PUBH 501 Biostatistics

STATA: ANOVA AND THE KRUSKAL-WALLIS TEST

ANOVA

Overview

- ANOVA (Analysis of variance)
  - Parametric
- Post hoc pairwise comparisons
- Kruskal-Wallis test
  - Non-parametric
- More comparisons

ANOVA

- Test whether the mean of the outcome is the same for all groups
- DV = interval/ratio level variable
- IV = categorical (grouping variable)
- Assumptions:
  - 3 or more groups that are mutually exclusive (meds, interventions, etc)
  - Dependent variable normally distributed
  - Equal variance in all groups
  - Random sample with independent observations

ANOVA

- H_0: there is no difference in means between groups
  - H0: mean1 = mean2 = mean3
- H_1: there is a difference in means between groups
  - H1: mean1 ≠ mean2 ≠ mean3

Example

- Using low birthweight data, look at the difference in birthweight between women of different races: white, black, and other
- Check for normality of birthweight within each group
- Check that the variance is equal between groups
- Run a one-way ANOVA

Normality of birthweight by race

- Tabstat
- Summ
- Histogram
- Box plot

Equality of variance (way1)

```
robvar bwt, by(race)

           |    Summary of birthweight (grams)
      race |        Mean    Std. Dev.       Freq.
-----------+------------------------------------------------------------
     white |     3,103.01    727.87244          96
     black |    2,719.692    638.68388          26
     other |    2,804.015    721.30115          67
-----------+------------------------------------------------------------
     Total |    2,944.286    729.01602         189

W0   =   0.44247899    df(2, 186)      Pr > F = 0.6431163

W50  =   0.46688372    df(2, 186)      Pr > F = 0.62768572

W10  =   0.45222731    df(2, 186)      Pr > F = 0.63690718
```

>> Outliers

Levene's test for homogeneity of variances

- W0: test centered at the mean
- W50: test centered at the median
- W10: test centered at the mean where the top 5% and bottom 5% of observations were trimmed (deleted)

Null is that variances are the same, fail to reject since p > 0.05. We have equal variance

ANOVA command

- There are two options for running a one-way ANOVA
- `anova bwt race`
- `oneway bwt race`
- The –anova- command will run a one-way ANVOA. However, can run more complex (two-way, three-way, MANOVA, ANCOVA…)
- -oneway- command
  - It includes Bartlett's test for equal variances (-anova- does not)
  - And you can include the option, "t" to add a tabulation to your output
- Results will be the same

```
anova bwt race
```

```
           Number of obs =       189    R-squared     =  0.0505
           Root MSE      =   714.169    Adj R-squared =  0.0403

    Source | Partial SS          df          MS         F    Prob>F
-----------+------------------------------------------------------
     Model |  5048361.1           2    2524180.5      4.95  0.0081
           |
      race |  5048361.1           2    2524180.5      4.95  0.0081
           |
  Residual |   94866938         186     510037.3
-----------+------------------------------------------------------
     Total |   99915299         188    531464.35
```

significant at $p<0.05$, we can reject our H_0

- Model is the between sums of squares (sum of all the squared differences between the individual means and the grand mean)
- Residual is the within sums of squares (sum of all the squared differences between the individual data and the group mean within each group)

```
oneway bwt race, t
             |   Summary of birthweight (grams)
        race |        Mean   Std. Dev.       Freq.
-------------+------------------------------------
       white |    3,103.01   727.87244          96
       black |   2,719.692   638.68388          26
       other |   2,804.015   721.30115          67
-------------+------------------------------------
       Total |   2,944.286   729.01602         189

                          Analysis of Variance
    Source                SS         df      MS            F     Prob > F
-----------------------------------------------------------------------------
Between groups         5048361.06      2    2524180.53     4.95      0.0081
 Within groups         94866937.5    186    510037.298
-----------------------------------------------------------------------------
    Total              99915298.6    188    531464.354

Bartlett's test for equal variances:  chi2(2) =   0.6560  Prob>chi2 = 0.720
```

Same result as –anova-. Also has the test for equal variance (way2) (do not reject H_0 of equal variance)

Post hoc pairwise comparisons

- Make group comparisons after running the ANOVA
  - Which group difference is driving the significant results?
- Must use this command right after the –anova- command
- Determine how to make the comparisons
  - We will use the Bonferroni method of adjustment for multiple comparisons
  - Use the –mcompare- option to designate the method of comparison
  - *pveffects* for p-values

```
pwcompare race, mcompare(bon) pveffects
```

```
. pwcompare race, mcompare(bon) pveffects

Margins      : asbalanced
-------------------------------------
             |    Number of
             |  Comparisons
-------------+-----------------------
        race |               3
-------------------------------------


------------------------------------------------------------------------------
|                                                 Bonferroni
|                  Contrast   Std. Err.       t     P>|t|
---------------------------- -------------------------------------------------
race
black vs white   -383.3181   157.8914     -2.43   0.048
other vs white   -298.9955   113.6899     -2.63   0.028
other vs black    84.32262   165.0131      0.51   1.000
------------------------------------------------------------------------------
```

Results show the mean differences between groups and the adjusted p-value

ANOVA results summary

- The one-way ANOVA was significant at $p<0.05$, so we rejected the H_0 that there was no difference in the means between groups
- We followed up with Bonferroni adjusted pairwise comparisons and found that mean birthweight for black women and women of another race was lower than the mean birthweight for white women by 383g and 299g respectively. There were no differences in the mean of birthweight for black women and women of other races.

ANOVA nuances

- anova bwt race if smoke==1
  - Subpopulation using if statement
- pwcompare race, mcompare(bon) pveffects
  - Followed by previous anova command, comparison will be within the subpopulation stated in the ANOVA

- quietly: anova bwt race
  - Runs ANOVA quietly, does not show output
- pwcompare race, mcompare(bon) pveffects
  - Runs post-hoc on quietly run ANOVA

Kruskal-Wallis test

Kruskal-Wallis

- Non-parametric alternative to the ANOVA
- No assumption of normality of dependent variable.
  - But does assume the distribution is the same shape between groups
- Groups must still be mutually exclusive and independent
- Need more than five observations in each group

Example

- We'll look at mother's weight by race
- H_0: the mean ranks of mother weight between groups are equal
- Check for assumptions in the same way we prepared to run the ANOVA

Check assumptions and run test

- Check for normality and equality of distributions
- Run test using

```
kwallis lwt, by(race)
```

- Run post hoc comparisons using kwallis2

```
findit kwallis2  kwallis2
lwt, by(race)
```

- Run post hoc comparisons using dunntest

```
findit dunntest  dunntest
lwt, by(race)
```

*both kwallis2 and dunntest will output the Kruskal-Wallis results

```
. kwallis lwt, by(race)

Kruskal-Wallis equality-of-populations rank test

  +---------------------------------+
  |  race | Obs | Rank Sum |
  |-------+-----+----------|
  | white |  96 |  9752.00 |
  | black |  26 |  3081.50 |
  | other |  67 |  5121.50 |
  +---------------------------------+

chi-squared =    13.909 with 2 d.f.
probability =     0.0010

chi-squared with ties =    13.929 with 2 d.f.
probability =     0.0009
```

Will use the version without ties to assess significance, which we have here

```
. kwallis2 lwt, by(race)

One-way analysis of variance by ranks (Kruskal-Wallis Test)

race      Obs    RankSum   RankMean
-------------------------------------
  1        96    9752.00     101.58
  2        26    3081.50     118.52
  3        67    5121.50      76.44

Chi-squared (uncorrected for ties) =     13.909 with     2 d.f. (p = 0.00095)
Chi-squared (corrected for ties)   =     13.929 with     2 d.f. (p = 0.00094)
```

First half of the -kwallis2- output is the same info as the –kwallis- output. Unadjusted for ties, we see a significant difference between ranking of mother weight by race.

```
 kwallis2 lwt, by(race)

(Adjusted p-value for significance is 0.008333)

Ho: lwt(race==1) = lwt(race==2)
    RankMeans difference =     16.94  Critical value =     28.95
    Prob = 0.080706 (NS)

Ho: lwt(race==1) = lwt(race==3)
    RankMeans difference =     25.14  Critical value =     20.85
    Prob = 0.001943 (S)

Ho: lwt(race==2) = lwt(race==3)
    RankMeans difference =     42.08  Critical value =     30.26
    Prob = 0.000436 (S)
```

The second half of the –kwallis2- output shows the pairwise comparisons. There is not a significant difference between race==1 (white) and race==2 (black). But there is between white and other and black and other.

Kruskal-Wallis results summary

- We reject the H_0, and conclude there is a difference in maternal weight by group.
- Post hoc pairwise comparison shows that there is a difference in mother weight for white women (higher weight) and women of another race, and black women (higher weight) and women of another race.

Tip of the day: formats

- Value formats: adding words to numbers
  - `label define YNFMT 0 “no” 1 “yes”`
- Can also format numbers to be displayed with a certain amount of detail
  - `format bwt2 %4.1f //creates a three-digit variable with one digit after the decimal place`
  - `format bwt %4.0gc //display commas in a numeric variable`
  - `format racenew %-14s //this make race 14 characters long, and justifies it to the left`
- Can format dates (this could be a whole class itself)
  - Use the date() function to transform to a date from original input: Jan 1, 2020 ; 1/1/2020; 1 Jan 2020
  - Because dates aren’t an easy interval variable (how many days are between January 1, 2020 and March 12, 2020?), Stata needs to mark them as dates to do calculations. They are stored in Stata as the number of days since January 1, 1960. You can format them so they look like dates

Note on =, ==, and if

- We've use =, ==, and if several times already
- = versus ==
  - = is often used to set something equal to something else. Creating a variable
    - `gen newvar=oldvar`
    - `gen newvar=.`
  - == is a test of equality. When we create a variable and set it equal if some other condition holds true
    - `gen newvar=. if oldvar==.`
- If is a logical operator, do something *if* the following condition is met. As in above example. Or
  - `anova bwt race if smoke==1`