Reshaping Data

The command **contract *varlist*** creates a dataset with an observation for each combination of the variables in ***varlist.*** The variable _freq is the frequency of each combination.

```
. use lbw1, clear
. contract race smoke
. list
. tab race smoke [fw=_freq]
```

Expand has the opposite effect

```
. expand _freq
. tab race smoke
```

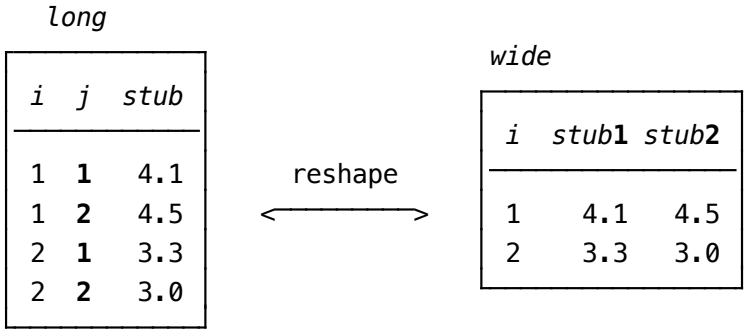
The **collapse** command can also be used to generate aggregated datasets

```
. use lbw1, clear
. collapse (mean) bwt, by(race smoke)
```

This generates a dataset that contains mean birthweight for each combination of smoking and race

Wide versus long data (**reshape**)

Overview



To go from long to wide:

```
                                    j existing variable
                                   /
        reshape wide stub, i(i) j(j)
```

To go from wide to long:

```
        reshape long stub, i(i) j(j)
                                   \
                                    j new variable
```

```
. sysuse bplong, clear
```

Long form:

```
     patient   sex   agegrp     when    bp
          1   Male    30-45   Before   143
          1   Male    30-45    After   153
          2   Male    30-45   Before   163
          2   Male    30-45    After   170
          3   Male    30-45   Before   153
          3   Male    30-45    After   168
```

```
. reshape wide bp, i(patient) j(when)
```

Wide form:

```
     patient  bp1   bp2    sex   agegrp
          1   143   153   Male    30-45
          2   163   170   Male    30-45
          3   153   168   Male    30-45
```

```
. sysuse bpwide, clear
```

Wide form:

```
     patient   sex   agegrp   bp_before   bp_after
          1   Male    30-45         143        153
          2   Male    30-45         163        170
          3   Male    30-45         153        168
```

```
. rename bp_before bp1
. rename bp_after bp2

. reshape long bp, i(patient) j(timeper)
```

Long form:

```
     patient   timeper   sex   agegrp    bp
          1         1   Male    30-45   143
          1         2   Male    30-45   153
          2         1   Male    30-45   163
          2         2   Male    30-45   170
          3         1   Male    30-45   153
          3         2   Male    30-45   168
```

<u>Descriptive Analysis</u>

Categorical Variables:

Two-way contingency tables (cross tables) using **tabulate** and **table**

```
. use lbw1, clear
. tab smoke ht, col chi2
```

```
. bysort race: tab smoke ht, col chi2
. tab smoke ht race, row col

. table smoke ht, c(mean bwt)
. table smoke ht, c(freq mean bwt max bwt)
```

Continuous Variables:

```
. summarize bwt, detail
. centile bwt, centile(25 50 75 99)
```

Summary statistics for numeric variables categorized by another variable

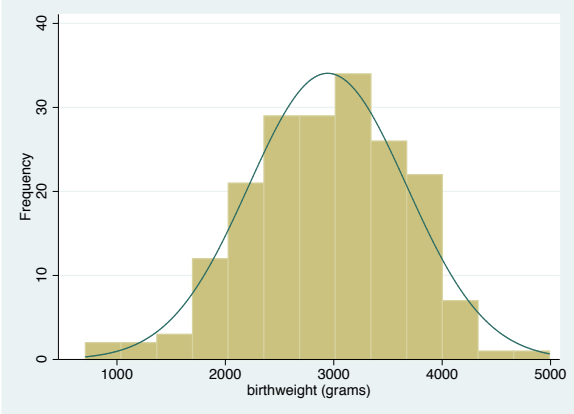
```
. tabstat age lwt bwt, by(race)
. tabstat age lwt bwt, by(smoke) stat(n mean sd semean median)
```

To display the statistics columnwise and control the display format:

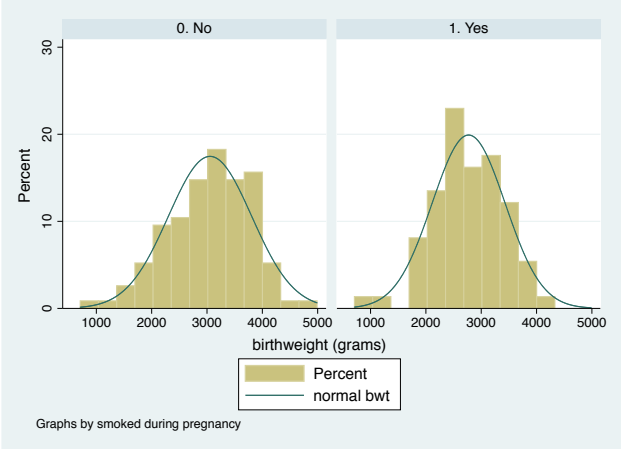
```
. tabstat age lwt bwt, by(smoke) stat(n mean sd p25 p75) col(stat)
format(%8.2f)
```

Histograms:

```
. histogram bwt
. hist bwt, frequency normal
```




```
. hist bwt, percent normal by(smoke)
```

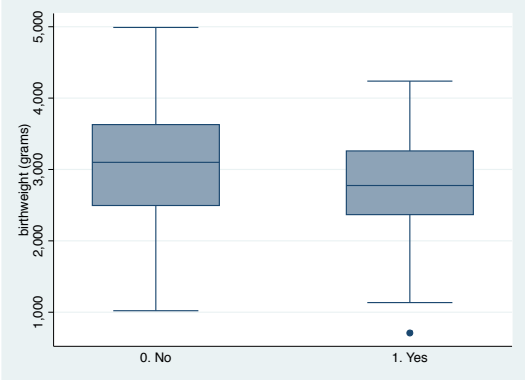


Q-Q plots:

```
. qnorm bwt
```

Boxplots:

```
. graph box bwt, over(smoke)
```



Oneway analysis of variance (comparison of means between two or more groups):

```
. oneway bwt race
```

T-tests to compare means of a normally distributed variable between two groups (under the assumption of equal and unequal variances):

```
. ttest bwt, by(smoke)
. ttest bwt, by(smoke) level(99)
. sdtest bwt, by(smoke)
. ttest bwt, by(smoke) unequal
```

Nonparametric tests:

```
. ranksum bwt, by(smoke)
```

Several other nonparametric tests are available under the menu:
Statistics → Summaries, tables and tests → Nonparametric tests of hypotheses

<u>Risk ratios and Odds Ratios:</u>

Cohort data (without censoring): estimate relative risk and risk difference using **cs** *case_var exp_var*

```
. cs low smoke
```

You can stratify and obtain the Mantel-Haenszel weighted risk-ratio estimates

```
. cs low smoke, by(ht)
```

Case-control data:

```
. cc ht smoke
. cc ht smoke, by(race)
```

tabodds to study the effect of multiple exposure levels in a case-control study

```
. tabodds ht race, or
. tabodds ht race, or base(2)

. tabodds ht race, adjust(smoke)
```

Immediate commands to perform calculations (does not use data in memory)

Treatment	Died	Survived	Total
A	4	10	14
B	7	3	10
Total	11	13	24

Using **tabi**, enter the data for each row separated by \

```
. tabi 4 10 \ 7 3, chi2 exact
```

csi is the immediate form of **ci** and **cci** the immediate form of **cc**

	Exposed	Unexposed
Event	7	12
No event	19	21

```
. csi 7 12 19 21
. cci 7 12 19 21
```

Using Stata as a calculator with **display**

```
. display 2*c(pi)*7
```

To find the p-value from a chi-squared test:

```
. display chi2tail(1, 3.84)
```

Confidence Intervals:

Use the **ci** and **cii** (the immediate form of **ci**) commands to calculate confidence intervals for means, proportions, and rates.

```
. ci bwt, level(99)
. ci smoke, binomial
```

Normal distribution: **cii** *N mean SD*

```
. cii 372 37.58 16.51
```

Binomial distribution: **cii** *N events, binomial*

```
. cii 153 40, binomial
```