# Running an Analysis of Variance

(Data Analysis Tools Week 1 Assignment)

## Expected Activities

- Run an analysis of variance.
- Need to analyze and interpret post hoc paired comparisons in instances where original statistical test was significant, and were examining more than two groups (i.e. more than two levels of a categorical, explanatory variable).
- Submit syntax used to run an ANOVA (copied and pasted from the program) along with corresponding output and a few sentences of interpretation.

## SAS Program

```
LIBNAME mydata "/courses/d1406ae5ba27fe300 " ACCESS=readonly;

DATA new;
      SET mydata.gapminder;
      KEEP country urbanrate lifeexpectancy urban;
      LABEL lifeexpectancy="Life Expectancy";
      LABEL urbanrate="Urbanisation Rate";
      LABEL urban="Urbanisation Groups";

      /* Delete records with missing data */
      IF urbanrate=. THEN
            delete;
      IF lifeexpectancy=. THEN
            delete;

      /* Data Management for variable urbanrate */
      IF urbanrate < 25 THEN
            urban="UR Group 1";
      IF urbanrate >=25 AND urbanrate < 50 THEN
            urban="UR Group 2";
      IF urbanrate >=50 AND urbanrate < 75 THEN
            urban="UR Group 3";
      IF urbanrate >=75 THEN
            urban="UR Group 4";

      /* Data Preparation or Management for variable lifeexpectancy */
PROC SORT;
      BY country;
```

```
PROC ANOVA;
      CLASS urban;
      MODEL lifeexpectancy=urban;
      MEANS urban;

PROC ANOVA;
      CLASS urban;
      MODEL lifeexpectancy=urban;
      MEANS urban/DUNCAN;

RUN;
```

# Output

The ANOVA Procedure

| Class Level Information | | |
|---|---|---|
| Class | Levels | Values |
| urban | 4 | UR Group 1 UR Group 2 UR Group 3 UR Group 4 |

| | |
|---|---|
| Number of Observations Read | 188 |
| Number of Observations Used | 188 |

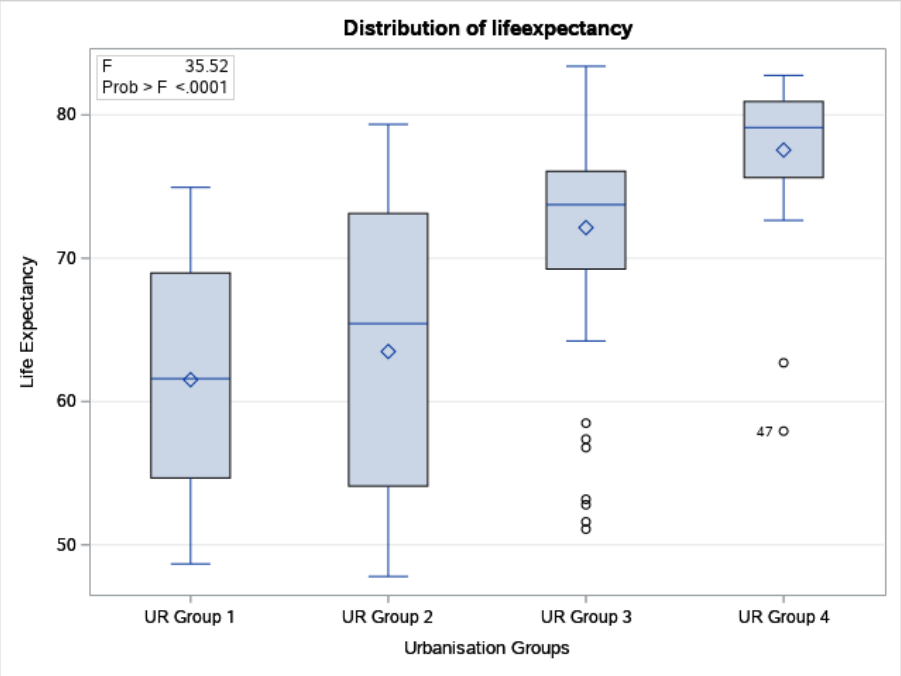
The ANOVA Procedure

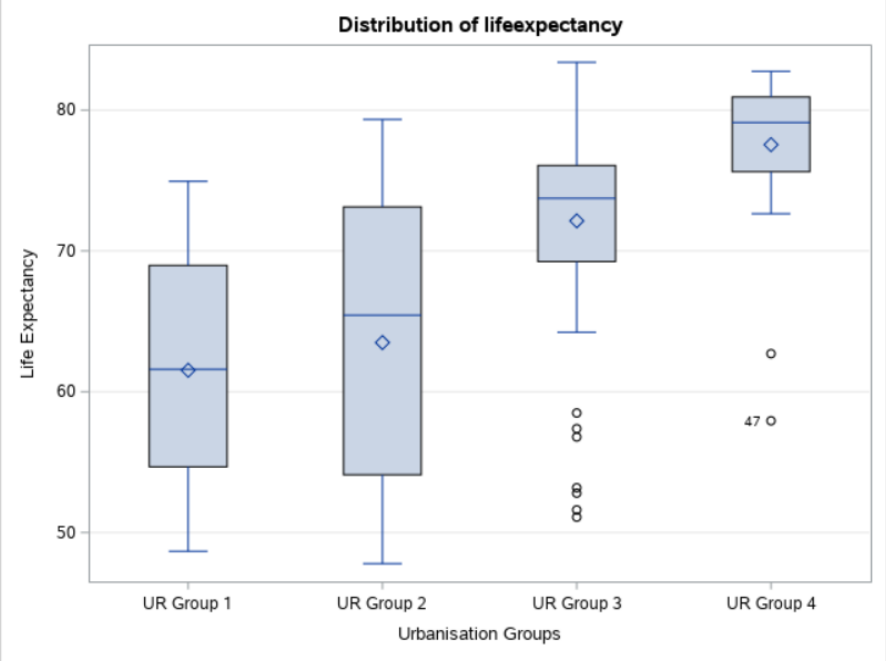
Dependent Variable: lifeexpectancy Life Expectancy

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 6463.63503 | 2154.54501 | 35.52 | <.0001 |
| Error | 184 | 11161.27993 | 60.65913 | | |
| Corrected Total | 187 | 17624.91496 | | | |

| R-Square | Coeff Var | Root MSE | lifeexpectancy Mean |
|---|---|---|---|
| 0.366733 | 11.19011 | 7.788397 | 69.60070 |

| Source | DF | Anova SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| urban | 3 | 6463.635034 | 2154.545011 | 35.52 | <.0001 |




| Level of urban | N | lifeexpectancy | |
|---|---|---|---|
| | | Mean | Std Dev |
| UR Group 1 | 21 | 61.5272857 | 8.48159754 |
| UR Group 2 | 55 | 63.4922000 | 9.77512626 |
| UR Group 3 | 71 | 72.1333662 | 7.17565443 |
| UR Group 4 | 41 | 77.5443659 | 4.89479774 |

**Post hoc comparisons**

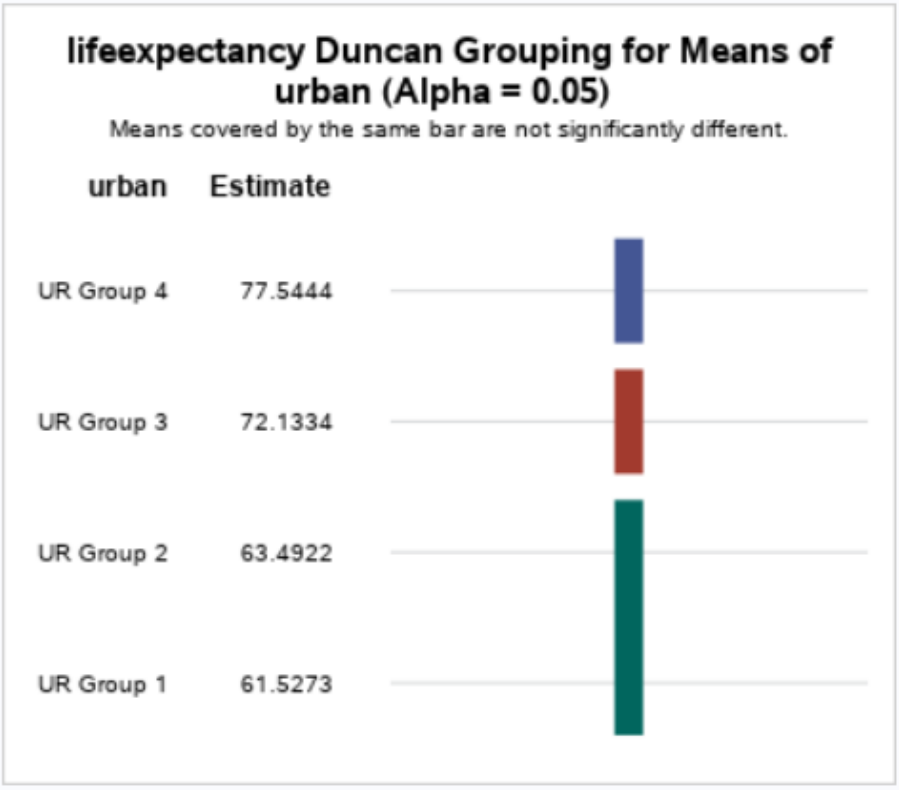
Duncan's Multiple Range Test for lifeexpectancy

**Note:** This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

| | |
|---|---|
| Alpha | 0.05 |
| Error Degrees of Freedom | 184 |
| Error Mean Square | 60.65913 |
| Harmonic Mean of Cell Sizes | 38.35988 |

**Note:** Cell sizes are not equal.

| Number of Means | 2 | 3 | 4 |
|---|---|---|---|
| Critical Range | 3.509 | 3.693 | 3.816 |



Model Interpretation for ANOVA

When examining the association between Life Expectancy (quantitative response) and Urbanisation Rate (categorical explanatory), an Analysis of Variance (ANOVA) revealed that countries with the highest Urbanisation Rate have higher Life Expectancy (Mean= 77.5443659, Stdev.= 4.89479774) compared to those with the lower Urbanisation Rates.

| Level of urban | N | lifeexpectancy | |
|---|---|---|---|
| | | Mean | Std Dev |
| UR Group 1 | 21 | 61.5272857 | 8.48159754 |
| UR Group 2 | 55 | 63.4922000 | 9.77512626 |
| UR Group 3 | 71 | 72.1333662 | 7.17565443 |
| UR Group 4 | 41 | 77.5443659 | 4.89479774 |

This information proves that there is a relationship between Life Expectancy and Urbanisation Rate.

Model Interpretation for Post Hoc ANOVA Results

ANOVA revealed that Urbanisation Rate (collapsed into 4 ordered categories, which is the categorical explanatory variable) and Life Expectancy (the quantitative response variable) were significantly associated, $F(3,184)=35.21$, $p<0.0001$.Post hoc comparisons of mean number of Life Expectancy revealed that the countries with Urbanization Rate higher than 75% reported significantly higher Life Expectancy compared to countries with urbanization rate between "50-75%", between "25-50%"and “lower than 25%". However, countries with employment rate between "25-50%" and countries with employment rate lower than "25%" had similar Life Expectancy.