ANOVA

ANA 500 – Foundations of Data Analytics

Module 2 - week 4C

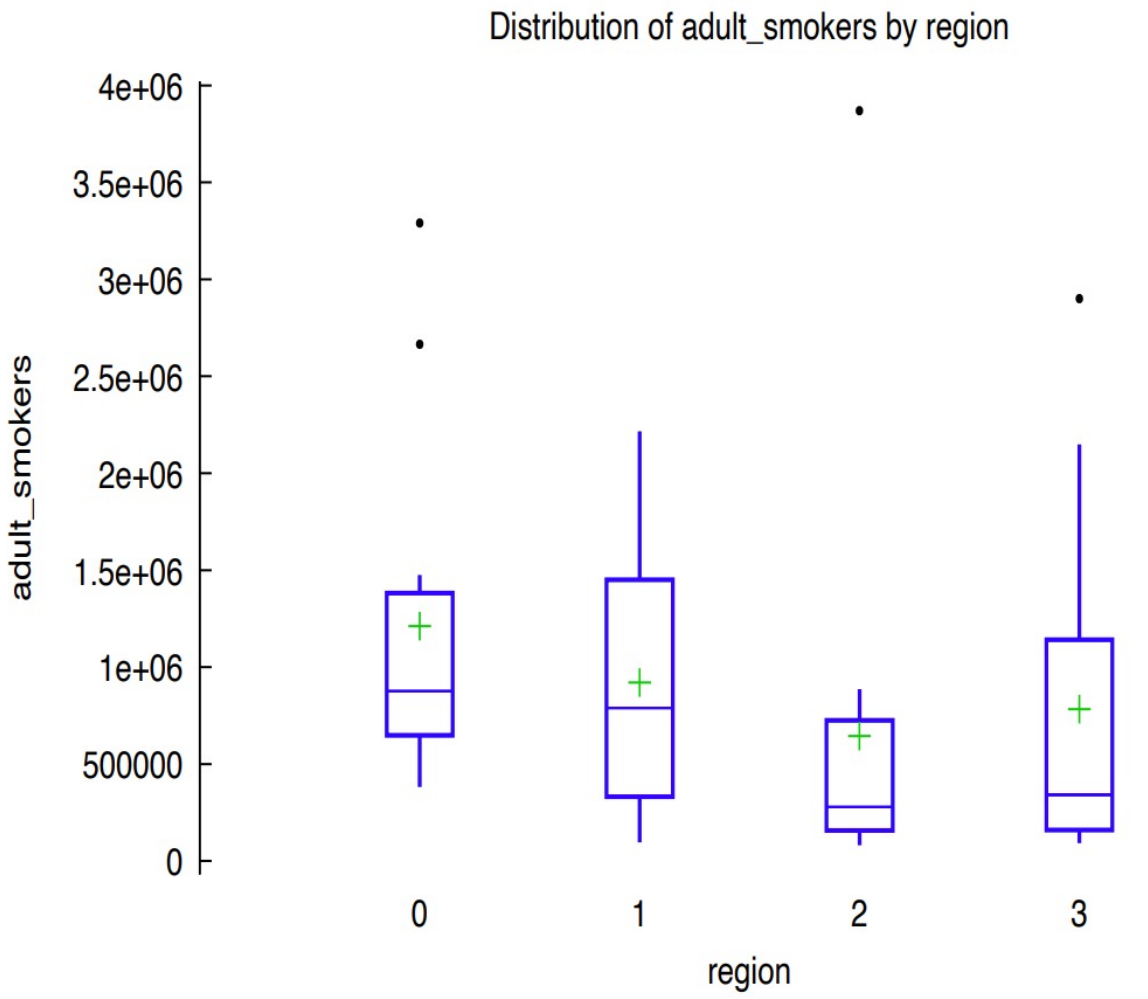
ANOVA

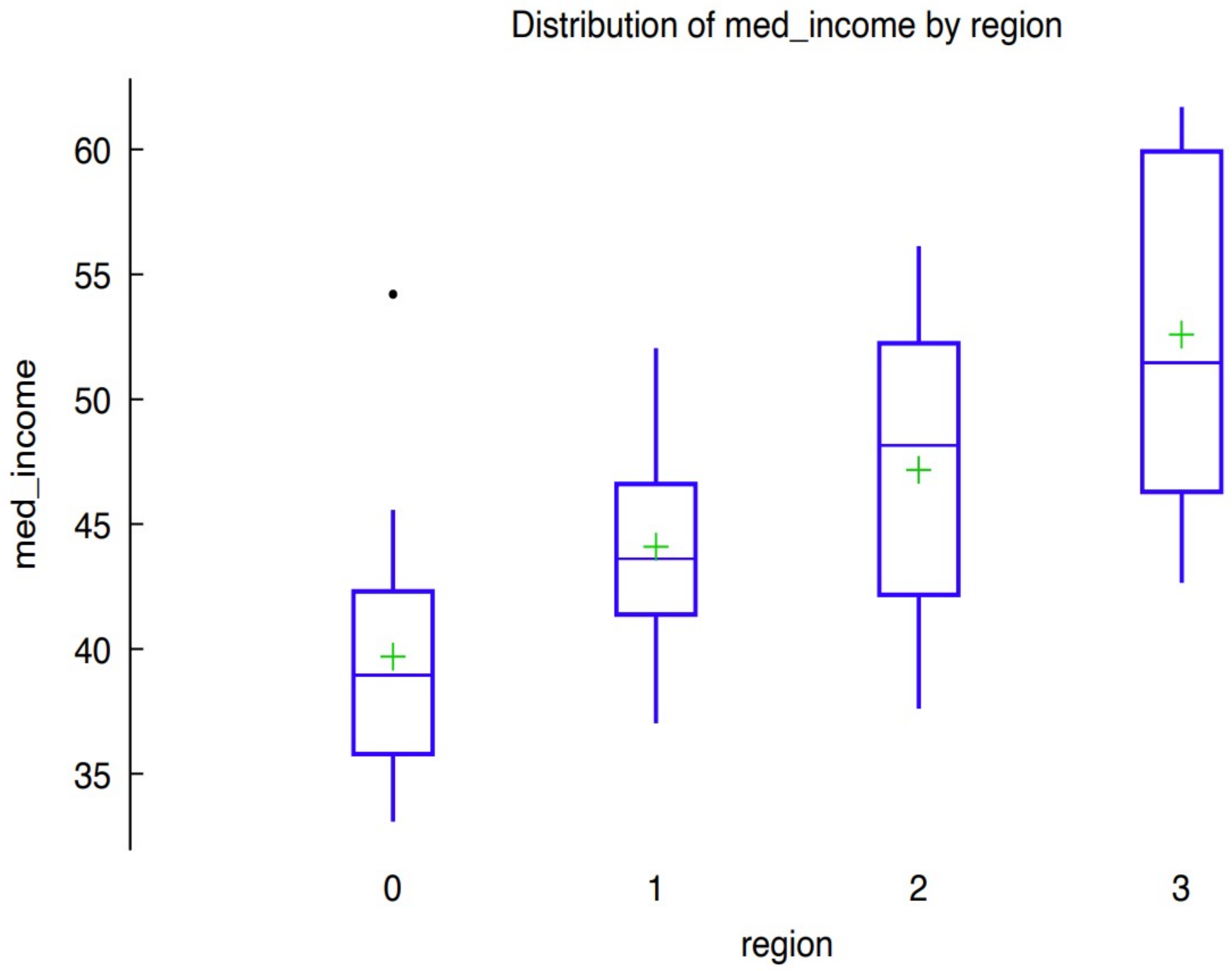
- **An**alysis **o**f **Va**riance
- Recall using a t-test to examine differences in means between two groups.
- We will now extend this to the case of k groups (more than 2 groups).
- The entire variation in the outcome of interest will be decomposed into separate components.
- Examples:
  - Is there a difference in average income across race?
  - Is there a difference in average MPG (miles per gallon) across different car types?
  - Is there a difference in GPA between 1st year, 2nd year, junior, and senior students?
- We will focus on the simple case of one-way ANOVA

ANOVA

- *Variances* are used to determine if *means* differ across groups.
- **Assumptions:**
  - Each population from which the sample is taken is normal
  - All samples are random and independent
  - Populations have equal variances
  - Each factor is categorical (e.g. profession, restaurant type)
  - Each response (outcome of interest) is numerical (e.g. income, revenue)
- H_o: all means are equal
- H_a: at least two means differ

ANOVA


Distribution of adult_smokers by region
adult_smokers
4e+06
3.5e+06
3e+06
2.5e+06
2e+06
1.5e+06
1e+06
500000
0
0
1
2
3
region


Distribution of med_income by region
med_income
60
55
50
45
40
35
0
1
2
3
region

ANOVA

- ANOVA uses the F-distribution
  - Derived from the Student’s t-distribution
- F-statistic is a ratio with numerator df and denominator df
- Variance between samples
  - An estimate of overall variance
  - Variance of the sample means from the overall mean
- Variance within samples
  - An estimate of overall variance
  - Variance of observations within a category from that category’s mean

ANOVA

- Sum of squares total (SST) = $\sum_i (xi - \bar{x})^2$
- Sum of squares between (SSB) = $\sum_k n_k (\bar{x}_k - \bar{x})^2$
  - $\bar{x}_k$ is each group mean; $\bar{x}$ is overall mean
  - Often called the explained or model sum of squares
- Sum of squares within (SSW) = $\sum_k (nk - 1)(sk^2)$
  - n_k is the number of observations in each group
  - ${s_k}^2$the variances within each group
  - Often called the sum of squares due to error (SSE)
- SST = SSB + SSW

ANOVA

- Mean squared within (MSW) = SSW/dfw = SSW/$n - k$
  - $df = n - k$ (the number of observations minus the number of categories)
  - Often denoted MSE for mean squared error

- Mean squared between (MSB) = SSB/dfw = SSB/$k - 1$
  - $df = k - 1$ (the number of categories minus one)

- F-test is all about comparing differences between groups relative to differences within groups.

$$ratio = \frac{Sum \text{ of squares between groups}}{\text{Sum of squares within groups}}$$

ANOVA

- MSB can be influenced by differences in population means among the different groups.
- MSW is **not** influenced by differences in population means among the different groups.
- Ho: populations all have the same normal distribution
  - Remember, we assume equal variances and normality, so if means are equal, the normal distributions for each group are the same.
  - If Ho is true, MSB and MSW should be about the same
- F-stat = MSB/MSW
  - If Ho is true, the F-stat ≈ 1

ANOVA

- E.g. Is there a difference in mean sales between McDonalds, Burger King, and Wendy's?
  - Suppose we have the following random sample of data on annual sales

McDonalds	Burger King	Wendy's
4.2	1.8	1.1
2.3	1.4	1.3
2.8	2.1	1.4
4.0	1.7	1.1
3.3	1.4	2.1
1.9	1.9	1.8
3.5	2.0	1.5
2.7	2.2	1.0

ANOVA

McDonalds	Burger King	Wendy's
4.2	1.8	1.1
2.3	1.4	1.3
2.8	2.1	1.4
4.0	1.7	1.1
3.3	1.4	2.1
1.9	1.9	1.8
3.5	2.0	1.5
2.7	2.2	1.0

- $\bar{x}$ =2.104, $\bar{x}_{\text{mcDon}}$=3.09, $\bar{x}_{\text{BK}}$=1.81, $\bar{x}_{\text{Wendy's}}$=1.41
- SST = 18.43 (SST) = $\sum_i (xi - \bar{x})^2$
- SSB = (8*(3.09 – 2.104)2) + (8*(1.81 – 2.104)2) + (8*(1.41 – 2.104)2) = 12.32
- SSW = SST – SSB = 18.43 – 12.32 = 6.11
- MSB = 12.32 / (3 – 1) = 6.16 MSW = 6.11 / (24 – 3) = 0.291
- F-stat = 6.16 / 0.291 = 21.17
- Numerator df = 2, denominator df = 21
- p-value = 0.000009
- 0.000009 < 0.05 --- reject Ho, there is a difference in mean sales between the three restaurants.

ANOVA