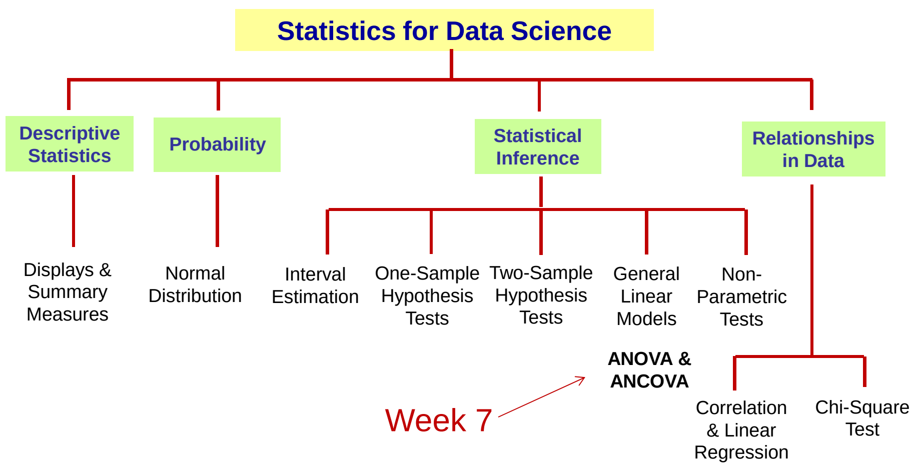
MATH 4044
Statistics for Data Science

Comparing Several Means
ANOVA

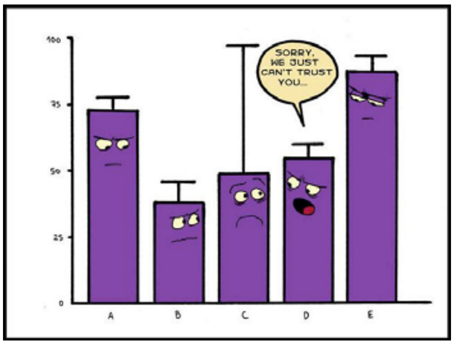
Course outline



Field, A & Miles, J, *Discovering Statistics Using SAS*, Chapter 10 (10.1-10.4, 10.6-10.7).

Topics to be covered

- Comparing several means:
  - Analysis of Variance (ANOVA)
  - Checking assumptions
  - Planned contrasts
  - Post-hoc tests



Multiple comparisons

- How do we make many comparisons at once with an overall measure of confidence in all our conclusions?
  - We can’t simply compare two parameters at a time.
- Usually done in two steps:
  - Step 1 is an overall test if there is good evidence of any differences among the parameters we want to compare.
  - Step 2 is a detailed follow up analysis to decide which of the parameters differ and to estimate how large the differences are.

Why not use lots of *t*-tests?

- Suppose there are three groups and we are interested in differences between these groups.
- If we were to carry out three separate t-tests at 5% significance level:
  - The probability of falsely rejecting the null hypothesis (Type I error) is 5% for each test.
    - The probability of avoiding a Type I error is therefore 95%.
  - If these tests are independent, the overall probability of no Type I errors is $(0.95)^3 = 0.857$.
  - The overall probability of at least one Type I error is then $1 – 0.857 = 0.143$ or $14.3\% > 5\%$.
- In general, experimentwise error rate is $1 – (0.95)^n$.

The **AN**alysis **O**f **VA**riance (ANOVA)

- We want to test the null hypothesis that there are *no differences* in mean response among groups:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$$

  All population means are equal

- The alternative hypothesis is that there is *a difference*.

  H_1: Not all population means are equal

- Comparing several means is the simplest form of ANOVA, called one-way ANOVA.

One-way ANOVA test statistic

$$F_{k-1,N-k} = \frac{\text{variation among the sample means}}{\text{variation among individuals in the same sample}}$$

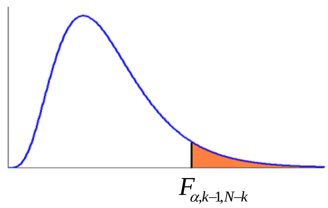
Between-group variance

Within-group variance

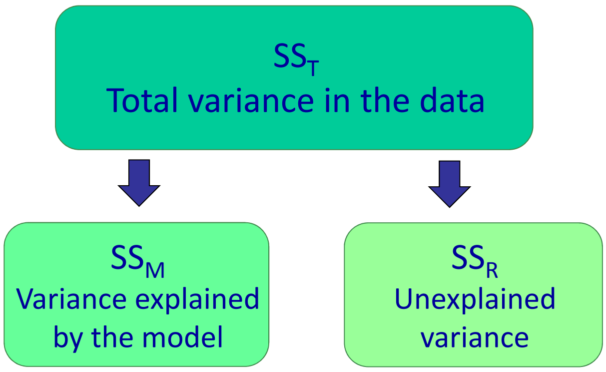
- If the null hypothesis that all k population means are equal is true, the ANOVA F statistic has the F distribution with $k-1$ degrees of freedom in the numerator and $N-k$ degrees of freedom in the denominator.

$$N = n_1 + n_2 + \ldots + n_k$$

$k =$ number of population groups



Theory of ANOVA



- If the experiment is successful, then the model will explain more variance than it can't:
  - SS_M will be greater than SS_R.

Conditions for applying ANOVA

- We have *k* independent samples.
- Each of the *k* populations has a Normal distribution with an unknown mean.
- The means may be different in the different populations.
- All of the populations have the same standard deviation σ, whose value is unknown.
- The results of the ANOVA *F* test are approximately correct when the largest sample standard deviation is *no more than twice as large* as the smallest sample standard deviation.

Example: Electronics sales

- The data set `store` contains the following variables:

Variable name	Description
Region	Region of the country (North, East, South, West)
Advertising	Advertising (Yes or No)
Gender	Gender of shopper (M or F)
Book_Sales	Amount spent on books
Music_Sales	Amount spent on music
Electronics_Sales	Amount spent ion electronics
Total_Sales	Total sales

Example: Electronics sales

- Suppose we want to determine whether the mean of electronics sales varies by region of the country.
- We will check the assumptions and then conduct one way ANOVA using PROC GLM (General Linear Model).

Example: Electronics sales

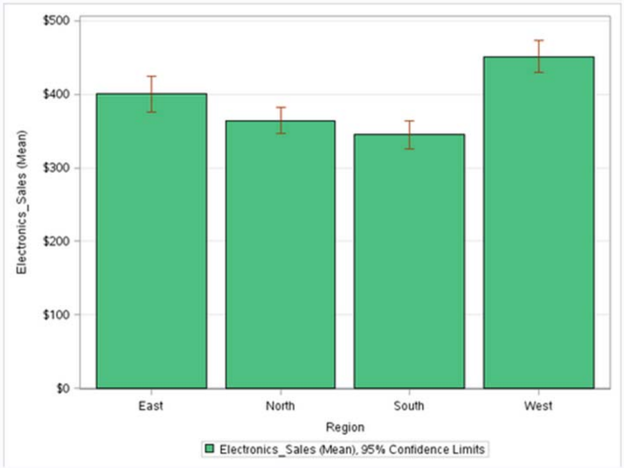
Descriptive Statistics

The MEANS Procedure

Analysis Variable : Electronics_Sales

Region	N Obs	N	Mean	Median	Std Dev	Lower Quartile	Upper Quartile	Minimum	Maximum
East	36	36	400.556	405.000	72.779	340.000	450.000	270.000	570.000
North	69	69	364.783	360.000	74.648	310.000	410.000	220.000	550.000
South	45	45	345.111	330.000	64.407	300.000	380.000	250.000	510.000
West	50	50	451.800	455.000	76.390	400.000	510.000	270.000	610.000

Example: Electronics sales



Sales in the West appear to be higher on average than in any other region.

Example: Electronics sales

- Assumptions:
  - Random and independent samples:
    - This requirement is assumed to be satisfied.
  - All four populations should be Normal:
    - We need to test samples for Normality.
  - All population standard deviations should be equal:
    - We will test for equality of variances.

Example: Electronics sales

North

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.985269	Pr < W	0.5941
Kolmogorov-Smirnov	D	0.071408	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.034502	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.244651	Pr > A-Sq	>0.2500

East

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.973376	Pr < W	0.5247
Kolmogorov-Smirnov	D	0.09497	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.0501	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.332182	Pr > A-Sq	>0.2500

South

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.95186	Pr < W	0.0598
Kolmogorov-Smirnov	D	0.12608	Pr > D	0.0727
Cramer-von Mises	W-Sq	0.103727	Pr > W-Sq	0.0980
Anderson-Darling	A-Sq	0.657236	Pr > A-Sq	0.0843

West

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.984083	Pr < W	0.7316
Kolmogorov-Smirnov	D	0.087876	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.05349	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.313183	Pr > A-Sq	>0.2500

Example: Electronics sales

- P-values for all Normality tests are greater than 0.05, which suggests that all four samples can be assumed to have come from Normal populations.
- Sample standard deviations appear to be quite similar.
  - Equality of variances will be tested formally later
- Therefore, we have all the requirements for one-way ANOVA.

Example: Electronics sales

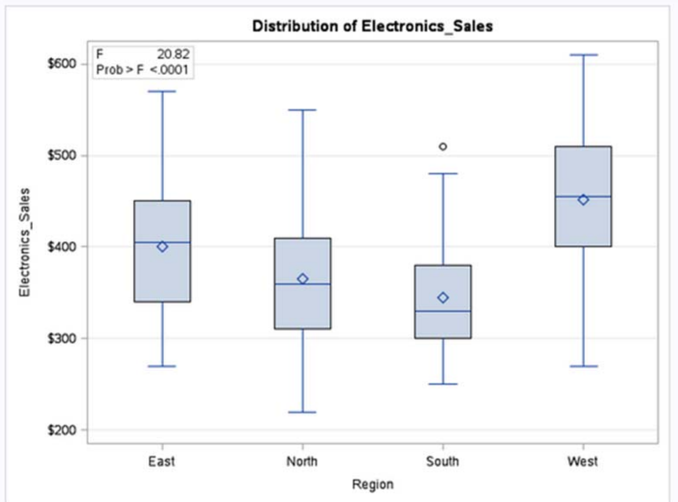
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	329106.428	109702.143	20.82	<.0001
Error	196	1032773.072	5269.250		
Corrected Total	199	1361879.500			

R-Square	Coeff Var	Root MSE	Electronics_Sales Mean
0.241656	18.68218	72.58960	388.5500

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Region	3	329106.4275	109702.1425	20.82	<.0001

There was a significant difference in mean electronics sales levels among the four regions, $F(3,196) = 20.82$, P-value < 0.0001.

Example: Electronics sales



GLM procedure also produces boxplots

Example: Electronics sales

Levene's Test for Homogeneity of Electronics_Sales Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Region	3	78036410	26012137	0.59	0.6199
Error	196	8.5892E9	43822583		

Since the P-value = 0.6199 > 0.05, the assumption of equal variances cannot be rejected.

Welch's ANOVA for Electronics_Sales			
Source	DF	F Value	Pr > F
Region	3.0000	20.46	<.0001
Error	99.4478		

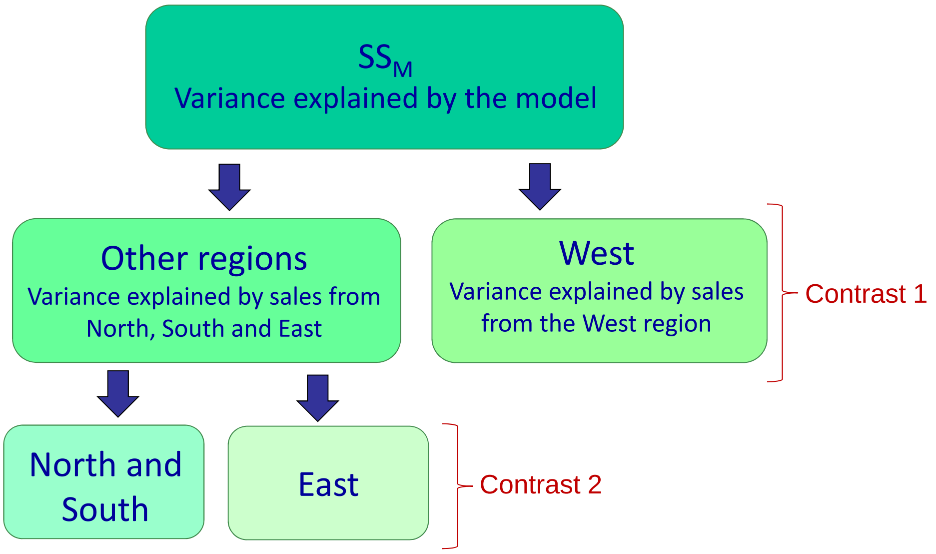
Correction for departures from homogeneity of variance.

We should look and report Welch's F-ratio instead of the one in the main table when the assumption of homogeneity of variance has been violated. We didn't actually need it for our data.

ANOVA: Follow-up tests

- The *F*-ratio tells us only that group means were different:
  - It does not tell us specifically which group means differ.
  - We need additional tests to find out where the group differences lie.
- Multiple *t*-tests:
  - We saw earlier that this is a bad idea.
- Orthogonal contrasts/comparisons:
  - Hypothesis driven, planned a priori.
- Post-hoc tests:
  - No hypothesis, all pairs of means are compared.

Partitioning variance



Defining contrasts using weights

- Rule 1: Choose sensible comparisons. If a group is singled out in one comparisons, it should be excluded from any subsequent contrasts.
- Rule 2: Groups coded with positive weights will be compared against those with negative weights.
- Rule 3: The sum of weights for a comparison should be zero.
- Rule 4: If a group is not involved in a comparison, it is assigned a weight of zero.
- Rule 5: For a given contrast, weights assigned to groups in one chunk of variance should be equivalent to the number of groups in the opposite chunk of variation.

Example: Electronics sales

- Contrast 1: West versus East, North and South
  - Required weights are -1, -1, -1, 3.
- Contrast 2: East versus North and South
  - Required weights are 2, -1, -1, 0.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
west vs other regions	244.950725	35.8928061	6.82	<.0001
east vs north and south	91.217391	27.9093652	3.27	0.0013

Planned comparisons revealed that electronics sales in the West region were significantly higher compared to the other regions, t(196) = 6.82, P-value < 0.0001, and that sales in the East region were significantly higher than in the North and South, t(196) = 3.27, P-value = 0.0013.

Post-hoc procedures

- Used when there is no specific a priori prediction about the data.
- Post-hoc tests consist of pairwise comparisons that are designed to compare all different combinations of treatment groups.
  - Experimentwise error is controlled by correcting the level of significance for each test such that the overall Type I error rate α across all comparisons remains at 0.05.
  - There may be loss of statistical power, i.e. the probability of rejecting an effect that actually exists (Type II error) may be increased.

Post-hoc procedures – points to consider

- Does the test control the Type I error rate?
  - Bonferroni (has more statistical power when the number of comparisons is small);
  - Tukey (more power when comparing a large number of means).
- Does the test also control the Type II error rate?
  - Ryan, Einor, Gabriel and Welsch Q (REGWQ) procedure (should not be used when group sizes are different).
- Is the test reliable when assumptions of ANOVA have been violated?
  - Most tests cope well with deviations from Normality, not so well when group sizes are unequal and when population variances are not equal.

Tukey's Studentized Range (HSD) Test

Comparisons significant at the 0.05 level are indicated by ***.

Region Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
West - East	51.24	10.13	92.36	***
West - North	87.02	52.08	121.95	***
West - South	106.69	68.04	145.34	***
East - West	-51.24	-92.36	-10.13	***
East - North	35.77	-2.90	74.44	
East - South	55.44	13.39	97.50	***
North - West	-87.02	-121.95	-52.08	***
North - East	-35.77	-74.44	2.90	
North - South	19.67	-16.37	55.71	
South - West	-106.69	-145.34	-68.04	***
South - East	-55.44	-97.50	-13.39	***
South - North	-19.67	-55.71	16.37	

Alpha	0.05
Error Degrees of Freedom	196
Error Mean Square	5269.25
Critical Value of Studentized Range	3.66452

The West region is significantly different from East, North and South.

East is significantly different from South.

Note that corresponding confidence intervals do not contain zero.

Example: Tukey-Kramer test

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Region	Electronics_Sales LSMEAN	LSMEAN Number
East	400.555556	1
North	364.782609	2
South	345.111111	3
West	451.800000	4

Least Squares Means for effect Region
Pr > |t| for H0: LSMean(i)=LSMean(j)
Dependent Variable: Electronics_Sales

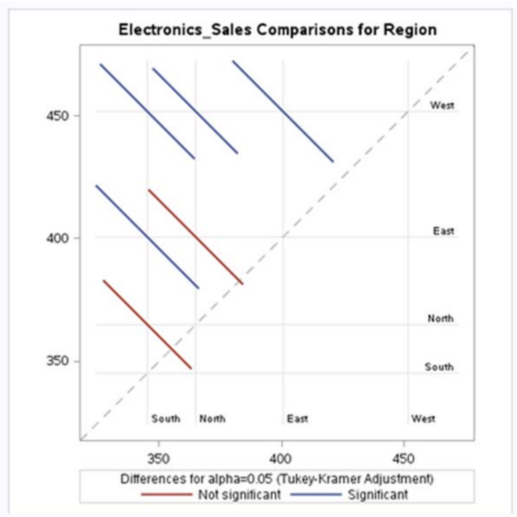
i/j	1	2	3	4
1		0.0810	0.0043	0.0079
2	0.0810		0.4919	<.0001
3	0.0043	0.4919		<.0001
4	0.0079	<.0001	<.0001	

In one-way ANOVA, least squares means are the same as 'ordinary' means.

Using P-values:
For the difference between East and North, P-value = 0.081 > 0.05 so these two means are not significantly different.

For the difference between South and West, P-value < 0.0001 so these two means are significantly different.

Example: Diffogram



Regions are listed on the x- and y-axes to form a matrix.

At the intersection of any two regions you see the difference between them.

The dashed diagonal line represents a difference of zero.

Red and blue lines correspond to confidence intervals for the differences.

If any of the confidence intervals crosses the main diagonal, those two means are not significantly different.

Example: SAS code

```
proc glm data=store plots=diagnostics;
    class Region;
    /* ANOVA */
    model Electronics_Sales=Region / solution;
    /* Planned contrasts */
    estimate 'west vs other regions' Region -1 -1 -1 3;
    estimate 'east vs north and south' Region 2 -1 -1 0;
    /* Equality of variance test, Welch's corrected
       F-test and post-hoc comparison using Tukey test */
    means Region / hovtest Welch Tukey;
    /* Diffogram of post-hoc comparisons using Tukey-
       Kramer method */
    lsmeans Region / pdiff adjust=Tukey;
run;
quit;
```


ANOVA as a General Linear Model (GLM)

- Is there a relationship between a numerical variable (measurement of interest) and a categorical variable (group membership)?
- Recall from linear regression:

outcome = (model) + error

$$\hat{y} = b_0 + \underbrace{b_1 x_1 + \ldots + b_p x_p}_{\text{model}} + error$$

Multiple linear regression

- In the sales example:

Predictors are dummy variables

$$Sales_i = b_0 + b_1 \times East_i + b_2 \times North_i + b_3 \times South_i + e_i$$

Response

Example: Electronics sales

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	329106	109702	20.82	<.0001
Error	196	1032773	5269.25037		
Corrected Total	199	1361880			

The model is statistically significant at 1% level.

Root MSE	72.58960	R-Square	0.2417
Dependent Mean	388.55000	Adj R-Sq	0.2300
Coeff Var	18.68218		

Region explains 23% of variability in electronics sales.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	451.80000	10.26572	44.01	<.0001
North	1	-87.01739	13.48150	-6.45	<.0001
South	1	-106.68889	14.91575	-7.15	<.0001
East	1	-51.24444	15.86673	-3.23	0.0015

All slopes are statistically significant at 1% level.

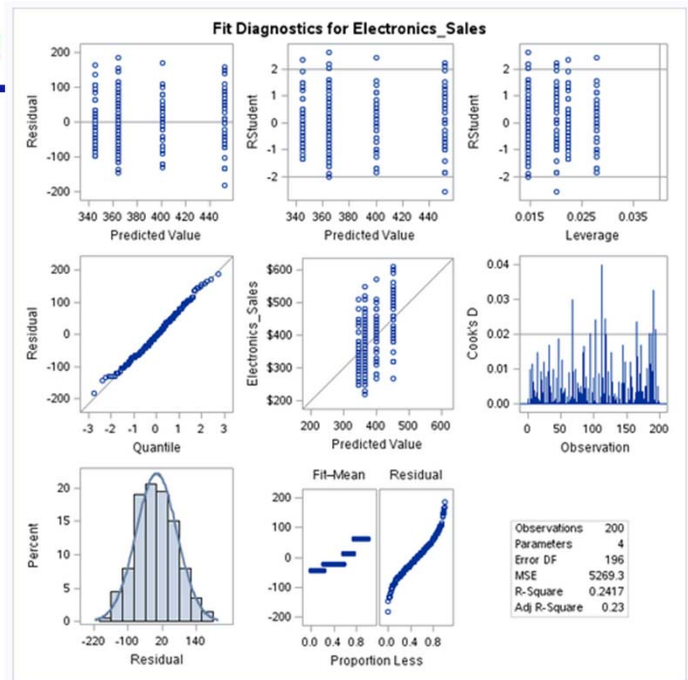
Example: Electronics sales

- Intercept = mean of the baseline group, West in this case:

$$\bar{x}_{West} = b_0 + b_1 \times 0 + b_2 \times 0 + b_3 \times 0$$
$$b_0 = \bar{x}_{West}$$

- For all other regions, slope = difference in means relative to the baseline group, e.g. for East:

$$\bar{x}_{East} = b_0 + b_1 \times 1 + b_2 \times 0 + b_3 \times 0$$
$$\bar{x}_{East} = \bar{x}_{West} + b_1$$
$$b_1 = \bar{x}_{East} - \bar{x}_{West}$$



How do we interpret these diagnostic plots?

Example: SAS code

```
/* Create dummy variable for level of region */
data work.store_dummies;
    set work.store;
    if Region='North' then North=1;
    else North=0;
    if Region='South' then South=1;
    else South=0;
    if Region='East' then East=1;
    else East=0;
run;

/* Fit a linear regresion model */
proc reg data=work.store_dummies plots=diagnostics;
    model Electronics_Sales=North South East;
    run;
quit;
```