what's the required sample size?

$$\text{power} = Pr\left(\left|\frac{\bar{x} - 24000}{\frac{4000}{\sqrt{n}}}\right| > 1.96 \,\middle|\, \mu = 23600\right)$$

$$= Pr(\tilde{Z} > 1.96 + 0.1\sqrt{n}) + Pr(\tilde{Z} < -1.96 + 0.1\sqrt{n})$$

(too small)

$= 0.16853 + 0.00154 = 0.17$

$\Rightarrow Pr(\tilde{Z} < -1.96 + 0.1\sqrt{n}) = 0.99$

$-1.96 + 0.1\sqrt{n} = \text{invNorm}(0.99, 0, 1) = 2.33$

$\Rightarrow n = 1840.41 \Rightarrow n = 1841$

Problem 5. $(5 + 5 + 15 = 25)$Imagine that you go to Las Vegas, to pay homage to the TV show CSI. Late one night in a bar you meet a guy who claims to know that in the casino at the Tropicana there are two sorts of slot machines: one that pays out 10% of the time, and one that pays out 20% of the time [note these numbers may not be very realistic]. The two types of machines are coloured red and blue. The only problem is, the guy is so drunk he can't quite remember which colour corresponds to which kind of machine. Unfortunately, that night the guy becomes the vic in the next CSI episode, so you are unable to ask him again when he's sober. Next day you go to the Tropicana to find out more. You find a red and a blue machine side by side.

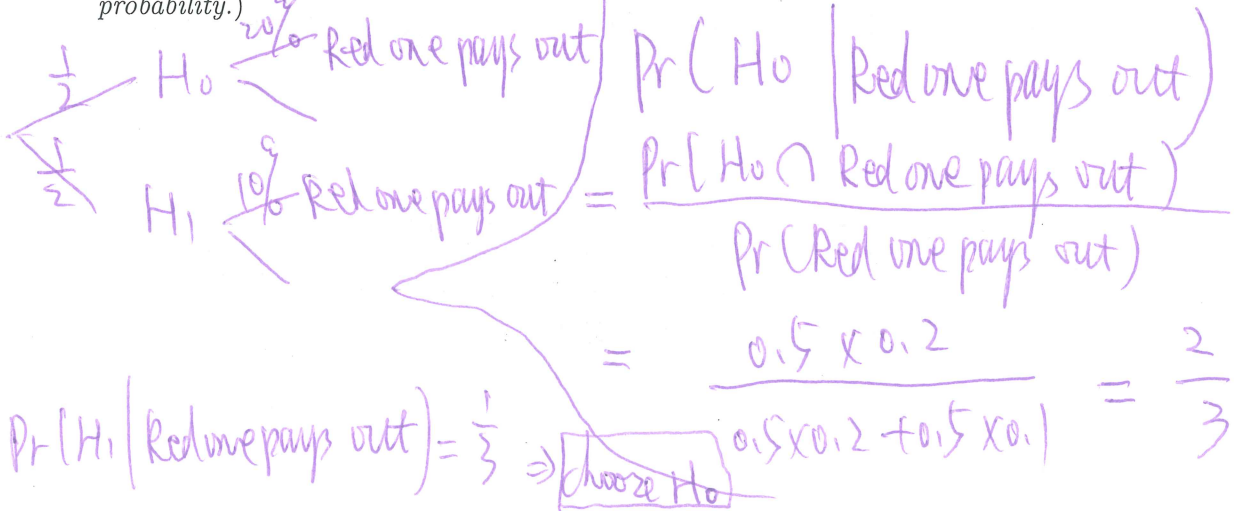
1. If H_0: red one pays out 20% of the time, and H_1: blue one pays out 20% of the time, what's your prior belief of these two hypotheses?

$$Pr(H_0) = Pr(H_1) = \frac{1}{2}$$

2. You toss a coin to decide which machine to try first; based on this you then put the coin into the red machine. If H_0 is true, what is the likelihood that the red one pays out?

$$Pr(\text{Red one pays out} \mid H_0) = 0.2$$

3. It actually indeed pays out. How should you make your decision about whether this red one is the one pays out 20% of the time? (*Hint: use the posterior probability of the two hypotheses to make your decision: choose the one with higher posterior probability.*)



$$Pr(H_0 \mid \text{Red one pays out}) = \frac{Pr(H_0 \cap \text{Red one pays out})}{Pr(\text{Red one pays out})} = \frac{0.5 \times 0.2}{0.5 \times 0.2 + 0.5 \times 0.1} = \frac{2}{3}$$

$Pr(H_1 \mid \text{Red one pays out}) = \frac{1}{3} \Rightarrow$ choose H_0

Problem 6. $(20 + 15 + 20 = 55)$To investigate the effect of two recent national Danish aquatic environment action plans the concentration of nitrogen (measured in g/m^3) have been measured in a particular river just before the national action plans were enforced (1998 and 2003) and in 2011. Each measurement is repeated 6 times during a short stretch of river. The result is shown in the following table: Further, the total variation

Year		
1998	2003	2011
5.01	5.59	3.02
6.23	5.13	4.76
5.98	5.33	3.46
5.31	4.65	4.12
5.13	5.52	4.51
5.65	4.92	4.42

$k = 3$

$n = 6$

in the data is SST = 11.4944. You got the following output from R corresponding to a one way analysis of variance: (where most of the information, however, is replaced by the letters A-E as well as U and V)

```
Analysis of Variance Table
Response: Nitrogen
          Df SumSq MeanSq Fvalue Pr(>F)
Year      A   B      C      U      V
Residuals D   4.1060 E
```

1. What numbers did the letters A-D substitute?

$A = 2$, $B = 11.4944 - 4.1060 = 7.3884$

$C = \frac{B}{A} = 3.6942$, $D = k(n-1) = 3 \times 5 = 15$

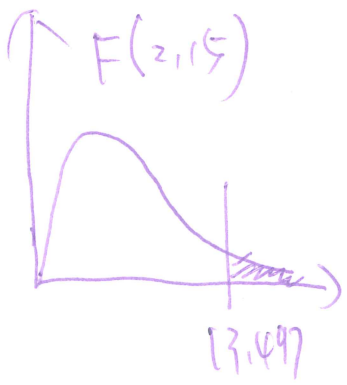
$E = \frac{4.1060}{15} = 0.2737$

2. What numbers did the letters U (F test statistic) and V (p-value for the test) substitute?

$U = \frac{C}{E} = \frac{3.6942}{0.2737} = 13.497$

$V = Fcdf(13.497, 9^9, 2, 15)$

$= 4.43 \times 10^{-4} = 0.000443$



3. Can you demonstrate statistically significant (at significance level $\alpha = 0.05$) differences in the mean Nitrogen values from year to year? Please write down the hypotheses with parameters of clear explanation. And state your conclusion with reasonable arguments.

$H_0: \mu_{1998} = \mu_{2003} = \mu_{2011}$, H_1: not all the same

μ_{1998} = the mean Nitrogen values in year 1998

μ_{2003}: ——— 2003

μ_{2011}: ——— 2011.

Since p-value $= 0.000443 < \alpha = 0.05$, the data show strong evidence to conclude that there is significant difference in the mean Nitrogen values from year to year.