An Introduction to *R*

Part 3: Comparing Groups (2)

Kjell Johnson

Overview

- Brief review of one-way analysis of variance (ANOVA)
- Overview of two-way ANOVA
- Fixed versus random factors
- How to model data with fixed and random factors?
  - Mixed models!
- How to analyze repeated measures data

ANOVA Review

- ANOVA is used to compare multiple groups

- Assumptions:
  - Data are independent
    - not repeated measures or measure replicates
  - Residuals are normally distributed
  - Group variances are similar

Example: Colon Cancer Chemoprevention

50 animals were randomly divided into 5 groups. Each group received a different treatment. Are the tumor diameters different among treatments?

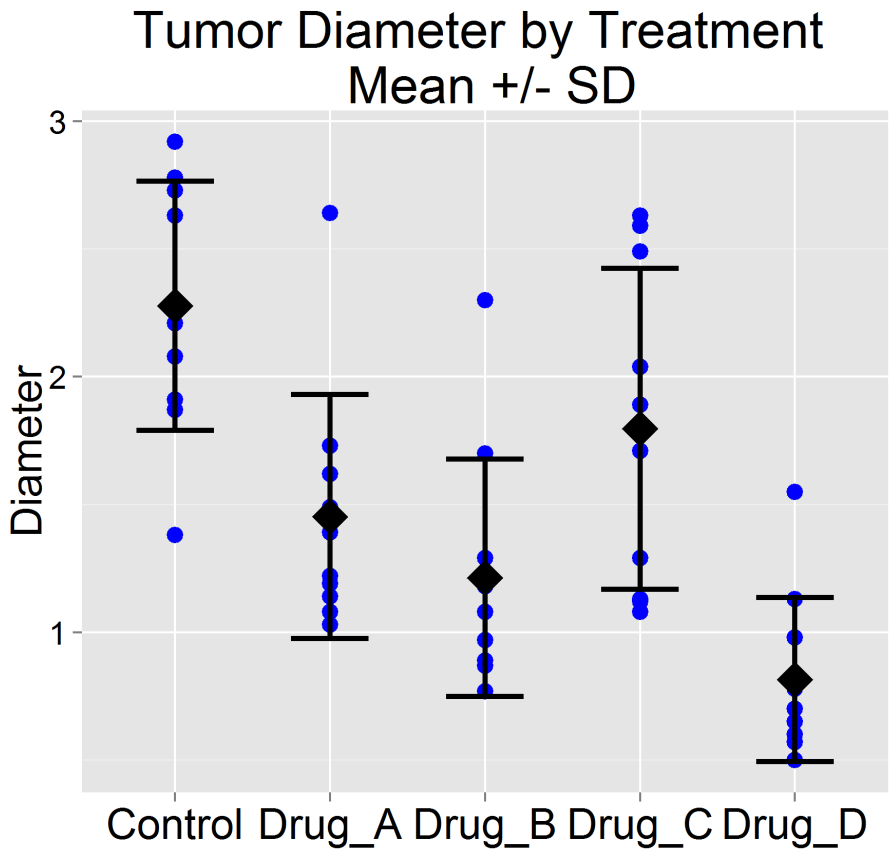
Animal	Control	Drug_A	Drug_B	Drug_C	Drug_D
1	2.27	1.73	0.97	1.29	0.50
2	1.38	1.19	1.08	1.13	0.70
3	1.91	1.39	0.77	1.12	1.55
4	2.21	1.08	1.29	1.08	0.98
5	2.63	1.14	1.08	1.71	0.65
6	2.73	1.22	1.18	2.49	0.70
7	2.08	1.62	0.87	2.04	1.13
8	2.92	1.03	0.89	2.59	0.60
9	2.78	2.64	1.70	2.63	0.57
10	1.87	1.49	2.30	1.89	0.78

The Statistical Model

$$Y_{i,j} = \mu_i + \epsilon_{i,j}$$

- $Y_{i,j}$ is the observed response value for the j^{th} subject on the i^{th} treatment
  - i = 1,…, # treatments (5 for the tumor diameter example)
  - j = 1,…,# of subjects (10 for the tumor diameter example)
- μ_i is the effect of the i^{th} treatment
- $\varepsilon_{i,j}$ is the random effect for the j^{th} subject on the i^{th} treatment that is not explained by the i^{th} treatment effect.
  - The errors are independent and follow a normal distribution with constant variance.

ANOVA Results




```
            Df Sum Sq Mean Sq F value Pr(>F)
Drug         5 126.67  25.334   107.3 <2e-16 ***
Residuals 45  10.62   0.236
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Fit: aov(formula = Diameter ~ Drug - 1, data = tumorNarrow)

Linear Hypotheses:
                       Estimate Std. Error t value Pr(>|t|)
Drug_A - Control == 0  -0.8250     0.2173  -3.797  0.00165 **
Drug_B - Control == 0  -1.0650     0.2173  -4.902  < 0.001 ***
Drug_C - Control == 0  -0.4810     0.2173  -2.214  0.10168
Drug_D - Control == 0  -1.4620     0.2173  -6.729  < 0.001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Two-way (factor) ANOVA

- One-way ANOVA is applied to data where one qualitative factor is systematically different across the experimental units
- Two-way ANOVA is applied to data where two qualitative factors are systematically changed across the units.
- Tumor diameter example (2 factors):
  1. Drug (Control, Drugs A, B, C, and D)
  2. Strain of mouse (B6C3F7, WT)

Why Two-way ANOVA?

- Account for different sources of variability in the *same* experiment
- Combined data for a more powerful analysis
  - Much better than analyzing treatment effect within each strain separately
- Learn if there is an interaction between factors
  - Do compounds affect B6C3F7 mice differently than WT mice?

Example: Colon Cancer Chemoprevention (2)

5 drugs and 2 strains (10 treatments) were used in this experiment. 10 mice were allocated to each treatment:

Animal	Drug	Strain	Diameter
1	Control	B6C3F7	2.27
2	Control	B6C3F7	1.38
3	Control	B6C3F7	1.91
4	Control	B6C3F7	2.21
5	Control	B6C3F7	2.63
…	…	…	…
96	Drug_D	WT	0.99
97	Drug_D	WT	1.52
98	Drug_D	WT	0.93
99	Drug_D	WT	1.08
100	Drug_D	WT	1.1

The Statistical Model

$$Y_{i,j,k} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{i,j,k}$$

- $Y_{i,j,k}$ is the observed response value for the k^{th} subject on the i^{th} level of α and j^{th} level of β.
  - i = 1,…, # levels of α (5 drugs)
  - j = 1,…,# levels of β (2 strains)
- $\mu_{..}$ is the overall average response
- $(\alpha\beta)_{ij}$ is the interaction between α and β
- $\varepsilon_{i,j,k}$ is the random variation of the k^{th} subject not explained by the i^{th} level of α and j^{th} level of β.
  - The errors are independent and follow a normal distribution with constant variance.

Bring the Data into R and Visualize

Set the working directory:

```
myLocation <- "c:/Documents and Settings/johns94/Desktop/Part3"
setwd(myLocation)
```

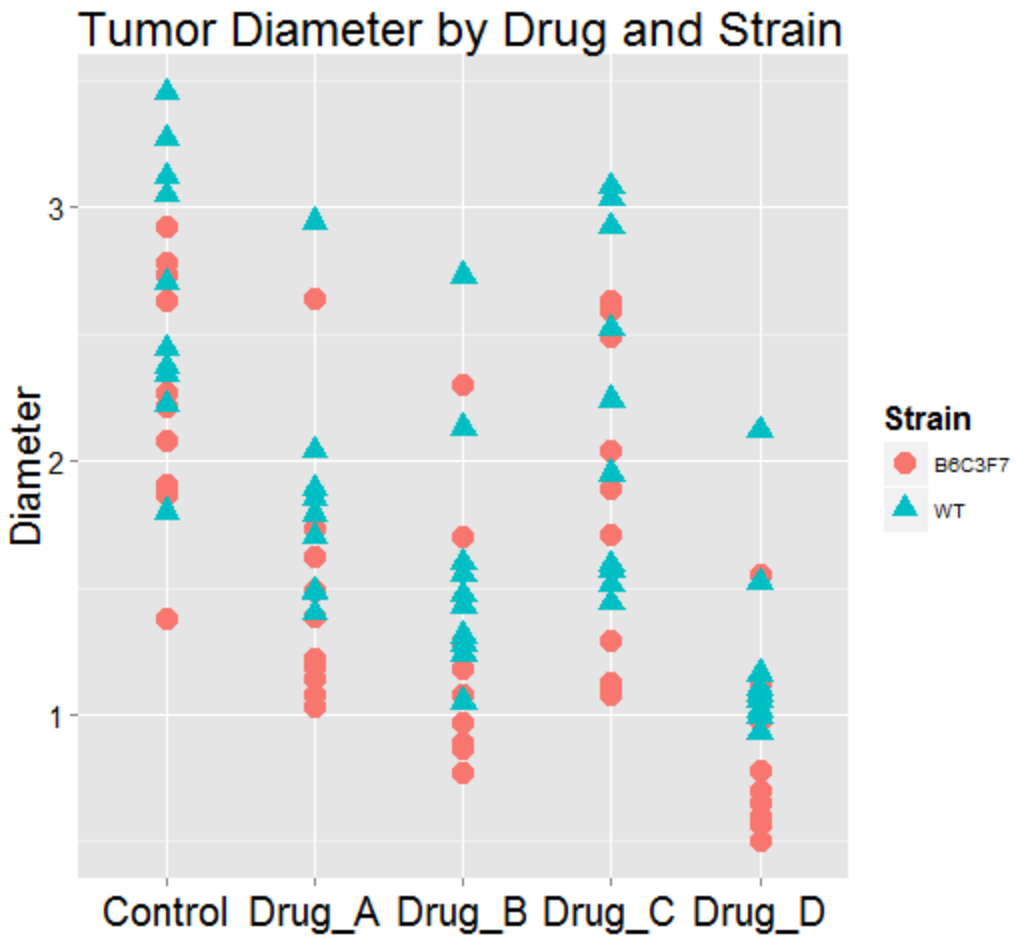
Get data:

```
tumor2 <- read.csv("tumor2.csv",header=TRUE)
```

Plot data:

```
ggplot(tumor2,
       aes(x=Drug,y=Diameter,color=Strain,shape=Strain)) +
       geom_point(aes(color=Strain,shape=Strain),size=5) +
       ggtitle("Tumor Diameter by Drug and Strain") +
       ylab("Diameter") +
       xlab("") +
       theme(axis.text.x = element_text(size=20,color="black"),
             axis.text.y = element_text(size=15,color="black"),
             axis.title.y = element_text(size=20),
             title = element_text(size=20))
```

Tumor Diameter Data

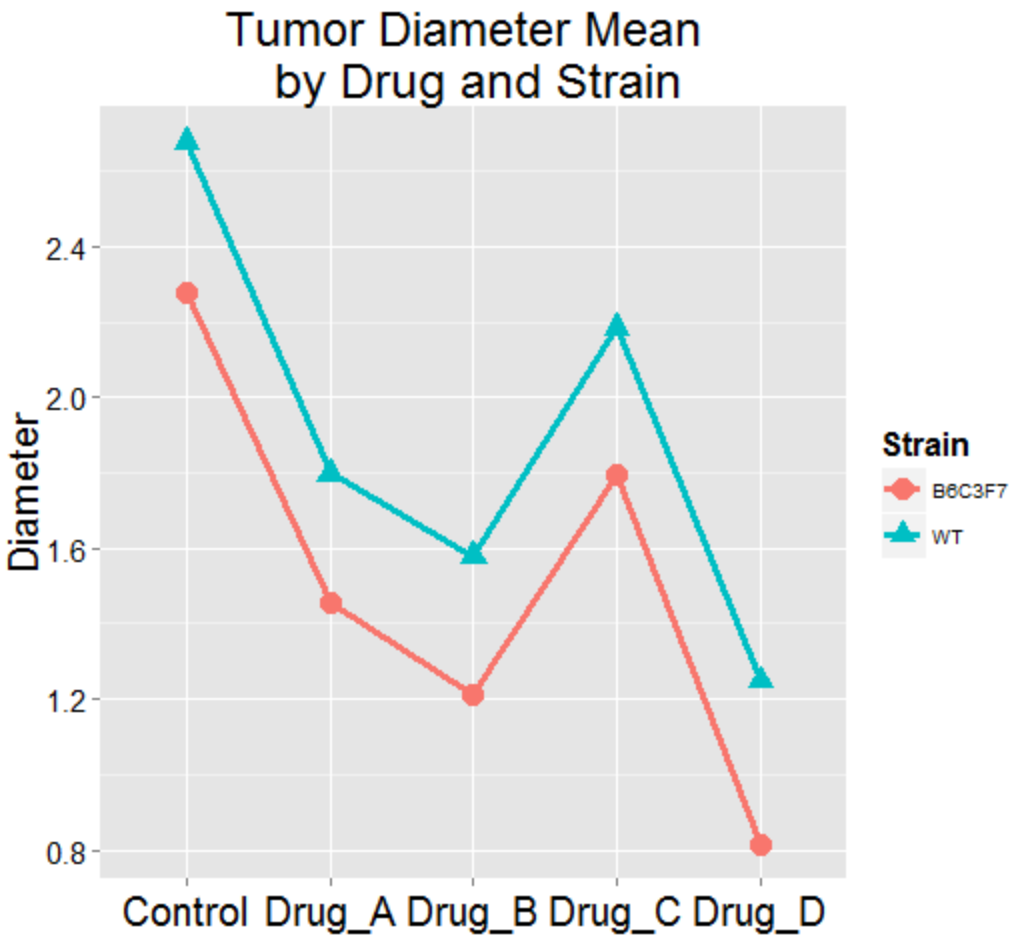


Visualize Treatment Means

Compute means of Diameter for each Drug and Strain:

```
library(doBy)
tumor2Summary <- summaryBy(Diameter ~ Drug + Strain,
                           data = tumor2,
                           FUN = mean)

ggplot(tumorSummary,
       aes(x=Drug,y=Diameter.mean,color=Strain,shape=Strain)) +
  geom_point(aes(color=Strain,shape=Strain),size=5) +
  geom_line(aes(group=Strain),size=1.2) +
  ggtitle("Tumor Diameter Mean \n by Treatment and Strain") +
  ylab("Diameter") +
  xlab("")+
  theme(axis.text.x = element_text(size=20,color="black"),
        axis.text.y = element_text(size=15,color="black"),
        axis.title.y = element_text(size=20),
        title = element_text(size=20))

ggsave(file = "tumor2Figure2.png")
```

Figure


Tumor Diameter Mean
by Drug and Strain
Diameter
2.4
2.0
1.6
1.2
0.8
Control
Drug_A
Drug_B
Drug_C
Drug_D
Strain
B6C3F7
WT

How to Perform Two-way ANOVA in R?

- Same as one-way ANOVA: the “aov” function

Perform two-way ANOVA:

```
tumor2ANOVA = aov(Diameter ~ Drug + Strain + Drug:Strain,
                  data=tumor2)

summary(tumor2ANOVA)
```

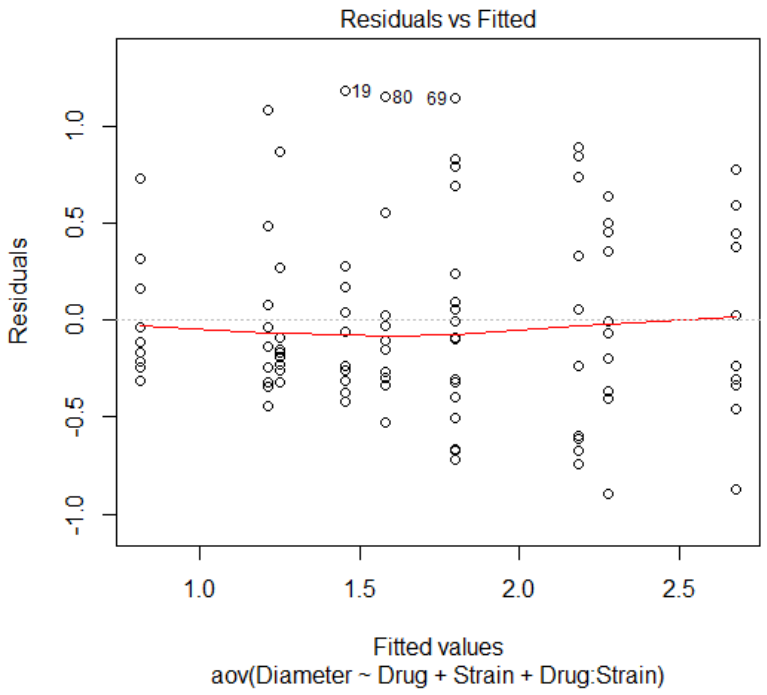
Interaction between Drug and Strain

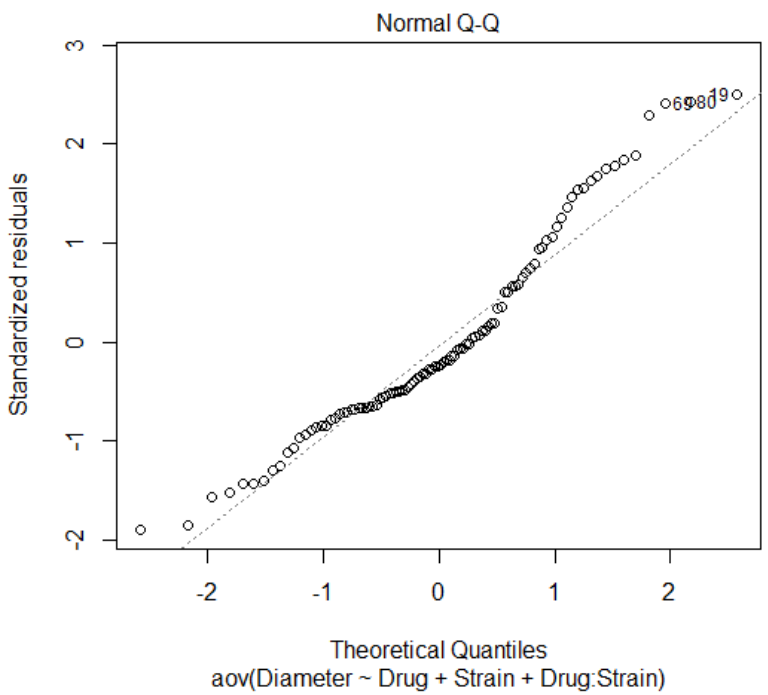
```
             Df Sum Sq Mean Sq F value   Pr(>F)
Drug          4 24.623   6.156  24.673 8.35e-14 ***
Strain        1  3.729   3.729  14.945 0.000209 ***
Drug:Strain   4  0.023   0.006   0.023 0.998978
Residuals    90 22.455   0.249
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The interaction between Drug and Strain is not significant. Drug and Strain each explain a significant amount of variability in tumor diameter.

Check Residuals

```
plot(tumor2ANOVA)
```





The residual plots look OK.

Examine Coefficients

Get factor coefficients:

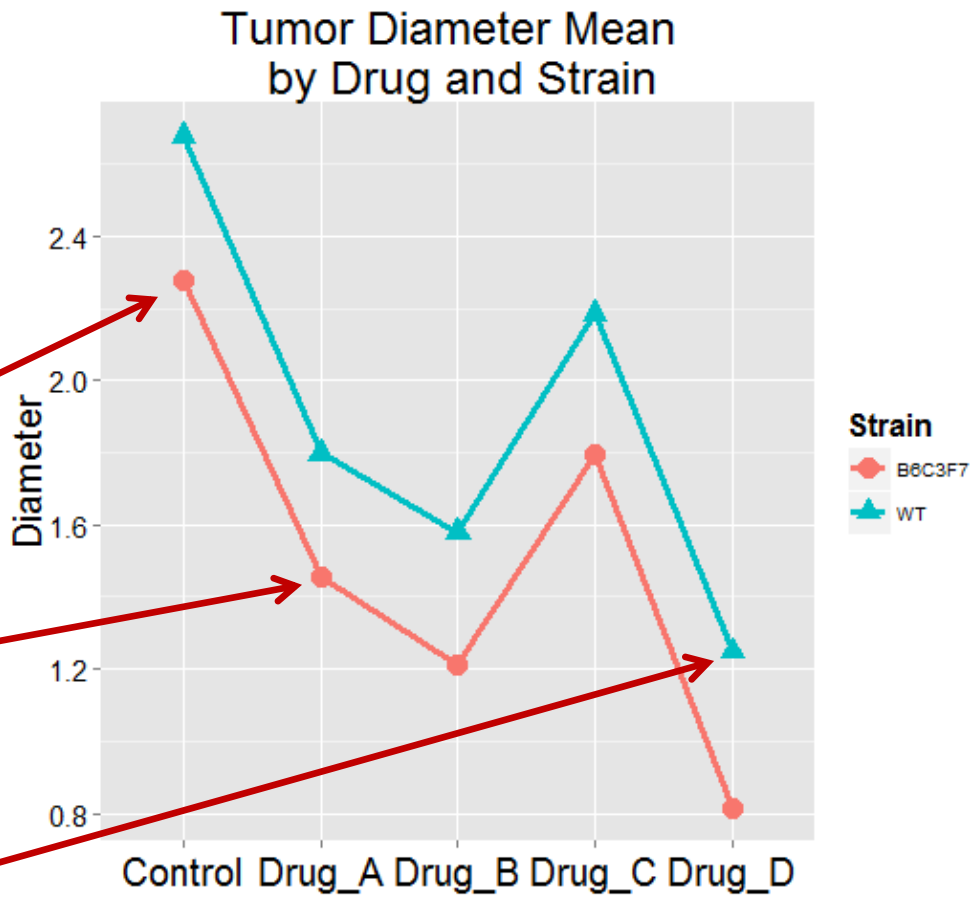
```
tumor2ANOVA$coefficients
```

```
Intercept =  2.28
Drug_A    = -0.83
Drug_B    = -1.07
Drug_C    = -0.48
Drug_D    = -1.46
WT        =  0.40
```

Intercept is the average effect of Control and B6C3F7 strain.

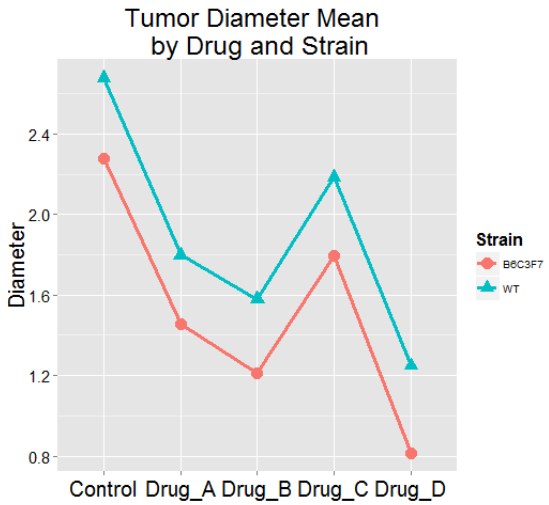
The average effect of Drug_A and B6C3F7 strain is:
2.28 – 0.83 = 1.45

The average effect of Drug_D and WT strain is:
2.28 – 1.46 + 0.40 = 1.22



Interaction Term is Not Significant

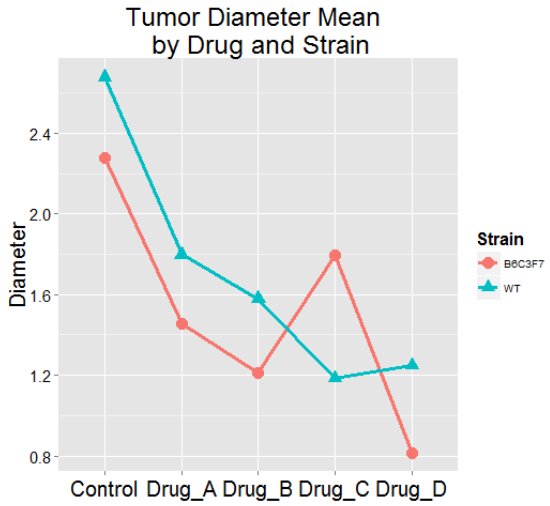
- The interaction term in the model was not significant
  - This indicates that the factors independently affect the response
  - Visually, the mean response profiles will run parallel to each other:



  - When this is the case, we can perform pairwise comparisons for each factor (like we did in one-way ANOVA) using all of the data.

Interaction Term is *Significant*

- If the interaction term is significant, then:
  - the factors work together to affect the response
  - the mean response profiles will not run parallel to each other:




```
             Df Sum Sq Mean Sq F value   Pr(>F)
Drug          4 22.893   5.723  22.940 4.35e-13 ***
Strain        1  0.867   0.867   3.474  0.06560 .
Drug:Strain   4  4.005   1.001   4.013  0.00485 **
Residuals    90 22.455   0.249
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

  - we must analyze the factor of interest within each level of the other factor (i.e. separate analyses).

Types of Factors

- When all of the levels we desire of a factor are included in the experiment, then the factor is called “fixed.”
  - If we are only interested in studying WT and B6C3F7 mice, then Strain is a fixed effect.
- When only a random selection of the levels of a factor are included in the experiment, then the factor is called “random.”
  - If we are interested in studying many strains of mice and randomly select WT and B6C3F7 mice for this study, then Strain is a random effect.

Analyzing Mixed Effects Models in R

- The aov function is only appropriate for fixed effects models.
- To analyze the mixed models, we need the lme4 package

Install and load the lme4 package:

```
install.packages("lme4", dependencies=TRUE)
library(lme4)
```

- We will use the *lmer* function
  - Fixed effects appear the same way in the formula
  - Random effects are listed as (1|*random effect*)

Strain as a Random Factor

Estimate the mixed model (terms in bold are random factors):

```
tumor2Mixed = lmer(Diameter ~ Drug +
                                (1|Strain) +
                                (1|Drug:Strain), data=tumor2)
summary(tumor2Mixed)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: Diameter ~ Drug + (1 | Strain) + (1 | Drug:Strain)
   Data: tumor2

REML criterion at convergence: 151.3986

Random effects:
 Groups       Name        Variance Std.Dev.
 Drug:Strain (Intercept) 0.00000  0.0000
 Strain      (Intercept) 0.06979  0.2642
 Residual                0.23912  0.4890
Number of obs: 100, groups: Drug:Strain, 10; Strain, 2

Fixed effects:
            Estimate Std. Error t value
(Intercept)   2.4770     0.2165  11.443
DrugDrug_A   -0.8515     0.1546  -5.507
DrugDrug_B   -1.0810     0.1546  -6.991
DrugDrug_C   -0.4860     0.1546  -3.143
DrugDrug_D   -1.4440     0.1546  -9.338
```

The estimate of the amount of between-strain variation.

Notice that we do not get an estimate of the effect of WT.

Comparison of Drug Effect between Models

Fixed effects model:

```
             Df Sum Sq Mean Sq F value   Pr(>F)
Drug          4 24.623   6.156  24.673 8.35e-14 ***
Strain        1  3.729   3.729  14.945 0.000209 ***
Drug:Strain   4  0.023   0.006   0.023 0.998978
Residuals    90 22.455   0.249
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mixed effects model:

```
anova(tumor2Mixed)
```

```
Analysis of Variance Table
     Df Sum Sq Mean Sq F value
Drug  4 24.623  6.1558  25.744
```

The sum of squares, and mean square values are the same. However the F values are different. This is a subtle but important difference, because the formula for the F value *changes* between a fixed effects and mixed effects model. For this data it doesn't make a difference, but in other data it can make a big difference.

Other Examples of Random Effects

- Repeated measures designs:
  - Each subject receives all treatments
    - In this case, we're interested in studying the effect of the treatment, while accounting for the within-subject variation
    - Cross-over designs are specific experimental designs in which all subjects receive all treatments following particular rules about treatment order
  - Each subject is randomly assigned to one treatment and is measured at two or more subsequent timepoints
    - In this layout, we're usually interested in understanding if treatment means change over time

Repeated Measures: Example 1

- Single cell patch clamping experiment in rat striatal neurons
  - 10 neurons were randomly selected
  - Each neuron received three treatments:
    - Control,
    - Drug_A, and
    - Drug_B
  - Resting membrane potential was measured for each neuron

Data

Neuron	Control	Drug_A	Drug_B
1	-61.5	-58.9	-61.7
2	-63.8	-58.0	-55.5
3	-48.6	-49.5	-43.4
4	-63.1	-65.7	-63.5
5	-69.6	-65.5	-68.8
6	-50.5	-48.7	-48.0
7	-49.9	-41.6	-36.9
8	-52.1	-39.0	-37.7
9	-50.8	-48.6	-49.8
10	-50.7	-43.0	-44.0

Statistical Model

$$Y_{i,j} = \mu_{..} + \alpha_i + \rho_j + \epsilon_{i,j}$$

- $Y_{i,j}$ is the observed response value for the i^{th} drug on the j^{th} subject
  - i = 1,…, # of drugs (3)
  - j = 1,…,# of subjects (10)
- $\mu_{..}$ is the overall average response
- ρ_j is random and follows a normal distribution
- $\varepsilon_{i,j}$ is the random variation of the j^{th} subject not explained by the i^{th} drug
- This is called an additive model because there is no interaction term between subjects and treatment

Bring the Data into R and Visualize

Get data:

```
patch <- read.csv("patchClamp.csv",header=TRUE)
```

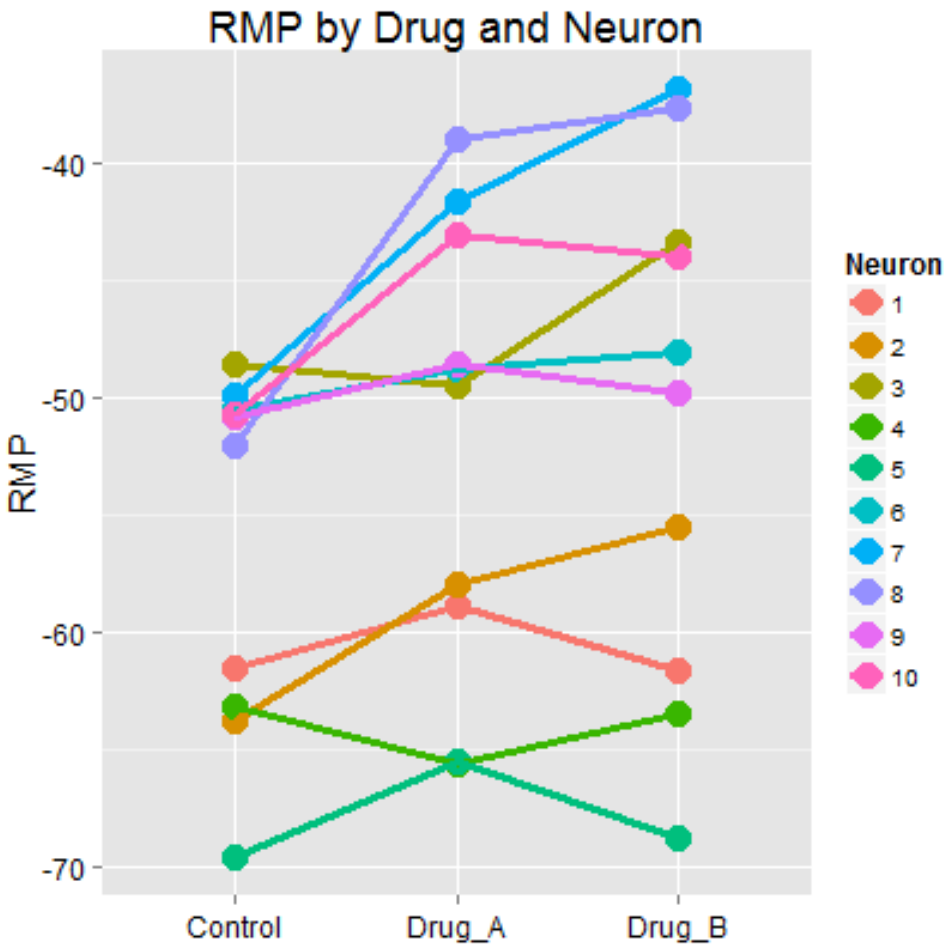
Change Neuron to a factor:

```
patch$Neuron <- factor(patch$Neuron)
```

Plot data:

```
ggplot(patch,
       aes(x=Drug,y=RMP,color=Neuron)) +
       geom_point(aes(color=Neuron),size=5) +
       geom_line(aes(group=Neuron),size=1.2) +
       ggtitle("RMP by Drug and Neuron") +
       ylab("RMP") +
       xlab("") +
       theme(axis.text.x = element_text(size=16,color="black"),
             axis.text.y = element_text(size=18,color="black"),
             axis.title.y = element_text(size=18),
             title = element_text(size=20))
```

Tumor Diameter Data

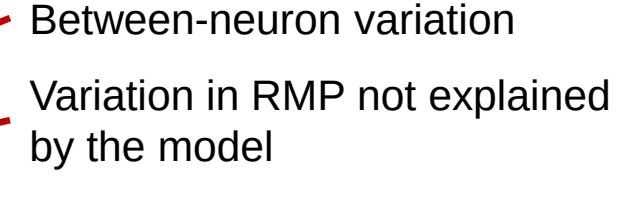


How to Analyze in R?

Perform mixed model analysis:

```
patchMixed <- lmer(RMP ~ Drug + (1|Neuron), data=patch)

summary(patchMixed)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: RMP ~ Drug + (1 | Neuron)
   Data: patch

REML criterion at convergence: 174.6891

Random effects:
 Groups   Name        Variance Std.Dev.
 Neuron   (Intercept) 81.050   9.003
 Residual              9.949   3.154
Number of obs: 30, groups: Neuron, 10

Fixed effects:
            Estimate Std. Error t value
(Intercept)  -56.057      3.017 -18.583
DrugDrug_A     4.197      1.411   2.975
DrugDrug_B     5.131      1.411   3.637
```



Expected control response:
-56.06
Expected Drug_A response:
-56.06 + 4.20
Expected Drug_B response:
-56.06 + 5.13

What We're Really Interested In:
Pairwise comparisons of drug effect

Load multcomp package:

```
library(multcomp)
```

Perform all pairwise comparisons, adjusting with Tukey:

```
glhtTukey <- glht(patchMixed, linfct = mcp(Drug = "Tukey")
summary(glhtTukey)
```

```
  Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lmer(formula = RMP ~ Drug + (1 | Neuron), data = patch)

Linear Hypotheses:
                     Estimate Std. Error z value Pr(>|z|)
Drug_A - Control == 0   4.1972     1.4106   2.975   0.0083 **
Drug_B - Control == 0   5.1309     1.4106   3.637   <0.001 ***
Drug_B - Drug_A == 0    0.9337     1.4106   0.662   0.7856
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Drug_A and Drug_B are significantly different from control
Drug_A and Drug_B are not significantly different from each other

Consequences of the Wrong Model

- What happens if we ignore the fact that each neuron received all three treatments?

Perform one-way ANOVA:

```
patchANOVA = aov(RMP ~ Drug, data=patch)

summary(patchANOVA)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Drug          2  149.4   74.69   0.821  0.451
Residuals    27 2457.0   91.00
```

In this model, drug *does not* explain a significant amount of response variation! WHY?

The between-neuron variation is now consolidated into the residual mean square

The two-factor mixed-model allows us to pull out the between-neuron variation:

```
Random effects:
 Groups   Name        Variance Std.Dev.
 Neuron   (Intercept) 81.050   9.003
 Residual              9.949   3.154
Number of obs: 30, groups: Neuron, 10
```

81.05 + 9.95 = 91.00
The residual becomes smaller, which increases the F value, making the effect of Drug significant.

Repeated Measures: Example 2

- Polysomnographic analysis of rat cortical EEG - wakefulness
  - 16 animals were randomly selected
    - 8 were randomly treated with vehicle; remaining 8 received Drug_A
  - Total time awake post treatment (0-3 hr) was measured for each animal at days 0, 1, 2, 4, 5, 8, 11, and 14

Data

Animal	Day	Drug	TTA
1	0	vehicle	27.0
1	1	vehicle	43.9
1	2	vehicle	40.2
1	4	vehicle	69.7
1	5	vehicle	52.4
1	8	vehicle	37.1
1	11	vehicle	37.5
1	14	vehicle	73.2
2	0	vehicle	54.2
…	…	…	…
15	8	DrugA	84.4
15	11	DrugA	60.9
15	14	DrugA	97.8
16	0	DrugA	65.5
16	1	DrugA	71.9
16	2	DrugA	68.8
16	4	DrugA	92.3
16	5	DrugA	81.5
16	8	DrugA	88.4
16	11	DrugA	71.5
16	14	DrugA	92.8

Statistical Model

$$Y_{i,j,k} = \mu_{...} + \alpha_i + \rho_{i,j} + \tau_k + (\alpha\tau)_{i,k} + \epsilon_{i,j,k}$$

- $Y_{i,j,k}$ is the observed response value for the j^{th} subject on the i^{th} drug at time k.
  - i = 1,…, # of drugs (2)
  - j = 1,…,# of subjects within drug (8)
  - k = 1,…,# of timepoints (8)
- $\mu_{...}$ is the overall average response
- $\rho_{i,j}$ is the effect corresponding to the j^{th} subject on the i^{th} drug and is random and follows a normal distribution [random]
- τ_k is the effect at time k [fixed]
- $(\alpha\tau)_{i,k}$ is the interaction between the i^{th} drug at time k [fixed]
- $\varepsilon_{i,,j,k}$ is the random variation of the j^{th} subject not explained by drug i at time k

Bring the Data into R

Get data:

```
wake <- read.csv("wakefulness.csv",header=TRUE)
```

Change Animal and Day to factors:

```
wake$Animal <- factor(wake$Animal)
wake$Day <- factor(wake$Day)
```

Compute means, standard deviations, and number of subjects per Drug and Day

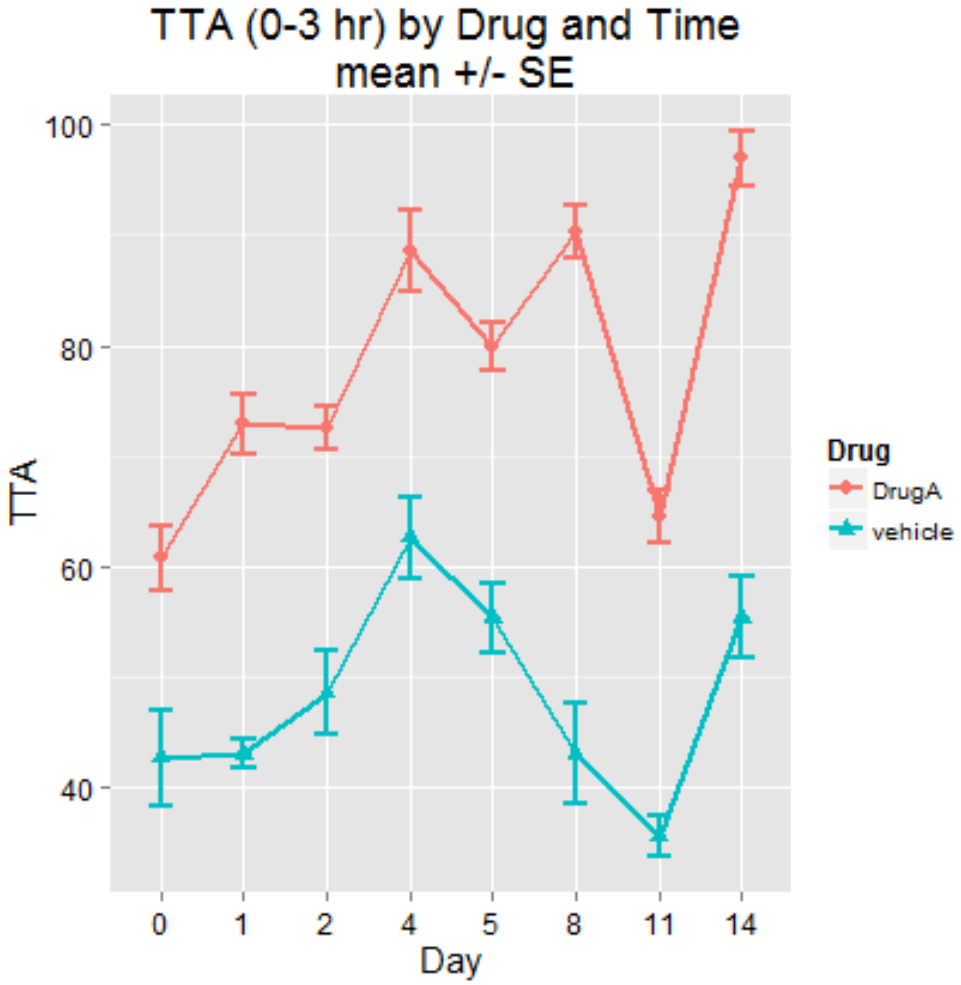
```
library(doBy)
forPlot <- summaryBy(TTA ~ Drug + Day, data=wake, FUN =
c(mean,sd,length))
forPlot$TTA.SEM <- forPlot$TTA.sd/sqrt(forPlot$TTA.length)
nDrug <- length(unique(wake$Drug))
```

Visualize Data

Plot data:

```
ggplot(forPlot,
       aes(x=Day, y=TTA.mean, color=Drug,shape=Drug)) +
  geom_point(aes(color=Drug,shape=Drug),size=3) +
  geom_line(aes(group=Drug),size=1) +
  geom_errorbar(aes(ymin=TTA.mean-TTA.SEM,
                    ymax=TTA.mean+TTA.SEM),width=0.3,size=1) +
  ggtitle("TTA (0-3 hr) by Drug and Time \n mean +/- SE") +
  xlab("Day") +
  ylab("TTA") +
  theme(axis.text.x = element_text(size=20,color="black"),
        axis.text.y = element_text(size=15,color="black"),
        axis.title.y = element_text(size=20),
        title = element_text(size=20))
```

Wakefulness Data



How to Analyze in R?

Perform mixed model analysis:

```
wakeMixed <- lmer(TTA ~ Drug + (1|Drug/Animal) +
                        Day + Drug:Day, data=wake)

summary(wakeMixed)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: TTA ~ Drug + (1 | Drug/Animal) + Day + Drug:Day
   Data: wake

REML criterion at convergence: 835.2642

Random effects:
 Groups      Name        Variance Std.Dev.
 Animal:Drug (Intercept)   1.31    1.144
 Drug        (Intercept) 110.97   10.534
 Residual                 74.16    8.612
Number of obs: 128, groups: Animal:Drug, 16; Drug, 2

Fixed effects:
                 Estimate Std. Error t value
(Intercept)        60.789     10.973   5.540
Drugvehicle       -18.136     15.518  -1.169
Day1               12.125      4.306   2.816
Day2               11.749      4.306   2.729
Day4               27.844      4.306   6.466
Day5               19.159      4.306   4.449
Day8               29.446      4.306   6.839
Day11               3.770      4.306   0.876
Day14              36.146      4.306   8.395
Drugvehicle:Day1  -11.764      6.089  -1.932
Drugvehicle:Day2   -5.854      6.089  -0.961
Drugvehicle:Day4   -7.892      6.089  -1.296
Drugvehicle:Day5   -6.461      6.089  -1.061
Drugvehicle:Day8  -29.042      6.089  -4.769
Drugvehicle:Day11 -10.857      6.089  -1.783
Drugvehicle:Day14 -23.394      6.089  -3.842
```

Animal nested within Drug

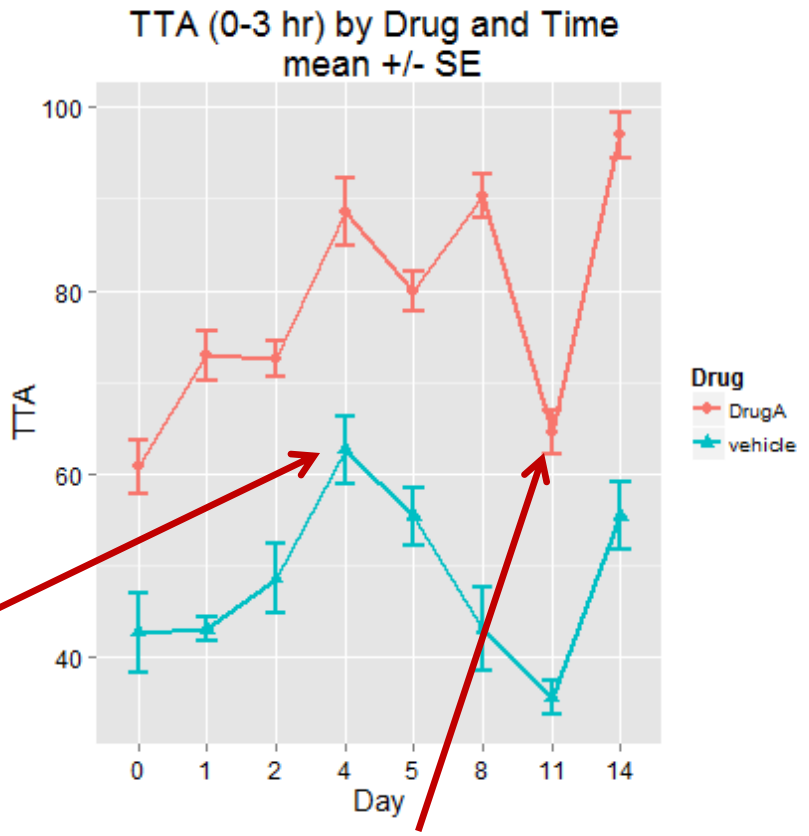
Interaction between Drug and Day

Random effects variation

Estimates of the contribution of each level of each factor

Interpreting Estimates

	Estimate	Std. Error	t value
(Intercept)	60.789	10.973	5.540
Drugvehicle	-18.136	15.518	-1.169
Day1	12.125	4.306	2.816
Day2	11.749	4.306	2.729
Day4	27.844	4.306	6.466
Day5	19.159	4.306	4.449
Day8	29.446	4.306	6.839
Day11	3.770	4.306	0.876
Day14	36.146	4.306	8.395
Drugvehicle:Day1	-11.764	6.089	-1.932
Drugvehicle:Day2	-5.854	6.089	-0.961
Drugvehicle:Day4	-7.892	6.089	-1.296
Drugvehicle:Day5	-6.461	6.089	-1.061
Drugvehicle:Day8	-29.042	6.089	-4.769
Drugvehicle:Day11	-10.857	6.089	-1.783
Drugvehicle:Day14	-23.394	6.089	-3.842



Expected response for vehicle at Day 4:
60.8 -18.1 + 27.8 – 7.9 = 62.6

Expected response for Drug_A at Day 11:
60.8 + 3.8 = 64.6

Are Fixed Effects Significant?

Install car package:

```
install.packages("car",dependencies=TRUE)
library(car)
```

Use Anova function to compute test statistics and p-values

```
Anova(wakeMixed,type=2,test.statistic="F")
```

```
Analysis of Deviance Table (Type II Wald F tests with Kenward-Roger df)

Response: TTA
               F Df Df.res    Pr(>F)
Drug      4.0193  1 100929   0.04498 *
Day      21.3549  7     98 < 2.2e-16 ***
Drug:Day  5.0150  7     98 6.929e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is a significant interaction between Drug and Day. This means that the time-course profiles between vehicle and Drug_A are not parallel. If we want to make inferences about treatments, we need to do that within specified days.

Upcoming Session

- Part 4: Covariance Structures in Mixed Models and Dimension Reduction and Classification
  - Principal component analysis (PCA), partial least squares (PLS), recursive partitioning (RPart), and random forests (RF)