Final Examination Review
CSSS/STAT/SOC 321: Case-Based Social Statistics I

Professor: Chris Adolph, Political Science and CSSS

Teaching Assistant: Aaron Erlich, Political Science

Fall Quarter 2012

December 7, 2012

1 Exam rules

- You may use a calculator, graphing calculator, or phone based calculator, but you may *not* store information in memory or use the internet or any other form of electronic communication. I strongly recommend you bring some sort of calculator to save time on arithmetic.
- You will be given a list of helpful formulas, reproduced below.
- If necessary, grades may be curved upwards, but not downwards.

2 What to Expect

In addition to a choice of short answers testing your knowledge of the concepts on the next page, you will need to solve several problems on the final. The range of possible problems includes:

1. How to *calculate* a z-score for a Normally distributed variable and report a percentile level from a table of critical values of z.
2. How to *calculate* a significance test and confidence interval for a sample mean.
3. How to *interpret* a significance test and confidence interval for a difference of two sample means.
4. How to read a cross-tabulation (i.e., by comparing column percentages) and how to *intepret* a χ^2 test of independence.
5. How to explain and *interpret* all the elements of a regression table, including coefficients, standard errors, t-statistics, p-values, confidence intervals, RMSE, and R^2.
6. How to explain graphical displays of regression results.

3 Concepts

In addition to the basic concepts of research design from the midterm, the final exam will assume your familiarity with the following concepts. You should be able to recognize and use these terms:

random variable
probability distribution
discrete distribution
continuous distribution
pdf
cdf
parameter
Bernoulli distribution
binomial distribution
uniform distribution
Normal distribution
χ^2 distribution
t distribution
central limit theorem
z-score
critical value
probability interval
standard error
standard error of the mean
hypothesis testing
null hypothesis
alternative hypothesis
degrees of freedom
proportional reduction in error
substantive significance
statistical significance

t-statistic
p-value
confidence interval
statistical independence
χ^2 test
fitted value
correlation coefficient
regression coefficient
population model
sample model
best fit line
least squares
residual
error term
standard error of β
goodness of fit
coefficient of determination
mean squared error
linear regression
multiple regression
specification
omitted variable bias
dummy variable
reference category
interaction term
log-transformation

4 Formulas

You should know when to apply the formulas below in order to solve problems on the final exam. You do *not* need to memorize these formulas; all required formulas in this section of the review sheet will be provided during the final.

In the equations below, x represents a random variable, n represents the number of observations of x, μ indicates the mean of x, and σ represents the standard deviation of x:

Concept	Formula	Definition
z-score	$z = \frac{x-\mu}{\sigma}$	Re-scaling of Normal variable to N$(0, 1)$
critical value of Normal variable	$x^{\star} = z^{\star}\sigma + \mu$	Value on original scale of x for a standardized $z^{\star}$ corresponding to some percentile of interest
Standard error of a mean	$\text{se}(\bar{x}) = \sigma/\sqrt{n}$	How much we expect the mean of a sample to differ, on average, from the population mean
t-statistic	$t = \frac{\text{Estimate}-\text{Null}}{\text{se}(\text{Estimate})}$	Measure of how unusual an estimate is given the null hypothesis
t-statistic of a sample mean	$t = \frac{\bar{x}-\text{Null}}{\sigma/\sqrt{n}}$	where Null is the value of the population mean we are trying to reject
t-statistic of a regression coefficient	$t = \frac{\beta-\text{Null}}{\text{se}(\beta)}$	where Null is the value of the population regression coefficient we are trying to reject; often this is 0
95% confidence interval	$\text{est} \pm \left(\text{se}(\text{est}) \times t^{\star}_{n-k}\right)$	where k is the number of estimated parameters; for a sample mean CI, $k = 1$, for a linear regression $k = 1 +$ number of covariates

5 Computing z-scores for a Normal variable

If we assume a random variable x follows a Normal distribution with known mean μ and variance σ^2, we can standardize that variable to have mean zero and unit variance, so that we have $z \sim \mathrm{N}(0, 1)$, using a z-score:

$$z = \frac{x - \mu}{\sigma}$$

To see how unusual a particular value of x, we can look up the quantile of z in a table of standard Normal probabilities. A table like the one below, including any values you need, will be provided on the exam:

z	probability
-3.0	0.0013
-2.5	0.0062
-2.0	0.0228
-1.5	0.0668
-1.0	0.1587
-0.5	0.3085
0.0	0.5000
0.5	0.6915
1.0	0.8413
1.5	0.9332
2.0	0.9772
2.5	0.9938
3.0	0.9987

Table 1. Standard Normal probabilities

6 Significance tests and confidence intervals for a sample mean

You should know how to use the t-statistic to perform a significance test. This involves two steps:

1. Calculate the appropriate t-statistic. For a sample mean, this is:

$$t = \frac{\bar{x} - \text{Null}}{\sigma/\sqrt{n}}$$

2. Look up that t-statistic in a table of p-values, given the appropriate level and degrees of freedom. For a sample mean, the degrees of freedom are $n - 1$.

A table like the one below, including any values you need, will be provided on the exam:

df / p-value	0.1	0.05	0.01	0.001
1	6.31	12.71	63.66	636.62
2	2.92	4.3	9.92	31.6
5	2.02	2.57	4.03	6.87
10	1.81	2.23	3.17	4.59
20	1.72	2.09	2.85	3.85
50	1.68	2.01	2.68	3.50
100	1.66	1.98	2.63	3.39
200	1.65	1.97	2.60	3.34
500	1.65	1.96	2.59	3.31
1000	1.65	1.96	2.58	3.30
2000	1.65	1.96	2.58	3.30
5000	1.65	1.96	2.58	3.29

Table 2. Critical values of t distribution, two-tailed

Finally, you should know how to calculate the confidence interval for either an estimated sample mean or an estimated regression coefficient:

$$95\%\ \text{Confidence Interval} = \text{estimate} \pm \left(\text{se(estimate)} \times \text{critical } t \text{ at } 0.05 \text{ level with } n - k \text{ df}\right)$$

7 Testing whether two categorical variables are independent

In a cross-tabulation of two categorical variables, we often want to know if the variable recorded in the columns is associated with the variable recorded in the rows. To check for independence of the row and column variable, use a χ^2 (chi-squared) test. When n_{ij} is the total observations falling in the cell at row i, column j, and $\hat{n}_{ij}$ is the predicted number under independence, we have the test statistic X^2:

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}},$$

which is distributed χ^2 with $(I-1)(J-1)$ degrees of freedom.

To see if we can *reject* the null hypothesis of independence at a given level, we look up the p-value of the observed X^2 in the χ^2 table. A table like the one below, including any values you need, will be provided on the exam:

df / *p*-value	0.1	0.05	0.01	0.001
1	2.71	3.84	6.63	10.83
2	4.61	5.99	9.21	13.82
5	9.24	11.07	15.09	20.52
10	15.99	18.31	23.21	29.59
20	28.41	31.41	37.57	45.31
50	63.17	67.50	76.15	86.66
100	118.5	124.34	135.81	149.45
200	226.02	233.99	249.45	267.54
500	540.93	553.13	576.49	603.45
1000	1057.72	1074.68	1106.97	1143.92
2000	2081.47	2105.15	2150.07	2201.16
5000	5128.58	5165.61	5235.57	5314.73

Table 3. Critical values of χ^2 distribution, one-tailed

8 The linear regression model

The linear regression model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

We interpret this model as follows:

Concept	Formula	Definition
Slope	$\beta_1, \beta_2, \ldots \beta_k$	The expected change in y given a 1-unit change in x_k
Intercept	β_0	The expected level of y when all x's are 0
Fitted value	$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_k x_{ki}$	The expected level of y_i given x_i; what the model predicts y_i should be
Error	ε_i	The discrepancy between the model's prediction of $\hat{y}_i$ and the actual y_i

Note that we can transform either x_i or y_i before including them in the model. Thus both of the following are valid regression models, but require special (e.g., graphical) tools to interpret:

Regression with a logged dependent variable: $\log(y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$

Regression with a logged independent variable $y_i = \beta_0 + \beta_1 \log(x_i) + \varepsilon_i$