Stata Recitation - Week 7 - Confidence Intervals, t-tests and Chi-squared tests

McCourt School of Public Policy, Georgetown University

Key Ideas:

- Calculate confidence intervals from data or from components (mean, sd, n).
- Calculate t-tests for single, paired, pooled, and non-pooled samples.
- Calculate tests for population proportions.
- Calculate statistics based on two variables: tab2 (chi2 option), corr, pwcorr

REVIEW PROBLEMS using the `citytemp.dta` example dataset

```
sysuse citytemp.dta, clear
```

- Generate a dummy variable, hightempJan, that is 1 for all cities with average January temperatures above 40 degrees and 0 for all those with average temperatures of 40 degrees or less.

```
gen hightempjan = 0
replace hightempjan = 1 if tempjan > 40
replace hightempjan = . if tempjan == .
```

- How many cities have high January temperatures?

```
tab hightempjan
```

Ans: 353 cities

- Create a bar graph showing mean cooling degree days and mean heating degree days, broken into two categories, for cities with high January temperatures and low January temperatures.

  ```
  graph bar (mean) heatdd (mean) cooldd, over(hightempjan)
  ```

- How many cities are in the "West" Census region and the "Mountain" Census division?

```
tab region
tab region, nolabel
tab division
tab division, nolabel
count  if region == 4 & division == 8
```

Ans: 61 cities in both the "West" Census region and the "Mountain" Census division

Confidence Intervals

`help ci`

- Syntax for ci: `ci var1name var2name`
- `level()` option, sets confidence level for ci: ci var1name var2name, level(99)
- Note where defaults are given (i.e. default for level is 95%)

Example: Compute 90% confidence intervals

```
clear
sysuse auto
ci weight length mpg , level(90)
```

- Syntax for cii, immediate command: `cii #obs mean sd`
- Kind of like a calculator for CI, can use to confirm your results of computing the formula
- Immediate commands are based on the inputs you type into the command.
- They are not dependent on the data that is in memory.
- You can clear current data set and get the same result

Example: Recompute 90% confidence intervals from components

- Careful about standard error vs. standard deviation
- – Don't use st. err. reported by ci, instead summarize to get st. dev.
- weight

  ```
  sum weight
  cii 74 3019.459 777.1936 , level(90)
  ```

- length

  ```
  sum length
  cii 74 187.9324 22.26634 , level(90)
  ```

T-Test

```
help ttest
```

- Types of Tests:

1. One sample test: One variable.
2. Paired test: Two variables, no options.
3. Unpaired, Pooled test:
    - One variable with `by()` option,
    - or two variables with unpaired option.
4. Unpaired, Nonpooled test:
    - One variable with `by()` option and unequal option,
    - or two variables with unpaired and unequal options

- For 3. and 4. data will usually be set up for the `by()` option version, not the two variable version.

Example 1: One-sample mean-comparison test

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

Review output:

- Null Hypothesis
- Test statistic
- Degrees of freedom
- Alternative hypotheses for one and two-sided tests
- P-values for each alternative hypothesis

Example 2: Paired test

- Two observations from each unit of observation, e.g. person, state, car.

  ```
  clear
  webuse fuel
  ```

- Assumption is that each line is a single car, tested with two types of fuel additive.
- That is why we use the paired test.

  ```
  ttest mpg1==mpg2
  ```

Example 3: Unpaired, Pooled test

- Different units sampled for each group, - not paired
- But, groups are from the same population (theoretical) - same standard deviation

```
clear
webuse fuel3
ttest mpg, by(treated)
```

Example 4: Unpaired, unpooled test

- Different units were sampled for each group - not paired
- Groups may not be from the same population (theoretical) - standard deviations may be different.

```
ttest mpg, by(treated) unequal
```

Proportion Tests

- Test population proportions: prtest - used for binary outcomes
- Examples:

```
clear
sysuse nlsw88.dta
tab married
tab union
tab collgrad
```

- Is the proportion of college graduates different for married and non-married populations?

```
prtest collgrad, by(married)
```

- Is it different for union and non-union populations?

```
prtest collgrad, by(union)
```

PRACTICE using the `lifeexp.dta` example dataset

```
sysuse lifeexp.dta, clear
```

1. Test the null hypothesis that the true value of world life expectancy is 74. What is the probability of seeing the data in this sample given that life expectancy is actually 74, with a two-tailed test?

   ```
   ttest lexp == 74
   ```

   Ans: p-value for two sided test is 0.0037

2. Report the 90 percent confidence interval for the measure of average annual percentage population growth.

   ```
   ci popgrowth, level(90)
   ```

   Ans: 90% confidence interval for average annual percentage population growth: 0.7837118 to 1.160406

3. A wild hypothesis appears: What are the chances of seeing this data, using a two-tailed test, given the null hypothesis that the value of the true mean life expectancy and the true mean water safety score are the same?

   ```
   ttest lexp == safewater, unpaired unequal
   ```

- Run an unpaired ttest since the question asks you to compare the MEANS of the two groups, rather than the difference between lexp and safewater for each observation, which you would use a paired test for.
- Also, add the 'unequal' option for unequal variances (unpooled test) since the sample standard deviations for lexp and safewater are very different. Ans: P-value for the two-sided test is 0.1927

Discrete/Categorical variables : Two-way tabulation and Pearson's chi-2 test

Two way tabulation

Produce counts of number of observations in each cell of a two-way table.

Examples

```
clear
sysuse nlsw88.dta
tab2 race married
```

- First variable goes in rows, second variable goes in columns.
- This is important when you have a variable with many categories:

  ```
  tab2 age married
  tab2 married age
  ```

- Many different options with tab:

  ```
  help tab2
  ```

- We've seen the missing option with one-way tabulations:

  ```
  tab2 union married
  tab2 union married , m
  ```

- Look at total number of observations for these two tables.
- Compare with Obs number from summarize:

  ```
  sum married union
  ```

Important

- We are moving into commands that take data from multiple variables
- If an observation has missing data for any of the variables, that observation is dropped from the calculation.
- With some commands, like `tab2`, we can avoid that behavior. But that will not be possible for other commands.

More options for tab2

- column : Gives percentage breakdown of row category within each column.

  ```
  tab2 race union, column
  tab2 race if union==0
  tab2 race if union==1
  ```

- row : Gives percentage breakdown of column category within each row.

```
tab2 race union, row
tab2 union if race==1
tab2 union if race==2
tab2 union if race==3
```

- cell : Gives percentage of observations in each cell.

```
tab2 race union, cell
```

- expected : Gives the expected number of observations in each cell based on marginal distributions of each variable

```
tab2 race union, row column
```

- expected number of observations in the white, nonunion cell:

```
display 0.7204*0.7545*1878
tab2 race union, expected
```

Report Chi2 test statistic:

- Test for independence of two categorical variables:
- This will be covered in class, but you should know how to calculate and find test statistic.

```
tab2 race union, chi2
```

- test statistic: Pearson chi2(2) = 13.0814
- P-value: Pr = 0.001
- Components of chi2 test statistic can be reported using option: `cchi2`
- Students should read through these options after learning about the Chi-2 test in class.