Lecture 5: Statistical analysis

BIOSTAT 212, Summer 2016

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

Review/ questions from last week?

Outline for today

- Descriptive Statistics
  - Means, SD, SE
  - Confidence intervals
  - Visualizations
- Statistical Hypothesis tests
  - T-test
  - ANOVA
  - Chi squared test

Descriptive Statistics

Descriptive: `tabstat`

```
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 [n, mean, standard deviation and variance] for baseline, post-study and one year-post study headache score

```
tabstat basehs posths oneyrhs, stat(n mean sd v)
```

- What is the mean headache score post-study?

- How many observations were there one year post-study?

Descriptive: `table`

`tabulate rowvar [colvar [supercolvar]], contents(clist)`

- Flexible table of summary statistics

`clist`	-contents of table cells, select up to 5, default is frequency
`mean varname`	
`sd varname`	-standard deviation
`semean varname`	-standard error of mean
`sum varname`	
`n varname`	-count
`max varname`	
`min varname`	
`median varname`	

Practice

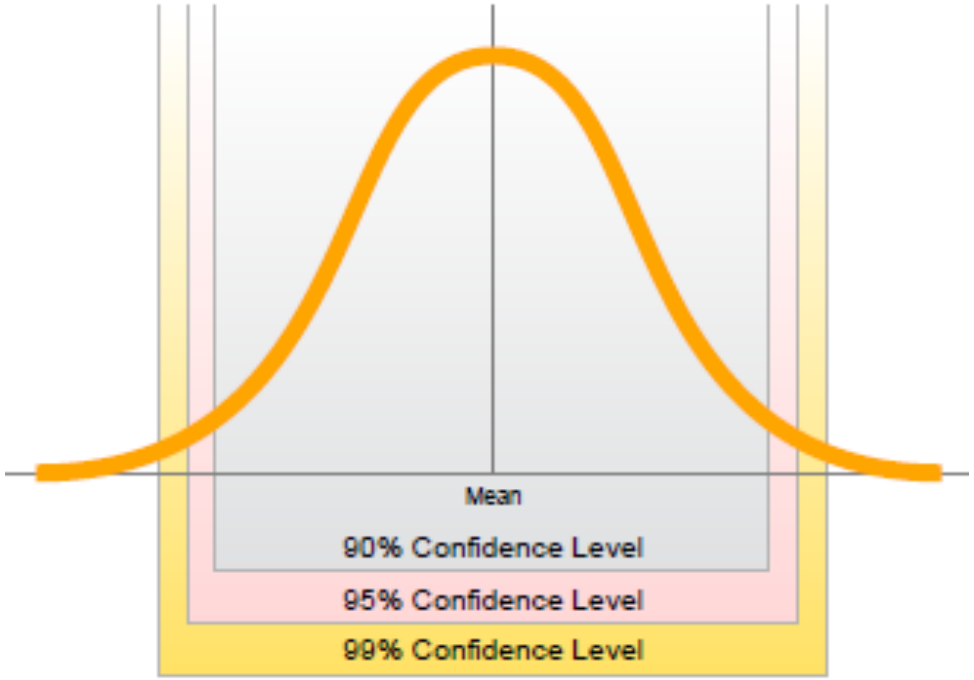
- What is the mean baseline headache score for males with no migraine diagnosis?

```
table sex migraine, c(mean basehs max basehs)
```

- What is the highest baseline headache score for females with a migraine diagnosis?

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



Descriptive: `ci`

`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

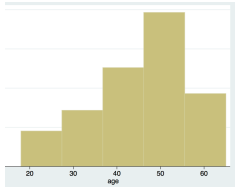
- What is the 95% CIs around the proportion of individuals who have migraine?

```
ci proportion migraine
```

- Is the 99% CI around the proportion of individuals with a migraine wider or narrower than the 95% CI?

```
ci proportion migraine, l(99)
```

Descriptive: `histogram`



`histogram varname` -draw histogram

options

,	`normal`	- overlay normal distribution
,	`kdensity`	-overlay smoothed estimate of distribution
,	`discrete`	-specify variable is dichotomous
,	`bin(#)`	-set number of bins
,	`frequency`	-draw using frequencies (default is density)
,	`by(varname)`	-histogram for every category of `varname`

Practice

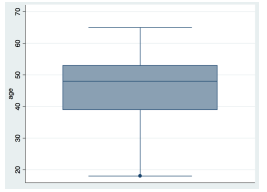
- Does baseline headache score look like it is normally distributed?

```
hist basehs, freq normal
```

- What about age?

```
hist age, freq normal
```

Descriptive: `graph box`



`graph box varname` - draws vertical box plots. (y axis is numerical, x axis is categorical)

`graph hbox varname` - draws horizontal box plots.

<u>options</u>

`, nooutsides` - don't display outliers

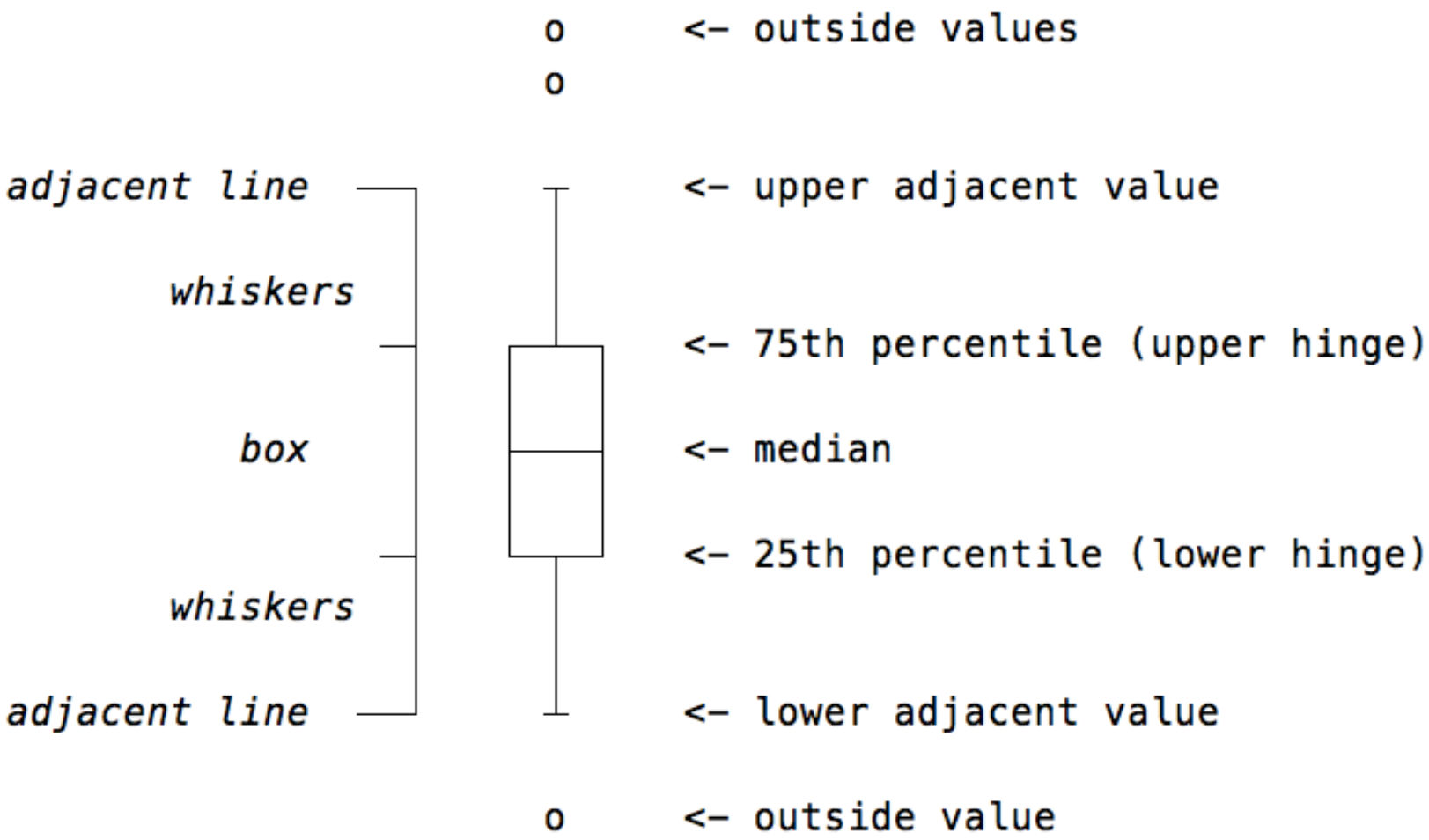
`, over(varname)` - boxplot for each category of `varname`

`, missing` - include missing as category of `varname`

`if` - subset data using logic operators

`in` -subset data

Box-whisker plot



Practice

- Is median post-study headache score lower in the acupuncture or control group?

```
graph box posths, over(group)
```

Statistical Hypothesis Tests

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`

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: `ttest`

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

Options

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

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.

Practice

- Is mean headache score post-intervention lower than mean headache score pre-intervention in the intervention group?

```
ttest basehs==posths if group==1
```

- What is the p-value for the difference in mean headache scores?

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`

`oneway contvar1 catvar2` – ANOVA test

Options

`,tabulate` - produces table of summary statistics

`,bonferroni` - Bonferroni multiple-comparison test

`,missing` - treat missing values as a category

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.*

Practice

- Is mean chronicity (years with headache) different by quartiles of age?
  - First, generate age quartile variable:

```
xtile age4 = age, nq(4)
```

  - Make sure it worked:

```
table age4, c(min age max age mean age)
```

  - Now, run ANOVA test:

```
oneway chronicity age4, tabulate
```

  - What is the P-value?

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: `tabulate`

`tab var1 var2, chi2` – chi squared test

Options

`,row` -gives row percentages

`,col` - gives column percentages

Chi squared test: annotated output

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

sex	migraine 0	migraine 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

Practice

- Is the proportion of people who completed the study similar in the control and intervention groups?

```
tab group completer, row chi2
```

- What is the p-value for the chi2 test?
- Bonus*! (prize) What does this p-value mean (in a sentence)?