Lecture 5: Statistical analyses

BIOSTAT 212, Summer 2017

Kristen Aiemjoy
Department of Epidemiology and Biostatistics
University of California San Francisco

Final Project

- Due 9/12 @ 9 PM
- 100 pts +10 Extra credit
- Create a publication quality table and graph
- Instructions and example on the course website
- Start looking for data

Doctoral Program in Epidemiology and Translational Sciences

Structural equation modeling and Bayesian Analysis Using Stata

Chuck Huber, PhD
Senior Statistician, StataCorp LP
Adjunct Associate Professor of Biostatistics,
Texas A&M School of Public Health

Structural Equation Modeling using Stata

In this talk I introduce the concepts and jargon of structural equation modeling (SEM) including path diagrams, latent variables, endogenous and exogenous variables, and goodness of fit. I describe the similarities and differences between Stata's -sem- and -gsem- commands. Then I demonstrate how to fit many familiar models such as linear regression, multivariate regression, logistic regression, confirmatory factor analysis, and multilevel models using -sem- and -gsem-. I wrap up by demonstrating how to fit structural equation models that contain both structural and measurement components.

Bayesian models using Stata

Bayesian analysis has become a popular tool for many statistical applications. Yet many data analysts have little training in the theory of Bayesian analysis and software used to fit Bayesian models. This talk will provide an intuitive introduction to the concepts of Bayesian analysis and demonstrate how to fit Bayesian models using Stata. No prior knowledge of Bayesian analysis is necessary and specific topics will include the relationship between likelihood functions, prior, and posterior distributions, Markov Chain Monte Carlo (MCMC) using the Metropolis-Hastings algorithm, and how to use Stata's Bayes prefix to fit Bayesian models.

August 29, 2017

10-12pm Structural Equation Modeling using Stata

12-1pm Bayesian models using Stata

MH-1401

We would appreciate RSVPs so we can ensure the appropriate room capacity

RSVP: http://bit.ly/2vCPe7N

Sponsored by the UCSF PhD program in Epidemiology and Translational Science

UCSF Department of Epidemiology & Biostatistics

Displaying numeric values for labels

`numlabel, add`

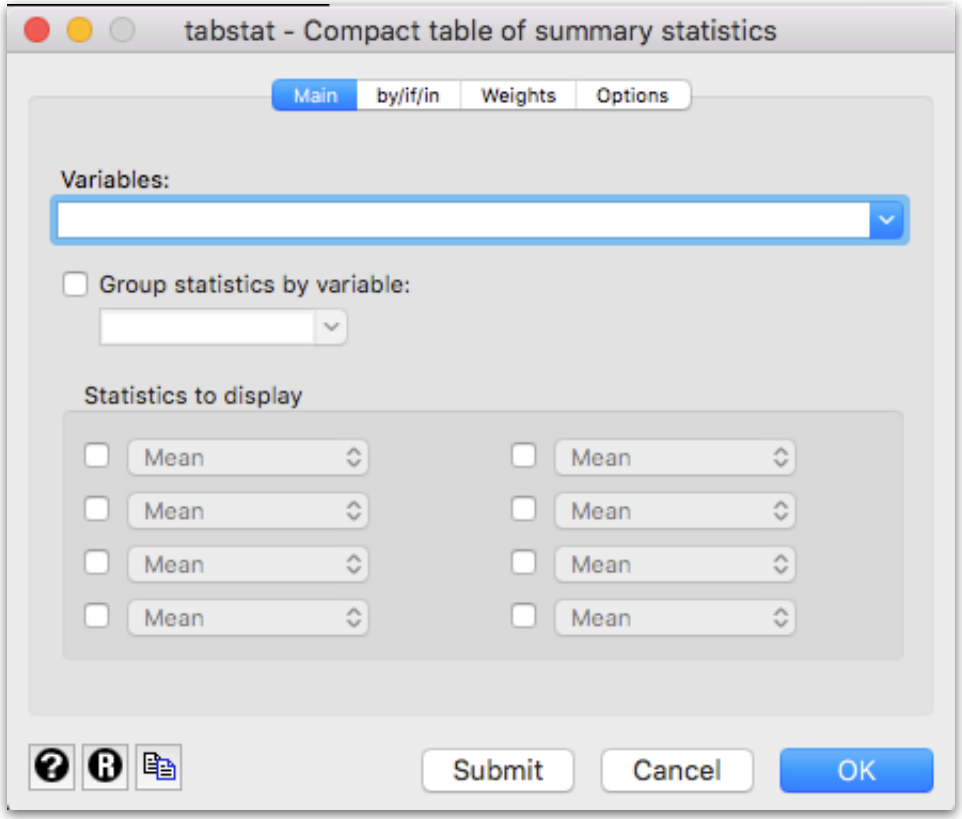
```
. tab sex

  1=male,
 2=female |      Freq.     Percent        Cum.
----------+-----------------------------------
   1. Male |      4,915       47.48       47.48
 2. Female |      5,436       52.52      100.00
----------+-----------------------------------
     Total |     10,351      100.00
```

`labelbook` – gets you a list of all numeric labels and their values

Outline

- **Descriptive Statistics**
- Confidence intervals
- Statistical Hypothesis tests and P-values
  - T-test
  - ANOVA
  - Chi squared test
- Linear Regression

Table of summary statistics: `tabstat`



foreign	variable	mean	sd	min	max
Domestic	price	6072.423	3097.104	3291	15906
	weight	3317.115	695.3637	1800	4840
	mpg	19.82692	4.743297	12	34
	rep78	3.020833	.837666	1	5
Foreign	price	6384.682	2621.915	3748	12990
	weight	2315.909	433.0035	1760	3420
	mpg	24.77273	6.611187	14	41
	rep78	4.285714	.7171372	3	5

```
sysuse auto, clear
tabstat price weight mpg rep78, by(foreign)
        stat(mean sd min max) nototal long col(stat)
```

Statistics > Summaries, tables, and tests > Other tables
> Compact table of summary statistics

`tabstat` syntax

`tabstat varlist, statistics(_____)`

-produces a summary table of statistics for all variables in `varlist`

Statistics

`mean`	-mean
`n`	-count of nonmissing observations
`sum`	-sum
`max`	-maximum
`min`	-minimum
`range`	-range = max - min
`sd`	-standard deviation
`variance`	-variance
`median`	median

*many more options-check help file

Practice

- Produce a table of [mean, standard deviation, min and max] for post-study headache score by intervention group

```
tabstat posths, by(group) stat(mean sd min max)
```

- What is the mean headache score post-study in the intervention group?

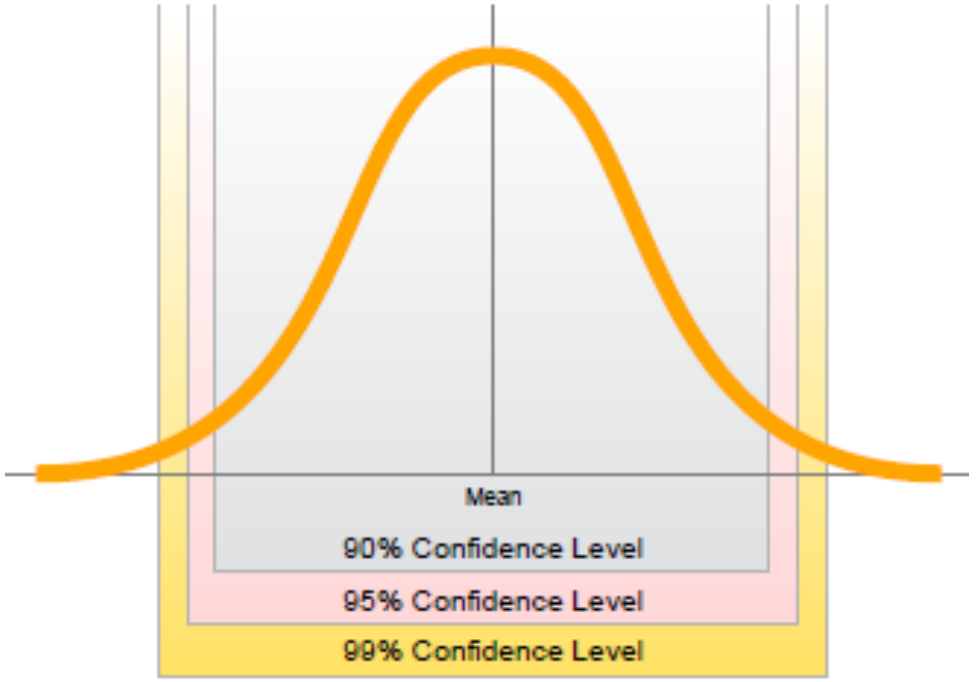
- What is the mean headache score post-study in the control intervention group?

Outline

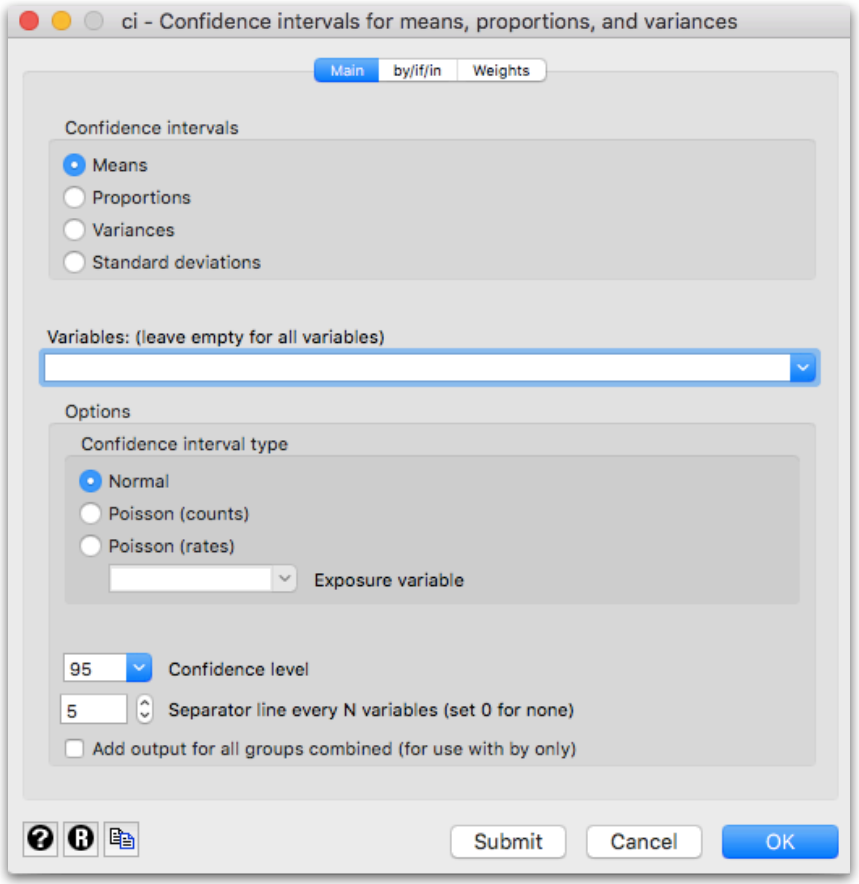
- Descriptive Statistics
- **Confidence intervals**
- Statistical Hypothesis tests and P-values
  - T-test
  - ANOVA
  - Chi squared test
- Linear Regression

Confidence intervals

- Over many repetitions of the study (with different samples) 95% of estimates will be within 1.96 SEs of the true mean
  - 99% will be within 2.33 SEs



Confidence intervals: `ci`




```
. ci means mpg price, level(90)

    Variable |        Obs        Mean    Std. Err.      [90% Conf. Interval]
-------------+---------------------------------------------------------------
         mpg |         74     21.2973    .6725511       20.17683    22.41776
       price |         74    6165.257    342.8719       5594.033     6736.48
```

```
sysuse auto
ci means mpg price, level(90)
```

Statistics > Summaries, tables, and tests
> Summary and descriptive statistics > Confidence intervals

`ci` syntax

`ci mean varname`	- calculate confidence interval for a mean
`ci proportion varname`	- calculate confidence interval for a proportion

<u>options</u>

`, level(#)`	-set confidence level (default is 95)
`if`	-subset data for CI using logic
`bysort varname:`	-CI for each level of `varname`

Practice

- What is the 95% CIs around the mean post-study headache score between the intervention group and the control group?

```
bysort group: ci mean posths
```

Outline

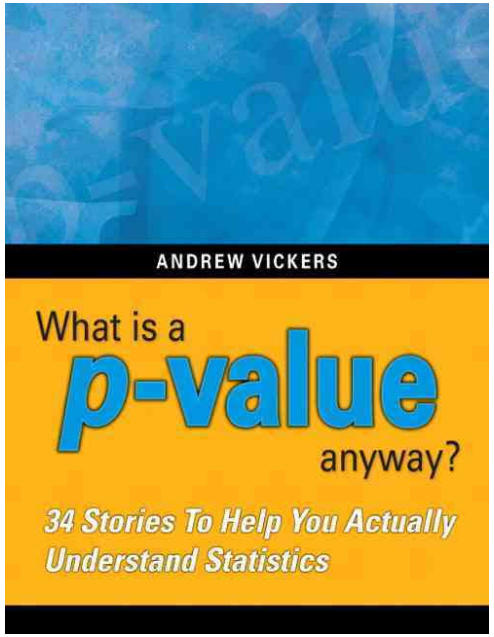
- Descriptive Statistics
- Confidence intervals
- **Statistical Hypothesis tests and P-values**
  - T-test
  - ANOVA
  - Chi squared test
- Linear Regression

Hypotheses

- Null hypothesis (H_0): No association between predictor and outcome in this population
- Alternative Hypothesis (H_a): There is an association between predictor and outcome in this population
  - One-sided: states specific direction of the association (larger or smaller)
  - Two-sided: does no specify direction of the association

P-values

- The probability of obtaining a value of the test statistic equal to or one more extreme than was obtained in the study if the null hypothesis is true.



Preview: Biostat 200: Hypothesis tests

Data and comparison type	**Parametric tests** Stata command	**Non-parametric tests** Stata command
Continuous; One mean	t-test `ttest varname==#`	

Continuous; Two means
paired data
	Paired t-test
	`ttest var1==var2`
		Sign test
		`signtest var1=var2`
		Wilcoxon Signed-Rank
		`signrank var1=var2`

Continuous; Two means independent data	T-test `ttest var1, by(var2)`	Wilcoxon rank-sum test `ranksum var1, by(var2)`
Continuous, Two or more means independent data	ANOVA `oneway var1 var2`	Kruskal Wallis test `kwallis var1, by(var2)`
Dichotomous; One proportion	Proportion test `prtest var1==#`	
Dichotomous; two proportions	Proportion test `prtest var1, by(var2)`	Chi-square test `tab var1 var2, chi exact`
Categorical by categorical		Chi-square test `tab var1 var2, chi exact`

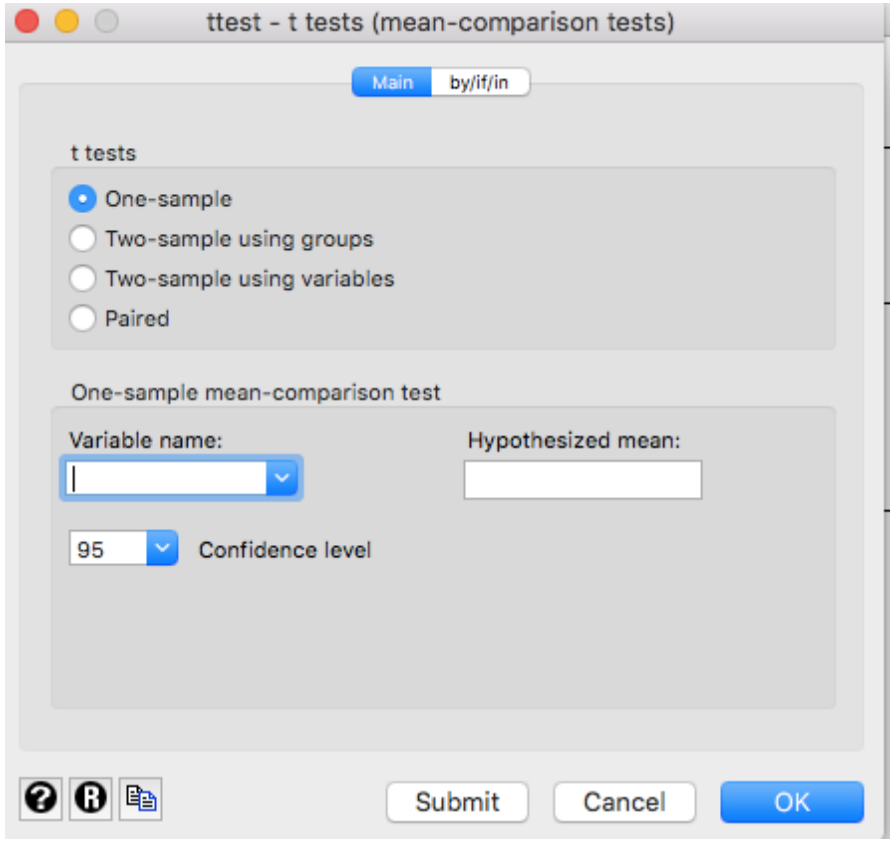
Outline

- Descriptive Statistics
- Confidence intervals
- Statistical Hypothesis tests and P-values
  - **T-test**
  - ANOVA
  - Chi squared test
- Linear Regression

t-test

1. Single sample t-test: tests the null hypothesis that the population mean is equal to a given mean
2. Paired t-test: compare 2 means when the observations are not independent of one another (ie weight before and after intervention)
3. Independent group t-test: compare means of same variable between two groups

t-test



One-sample t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
mpg	74	21.2973	.6725511	5.785503	19.9569	22.63769

mean = mean(**mpg**) t = **1.9289**
Ho: mean = **20** degrees of freedom = **73**

Ha: mean < **20** Ha: mean != **20** Ha: mean > **20**
Pr(T < t) = **0.9712** Pr(|T| > |t|) = **0.0576** Pr(T > t) = **0.0288**

```
sysuse auto
ttest mpg==20
```

Statistics > Summaries, tables, and tests > Classical tests of hypotheses > t test (mean-comparison test)

`ttest` syntax

`ttest contvar==#`	- single sample t-test
`ttest contvar1==contvar2`	- paired t-test
`ttest contvar1, by(catvar2)`	- independent group t-test

Options

`,level(#)`	– set confidence level; default is 95
`,if`	– subset data using logic operators

Practice

- Is mean baseline headache score different between males and females?

  - STEP 1: Visualize

```
graph bar (mean) basehs, over(sex)
```

  - Step 2: t-test

```
ttest basehs, by(sex)
```

t-test: annotated output

```
ttest basehs, by(sex)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group[a] |     Obs[b]        Mean[c]    Std. Err.[d]   Std. Dev.[e]   [95% Conf. Interval][f]
---------+--------------------------------------------------------------------
       0 |        64    27.22656    2.317553    18.54043     22.5953    31.85782
       1 |       337    26.37389    .8551513    15.69849    24.69176    28.05601
---------+--------------------------------------------------------------------
combined |       401    26.50998    .8071529    16.16322    24.92318    28.09677
---------+--------------------------------------------------------------------
    diff[g]|            .8526752    2.206264               -3.48468     5.19003
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   0.3865[h]
Ho: diff = 0                     degrees of freedom =      399[i]

    Ha: diff < 0[k]                 Ha: diff != 0[j]                 Ha: diff > 0[k]
 Pr(T < t) = 0.6503         Pr(|T| > |t|) = 0.6993          Pr(T > t) = 0.3497
```

t-test: annotated output, cont.

a **Group** - This column gives categories of the predictor variable (sex). This variable is specified by the `by(sex)` statement.

b **Obs** - This is the number of valid (i.e., non-missing) observations in each group.

c **Mean** - This is the mean of the outcome variable (basehs) for each level of the predictor variable (sex).

d **Std Err** - This is the standard error of the mean for each level of the predictor variable (sex).

e **Std Dev** - This is the standard deviation of the outcome variable (basehs) for each of the levels of the predictor(sex). On the last line the standard deviation for the difference is given.

f **[95% Conf. Interval]** - These are the lower and upper confidence limits of the means.

t-test: annotated output, cont.

g diff - The difference in mean basehs between males and females.

h t - This is the t-statistic. It is the ratio of the mean of the difference to the standard error of the difference: (.8526752/2.206264)

i degrees of freedom - The degrees of freedom for the paired observations is simply the number of observations minus 2. We use one degree of freedom for estimating the mean of each group, and because there are two groups, we subtract two degrees of freedom.

j Pr (|T| > |t|) - This is the **two-tailed p-value** computed using the t distribution. It is the probability of observing a greater absolute value of t under the null hypothesis.

k Pr(T < t), Pr(T>t) - These are the one-tailed p-values for the alternative hypotheses (mean difference < 0) and (mean difference > 0), respectively.

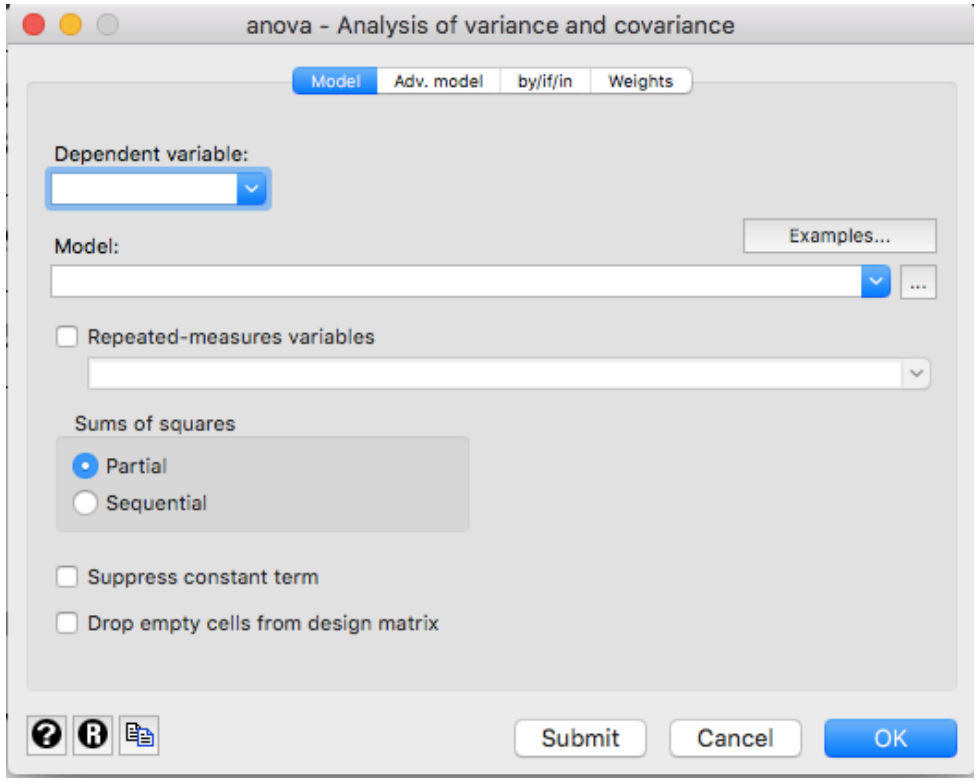
Outline

- Descriptive Statistics
- Confidence intervals
- Statistical Hypothesis tests and P-values
  - T-test
  - **ANOVA**
  - Chi squared test
- Linear Regression

ANOVA

- The extension of the t-test to several independent groups is called analysis of variance or ANOVA (Analysis of Variance)
- The null hypothesis is:
  - H_0: all equal means $\mu_1=\mu_2=\mu_3=\ldots$
- The alternative H_A is:
  - at least one of the means differs from the others

ANOVA




```
. oneway systolic drug, tabulate

           |  Summary of Increment in Systolic B.P.
Drug Used  |      Mean   Std. Dev.       Freq.
-----------+------------------------------------
         1 |  26.066667   11.677002          15
         2 |  25.533333    11.61813          15
         3 |      8.75     10.0193          12
         4 |      13.5   9.3238047          16
-----------+------------------------------------
     Total |   18.87931   12.800874          58

                    Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      3133.23851      3   1044.41284      9.09     0.0001
 Within groups      6206.91667     54   114.942901
------------------------------------------------------------------------
     Total          9340.15517     57   163.862371
```

```
webuse systolic
oneway systolic drug, tabulate
```

Statistics > Linear models and related >
ANOVA/MANOVA > Analysis of variance and covariance

ANOVA: `oneway`

`oneway contvar1 catvar` – ANOVA test

Options

`,tabulate` - produces table of summary statistics

`,bonferroni` - Bonferroni multiple-comparison test

`,missing` - treat missing values as a category

Practice

- Is mean baseline headache score different by terciles of chronicity?
  - Step 1, generate chronicity tercile variable:

```
xtile chron_3 = chronicity, nq(3)
```

  - Step 2: Visualize

```
graph bar (mean) basehs, over(chron_3)
```

  - Now, run ANOVA:

```
oneway basehs chron_3, tabulate
```

  - What is the P-value?

ANOVA: annotated output

oneway basehs chron_3, tabulate

3 quantiles chronicity	Summary of basehs Mean[a]	Std. Dev.[b]	Freq.[c]
1	25.229167	14.606128	136
2	25.943813	16.918058	132
3	28.381579	16.848916	133
Total	26.509975	16.163224	401

Analysis of Variance

Source	SS[d]	df[e]	MS[f]	F[g]	Prob > F[h]
Between groups	731.300961	2	365.65048	1.40	0.2472
Within groups	103768.618	398	260.72517		
Total	104499.919	400	261.249797		

Bartlett's test for equal variances[i]: chi2(2) = 3.6110[i] Prob>chi2 = 0.164

ANOVA: annotated output, cont.

a **Mean** - the mean of the outcome variable (basehs) for each category of the predictor variable (chronicity quantile)

b **Std Dev** - the standard deviation of the outcome variable (basehs) for each of the levels of the predictor variable (chronicity quantile)

c **Freq** - the number of observation in each category of the predictor variable (chronicity quantile)

ANOVA: annotated output, cont.

d **SS** – Sum of squares [SS between = variability around the overall mean] [SS within = weighted average of variance within each group]

e **df**- degrees of freedom [Between group: number of groups(k) – 1] [within group: total number of observations(n) – number of groups(k)]

f **MS**- mean squares is equal to the SS/df

g **F**- The F statistic is the ratio of between-group variability [MS between] to within-group variability [MS within]

h **Prob > F-** p-value for the null hypothesis test that means are the same

i ***Bartlett's test for equal variance-*** *ANOVA assumes that the variance is equal across groups. This test verifies that assumption. A low p-value (<0.05) indicates that the variance is not equal between groups.*

Outline

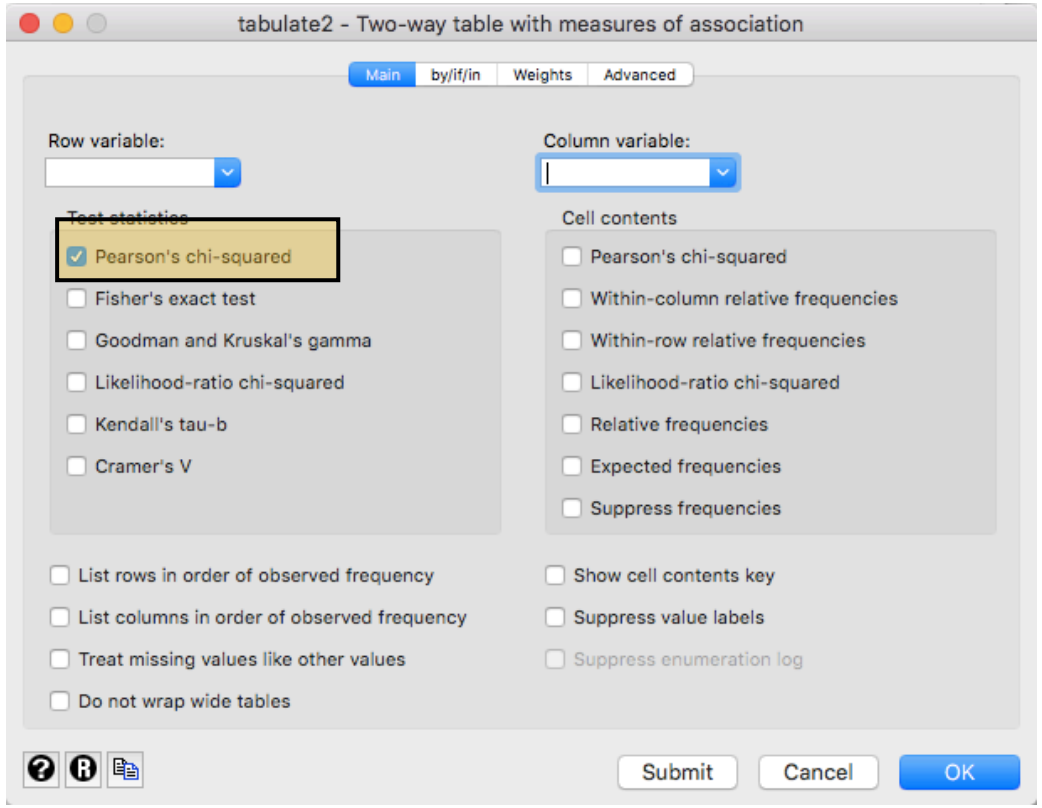
- Descriptive Statistics
- Confidence intervals
- Statistical Hypothesis tests and P-values
  - T-test
  - ANOVA
  - **Chi squared test**
- Linear Regression

Chi-squared test

- 2 categorical variables
- compares the observed frequency (O) in each cell with the expected frequency (E) under the null hypothesis of no difference
- The differences O-E are squared, divided by E, and added up over all the cells
- The sum of this is the test statistic

$$X_1^2 = \sum_{i=1}^{4} \left(\frac{(O_i - E_i)^2}{E_i} \right)$$

Chi-squared test



Drug type (1=placebo)	1 if patient died 0	1	Total
1	1 5.00	19 95.00	20 100.00
2	8 57.14	6 42.86	14 100.00
3	8 57.14	6 42.86	14 100.00
Total	17 35.42	31 64.58	48 100.00

Pearson chi2(2) = 13.8678 Pr = 0.001

```
sysuse cancer
tab drug died, row chi2
```

Statistics > Summaries, tables, and tests > Frequency tables > Two-way table with measures of association

Chi squared test syntax

`tab var1 var2, chi2` – chi squared test

Options

`,row` -gives row percentages

`,col` - gives column percentages

Practice

- Is the proportion men with migraine different from the proportion of women with migraine?
  - Step 1: Visualize

```
graph bar (mean) migraine, over(sex)
```

  - Step 2: Chi-squared test

```
tab sex migraine, row chi2
```

Chi squared test: annotated output

```
tab sex migraine, row chi2

           |       migraine
       sex |         0          1 |     Total
-----------+----------------------+----------
         0 |         5         59 |        64
           |      7.81      92.19 |    100.00
-----------+----------------------+----------
         1 |        19        318 |       337
           |      5.64      94.36 |    100.00
-----------+----------------------+----------
     Total |        24        377 |       401
           |      5.99      94.01 |    100.00

          Pearson chi2(1)[a] =   0.4520[b]   Pr = 0.501[c]
```

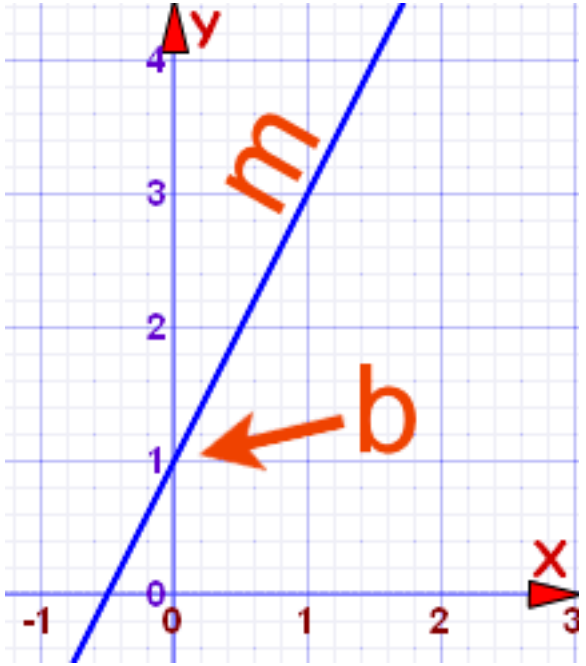
a (degrees of freedom) = (row-1)*(col-1)
b chi squared test statistic = sum of [(O-E)^2/E]
c p-value

Outline

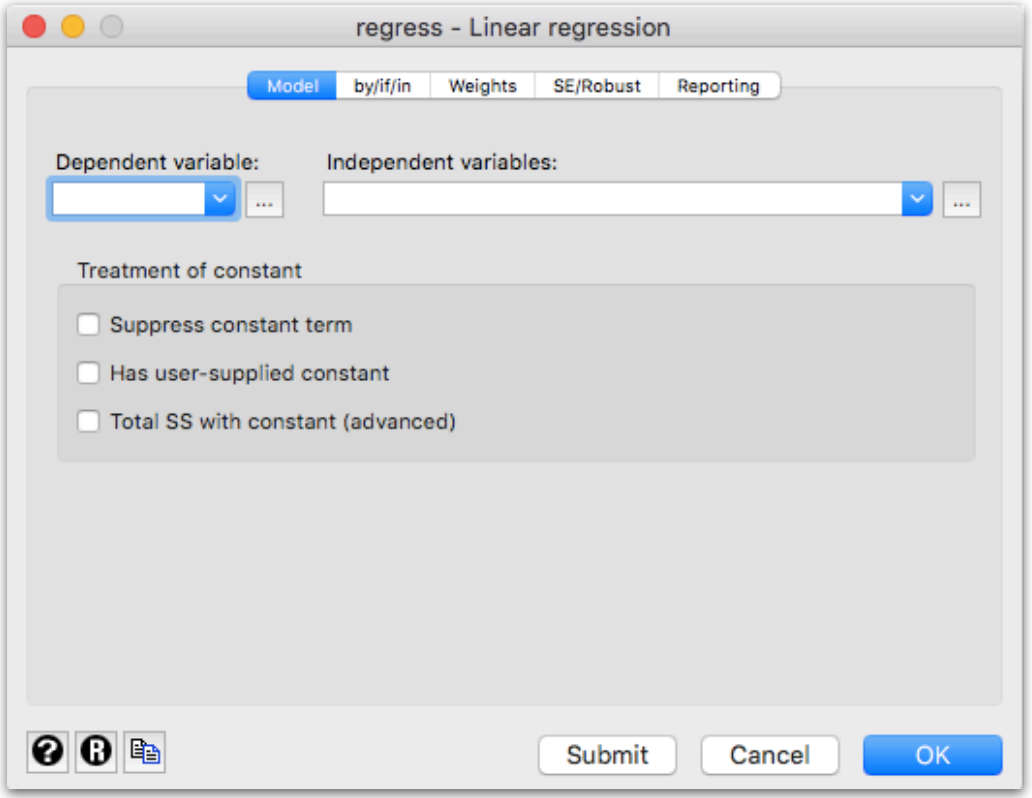
- Descriptive Statistics
- Confidence intervals
- Statistical Hypothesis tests and P-values
  - T-test
  - ANOVA
  - Chi squared test
- **Linear Regression**

What is a linear **regression**?

- Fit a linear equation to your data
- $y = mx + b$
  - m=slope
  - b=intercept
  - y=outcome / dependent
  - x=predictor / independnet



Linear regression




```
. regress mpg weight
```

Source	SS	df	MS	Number of obs	=	74
				F(1, 72)	=	134.62
Model	1591.9902	1	1591.9902	Prob > F	=	0.0000
Residual	851.469256	72	11.8259619	R-squared	=	0.6515
				Adj R-squared	=	0.6467
Total	2443.45946	73	33.4720474	Root MSE	=	3.4389

mpg	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
weight	-.0060087	.0005179	-11.60	0.000	-.0070411	-.0049763
_cons	39.44028	1.614003	24.44	0.000	36.22283	42.65774

```
sysuse auto
regress mpg weight
```

Statistics > Linear models and related
> Linear regression

Practice

- Do age and chronicity (years with headache) have a linear relationship?

  - Step 1: Visualize

```
scatter chronicity age || lfit  chronicity age
```

  - Step 2: Linear regression

```
regress chronicity age
```

Linear regression annotated output

```
regress chronicity age

      Source[a] |       SS[b]           df[c]       MS[d]      Number of obs[e]   =       401
-------------+----------------------------------   F(1, 399)[f]       =    124.44
       Model |  18021.6443             1  18021.6443   Prob > F[f]        =    0.0000
    Residual |  57783.9268           399  144.821872   R-squared[g]       =    0.2377
-------------+----------------------------------   Adj R-squared[h]   =    0.2358
       Total |  75805.5711           400  189.513928   Root MSE[i]        =    12.034

------------------------------------------------------------------------------
  chronicity[j] |      Coef.[k]   Std. Err.[l]      t[m]    P>|t|[m]     [95% Conf. Interval][n]
--------------+---------------------------------------------------------------
          age |   .6074257   .0544519    11.16   0.000     .5003772    .7144742
        _cons |  -6.202495    2.55145    -2.43   0.015    -11.21846   -1.18653
------------------------------------------------------------------------------
```

Linear regression annotated output

a. Source – This is the source of variance, Model, Residual, and Total. The Total variance is partitioned into the variance which can be explained by the independent variables (Model) and the variance which is not explained by the independent variables (Residual, sometimes called Error). Note that the Sums of Squares for the Model and Residual add up to the Total Variance, reflecting the fact that the Total Variance is partitioned into Model and Residual variance.

b. SS – These are the Sum of Squares associated with the three sources of variance, Total, Model and Residual. These can be computed in many ways. Conceptually, these formulas can be expressed as:

SSTotal The total variability around the mean.
SSResidual The sum of squared errors in prediction.
SSModel The improvement in prediction by using the predicted value of Y over just using the mean of Y.

- Note that the SSTotal = SSModel + SSResidual.
- Note that SSModel / SSTotal is equal to 0.2377, the value of R-Squared.

c. df – These are the degrees of freedom associated with the sources of variance. The total variance has N-1 degrees of freedom. In this case, there were N=401 study participants, so the DF for total is 400. The model degrees of freedom corresponds to the number of predictors + the intercept minus 1 (K-1). Including the intercept, there are 2 predictors, so the model has 2-1=1 degrees of freedom. The Residual degrees of freedom is the DF total minus the DF model, 199 – 4 is 195.

d. MS – These are the Mean Squares, the Sum of Squares divided by their respective DF. For the Model, 18021.6443 / 1 = 18021.6443. These are computed so you can compute the F ratio, dividing the Mean Square Model by the Mean Square Residual to test the significance of the predictors in the model.

https://stats.idre.ucla.edu/stata/output/regression-analysis/

Linear regression annotated output, cont.

e. Number of obs – This is the number of observations used in the regression analysis.

f. F and **Prob > F** – The F-value is the Mean Square Model divided by the Mean Square Residual. The p-value is compared to your alpha level (typically 0.05) and, if smaller, you can conclude "Yes, the independent variables reliably predict the dependent variable". Note that this is an overall significance test assessing whether the group of independent variables when used together reliably predict the dependent variable, and does not address the ability of any of the particular independent variables to predict the dependent variable.

g. R-squared – R-Squared is the proportion of variance in the dependent variable (**chronicity**) which can be predicted from the independent variable (**age**). This value indicates that 23.77% of the variance in chronicity can be predicted from the variables **age**. Note that this is an overall measure of the strength of association, and does not reflect the extent to which any particular independent variable is associated with the dependent variable.

h. Adj R-squared – Adjusted R-square. As predictors are added to the model, each predictor will explain some of the variance in the dependent variable simply due to chance. One could continue to add predictors to the model which would continue to improve the ability of the predictors to explain the dependent variable, although some of this increase in R-square would be simply due to chance variation in that particular sample. The adjusted R-square attempts to yield a more honest value to estimate the R-squared for the population.

i. Root MSE – Root MSE is the standard deviation of the error term, and is the square root of the Mean Square Residual (or Error).

https://stats.idre.ucla.edu/stata/output/regression-analysis/

Linear regression annotated output, cont.

j. chronicity – This column shows the dependent variable at the top (**chronicity**) with the predictor variables below it (**age** and **_cons**). The last variable (**_cons**) represents the constant, also referred to in textbooks as the Y intercept, the height of the regression line when it crosses the Y axis. In other words, this is the predicted value of **chronicity** when all other variables are 0.

k. Coef. – These are the values for the regression equation for predicting the dependent variable from the independent variable. These estimates tell the amount of increase in science scores that would be predicted by a 1 unit increase in the predictor.

Chronicity_Predicted = -6.2 +.61*age

For every year increase in age, there is a predicted .61 increase in chronicity

l. Std. Err. – These are the standard errors associated with the coefficients. The standard error is used for testing whether the parameter is significantly different from 0 by dividing the parameter estimate by the standard error to obtain a t-value. The standard errors can also be used to form a confidence interval for the parameter, as shown in the last two columns of this table.

m. t and **P>|t|** – These columns provide the t-value and 2-tailed p-value used in testing the null hypothesis that the coefficient (parameter) is 0.

n. [95% Conf. Interval] – This shows a 95% confidence interval for the coefficient.

https://stats.idre.ucla.edu/stata/output/regression-analysis/

Linear regression output interpretation

Coefficient for age (*m*) = 0.61

Intercept (*b*) = -6.2

Equation: chronicity = 0.61**age* – 6.2

For every year increase in age, there is a predicted .61 increase in chronicity.

For 50 year olds, the predicted chronicity is:

y = 0.61*50 -6.2

= 24.3

```
. sum chronicity if age==50
```

Variable	Obs	Mean	Std. Dev.	Min	Max
chronicity	15	18.93333	14.22506	2	38