STAT 401A - Statistical Methods for Research Workers

Two-way ANOVA

Jarad Niemi (Dr. J)

Iowa State University

last updated: December 1, 2014

Data

An experiment was run on tomato plants to determine the effect of

- 3 different varieties (A,B,C)

Data

An experiment was run on tomato plants to determine the effect of

- 3 different varieties (A,B,C) and
- 4 different planting densities (10,20,30,40)

Data

An experiment was run on tomato plants to determine the effect of

- 3 different varieties (A,B,C) and
- 4 different planting densities (10,20,30,40)

on yield.

Data

An experiment was run on tomato plants to determine the effect of

- 3 different varieties (A,B,C) and
- 4 different planting densities (10,20,30,40)

on yield.

There is an expectation that planting density will have a different effect depending on the variety.

Data

An experiment was run on tomato plants to determine the effect of

- 3 different varieties (A,B,C) and
- 4 different planting densities (10,20,30,40)

on yield.

There is an expectation that planting density will have a different effect depending on the variety. Therefore a balanced, complete, randomized design was used.

Data

An experiment was run on tomato plants to determine the effect of

- 3 different varieties (A,B,C) and
- 4 different planting densities (10,20,30,40)

on yield.

There is an expectation that planting density will have a different effect depending on the variety. Therefore a balanced, complete, randomized design was used.

- complete: each treatment (variety $\times$ density) is represented in the experiment

Data

An experiment was run on tomato plants to determine the effect of

- 3 different varieties (A,B,C) and
- 4 different planting densities (10,20,30,40)

on yield.

There is an expectation that planting density will have a different effect depending on the variety. Therefore a balanced, complete, randomized design was used.

- complete: each treatment (variety $\times$ density) is represented in the experiment
- balanced: each treatment in the experiment has the same number of replications

Data

An experiment was run on tomato plants to determine the effect of

- 3 different varieties (A,B,C) and
- 4 different planting densities (10,20,30,40)

on yield.

There is an expectation that planting density will have a different effect depending on the variety. Therefore a balanced, complete, randomized design was used.

- complete: each treatment (variety $\times$ density) is represented in the experiment
- balanced: each treatment in the experiment has the same number of replications
- randomized: treatment was randomly assigned to the plot

Data

An experiment was run on tomato plants to determine the effect of

- 3 different varieties (A,B,C) and
- 4 different planting densities (10,20,30,40)

on yield.

There is an expectation that planting density will have a different effect depending on the variety. Therefore a balanced, complete, randomized design was used.

- complete: each treatment (variety $\times$ density) is represented in the experiment
- balanced: each treatment in the experiment has the same number of replications
- randomized: treatment was randomly assigned to the plot

This is also referred to as a full factorial or fully crossed design.

Hypotheses

- Does variety affect mean yield?

Hypotheses

- Does variety affect mean yield?

- Does density affect mean yield?

Hypotheses

- Does variety affect mean yield?
- Does density affect mean yield?
- Does density affect yield differently for each variety?

Hypotheses

- Does variety affect mean yield?
  - Is the mean yield for variety A different from B on average?
- Does density affect mean yield?
- Does density affect yield differently for each variety?

Hypotheses

- Does variety affect mean yield?
  - Is the mean yield for variety A different from B on average?
  - Is the mean yield for variety A different from B at a particular value for density?
- Does density affect mean yield?

- Does density affect yield differently for each variety?

Hypotheses

- Does variety affect mean yield?
  - Is the mean yield for variety A different from B on average?
  - Is the mean yield for variety A different from B at a particular value for density?
- Does density affect mean yield?
  - Is the mean yield for density 10 different from density 20 on average?
- Does density affect yield differently for each variety?

Hypotheses

- Does variety affect mean yield?
  - Is the mean yield for variety A different from B on average?
  - Is the mean yield for variety A different from B at a particular value for density?
- Does density affect mean yield?
  - Is the mean yield for density 10 different from density 20 on average?
  - Is the mean yield for density 10 different from density 20 at a particular value for variety?
- Does density affect yield differently for each variety?

Hypotheses

- Does variety affect mean yield?
  - Is the mean yield for variety A different from B on average?
  - Is the mean yield for variety A different from B at a particular value for density?
- Does density affect mean yield?
  - Is the mean yield for density 10 different from density 20 on average?
  - Is the mean yield for density 10 different from density 20 at a particular value for variety?
- Does density affect yield differently for each variety?

For all of these questions, we want to know

- is there any effect and
- if yes, what is the nature of the effect.

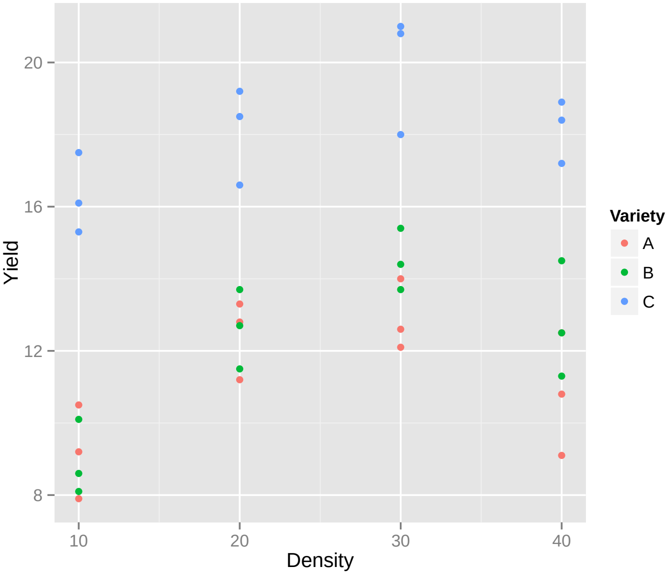
Hypotheses

- Does variety affect mean yield?
  - Is the mean yield for variety A different from B on average?
  - Is the mean yield for variety A different from B at a particular value for density?
- Does density affect mean yield?
  - Is the mean yield for density 10 different from density 20 on average?
  - Is the mean yield for density 10 different from density 20 at a particular value for variety?
- Does density affect yield differently for each variety?

For all of these questions, we want to know

- is there any effect and
- if yes, what is the nature of the effect.

Confidence intervals can answer these questions.


20
16
12
8
Yield
10
20
30
40
Density
Variety
A
B
C

Summary statistics

Number of replicates

```
  Variety 10 20 30 40
1       A  3  3  3  3
2       B  3  3  3  3
3       C  3  3  3  3
```

Mean Yield

```
  Variety        10       20       30       40
1       A  9.200000 12.43333 12.90000 10.80000
2       B  8.933333 12.63333 14.50000 12.76667
3       C 16.300000 18.10000 19.93333 18.16667
```

Standard deviation of yield

```
  Variety       10       20        30        40
1       A 1.300000 1.096966 0.9848858 1.7000000
2       B 1.040833 1.101514 0.8544004 1.6165808
3       C 1.113553 1.345362 1.6772994 0.8736895
```

Two-way ANOVA

- Setup: Two categorical explanatory variables with I and J levels

Two-way ANOVA

- Setup: Two categorical explanatory variables with I and J levels
- Model:

$$Y_{ijk} \stackrel{ind}{\sim} N(\mu_{ij}, \sigma^2)$$

  where Y_{ijk} is the

Two-way ANOVA

- Setup: Two categorical explanatory variables with I and J levels
- Model:

$$Y_{ijk} \stackrel{ind}{\sim} N(\mu_{ij}, \sigma^2)$$

  where Y_{ijk} is the
  - kth observation at the

Two-way ANOVA

- Setup: Two categorical explanatory variables with I and J levels
- Model:
$$Y_{ijk} \stackrel{ind}{\sim} N(\mu_{ij}, \sigma^2)$$
  where Y_{ijk} is the
  - kth observation at the
  - ith level of variable 1 (variety) with $i = 1, \ldots, \mathrm{I}$ and the

Two-way ANOVA

- Setup: Two categorical explanatory variables with I and J levels
- Model:

$$Y_{ijk} \stackrel{ind}{\sim} N(\mu_{ij}, \sigma^2)$$

  where Y_{ijk} is the
  - kth observation at the
  - ith level of variable 1 (variety) with $i = 1, \ldots, \mathrm{I}$ and the
  - jth level of variable 2 (density) with $j = 1, \ldots, \mathrm{J}$.

Two-way ANOVA

- Setup: Two categorical explanatory variables with I and J levels
- Model:

$$Y_{ijk} \stackrel{ind}{\sim} N(\mu_{ij}, \sigma^2)$$

  where Y_{ijk} is the
  - kth observation at the
  - ith level of variable 1 (variety) with $i = 1, \ldots, \mathrm{I}$ and the
  - jth level of variable 2 (density) with $j = 1, \ldots, \mathrm{J}$.

Two-way ANOVA

- Setup: Two categorical explanatory variables with I and J levels
- Model:

$$Y_{ijk} \stackrel{ind}{\sim} N(\mu_{ij}, \sigma^2)$$

  where Y_{ijk} is the
  - kth observation at the
  - ith level of variable 1 (variety) with $i = 1, \ldots, \mathrm{I}$ and the
  - jth level of variable 2 (density) with $j = 1, \ldots, \mathrm{J}$.

  Consider the models:
  - Additive: $\mu_{ij} = \mu + \alpha_i + \eta_j$

Two-way ANOVA

- Setup: Two categorical explanatory variables with I and J levels
- Model:

$$Y_{ijk} \stackrel{ind}{\sim} N(\mu_{ij}, \sigma^2)$$

  where Y_{ijk} is the
  - kth observation at the
  - ith level of variable 1 (variety) with $i = 1, \ldots, \mathrm{I}$ and the
  - jth level of variable 2 (density) with $j = 1, \ldots, \mathrm{J}$.

  Consider the models:
  - Additive: $\mu_{ij} = \mu + \alpha_i + \eta_j$
  - Cell-means: $\mu_{ij} = \mu + \alpha_i + \eta_j + \gamma_{ij}$

Two-way ANOVA

- Setup: Two categorical explanatory variables with I and J levels
- Model:
$$Y_{ijk} \stackrel{ind}{\sim} N(\mu_{ij}, \sigma^2)$$
  where Y_{ijk} is the
  - kth observation at the
  - ith level of variable 1 (variety) with $i = 1, \ldots, \mathrm{I}$ and the
  - jth level of variable 2 (density) with $j = 1, \ldots, \mathrm{J}$.

  Consider the models:
  - Additive: $\mu_{ij} = \mu + \alpha_i + \eta_j$
  - Cell-means: $\mu_{ij} = \mu + \alpha_i + \eta_j + \gamma_{ij}$

Two-way ANOVA

- Setup: Two categorical explanatory variables with I and J levels
- Model:

$$Y_{ijk} \stackrel{ind}{\sim} N(\mu_{ij}, \sigma^2)$$

  where Y_{ijk} is the
  - kth observation at the
  - ith level of variable 1 (variety) with $i = 1, \ldots, \mathrm{I}$ and the
  - jth level of variable 2 (density) with $j = 1, \ldots, \mathrm{J}$.

Consider the models:

- Additive: $\mu_{ij} = \mu + \alpha_i + \eta_j$
- Cell-means: $\mu_{ij} = \mu + \alpha_i + \eta_j + \gamma_{ij}$

	10	20	30	40
A	μ_{11}	μ_{12}	μ_{13}	μ_{14}
B	μ_{21}	μ_{22}	μ_{23}	μ_{24}
C	μ_{31}	μ_{32}	μ_{33}	μ_{34}

As a regression model

1. Assign a reference level for both variety (C) and density (40).

As a regression model

1. Assign a reference level for both variety (C) and density (40).
2. Let V_i and D_i be the variety and density for observation i.

As a regression model

1. Assign a reference level for both variety (C) and density (40).
2. Let V_i and D_i be the variety and density for observation i.
3. Build indicator variables, e.g. $\mathrm{I}(V_i = A)$ and $\mathrm{I}(D_i = 10)$.

As a regression model

1. Assign a reference level for both variety (C) and density (40).
2. Let V_i and D_i be the variety and density for observation i.
3. Build indicator variables, e.g. $\mathrm{I}(V_i = A)$ and $\mathrm{I}(D_i = 10)$.
4. The additive model:

$$\begin{aligned}\mu_{ij} = \quad & \beta_0 + \beta_1 \mathrm{I}(V_i = A) + \beta_2 \mathrm{I}(V_i = B) \\ & + \beta_3 \mathrm{I}(D_i = 10) + \beta_4 \mathrm{I}(D_i = 20) + \beta_5 \mathrm{I}(D_i = 30).\end{aligned}$$

β_1 is the expected difference in yield between varieties A and C at any fixed density

As a regression model

1. Assign a reference level for both variety (C) and density (40).
2. Let V_i and D_i be the variety and density for observation i.
3. Build indicator variables, e.g. $\mathrm{I}(V_i = A)$ and $\mathrm{I}(D_i = 10)$.
4. The additive model:

$$\begin{aligned}\mu_{ij} = \quad & \beta_0 + \beta_1 \mathrm{I}(V_i = A) + \beta_2 \mathrm{I}(V_i = B) \\ & + \beta_3 \mathrm{I}(D_i = 10) + \beta_4 \mathrm{I}(D_i = 20) + \beta_5 \mathrm{I}(D_i = 30).\end{aligned}$$

   β_1 is the expected difference in yield between varieties A and C at any fixed density

5. The cell-means model:

$$\begin{aligned}\mu_{ij} = \quad & \beta_0 + \beta_1 \mathrm{I}(V_i = A) + \beta_2 \mathrm{I}(V_i = B) \\ & + \beta_3 \mathrm{I}(D_i = 10) + \beta_4 \mathrm{I}(D_i = 20) + \beta_5 \mathrm{I}(D_i = 30) \\ \\ & + \beta_6 \mathrm{I}(V_i = A)\mathrm{I}(D_i = 10) + \beta_7 \mathrm{I}(V_i = A)\mathrm{I}(D_i = 20) + \beta_8 \mathrm{I}(V_i = A)\mathrm{I}(D_i = 30) \\ & + \beta_9 \mathrm{I}(V_i = B)\mathrm{I}(D_i = 10) + \beta_{10} \mathrm{I}(V_i = B)\mathrm{I}(D_i = 20) + \beta_{11} \mathrm{I}(V_i = B)\mathrm{I}(D_i = 30)\end{aligned}$$

   β_1 is the expected difference in yield between varieties A and C at a density of 40

ANOVA Table

ANOVA Table - Additive model

Source	SS	df	MS	F
Factor A	SSA	(I-1)	SSA/(I-1)	MSA/MSE
Factor B	SSB	(J-1)	SSB/(J-1)	MSB/MSE
Error	SSE	n-I-J+1	SSE/(n-I-J+1)	
Total	SST	n-1		

ANOVA Table

ANOVA Table - Additive model

Source	SS	df	MS	F
Factor A	SSA	(I-1)	SSA/(I-1)	MSA/MSE
Factor B	SSB	(J-1)	SSB/(J-1)	MSB/MSE
Error	SSE	n-I-J+1	SSE/(n-I-J+1)	
Total	SST	n-1		

ANOVA Table - Cell-means model

Source	SS	df	MS	
Factor A	SSA	I-1	SSA/(I-1)	MSA/MSE
Factor B	SSB	J-1	SSB/(J-1)	MSB/MSE
Interaction AB	SSAB	(I-1)(J-1)	SSAB /(I-1)(J-1)	MSAB/MSE
Error	SSE	n-IJ	SSE/(n-IJ)	
Total	SST	n-1		

Additive vs cell-means

Opinions differ on whether to use an additive vs a cell-means model when the interaction is not significant.

Additive vs cell-means

Opinions differ on whether to use an additive vs a cell-means model when the interaction is not significant. Remember that an insignificant test does not prove that there is no interaction.

Additive vs cell-means

Opinions differ on whether to use an additive vs a cell-means model when the interaction is not significant. Remember that an insignificant test does not prove that there is no interaction.

	Additive	Cell-means
Interpretation	Direct	Complicated
Estimate of σ^2	Biased	Unbiased

Additive vs cell-means

Opinions differ on whether to use an additive vs a cell-means model when the interaction is not significant. Remember that an insignificant test does not prove that there is no interaction.

	Additive	Cell-means
Interpretation	Direct	Complicated
Estimate of σ^2	Biased	Unbiased

We will continue using the cell-means model to answer the scientific questions of interest.

Two-way ANOVA using PROC GLM

```
DATA tomato;
  INFILE 'Ch13-tomato.csv' DSD FIRSTOBS=2;
  INPUT variety $ density yield;

PROC GLM DATA=tomato PLOTS=all;
  CLASS variety density;
  MODEL yield = variety|density / SOLUTION;
  LSMEANS variety / cl adjust=tukey;
  LSMEANS density / cl adjust=tukey;
  LSMEANS variety*density / cl adjust=tukey;
  RUN;
```

Two-way ANOVA using PROC GLM

```
                                   The GLM Procedure

Dependent Variable: yield

                                               Sum of
       Source                      DF         Squares     Mean Square    F Value    Pr > F
       Model                       11     422.3155556      38.3923232      24.22    <.0001
       Error                       24      38.0400000       1.5850000
       Corrected Total             35     460.3555556

                      R-Square     Coeff Var      Root MSE    yield Mean
                      0.917368      9.064568      1.258968      13.88889

       Source                      DF       Type I SS     Mean Square    F Value    Pr > F
       variety                      2     327.5972222     163.7986111     103.34    <.0001
       density                      3      86.6866667      28.8955556      18.23    <.0001
       variety*density              6       8.0316667       1.3386111       0.84    0.5484

       Source                      DF     Type III SS     Mean Square    F Value    Pr > F
       variety                      2     327.5972222     163.7986111     103.34    <.0001
       density                      3      86.6866667      28.8955556      18.23    <.0001
       variety*density              6       8.0316667       1.3386111       0.84    0.5484
```

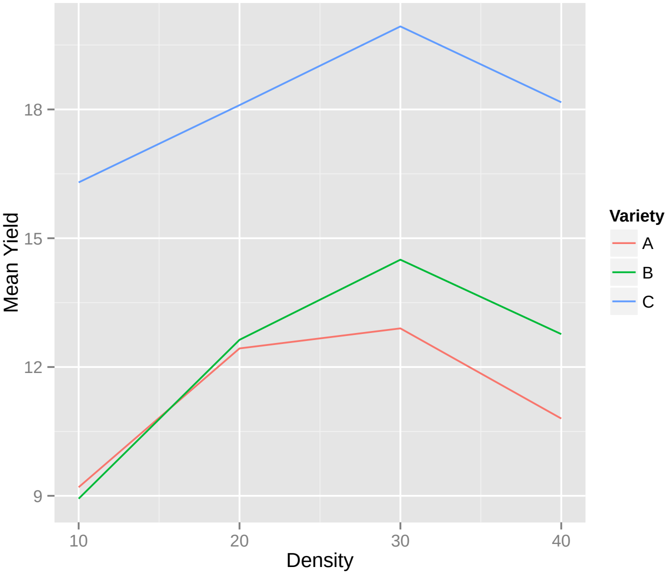
The Type I and Type III SS are equal because the design is balanced.

Two-way ANOVA using PROC GLM

```
MODEL yield = variety|density / SOLUTION;
                                   The GLM Procedure

                                                  Standard
        Parameter                  Estimate          Error    t Value    Pr > |t|

        Intercept               18.16666667 B     0.72686542      24.99      <.0001
        variety           A     -7.36666667 B     1.02794293      -7.17      <.0001
        variety           B     -5.40000000 B     1.02794293      -5.25      <.0001
        variety           C      0.00000000 B      .                .          .
        density           10    -1.86666667 B     1.02794293      -1.82      0.0819
        density           20    -0.06666667 B     1.02794293      -0.06      0.9488
        density           30     1.76666667 B     1.02794293       1.72      0.0986
        density           40     0.00000000 B      .                .          .
        variety*density A 10     0.26666667 B     1.45373083       0.18      0.8560
        variety*density A 20     1.70000000 B     1.45373083       1.17      0.2537
        variety*density A 30     0.33333333 B     1.45373083       0.23      0.8206
        variety*density A 40     0.00000000 B      .                .          .
        variety*density B 10    -1.96666667 B     1.45373083      -1.35      0.1887
        variety*density B 20    -0.06666667 B     1.45373083      -0.05      0.9638
        variety*density B 30    -0.03333333 B     1.45373083      -0.02      0.9819
        variety*density B 40     0.00000000 B      .                .          .
        variety*density C 10     0.00000000 B      .                .          .
```



Mean Yield
18
15
12
9
10
20
30
40
Density
Variety
A
B
C

Is the mean yield for variety A different from B on average?

```
LSMEANS variety / cl adjust=tukey;
                                 Least Squares Means
                        Adjustment for Multiple Comparisons: Tukey

...

                          Least Squares Means for effect variety
                           Pr > |t| for H0: LSMean(i)=LSMean(j)

                                Dependent Variable: yield

                      i/j              1              2              3
                         1                       0.2249         <.0001
                         2        0.2249                        <.0001
                         3        <.0001         <.0001

                     variety     yield LSMEAN      95% Confidence Limits
                     A              11.333333       10.583245     12.083422
                     B              12.208333       11.458245     12.958422
                     C              18.125000       17.374912     18.875088

                           Least Squares Means for Effect variety
                                Difference          Simultaneous 95%
                                   Between       Confidence Limits for
                      i    j         Means        LSMean(i)-LSMean(j)
                      1    2     -0.875000       -2.158534      0.408534
                      1    3     -6.791667       -8.075201     -5.508132
                      2    3     -5.916667       -7.200201     -4.633132
```

Is the mean yield at density 10 different from density 20 on average?

```
LSMEANS density / cl adjust=tukey;
                                 Least Squares Means
                        Adjustment for Multiple Comparisons: Tukey
                                         ...

                      density    yield LSMEAN      95% Confidence Limits
                      10            11.477778       10.611650    12.343905
                      20            14.388889       13.522762    15.255016
                      30            15.777778       14.911650    16.643905
                      40            13.911111       13.044984    14.777238

                              Least Squares Means for Effect density
                                   Difference          Simultaneous 95%
                                      Between      Confidence Limits for
                       i    j           Means       LSMean(i)-LSMean(j)

                       1    2       -2.911111       -4.548299    -1.273923
                       1    3       -4.300000       -5.937188    -2.662812
                       1    4       -2.433333       -4.070521    -0.796145
                       2    3       -1.388889       -3.026077     0.248299
                       2    4        0.477778       -1.159410     2.114966
                       3    4        1.866667        0.229479     3.503855
```

Is mean yield different for particular combinations?

```
LSMEANS variety*density / cl adjust=tukey;
```

variety	density	yield LSMEAN	95% Confidence Limits	
A	10	9.200000	7.699824	10.700176
A	20	12.433333	10.933157	13.933510
A	30	12.900000	11.399824	14.400176
A	40	10.800000	9.299824	12.300176
B	10	8.933333	7.433157	10.433510
B	20	12.633333	11.133157	14.133510
B	30	14.500000	12.999824	16.000176
B	40	12.766667	11.266490	14.266843
C	10	16.300000	14.799824	17.800176
C	20	18.100000	16.599824	19.600176
C	30	19.933333	18.433157	21.433510
C	40	18.166667	16.666490	19.666843

Is mean yield different for particular combinations?

```
LSMEANS variety*density / cl adjust=tukey;

                    Least Squares Means for Effect variety*density

                          Difference         Simultaneous 95%
                             Between      Confidence Limits for
             i      j          Means        LSMean(i)-LSMean(j)
             1      2      -3.233333       -6.939704     0.473037
             1      3      -3.700000       -7.406371     0.006371
             1      4      -1.600000       -5.306371     2.106371
             1      5       0.266667       -3.439704     3.973037
             1      6      -3.433333       -7.139704     0.273037
             1      7      -5.300000       -9.006371    -1.593629
             1      8      -3.566667       -7.273037     0.139704
             1      9      -7.100000      -10.806371    -3.393629
             1     10      -8.900000      -12.606371    -5.193629
             1     11     -10.733333      -14.439704    -7.026963
             1     12      -8.966667      -12.673037    -5.260296
             2      3      -0.466667       -4.173037     3.239704
             2      4       1.633333       -2.073037     5.339704
             2      5       3.500000       -0.206371     7.206371
             2      6      -0.200000       -3.906371     3.506371
             2      7      -2.066667       -5.773037     1.639704
             2      8      -0.333333       -4.039704     3.373037
             2      9      -3.866667       -7.573037    -0.160296
             2     10      -5.666667       -9.373037    -1.960296
             2     11      -7.500000      -11.206371    -3.793629
             2     12      -5.733333       -9.439704    -2.026963
             3      4       2.100000       -1.606371     5.806371
             3      5       3.966667        0.260296     7.673037
             3      6       0.266667       -3.439704     3.973037
```

Summary

- Use LSMEANS to answer questions of scientific interest.

Summary

- Use LSMEANS to answer questions of scientific interest.
- Check model assumptions

Summary

- Use LSMEANS to answer questions of scientific interest.
- Check model assumptions
- Consider alternative models, e.g. treating density as continuous