# Lecture 15: ANOVA

## Criminology 250

Prof Maria Cuellar

University of Pennslyvania

# The basic ANOVA question

- Two variables: One quantitative (y) and one categorical (x).

- For example, y: birth weight, and x: race (white, Black, other).

- Main question: Are the values of y different for the groups? e.g., do people of different races have different birth weights?

- Note this is not a causal question necessarily.

- If categorical variable only has 2 values: use a 2-sample t-test.

- ANOVA allows for 3 or more groups.

# Analysis of variance

- We saw how to use a $t$-test to see whether two groups have equal means.

- If you have more groups, you can't just look at differences in the means.

- We can build a hypothesis test to check whether the variation in the means is bigger than we'd expect it to be just from random fluctuations.

- We'll need a new sampling model, called the $F$-model.

# Note: Regression vs. table

- You can think of ANOVA (analysis of variance) as a more general version of the t-test, or a special case of linear regression in which all covariates are factors (i.e. categorical variables).

We will see that you can write the results from an ANOVA test in a table. But they are no different than what we have already seen in linear regression.

# History

- The ANOVA test was designed by R.A. Fisher.

- But it had been used since the 1800s by Laplace, with some foundations from Gauss.

- RA Fisher was a mathematician and a eugenist, which was common at the time. Eugenics were later used by Nazis, but at the time this was not clearly a racist endeavor.

- RA Fisher did not think race determined psychological attributes. He thought it was "common historical and social background". - From "The Race Concept: Results of an Inquiry" (PDF). UNESCO. 1952.

- Famous for many things, but

# Are the means of these four groups equal?

First, an informal graphical investigation (EDA): side-by-side boxplots, or multiple histograms.
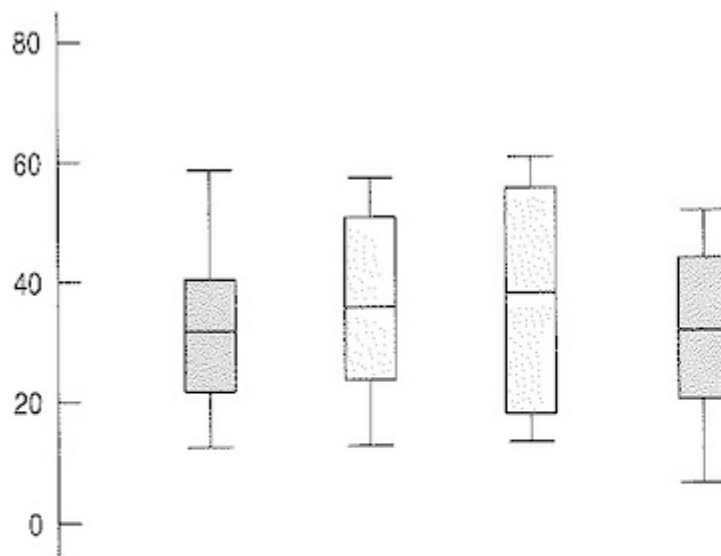


**Figure 26.2**

It's hard to see the difference in the means in these boxplots because the spreads are large relative to the differences in the means.
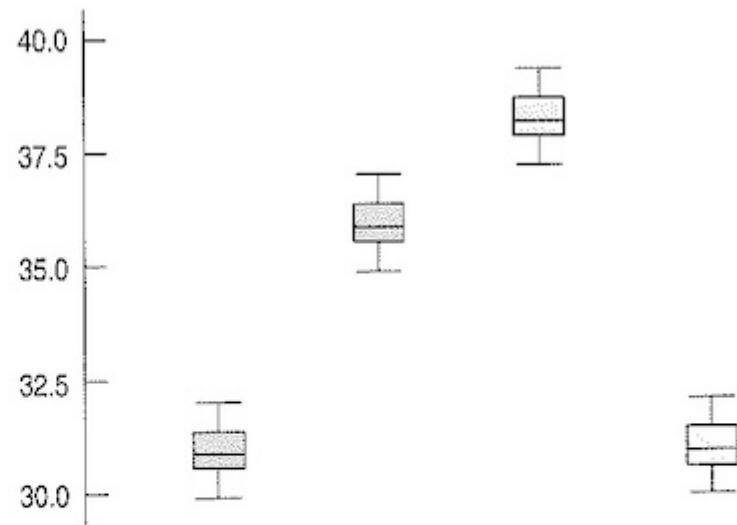
**Figure 26.3**

In contrast with Figure 26.2, the smaller variation makes it much easier to see the differences among the group means. (Notice also that the scale of the y-axis is considerably different from the plot on the left.)

Whether the differences between groups are signficant depends on: 1) the differences in the means, 2) the standard deviations of each group, and 3) the sample sizes.

# Four groups

- Actually, the sets of means in both figures (left and right) are the same! (they're: 31, 36, 38, 31).

- Right: The variation *within* each group is so small that the differences *between* the means stand out.

- Left: The variation *within* each group is large, so that the differences *between* the means seems small.

- This is the central idea of the $F$-test: We compare the differences *between* the means of the groups with the variation *within* the groups.

- When the differences between means are large compared with the variation within the groups, we reject the null hypothesis and conclude that the means are not equal.

# $F$-test

Let the null hypothesis be:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu.$$

And let the alternative hypothesis be:

$$H_A : \text{Not all the means are equal.}$$

The alternative hypothesis does not say how or which ones differ. To answer more specific questions, we can follow up with multiple comparisons.

# The F-statistic

The F-statistic is a ratio of the between-group variation and the within-group variation:

$$F = \frac{\text{Between}}{\text{Within}} = \frac{MSG}{MSE}.$$

A large value of F indicates relatively more difference between groups than within groups (evidence against the null hypothesis).

To get the p-value, we compute the F distribution based on certain numbers of degrees of freedom:

- I-1 df in numerator (# groups - 1)

- n-I df in denominator (rest of df)

# Multiple comparisons

- Once ANOVA indicates that the groups do not all have the same means, we can compare them two by two by using 2-sample t-test.

- But we need to adjust our p-value threshold because we are doing multiple tests with the same data: the Bonferroni correction (can do this in R with p.adjust from the stats package).

- If we really just want to test the difference between one pair of treatments, then *just use the t-test.*

- We can use the Tukey HSD test.

# Assumptions of ANOVA

- Each group is approximately normal.

  - Check this by looking at histograms and normal quantile plots.

  - Can handle some non-normality, but no severe outliers.

- The standard deviations of each group are approximately equal.

  - Rule of thumb: ratio of the largest to smallest sample standard deviation must be less than 2:1.

# Example: Birth weight vs. race

- Question: Is there a significant association between race and birth weight?

```
## # A tibble: 3 × 3
##    race  mean.bwt se.bwt
##    <fct>    <dbl>  <dbl>
## 1 white     3103     74
## 2 Black     2720    125
## 3 other     2805     88
```

- It looks like there's some association, but we don't yet know if it's statistically significant.

- Note that if we had just two racial categories in our data, we could run a t-test. Since we have more than 2, we need to run a 1-way analysis of variance (ANOVA).
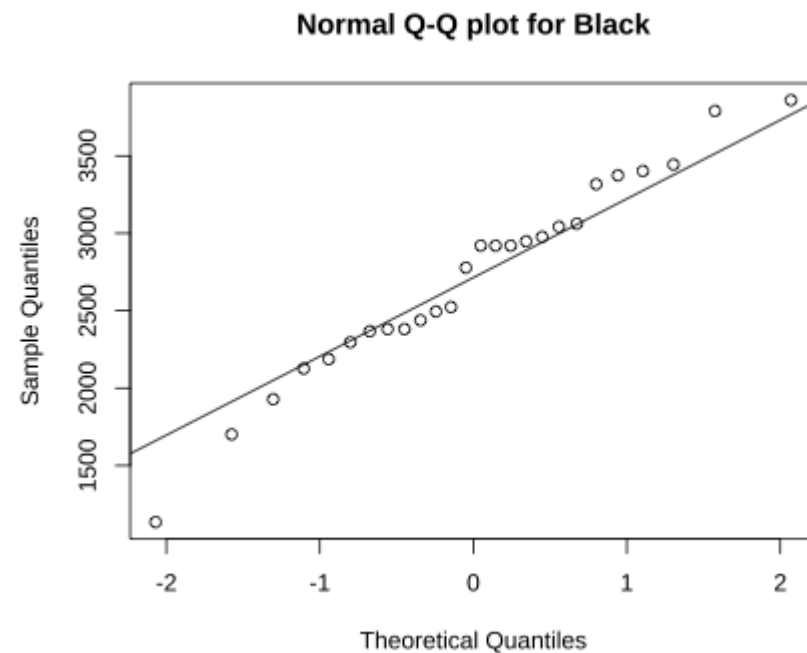
# Test assumptions in birthwt example: Normality
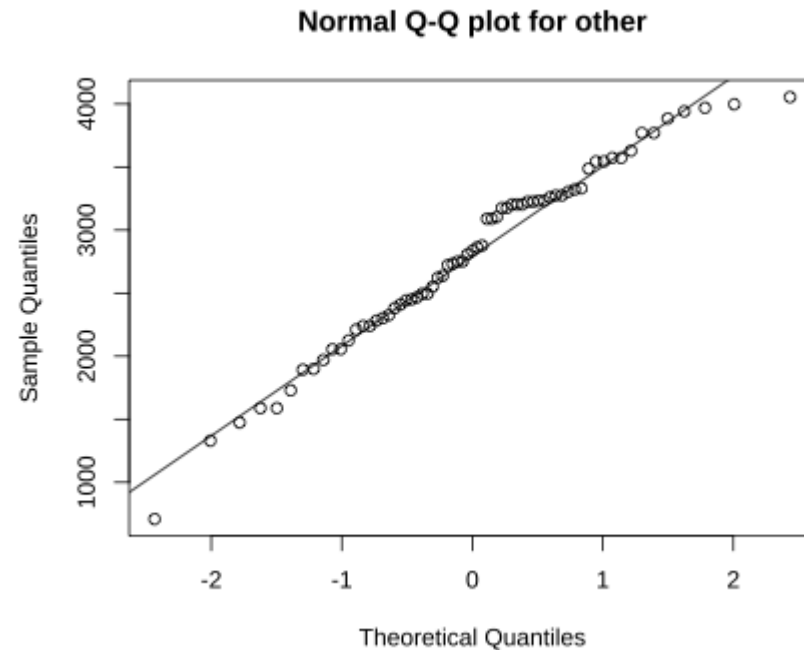
QQ plot for the individuals who are white
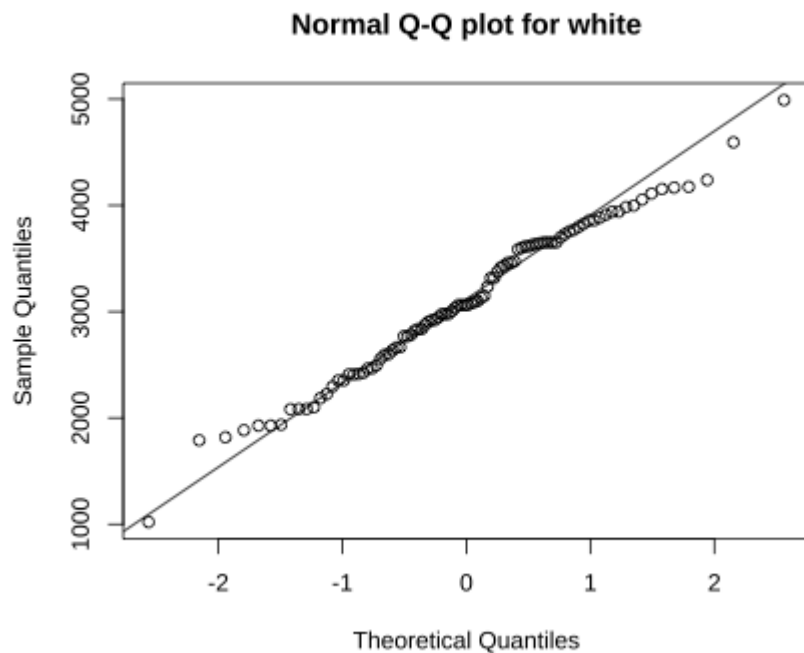
QQ plot for the individuals who are Black

# Test assumptions in birthwt example: Normality

QQ plot for the individuals who are of another race

# Testing assumptions in birthwt example: Standard deviations are approx equal

Compare the largest and smallest standard deviations:

- Largest: 125

- Smallest: 74

- Ratio: 125/74: 1.69. (Check, it's less than 2).

# Brief interlude to discuss how to use factors in R

It is worth noting that your categorical variable in the ANOVA analysis in R, aov(), needs to be a factor.

This website gives a nice overview of how to use factors in R:

https://www.gormanalysis.com/blog/r-introduction-to-factors-tutorial/

# ANOVA: Running this test in R

Actually, you can do this by using the same lm command we've been using!

```
##
## Call:
## lm(formula = birthwt.grams ~ race, data = birthwt)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -2096.28   -502.72    -12.72    526.28   1887.28
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3102.72       72.92  42.548  < 2e-16 ***
## raceBlack     -383.03      157.96  -2.425  0.01627 *
## raceother     -297.44      113.74  -2.615  0.00965 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 714.5 on 186 degrees of freedom
## Multiple R-squared:  0.05017,    Adjusted R-squared:  0.03996
## F-statistic: 4.913 on 2 and 186 DF,  p-value: 0.008336
```

# ANOVA: Running this test in R

Can do this two ways: Using an lm model:

```
reg.output <- lm(birthwt.grams ~ race, data = birthwt)
anova.output <- aov(reg.output)
summary(anova.output)
```

```
##               Df    Sum Sq Mean Sq F value  Pr(>F)
## race          2   5015725 2507863   4.913 0.00834 **
## Residuals   186  94953931  510505
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

or just doing aov directly:

```
anova.output1 <- aov(birthwt.grams ~ race, data = birthwt)
summary(anova.output1)
```

```
##               Df    Sum Sq Mean Sq F value  Pr(>F)
## race          2   5015725 2507863   4.913 0.00834 **
## Residuals   186  94953931  510505
```

# How to read ANOVA table in R

Coefficients are the same as in lm: anova.output$coefficients.

- Df (race): The degrees of freedom for the variable race. This is calculated as # groups - 1. In this case, there were 3 groups, so this value is 3-1=2.

- Df (Residuals): The degrees of freedom for the residuals. This is calculated as # total observations - # groups. In this case there were 189 observations and 3 groups: 189-3=186.

- Sum Sq.: The sum of squares associated with the variables race and the residuals.

- Mean Sq.: This is calculated as Sum Sq. / Df, for each of the two variables, race and residuals. For race, this is calculated as: 5015725 / 2 = 2507863.

- F value: The overall F-statistic of the ANOVA model. This is calculated as Mean Sq. race / Mean sq. Residuals. In this case it's 2507863/510505 = 4.913.

- Pr(>F): The p-value associated with the F-statistic with numerator df = 2 and denominator df = 186. In this case, the p-value is 0.00834, which is statistically significant at the 0.05 level.

- Signif. codes: The same as in the lm model.

# Tukey Honest Significant Differences (HSD) test

The Tukey HSD procedure will run a pairwise comparison of all possible combinations of groups and test these pairs for significant differences between their means, all while adjusting the p-value to a higher threshold for significance in order to compensate for the fact that many statistical tests are being performed and the chance for a false positive increases with increasing numbers of tests.



95% family-wise confidence level

Differences in mean levels of race
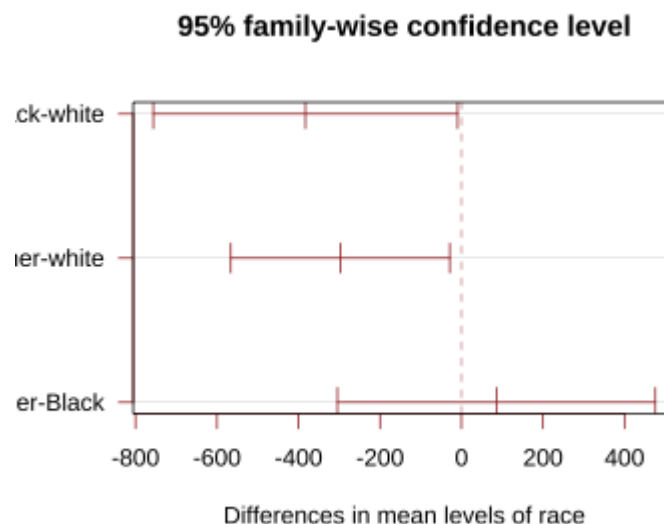
# Final project

- Make groups of 3-4 people.

- Select a final project topic.

- You'll have to find a question of interest in criminology and data that can answer it (probably easier with continuous, quantitative variables or a categorical x.

--

- First challenge: find data.