6.12 Exercise: Inference with iNZight

This exercise will enable you to make comparisons between sub-groups allowing for sampling error.

Background understanding: see Step 6.9

The skills addressed in this Exercise are:

- Use iNZight to get inferential mark-ups of plots so that you can make visual comparisons between sub-groups allowing for sampling error.
- To obtain numerical confidence limits for true between-group differences.

[*iNZight Lite version linked here*]

INSTRUCTIONS

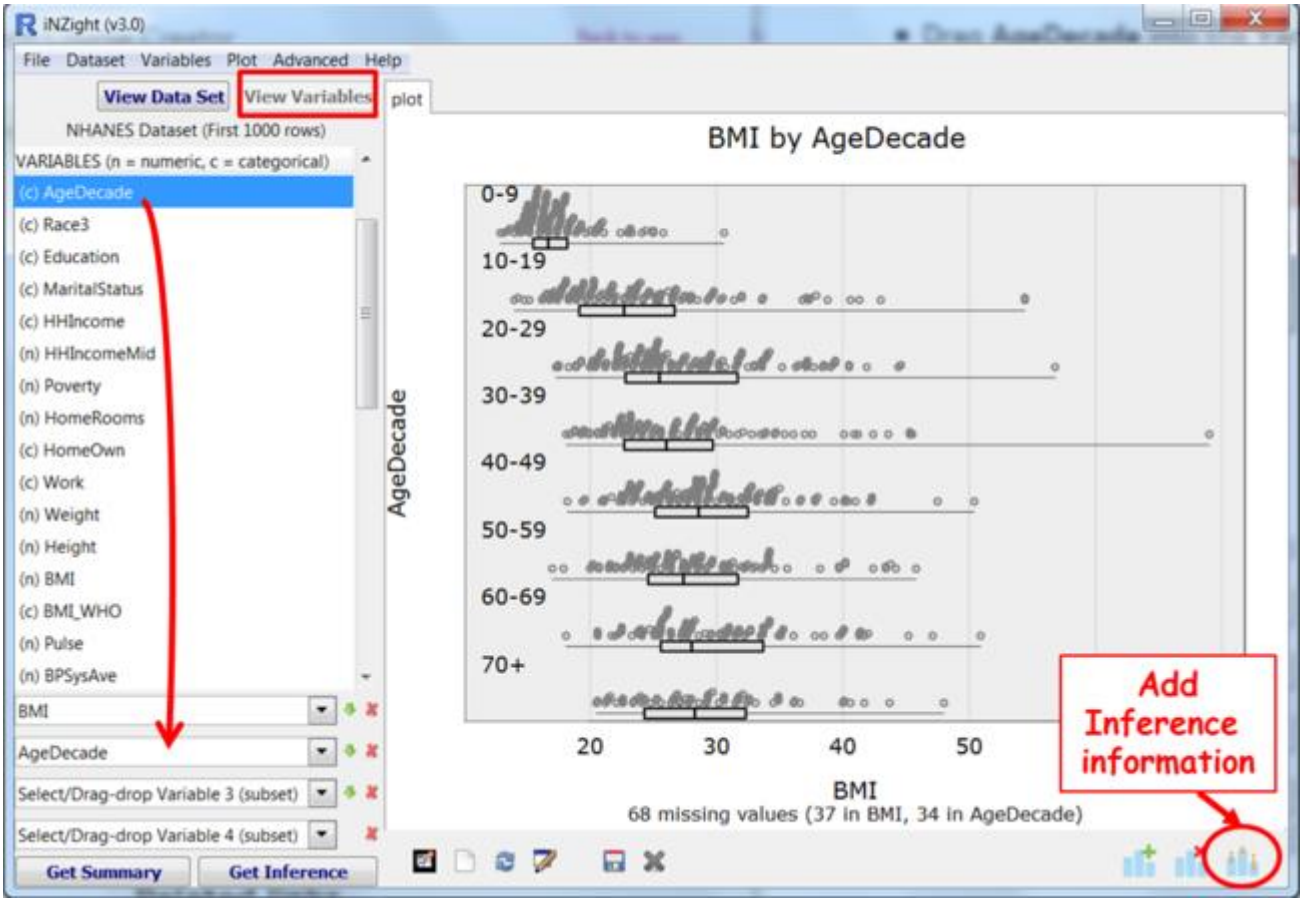
Follow the instructions below to generate the graphs. Or you may prefer to print the instructions. If you have a problem doing the exercise, scroll down to **Common questions**.

Start **iNZight** (not VIT) and load the **nhanes_1000** dataset into iNZight using **File > Example data ….** You will find the data set in **Module (package)** *FutureLearn*.

Construct confidence intervals for sub-groups of a numeric outcome

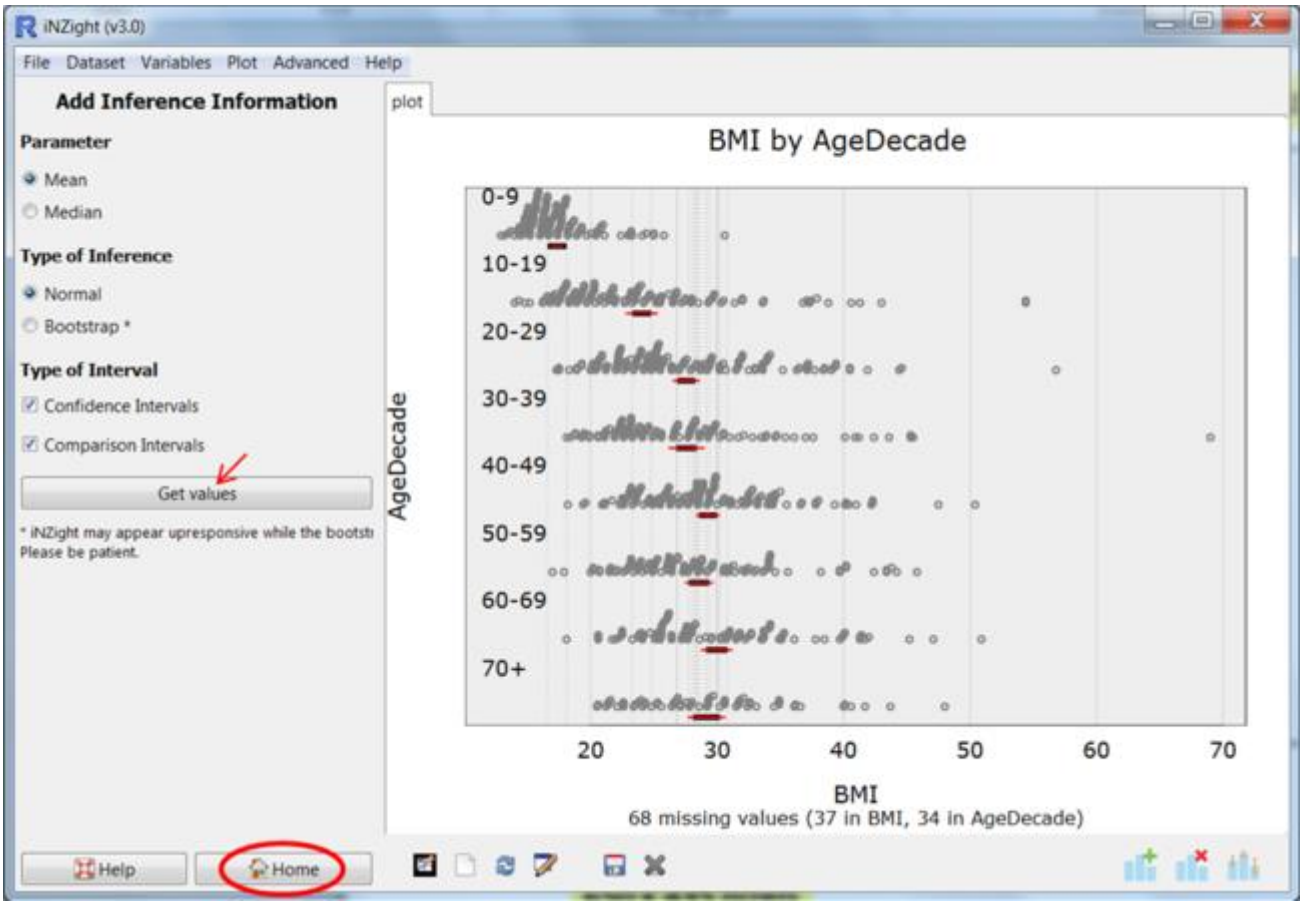
We are revisiting **BMI** (Body Mass Index) for different age groups and using the nhanes-1000 data, a sample of the American population. We will use our sample to estimate the mean BMI for different age groups. To do this we need to construct intervals around our estimates in order to allow for sampling error.

- Drag **BMI** into the **Variable 1** slot
- Drag **AgeDecade** into the **Variable 2** slot.



You should have a series of dot plots of BMI for different age groups in your plot window.

- Click **Add Inference Information** icon (or go *Plot > Add Inference …*)



You can squash your plot window vertically so that it is easier to see how much overlap there is between each age group.

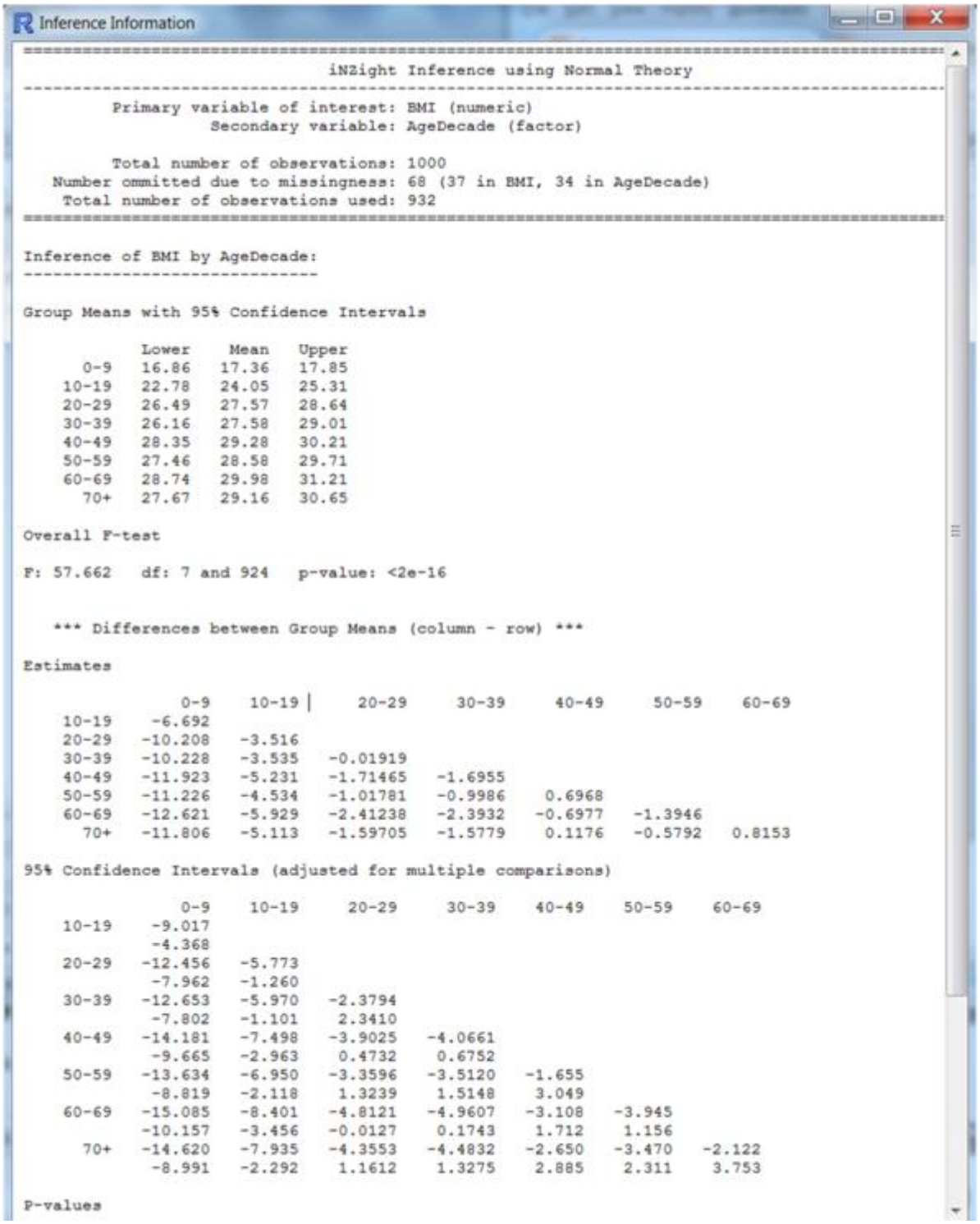
What do you see here? The thick black lines are called 'comparison intervals' and are the lines that we look at when observing any overlap. The thin red lines are the individual confidence intervals for each mean/median. [The endpoints of the lines are available (if you really need them) by clicking *Get values*.]

Post a comment if you have any interesting observations about the estimation of the population means for each age group and the differences between them.

For the actual **confidence intervals for the true differences** in means/medians:

- Return to the **Home** command panel
- Click **Get Inference**
- Select **Normal** inference
- Click **OK**.

The **Summary window** will pop up with the group means estimate and the 95% Confidence Interval around each estimate. You will also see differences between each group mean and their respective confidence intervals as well. There is a lot of extra information in the Summary window. Further study of statistics will help you understand this content.



Inference Information

```
================================================================================
                          iNZight Inference using Normal Theory
--------------------------------------------------------------------------------
         Primary variable of interest: BMI (numeric)
                   Secondary variable: AgeDecade (factor)

         Total number of observations: 1000
   Number ommitted due to missingness: 68 (37 in BMI, 34 in AgeDecade)
    Total number of observations used: 932
================================================================================

Inference of BMI by AgeDecade:
------------------------------

Group Means with 95% Confidence Intervals

           Lower    Mean   Upper
     0-9   16.86   17.36   17.85
   10-19   22.78   24.05   25.31
   20-29   26.49   27.57   28.64
   30-39   26.16   27.58   29.01
   40-49   28.35   29.28   30.21
   50-59   27.46   28.58   29.71
   60-69   28.74   29.98   31.21
     70+   27.67   29.16   30.65

Overall F-test

F: 57.662   df: 7 and 924   p-value: <2e-16


   *** Differences between Group Means (column - row) ***

Estimates

               0-9    10-19 |    20-29     30-39     40-49     50-59    60-69
   10-19    -6.692
   20-29   -10.208   -3.516
   30-39   -10.228   -3.535   -0.01919
   40-49   -11.923   -5.231   -1.71465   -1.6955
   50-59   -11.226   -4.534   -1.01781   -0.9986    0.6968
   60-69   -12.621   -5.929   -2.41238   -2.3932   -0.6977   -1.3946
     70+   -11.806   -5.113   -1.59705   -1.5779    0.1176   -0.5792   0.8153

95% Confidence Intervals (adjusted for multiple comparisons)

               0-9    10-19     20-29     30-39    40-49    50-59    60-69
   10-19    -9.017
            -4.368
   20-29   -12.456   -5.773
            -7.962   -1.260
   30-39   -12.653   -5.970   -2.3794
            -7.802   -1.101    2.3410
   40-49   -14.181   -7.498   -3.9025   -4.0661
            -9.665   -2.963    0.4732    0.6752
   50-59   -13.634   -6.950   -3.3596   -3.5120   -1.655
            -8.819   -2.118    1.3239    1.5148    3.049
   60-69   -15.085   -8.401   -4.8121   -4.9607   -3.108   -3.945
           -10.157   -3.456   -0.0127    0.1743    1.712    1.156
     70+   -14.620   -7.935   -4.3553   -4.4832   -2.650   -3.470   -2.122
            -8.991   -2.292    1.1612    1.3275    2.885    2.311    3.753

P-values
```

EXPLORE (~5 min)
Find another numeric variable in the NHANES-1000 dataset that you might like to explore across age groups, e.g. **Height** and move it to the **Variable 1** slot. Use iNZight to explore the estimates for the American population and build intervals around your estimates.

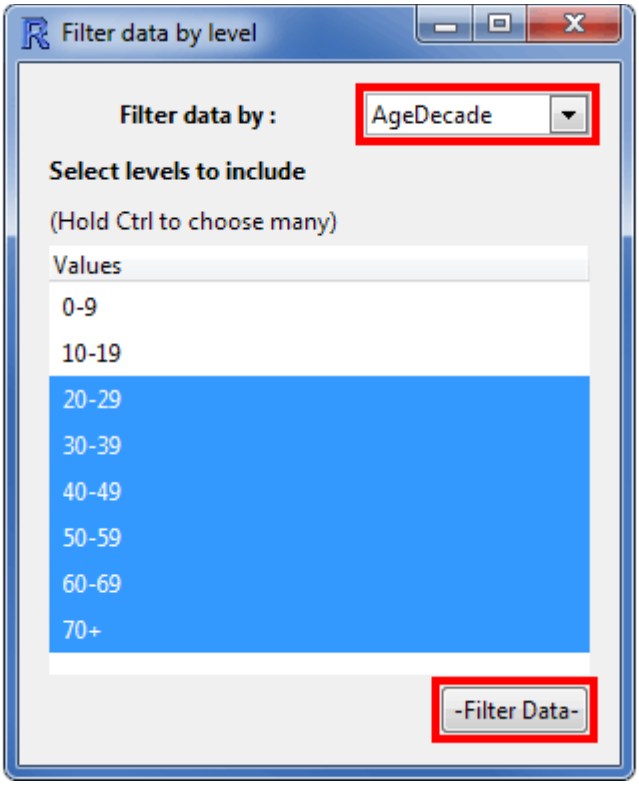
Post a comment if you see anything interesting.

Construct confidence intervals for sub categories of a categorical outcome

Now we'll use the NHANES-1000 dataset to form confidence intervals around estimates of how people in the American population rate their general health at the time of the survey. Is there any difference between age groups?

First, filter out people below the ages of 20 (because "general health" was not recorded for them):

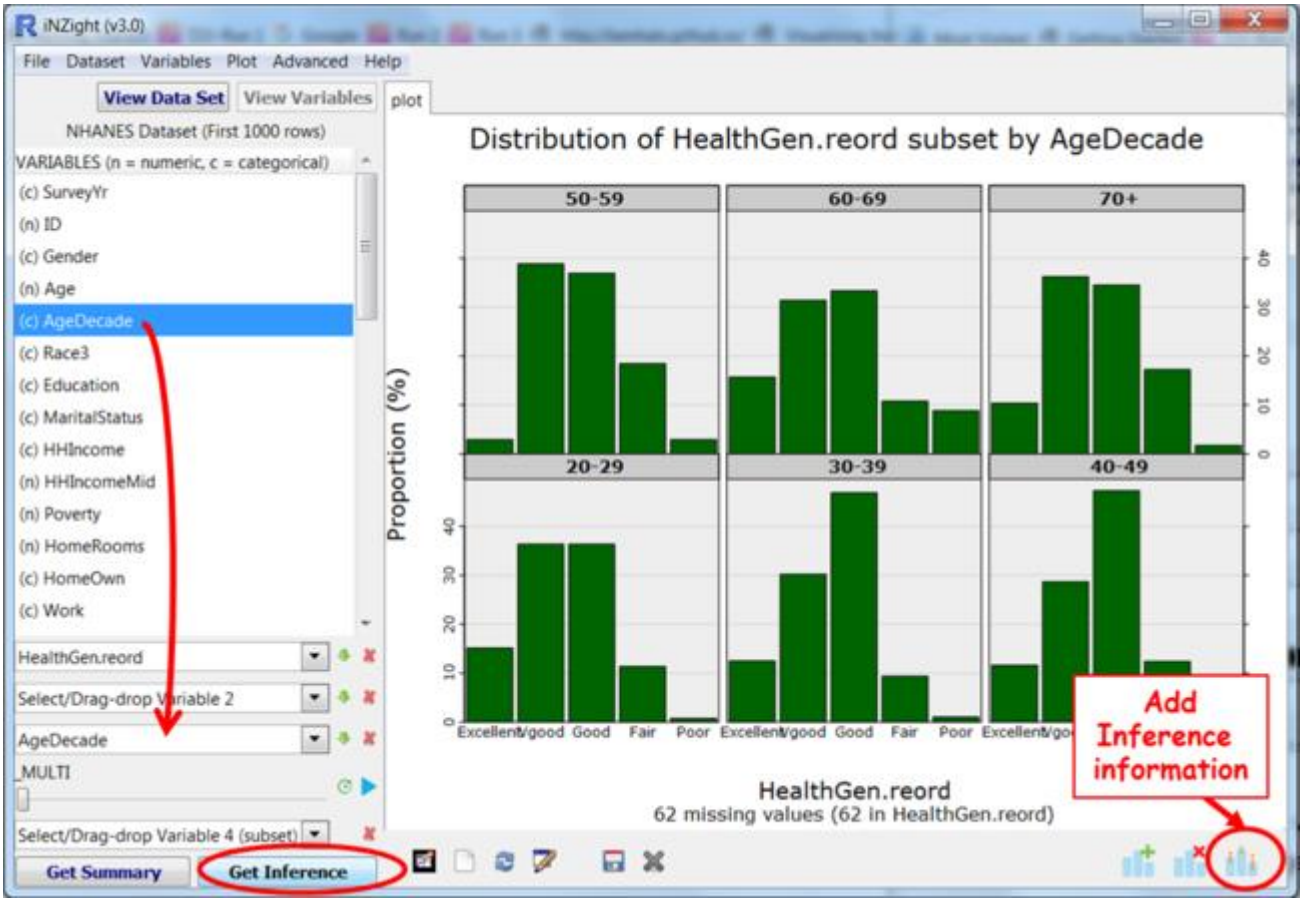
- Click **Dataset > Filter Dataset** and select **levels of a categorical variable**. Then click **Proceed**.
- **Filter data by:** Select **AgeDecade**
- Using the **Ctrl** or **Shift** keys select all of the age groups you want to include (ages 20 and above)
- Click **-Filter Data-**.



Now re-order the values for **HealthGen** into a sensible order (it is currently alphabetic). We will use the order from 1=Excellent to 5=Poor using **Variables > Categorical Variables > Reorder Levels** to produce the new variable **HealthGen.reord**.

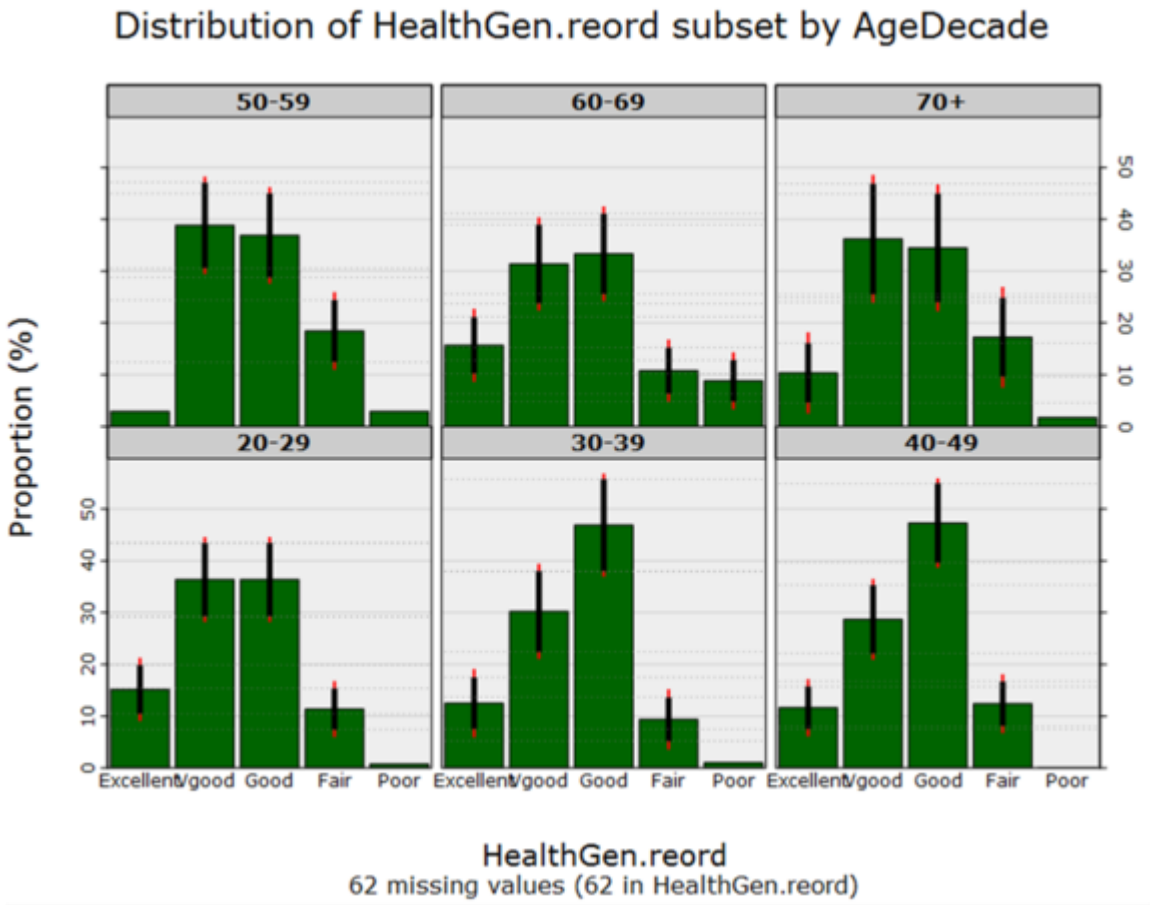
Now construct the series of plots for **HealthGen.reord** and **AgeDecade**:

- Drag **HealthGen.reord** into the **Variable 1** slot
- Drag **AgeDecade** into the **Variable 3 (subset)** slot to create separate bar charts for each age group.



A series of graphs will appear in the plot window. It will help you see the detail if you enlarge your plot window.

- Click **Add Inference Information** icon (or go *Plot > Add Inference …*)



[Again, the endpoints of the lines are available (if you really need them) by clicking *Get values*.]

What does this graph tell you? What can you infer about the true differences between the percentages in each health category?

For the actual **confidence intervals for the true differences** in the proportions, go back to the **Home** command panel and click the **Get Inference** button.

Now construct side-by-side bar charts for **HealthGen.reord** by **AgeDecade**

- use the little **green** "switch" **arrow** by slot 2 so that **HealthGen.reord** remains in the **Variable 1** slot and **AgeDecade** is switched into the **Variable 2** slot.
- Click the **Add Inference Information** icon to add the intervals.

What does this graph tell you about age-differences in the percentages for each health category (e.g. differences between age groups in the percentages saying they are in very good health) ?

Again, if you want the actual **confidence intervals for the true differences** in the proportions, go back to the **Home** command panel and click the **Get Inference** button.