

Lecture 2: January 29 and Feb 3

Information

Textbook issues resolved?

Class Survey. Extra sessions?

Lecture

Measurement issues

How are we going to measure (quantify) what we are interested in? First we need a good definition.

Some possible problems: student learning, bad behavior of children, what it means to be poor. What does it mean to be food insecure? How do we define an economic slowdown?

Measurement

Define the concept and looks for ways to quantify it. Measure the degree of a characteristic, intensity, frequency, etc,

Types of Scales: Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, 103(2684), 677-680.

Nominal Scale- The numbers serve only as labels, like players jersey's
Ex. Code men=1, female=2

Ordinal scale- ranking (ordering)
Ex. Excellent, good, fair

Interval Scale- Orders the objects according to magnitude, and distinguishes this order into equal intervals
Ex. Temperature scale, because 40 deg is not 2times as warm as 20deg.

Ratio scale- A scale having absolute rather than relative quantities, where zero means an absence of that attribute
Ex. Weight

Examples: GPA
SEI

Statistical analysis needs to consider the level of measurement of the variable.

Velleman, P. F., & Wilkinson, L. (1993). Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading. *The American Statistician*, 47(1), 65-72.

Representation problem
Uniqueness problem
Meaningfulness problem

Types of data

Cross section –observation at a given point in time on multiple units (NHLSL)
Time series – observations at different points in time over the same unit (GDP data)
Panel (GDP for multiple countries, PSID)

Longitudinal (NLSY)

Repeated cross section (same data, but different cross sectional units at each time period. Ex CPS, GSS)

I'll start with some basic terminology that's relevant to longitudinal data. First, the term 'longitudinal data' is somewhat vague. Generally, the term implies that one has panel data, that is, data collected on multiple units across multiple points in time (like the PSID). However, it is often also used to refer to repeated cross-sectional data, that is, data collected on multiple *different* units at multiple points in time (like the GSS).

A basic model:

Exam Score = f(X??) hours spent studying, IQ, ACT, Gender?, Age? Major?

GDP Growth = f (x??) Investment growth, Democracy, Political Stability, Corruption,

Endogenous (explained) = Function of Exogenous (explanatory)

Left Hand Side (LHS) = Right hand side (RHS)

Some other ways of classifying data: For the most part RHS data can be of any type, but different left hand side variables often dictate different models. Some common types of data and their models.

Interval and ratio data as LHS variable are often candidates for Linear Regression

Dichotomous Choice, Binomial, discrete. Taking on just two values, 1 or 0, yes or no are often estimated using logit or probit models. These are often nominal variables, which take only two values.

Multinomial Logit models are often used for dependent variables that are nominal but include more than two choices. For example if you were trying to estimate (explain) a person's choice of transportation, where the choices could be bike, car, bus, subway.

Ordinal dependent variables are often estimated with Ordered Probit or Ordered Logit

Count Data. The number of events, for example the number of car accident someone has had or the number of drinks. Ie the data are not continuous. Typically there are respondents with 0 events. One method of estimating these models is the Poisson Regression, since a Poisson distribution better represents the data. But there are some crucial aspects to this such as the degree of dispersion and the number of zeros. There are other methods of Poisson isn't the best fitting like negative binomial or a Hurdle model.

Interval Regression. A better way of handling variables that are in fact intervals. For example: Lets say you ask

"What is your annual gross income?" and offer the following categories:

1. \$0-10,000

2. \$10,001 - \$50,000

3. \$50,001-100,000

4. \$100,001 +

Duration Data the time it takes for something to happen, such as an event. Related to the probit and logit models in those models you predict if the event happens or does not.

Data Censoring and Truncation.

Censoring – recording a value that is not the true value, through setting an upper or lower limit. The income variable above is right censored if we recorded 100,000 for everyone who said 100,000 +

Truncating is when data is not recorded at all, say due to inability to measure it at certain levels.

For your projects you want to find an interval or ration scaled variable to explain, in other words to be the LHS variable.

Data Sources:

This link provides a good list <http://www.oswego.edu/~economic/data.htm>

Review Basic Stats

I will be using data from DeMaris (2004) on faculty salaries.

Descriptive vs inferential stats

Descriptive Statistics

Statistics is broken into two branches

Descriptive Statistics-describe the data collected

Inferential Statistics – draw inferences about the population from which the sample was drawn

Some quick notes:

Deciding on the appropriate statistical test requires understanding the level of measurement and the type of variable.

categorical(discreet) vs. continuous

nominal, ordinal, interval and ratio

Conventions:

I will try and use Latin letters to represent sample statistics, I will also use Greek letters to represent population parameters.

Estimates of population parameters are often represented with a ^ (pronounced hat) over the letter.

Descriptive Statistics

Describing the data you've collected

Univariate single variable

Frequency distributions (categorical)

count

Relative frequency (percentage) distributions

valid percent

total percent

Proportion

Other ways of describing the distribution

3 Measures of Central tendency

1. Mean -sometime called the first moment

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{(x_1 + x_2 + \dots + x_n)}{n}$$

2. Median – When the data is ordered largest to smallest it is the middles number if there are an odd number, and the mean of the middle two if there are an even number. The 50th percentile

3. Mode – the most frequently occurring

4. Trimmed mean (delete upper and lower % and calculate mean)

When are the mean and median different why might you prefer one to the other. When might you use the mode?

SPSS Demo

Measures of Dispersion (spread)

Range – highest – lowest value

Variance - sometimes called the second moment

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

Standard Deviation

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

Tchebysheff's theorem: $1 - \frac{1}{k^2}$ for $k > 1$ for any distribution. Where k represents the number of standard deviations. So 3/4 of the values lie within 2 standard deviations

Empirical rule: 68% lie +/- 1 s.d., 95% lie +/- 2 s.d.

Measures of Shape

Skewness (third moment) Positive Skew long tail to right

$$\frac{\sum (X - \mu)^3}{N\sigma^3}$$

Kurtosis (fourth moment) measure of the size of tails Leptokurtic, fat tails > 0

Platykurtic small tails < 0

$$\frac{\sum (X - \mu)^4}{N\sigma^4} - 3$$

Graphical Representation of Univariate descriptive stats

Categorical

Bar Chart

Pie Chart

Continuous

Histogram

Line Chart

Box and Whiskers

The following example uses SPSS syntax to generate some descriptive statistics from above. This is done using the DeMaris (2004) dataset for faculty salaries.

salary: Academic year (9 month) salary in US dollars

market: the ratio of average national salary for the discipline to average salary of all disciplines.

male: dummy variable (indicator variable) 1=male 0=female

yearsdeg: time since degree in years

GET

FILE='C:\Documents and Settings\brooks.tagg\My Documents\Classes\ECO
307\data\faculty.sav'.

DATASET NAME DataSet2 WINDOW=FRONT.

DESCRIPTIVES VARIABLES=salary

/STATISTICS=MEAN STDDEV MIN MAX SKEWNESS KURTOSIS.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
salary	514	29000.00	96156.00	50863.8734	12672.77130
Valid N (listwise)	514				

Descriptive Statistics

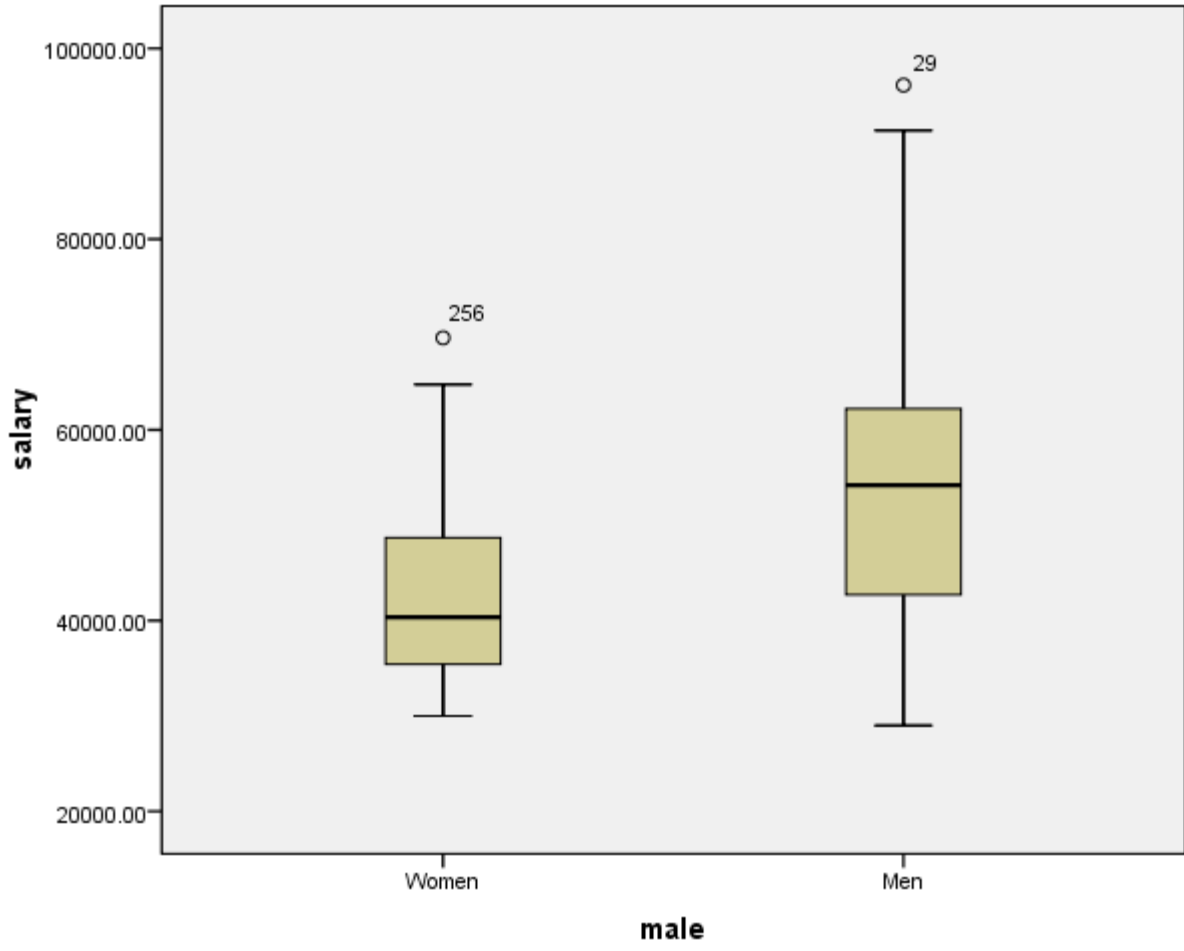
	N	Skewness		Kurtosis	
	Statistic	Statistic	Std. Error	Statistic	Std. Error
salary	514	.449	.108	-.235	.215
Valid N (listwise)	514				

EXAMINE VARIABLES=salary BY male

/PLOT=BOXPLOT

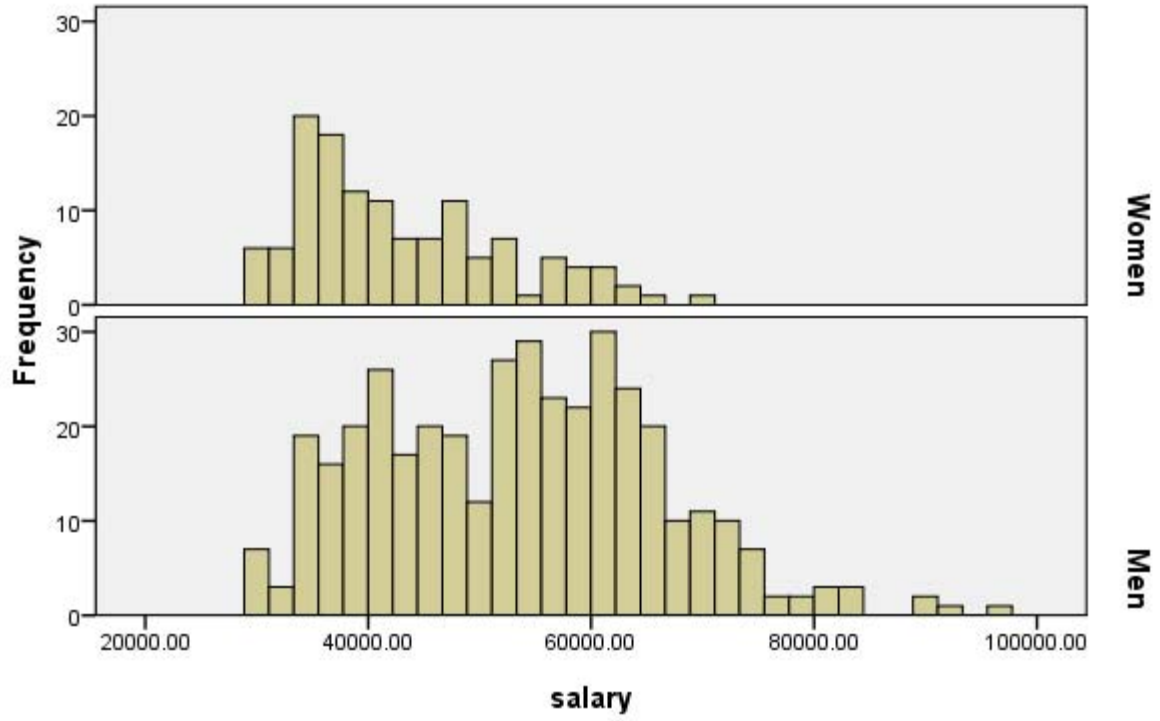
/STATISTICS=NONE

/NOTOTAL.



```
GRAPH  
/TITLE='Histogram of Academic Salary'  
/FOOTNOTE='Footnote Data are from DeMaris (2004)'  
/HISTOGRAM=salary  
/PANEL ROWVAR=male ROWOP=CROSS.
```

Histogram of Academic Salary



Footnote: Data are from DeMaris (2004)

Lecture 3:

Univariate and Bivariate

Bivariate Descriptive statistics

2 variables

3 possible combinations

Categorical/Categorical

Crosstabulations (2 way frequency tables, Crosstabs, Bivariate distributions)

Example:

Smoke\Gender	Male	Female	Row total
Yes	30	25	55
No	20	25	45
column total	50	50	100

Categorical/Continuous

Any statistic that applied to cont. variables done for each category

Continuous/Continuous

Simple Correlation coefficient (Pearson's product-moment correlation coefficient, Covariance)

$$r_{xy} = r_{yx} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

this ranges from +1 to -1

Graphical Representations

Bar Charts pie charts etc.

histogram, box plots

scatter plots

Inferential Statistics

http://www.ruf.rice.edu/~lane/stat_sim/sampling_dist/index.html

To draw inference from a sample about the properties of a population

Population distribution: The distribution of a given variable(parameter) for the entire population

Sample distribution: A sample of size n, is drawn from the population and the variable's distribution is called the sample distribution.

Sampling distribution: This refers to the properties of a particular test statistic. The sampling distribution draws the distribution of the test statistic if it were calculated from a sample of size n , then resample using n observation to calculate another test statistic. Collect these into the sampling distribution.

Central Limit Theorem: The central limit theorem states that given a distribution with a mean m and variance s^2 , the sampling distribution of the mean approaches a normal distribution with a mean (m) and a variance s^2/N as N , the sample size increases. The amazing and counter- intuitive thing about the central limit theorem is that no matter what the shape of the original distribution, the sampling distribution of the mean approaches a normal distribution. Furthermore, for most distributions, a normal distribution is approached very quickly as N increases.

How can we use this information? We can use our knowledge of the sampling distribution of a test statistic, a single realization of that test statistic to infer the probability that it came from a certain population

Lecture

Z versus T

We estimate the standard deviation of the population using the sample data, and sample standard deviation, which we then call the standard error.

Confidence intervals for the mean/proportion

$$CI = \bar{x} \pm Z_{C.L.} S_{\bar{x}}$$

Where $Z_{C.L.}$ is the appropriate std. normal value for the associate confidence level.

95% C.L. = 1.96

99% C.L. = 2.57

90% C.L. = 1.65

http://www.math.csusb.edu/faculty/stanton/m262/confidence_means/confidence_means.html

and $S_{\bar{x}} = \frac{S}{\sqrt{n}}$ the standard error of the mean (based on the C.L.T)

The population mean lies within the range.

General formula for a confidence interval for any statistic.

$CI = STAT \pm Z_{C.L.} \sigma_{STAT}$ if the normal distribution describes the sampling distribution of the stat.

Hypothesis testing.

Develop a hypothesis about the population, then ask does the data in our sample support the hypothesized population characteristic.

Ho: Null hypothesis

Ha: Alternative hypothesis

Significance level. The critical point where the probability of realizing this sample when pulled from a population as hypothesized under the under the null

Type I and II Errors (Innocent until proven Guilty)

What if Ho = innocent

State of Ho in pop	Accept Ho	Reject Ho
Ho is true	Correct	Type I error
Ho is false	Type II error	Correct

alpha = the nominal size of the test (probability of a type I error)

Beta = probability of a type II error

1-beta= the power of a test (ability to reject a false null)

Trade off between type one and type two errors, for a given sample size. More information will reduce both type I and type II errors.

One sample T test

$$Z = \frac{\bar{x} - \mu}{S_{\bar{x}}} \text{ where } \mu \text{ is the hypothesized mean.}$$

If the calculated Z statistic is larger than the critical value (C.L.) then we reject the null hypothesis, we can also use p-values. That is the exactly probability of drawing this sample (or one at least as different as this sample) from a population as is hypothesized under the null distribution. If the p-value is large (generally larger than .05 (5%)), we fail to reject the null, if it is small we reject the null. As the probability of realizing a sample like this one, from a population as is hypothesized under the null is quite small. Small enough that we a fairly confident it didn't come from this population.

Z distribution (standard normal) vs. t-distribution (students t)

The t distribution is used in situations where the population variance is unknown and the sample size is less than 30.

Example

Population proportion

$$Z = \frac{p - \pi}{S_p} \text{ where } \pi \text{ is the hypothesized population proportion}$$

$$S_p = \sqrt{\frac{p(1-p)}{n}}$$

Example

Males represent 47.9% of the population over the age of 18.

$$H_0: \pi = .479$$

$$H_a: \pi \neq .479$$

Bivariate inferential stats

continuous/categorical

compare means (z test of differences, independent sample) Most often we will have independent groups, one observation on the variable of interest per individual, with the individuals divided into groups. Tests fro differences in means with dependent groups would require you observe the same individual twice, once pre treatment and once post (sometimes called a related pairs test).

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{S_{\bar{x}_1 - \bar{x}_2}}$$

$$H_0 : \mu_1 - \mu_2 = 0; \mu_1 = \mu_2$$

$$H_a : \mu_1 - \mu_2 \neq 0; \mu_1 \neq \mu_2$$

Where $S_{\bar{x}_1 - \bar{x}_2}$ is the pooled estimate of the standard error of the mean, assuming the underlying population variances are equal (homoscedastic).

$$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}} \quad \text{Pooled estimate of the standard error (population variances equal)}$$

This test can be adapted for comparing population proportions

$$Z = \frac{p_1 - p_2}{S_{p_1 - p_2}}$$

$$H_0 : \pi_1 - \pi_2 = 0; \pi_1 = \pi_2$$

$$H_a : \pi_1 - \pi_2 \neq 0; \pi_1 \neq \pi_2$$

$$S_{p_1 - p_2} = \sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Where $\bar{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

And $\bar{q} = 1 - \bar{p}$

```
T-TEST GROUPS=male(1 0)
/MISSING=ANALYSIS
/VARIABLES=salary
/CRITERIA=CI(.9500).
```

Group Statistics

	male	N	Mean	Std. Deviation	Std. Error Mean
salary	Men	386	53499.2370	12583.47833	640.48218
	Women	128	42916.6048	9161.60959	809.77953

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	95% Confidence Interval of the Difference	
									Lower	Upper
salary	Equal variances assumed	15.146	.000	8.772	512	.000	10582.63224	1206.34541	8212.63627	12952.62821
	Equal variances not assumed			10.250	297.227	.000	10582.63224	1032.45354	8550.78703	12614.47745

Continuous/Continous

pearson's correlation coefficient

$$r_{xy} = r_{yx} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$Z = \frac{r_{xy} - 0}{S_r}$$

$$S_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

$$H_0 : \rho_{xy} = 0$$

$$H_a : \rho_{xy} \neq 0$$

CORRELATIONS

```

/VARIABLES=salary yearsdg market
/PRINT=TWOTAIL NOSIG
/MISSING=PAIRWISE.

```

Correlations

		salary	yearsdg	market
salary	Pearson Correlation	1.000	.680**	.407**
	Sig. (2-tailed)		.000	.000
	N	514.000	514	514
yearsdg	Pearson Correlation	.680**	1.000	-.083
	Sig. (2-tailed)	.000		.059
	N	514	514.000	514
market	Pearson Correlation	.407**	-.083	1.000
	Sig. (2-tailed)	.000	.059	
	N	514	514	514.000

** . Correlation is significant at the 0.01 level (2-tailed).

Lecture :

Categorical/Categorical

Chi-squared test of independence

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

with degrees of freedom (R-1)(C-1)

where R = number of rows and

C= number of columns

Where $E_{ij} = \frac{R_i C_j}{n}$

Smoke\Gender	Male	Female	Row total
Yes	30 (27.5)	25 (27.5)	55
No	20 (22.5)	25 (22.5)	45
column total	50	50	100

$\chi^2 = 1.01$ and the critical value with 1 degree of freedom at the 5% level is 3.84 fail to reject H_0

H_0 :The variables are independent, that is to say knowledge of one will not help to predict the outcome of the other

H_a : The variables are related, that is to say knowing one variable will help you predict the other

This is not the most powerful test (ability to reject a false null) primarily because the underlying data can be ordinal in nature and the chi-squared test does not exploit this information

CAUTION For this test to be valid, every cell must have at least 5 observations, if they don't collapse categories where it makes sense.

CROSSTABS

```
/TABLES=rank BY male  
/FORMAT=AVALUE TABLES  
/STATISTICS=CHISQ  
/CELLS=COUNT EXPECTED  
/COUNT ROUND CELL.
```

rank * male Crosstabulation

			male		
			Women	Men	Total
rank	Assistant	Count	60	83	143
		Expected Count	35.6	107.4	143.0
	Associate	Count	49	111	160
		Expected Count	39.8	120.2	160.0
	Full	Count	19	192	211
		Expected Count	52.5	158.5	211.0
Total		Count	128	386	514
		Expected Count	128.0	386.0	514.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	53.560 ^a	2	.000
Likelihood Ratio	57.596	2	.000
Linear-by-Linear Association	51.919	1	.000
N of Valid Cases	514		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 35.61.