POLS 205: Concepts for Final Exam

Christopher Adolph

June 2, 2010

1 Research Design Concepts

In the first half of the course, we worked methodically through the elements of a well-designed research project: development of a *research question*, choice of a *unit of analysis* and *data collection strategy*, definition and operationalization of *variables*, and elaboration of testable *hypotheses*.

These concepts allowed us to discuss qualitative strategies for analyzing data to accept or reject our hypotheses. The same concepts provide the foundation for the second half of the course, on quantitative data analysis. All the concepts from the first half of the course remain relevant to the final, and will help you understand and answer questions about quantitative research.

2 Statistical 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 these terms and use them to interpret brief selections from research papers, including tables and figures.

- continuous variable
- discrete variable
- nominal variable
- ordered variable
- binary variable
- additive measure
- ratio measure
- expected value
- mean
- median
- mode
- range
- quantile
- standard deviation
- variance
- histogram
- density plot
- box-and-whisker plot
- scatterplot
- Normal distribution
- t distribution
- random sample
- stratified sample
- convenience sample
- statistical significance
- substantive significance
- t-statistic
- p-value
- confidence interval
- cross-tabulation
- statistical independence
- χ^2 test
- fitted value
- covariance
- correlation coefficient
- regression coefficient
- population model
- sample model
- best fit line
- least squares
- residual
- error term
- standard error
- goodness of fit
- coefficient of determination
- mean squared error
- linear regression
- multiple regression
- specification
- omitted variable bias
- dummy variable
- reference category
- interaction term
- logarithm
- log-transformation
- explanatory model
- predictive model
- out-of-sample test
- cross-validation

3 Statistics and Math

Three problems you need to be able to solve on the final:

1. How to calculate a significance test and confidence interval for a sample mean
2. How to read a cross-tabulation (i.e., by comparing column percentages), and how to intepret a χ^2 test of independence
3. How to interpret all the elements of a regression table

To solve these problems, 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. No other formulas will be required on the final exam.

3.1 Measures of central tendency

You should know the definitions of the mode and median, as well as the formula for the mean of a variable. In the equations below, x represents a random variable, and n represents the number of observations of x:

$$\text{mean}(x) = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

3.2 Variance, standard deviation, and standard error

You should know the definition of the variance, the standard deviation, and the standard error, as well as their application in specific cases:

Concept	Formula	Definition
Variance	$\sigma^2 = \mathrm{E}\left((x - \mathrm{E}(x))^2\right)$	The square of the standard deviation
Standard devation	$\sigma = \sqrt{\mathrm{E}\left((x - \mathrm{E}(x))^2\right)}$	How much we expect a random draw from x to differ, on average, from its expected value
Standard deviation of a sample	$\hat{\sigma}_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{n}{n-1}\bar{x}^2}$	How much we expect a random draw from a sampled variable to differ, on average, from its sample mean
Standard error of a mean	$\hat{\sigma}_{\bar{x}} = \hat{\sigma}_x/\sqrt{n}$	How much we expect the mean of a sample to differ, on average, from the population mean
Standard error of a regression coefficient	$\hat{\sigma}_{\hat{\beta}}$ or $\mathrm{se}(\hat{\beta})$	How much we expect the regression coefficient estimated from the sample to differ from the true, or population regression coefficient

3.3 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:

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

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:

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 1: Critical values of χ^2 distribution, one-tailed

3.4 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 $y_i = \beta_0 + \beta_1 \log(x_i) + \varepsilon_i$

3.5 Significance tests and confidence intervals

You should know the definition of the t-statistic, as well as how to calculate the t-statistic for the estimate of a sample mean, and the how to calculate the t-statistic for the estimate of a regression coefficient. In the equations below, μ_0 represents the null hypothesis:

Concept	Definition	Formula
t-statistic	The ratio of an estimate to its standard error	$t = \frac{\text{estimate}}{\text{standard error}}$
t-statistic of a sample mean	The ratio of the sample mean to the standard error of the sample mean	$t = \frac{\bar{x}-\mu_0}{\hat{\sigma}_x/\sqrt{n}}$
t-statistic of a regression coefficient	The ratio of a regression coefficient to its standard error	$t = \hat{\beta}_1/\text{se}(\hat{\beta}_1)$

You should know how to use the t-statistic to perform a significance test. This involves two steps: 1.) calculating the appropriate t-statistic, and 2.) looking up that t-statistic in the following table of p-values, given the appropriate level and degrees of freedom:

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:

$$\text{95\% Confidence Interval} = \text{estimate} \pm \text{se(estimate)} \times \text{critical } t \text{ at 0.05 level with } n-1 \text{ df}$$