K300 (4392) Statistical Techniques (Fall 2007)

Final Exam, Friday, December 14 (10:30-12:30)

Instructor: Hun Myoung Park
kucc625@indiana.edu, (317) 274-0573

Read instruction and questions carefully and do not skip any question.

- The p-value approach is recommended.
- Use the .05 significance level unless otherwise specified.
- **Report test statistics (t, F, or chi-squared) and their p-values** as well in step 4.
- You should write answers clearly; **DO NOT scribble.**

1. (5 points) Determine the level of measurement of the following variables. Choose **only one** that is most likely.

1) The number of customers who arrived per hour (ratio/interval)
2) Personal income of an Indiana resident (ratio/interval)
3) Insurance company that an adult prefers (nominal)
4) Celsius temperature measured at the IUPUI Library (interval)
5) Junior, senior, sophomore, and freshman (oridnal)

2. (3 points) What is a *hypothesis*? Do you think that $\bar{x} = 0$ is a relevant hypothesis? Tell me why or why not. A conjecture about an aspect of a population. Not relevant because it is a sample statistic, which is already known.

3. (5 points) Suppose the probability that an IUPUI student owns an iPod, *P(iPod)*, is .6, the probability that an IUPUI student owns a laptop, *P(laptop)* is .7, and the probability that the probability an IUPUI student owns both iPod and laptop is .42. Are *owning an iPod* and *owning a laptop* independent? Show your reasoning clearly.
Since P(iPod|Laptop)=.42/.7=.6=P(iPod), they are independent.

4. (3 points) What is the *Central Limit Theorem*?
As N goes infinity (large), the sample mean follow a normal probability distribution.

5. (4 points) What is a *p-value*? How would you explain the p-value to your colleague, who do not know much about statistics? A p-value is the probability that you will get unlikely observation or amount of risk you have to take if you reject the null hypothesis. If you have a small p-value, it is not risky to reject the null hypothesis since there is only tiny change of being wrong when reject the null hypothesis.

6. (5 points) Researchers measured a dependent variable of 12 subjects before (`pre`) and after (`post`) a treatment. Your boss wants to know if the treatment was effective. SPSS output is provided as follows. Test the null hypothesis by following all five steps. And then write down a **complete sentence** in order to report the result to your boss, who knows what the p-value is. Do not forget to add *(p<…)* at the end of the sentence.

(1) $H_0 : \mu_{\bar{d}} = 0$, $H_a : \mu_{\bar{d}} \neq 0$,(2) alpha=.05, (3) t=-1.09, p<.2992, (4) do not reject the null hypothesis since p<.05. (5) $\mu_{\bar{d}} = 0$

The treatment does not make any mean difference (p<.2992).

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Upper	Lower			
Pair 1	pre - post	-1.833	5.8284	1.6825	-5.563	1.8698	-1.09	11	.2992

7. (10 points) The following independent sample t-test uses fake data. Your boss wants to know if there is any mean difference in Y between male (=1) and female (=0). Male has a smaller mean than female (see the sign of mean difference).

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
Y	Equal variances assumed	32.859	.000	-1.834	23	.080	-.859	.468
	Equal variances not assumed			-2.235	15.258	.041	-.859	.384

- 7.1 Test if two variables have the equal variance. Do not skip any of five steps. Do they have the equal variance? (1) $H_0 : \sigma_1^2 = \sigma_2^2, H_a : \sigma_1^2 \neq \sigma_2^2$, (2) alpha=.05, (3) F=32.859 P<.000, (4) reject the null hypothesis since p<.05, (5) $\sigma_1^2 \neq \sigma_2^2$
- 7.2 Which t statistic would you report to your boss? Show him how it was computed. Do not copy a complicate formula. Just use the information provided. -2.235=-.859/.384
- 7.3 Now, test if two variables have the same mean. Follow all five steps. (1) $H_0 : \mu_1 = \mu_2$, $H_a : \mu_1 \neq \mu_2$, (2) alpha=.05, (3) t=-2.235, p<.041, (4) reject the null hypothesis since p<.05, (5) $\mu_1 \neq \mu_2$
- 7.4 Write down a **complete sentence** to summarize the conclusion. Do not forget to add **(**p<…) at the end of the sentence. There is significant mean difference in Y between male and female (p<.041).

8. (10 points) Your boss wants to know if there is a significant mean difference of Y among groups. Fill the all five blank cells from (1) through (5). How many groups are compared in this ANOVA? Test the null hypothesis at the **.01 significance level**. Follow all five steps. And then write a complete sentence in order to report the result to your boss.

ANOVA

Y

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	5.5988	7	(3)	(5)	.0159
Within Groups	(1)	(2)	(4)		
Total	11.6388	31			

(1) 6.0400=11.6388-5.5988, (2) 24=32-8, (3) .7998=5.5988/7, (4) .2517=6.0400/24, (5) 3.1781=.7998/.2517. 8 groups are compared.
(1) H0: all groups have the same mean, Ha: at least one group has a different mean, (2) alpha=.01, (3) F=3.1781, p<.0159, (4) do not reject the null hypothesis since p>.01, (5) all groups have the same mean.
There is no significant mean difference among 8 groups (p<.0159).

9. (5 points) Your boss brought two categorical variables *R* and *C*, and ask you if two variables are related somehow. See the contingency table below. Test the null hypothesis by following all five steps. And then write down a complete sentence as a conclusion.

drug * effect Crosstabulation

			C		
			1	2	Total
R	1	Count	5	8	13
		Expected Count	8.1	4.9	13.0
	2	Count	20	7	27
		Expected Count	16.9	10.1	27.0
Total		Count	25	15	40
		Expected Count	25.0	15.0	40.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4.7483	1	.029		

(1) H_0: R and C are independent, H_a: R and C are not independent; (2) alpha=.05, Since the expected frequency of a cell is less than 5, stop here. We cannot not draw conclusive inference from data provided; chi-square statistic is not reliable.

Bonus. (2 points) Your boss wants to know the strength of association of two variables *R* and *C*. Interpret lambda (C|R) of .2 to your boss; ignore the result of Q9 here.

Knowing information of R will reduce the error by 20 percent when predicting C.

10. (5 points) Your boss also wants to know if two continuous variables *X* and *Y* are correlated. Report *r* and describe their relationship (e.g., positive and negative) by

drawing a simple plot (ignore scales but put labels on X- and Y-axis). Test the null hypothesis and check if *r* is reliable (not zero).

Correlations

		X	Y
X	Pearson Correlation	1	-.8684(*)
	Sig. (2-tailed)		.0001
	N	28	28
Y	Pearson Correlation	-.8684(*)	1
	Sig. (2-tailed)	.0001	
	N	28	28

* Correlation is significant at the 0.05 level (2-tailed).

r is -.8684. It is a strong negative relationship.
(1) $H_0 : \rho = 0$ $H_a : \rho \neq 0$, (2) alpha=.05, (3) p<.0001, (4) reject the null hypothesis since p<.05, (5) $\rho \neq 0$. *r* is reliable.

11. (5 points) What is the least squares method for the ordinary least squares (OLS)?

The least squares method estimate parameters that minimize the sum of squared due to error.

12. (20 points) Your boss has a regression model of $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 D + \varepsilon$, where D is a dummy variable. Let us call the baseline (D=0) "group 0."

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-5.664	.383		-14.79	.000
	X1	(1)	.028		16.95	.000
	X2	.050	.628		.08	.937
	D	-.277	.063		(2)	.000

a Dependent Variable: Y

- 12.1 Compute the parameter estimator (coefficient) of X_1 labeled as "(1)". Write down the regression equation. ε is not needed here (an empirical world). Y=-5.664+.4746X_1+.050X_1-.277D. .4746=16.95*.028
- 12.2 Test the hypothesis that the parameter of X_1 is zero. Again, follow all five steps. Is the parameter significant (or having discernable influence on Y)?
  (1) $H_0 : \beta_1 = 0$ $H_a : \beta_1 \neq 0$, (2) alpha=.05, (3) t=16.95, p<.000, (4) reject the null hypothesis since p<.05, (5) $\beta_1 \neq 0$. Significant.

- 12.3 Interpret the coefficient of X_1. Do not forget to add the *ceteris paribus* phrase. For a unit increase in X1, Y will increase by .4746, holding all other variables constant.
- 12.4 Test the hypothesis that the parameter of the dummy variable D is zero. Is the parameter significant (or having discernable influence on Y)? (1) $H_0 : \beta_3 = 0$, $H_a : \beta_3 \neq 0$, (2) alpha=.05, (3) t=-4.3968=-.277/.063, p<.000, (4) reject the null hypothesis since p<.05, (5) $\beta_3 \neq 0$. Significant.
- 12.5 What does the coefficient of dummy variable D mean? Interpret it substantively. Group 1 has on average .277 smaller value of Y than Group 0 (baseline), holding all other variables constant.

13. (10 points) The following is the ANOVA table of the regression model in Q12 above). Fill the three blank cells. How many observations were used in this regression model? Test the null hypothesis that all parameters are zero. Again, you must follow all five steps.

ANOVA(b)

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	(1)	3	(2)	123.18	.000(a)
	Residual	4.784	86	.056		
	Total	25.340	(3)			

a Predictors: (Constant), D, X1, X2
b Dependent Variable: Y

(1) 20.556=25.340-4.784, (2) 6.852=20.556/3, (3) 89=90-1, K=4, N=90.
(1) H_0: all parameters are zero, H_a: at least one parameter is not zero, (2) alpha=.05, (3) F=123.18, p<.000, (4) reject the null hypothesis since p<.05, (5) at least one parameter is not zero.

14. (10 points), From the following SPSS output, report R^2 and show how R^2 is computed. See Q13 for necessary information. Explain the following R^2 to your boss, who do not know what that means. Finally, evaluate the regression model on the basis of Q13 and Q14. Is it a lemon or peach?

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.901(a)	.811	.805	.236

a Predictors: (Constant), D, X1, X2

R^2=.811=20.566/25.340. This regression model can explain 81 percent of variation of Y. Because of large F statistic (small p-value) and R^2, this model appears to be a peach.

Bonus. (3 points) What is the *Type I Error*?
Rejecting the true null hypothesis.

Bonus. (3 points) What is a *standard error*? How does it differ from the *sample standard deviation* of a variable *x*? Standard error is the standard deviation of a sample mean not of an individual variable *x*.

Have a great Christmas holiday and happy New Year!