DSA 8010 - one way analysis of variance (ANOVA)

Inference on several population means

Inference on several population means

Inference on several means

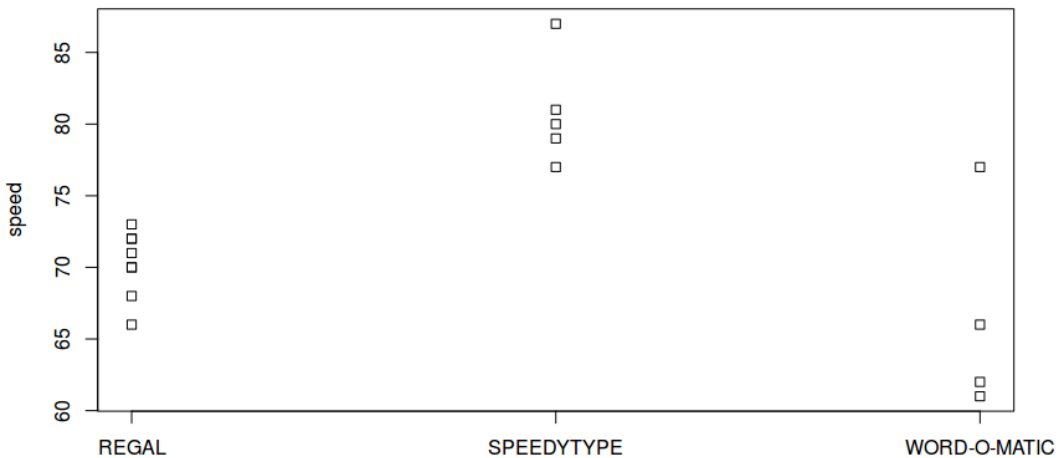
Method	types of variables
inference on two means	numeric & binary
inference on the difference between proportions	binary & binary
simple linear regression	numeric & numeric
one way ANOVA	numeric and categorical (more than 2 groups)

Descriptive analysis of numeric and categorical variables

- Make grouped boxplots and histograms.
- Compare summary statistics of the numeric variable across the groups.
- If the average values of the numeric variable are very different, relative to the sampling variability, there might be evidence of association between the numeric (response) variable and the categorical (grouping) variable.

Example (typing speeds)

Seventeen subjects were trained in typing using three different programs, "Regal," "Speedtype" and "Word-o-Matic." After completing the training, the subjects' typing speed was recorded (words per minute.) The speed data are shown below and are in Canvas under `typing.csv`.



Do the speeds appear to differ significantly across the three training programs?

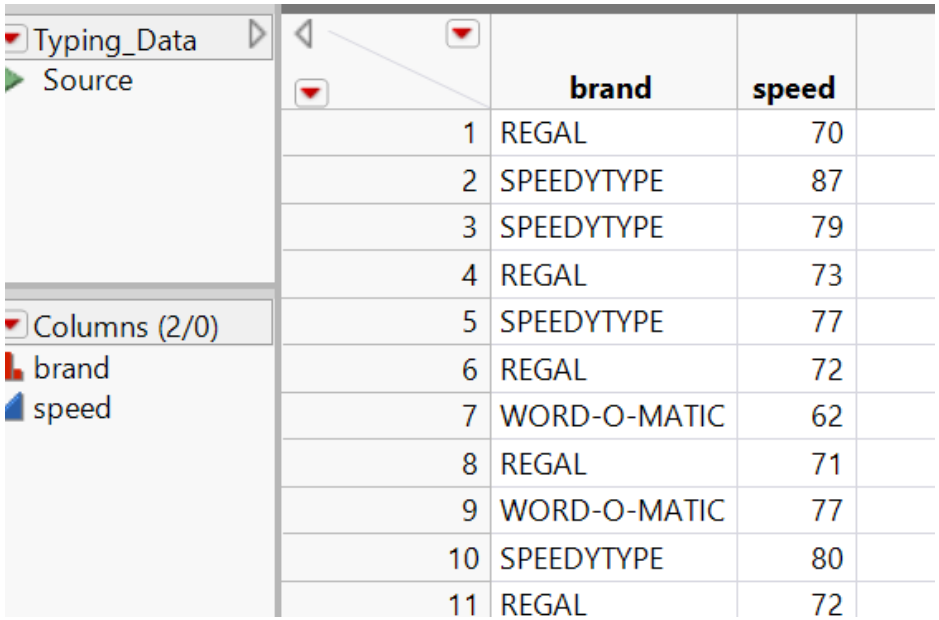
Example (typing speeds)

	average speed	standard deviation
REGAL	70.25	2.31
SPEEDYTYPE	80.80	3.77
WORD-O-MATIC	66.50	7.33

Do the speeds appear to differ significantly across the three training programs?

Inference on several population means: data and notation

Data. One numeric (response) variable and one categorical (grouping) variable. The grouping variable has at least three possible values.



	brand	speed
1	REGAL	70
2	SPEEDYTYPE	87
3	SPEEDYTYPE	79
4	REGAL	73
5	SPEEDYTYPE	77
6	REGAL	72
7	WORD-O-MATIC	62
8	REGAL	71
9	WORD-O-MATIC	77
10	SPEEDYTYPE	80
11	REGAL	72

We typically want to know: does the mean of response variable differ significantly across groups?

Inference on several population means: data and notation

Data. One numeric (response) variable and one categorical (grouping) variable.

Notation. y_{ij} : the jth measurement from group i.
t : the number of groups.
n_i : the number of observations in group i.
n_T : the total number of observations. $n_T = \sum_{i=1}^{t} n_i$.

Inference on several population means: data and notation

Data. One numeric (response) variable and one categorical (grouping) variable.

Notation. y_{ij} : the jth measurement from group i.
t : the number of groups.
n_i : the number of observations in group i.
n_T : the total number of observations. $n_T = \sum_{i=1}^{t} n_i$.

$$\bar{y}_{i.} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} \text{ : the mean of observations from group } i.$$

$$\bar{y}_{..} = \frac{\sum_{i=1}^{t} \sum_{j=1}^{n_i} y_{ij}}{n_T} \text{ : the mean of all observations.}$$

Inference on several population means: data and notation

Data. One quantitative (response) variable is measured for three or more groups.

Notation. y_{ij} : the jth measurement from group i.
t : the number of groups.
n_i : the number of observations in group i.
n_T : the total number of observations. $n_T = \sum_{i=1}^{t} n_i$.
$\bar{y}_{i\cdot}$: the sample mean of observations from group i.
$\bar{y}_{\cdot\cdot}$: the sample mean of all observations.

Statistical model. $y_{ij} = \mu_i + \epsilon_{ij}$, where the ϵ_{ij} are i.i.d. and approximately $N(0, \sigma^2)$.

Sources of variability

Recap: sample variance / standard deviation

The first tool we learned to measure the variability in a data set is the sample variance.

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

- Variability is captured by measuring squared differences between individual data points and the average.

Sums of squares

Total sum of squares. $TSS = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$.

Total variability in the sample.

Between-group sum of squares.

Within-group sum of squares/error sum of squares.

Sums of squares

Total sum of squares. $TSS = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$.

Total variability in the sample.

Between-group sum of squares. $SSB = \sum_{i=1}^{t} n_i (\bar{y}_{i.} - \bar{y}_{..})^2$

Variability among the group means.

Within-group sum of squares/error sum of squares.

Sums of squares

Total sum of squares. $TSS = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$.

Total variability in the sample.

Between-group sum of squares. $SSB = \sum_{i=1}^{t} n_i (\bar{y}_{i.} - \bar{y}_{..})^2$

Variability among the group means.

Within-group sum of squares/error sum of squares.
$SSW = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$

Left-over variability after accounting for the group means.

Sums of squares

Total sum of squares. $TSS = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$.

Total variability in the sample.

Between-group sum of squares. $SSB = \sum_{i=1}^{t} n_i (\bar{y}_{i.} - \bar{y}_{..})^2$

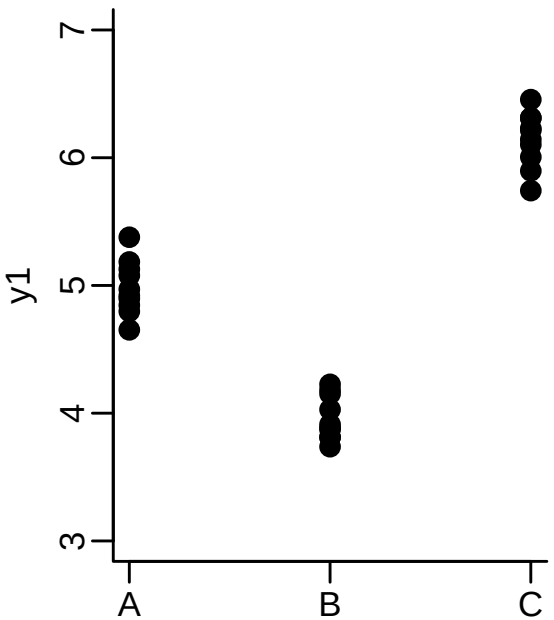
Variability among the group means.

Within-group sum of squares/error sum of squares.
$SSW = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$

Left-over variability after accounting for the group means.

Partitioning of sums of squares: $TSS = SSB + SSW$.

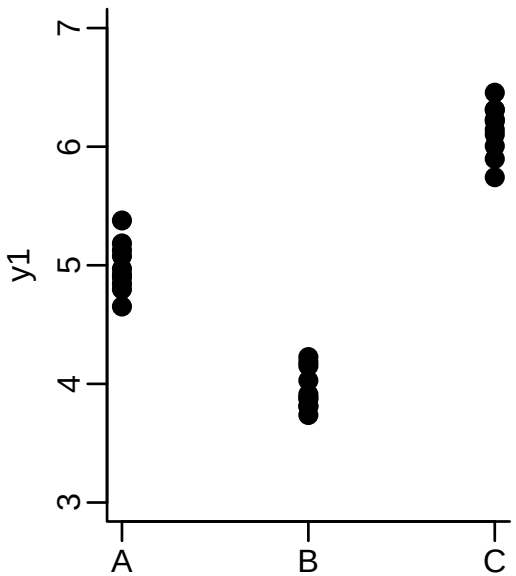
Sums of squares



Mean squares

Between-group mean squares. $s_B^2 = \frac{SSB}{t-1}$

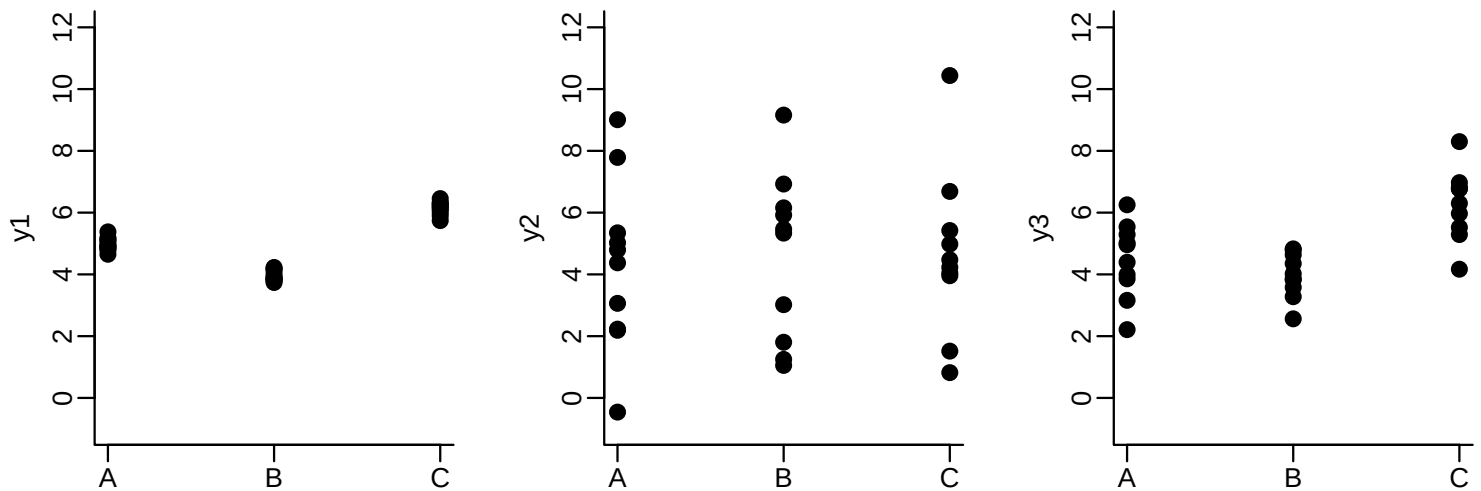
Within-group mean squares/error mean squares. $s_W^2 = \frac{SSW}{n_T - t}$



Mean squares

Between-group mean squares. $s_B^2 = \frac{SSB}{t-1}$

Within-group mean squares/error mean squares. $s_W^2 = \frac{SSW}{n_T - t}$



Are there significant differences among the group means?

Mean squares

Between-group mean squares. $s_B^2 = \frac{SSB}{t-1}$

Within-group mean squares/error mean squares. $s_W^2 = \frac{SSW}{n_T - t}$

- If the ratio $\frac{s_B^2}{s_W^2}$ is large, the across-group variability is large compared to the between group variability.
- Large values of $\frac{s_B^2}{s_W^2}$ provide evidence that the population means differ.

One-way ANOVA F test

Analysis of variance - F test

Hypotheses. $H_0 : \mu_1 = \mu_2 = \ldots = \mu_t$ (all population means equal)

H_A : Not all population means are equal. (At least one mean is different.)

Analysis of variance - F test

Hypotheses. $H_0 : \mu_1 = \mu_2 = \ldots = \mu_t$ (all population means equal)

H_A : Not all population means are equal

Test statistic. $f_0 = \frac{s_B^2}{s_W^2}$

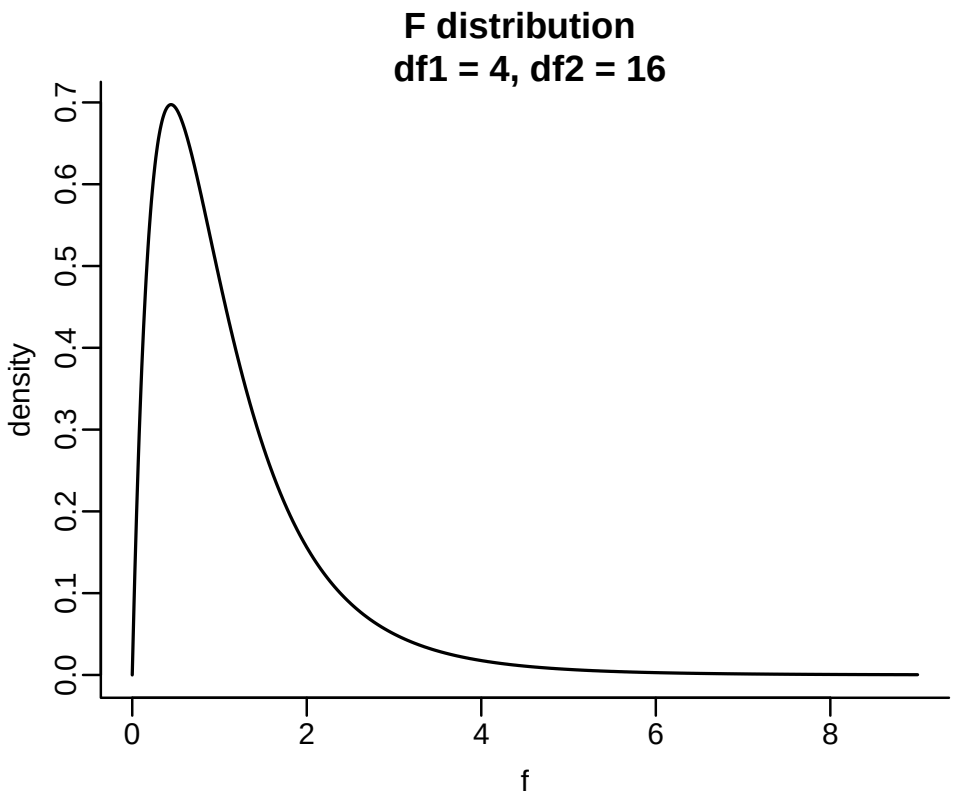
p-value and decision. Use the F distribution with df1 $= t - 1$ and df2 $= n_T - t$ to find the right-tail probability associated with f_0.

R code: `pf(f0, df1=t-1, df2=nT-t, lower.tail=FALSE)`.

The F distribution

Shape. Right-skewed, always positive.

Parameters. numerator degrees of freedom (df1) and denominator degrees of freedom (df2).



Analysis of variance - F test

Statistical model. $y_{ij} = \mu_i + \epsilon_{ij}$, where ϵ_{ij} are approx. $N(0, \sigma^2)$.

Hypotheses. $H_0 : \mu_1 = \mu_2 = \ldots = \mu_t$ (all population means equal)

H_A : Not all population means are equal

Test statistic. $f_0 = \frac{s_B^2}{s_W^2}$

p-value and decision. Use the F distribution with df1 $= t - 1$ and df2 $= n_T - t$ to find the right-tail probability associated with f_0.

If H_0 is rejected, the data show evidence of differences among at least two of the μ_i. This can be interpreted as evidence of an *association* between the grouping variable and the response.

If H_0 is not rejected, there is insufficient evidence to conclude that the group means differ.

The ANOVA table

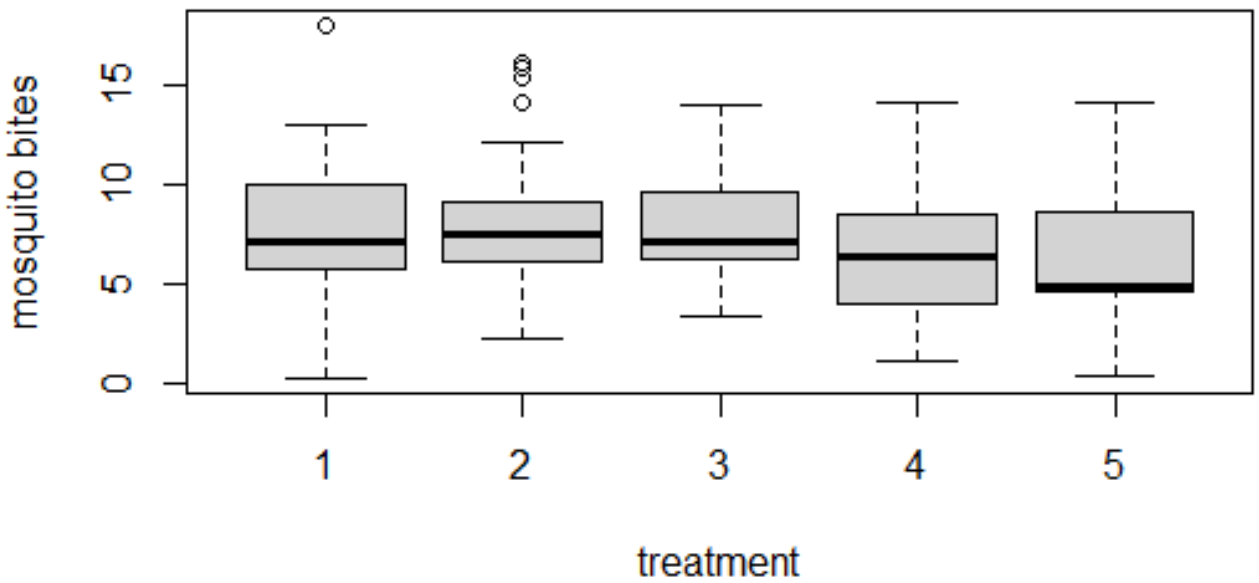
The ANOVA table is a standard method for presenting results of analysis of variance.

Source	Sum of squares	df	mean square	F	p-value
between groups (Model) within groups (Error, residual) total					

Example (mosquito repellents)

Five different mosquito repellents were tested on 150 subjects, who were randomly assigned to use one of the repellents. The number of mosquito bites in a pre-specified time period was measured. Is there a significant difference among the mean number of bites across the five treatments? Use $= 0.05$.

Here is a boxplot of the bite rates across the groups.



Example (mosquito repellents)

Is there a significant difference among the mean number of bites across the five treatments? Use $= 0.05$.

1. Write the null and alternative hypotheses.
2. Calculate the sums of squares.
3. Complete the ANOVA table.
4. Compare the p-value to and draw a conclusion.

Example (mosquito repellents)

The following sums of squares can be calculated (see R code):

between-group sum of squares $= SSB = 113.06$;
within-group sum of square $= SSW = 1466.55$

Example (mosquito repellents)

ANOVA table

Source	Sum of squares	df	mean square	F	p-value

Is there a significant difference among the mean number of bites across the five treatments? Use $= 0.05$.

Example (typing speeds)

Do the typing speeds differ significantly across the groups? Answer using the software output from the analysis of variance.

```
Analysis of Variance Table

Response: speed
          Df Sum Sq Mean Sq F value    Pr(>F)
brand      2 528.94 264.468  14.503 0.0003875 ***
Residuals 14 255.30  18.236
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
```

Residual analysis

Residual analysis

Notation: y_{ij} is the jth measurement from group i.

Statistical model for ANOVA: $y_{ij} = \mu_i + \epsilon_{ij}$, with ϵ_{ij} i.i.d. and approximately $N(0, \sigma^2)$.

Features of the model:
-equality of variances
-normality

Residual analysis

Statistical model for ANOVA: $y_{ij} = \mu_i + \epsilon_{ij}$, with ϵ_{ij} i.i.d. and approximately $N(0, \sigma^2)$.

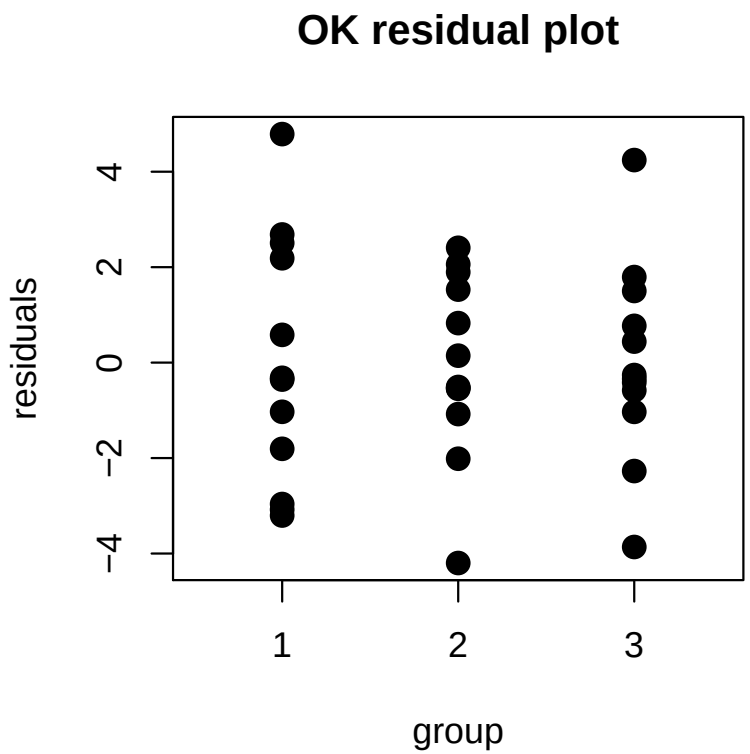
The ijth residual for one-way ANOVA is

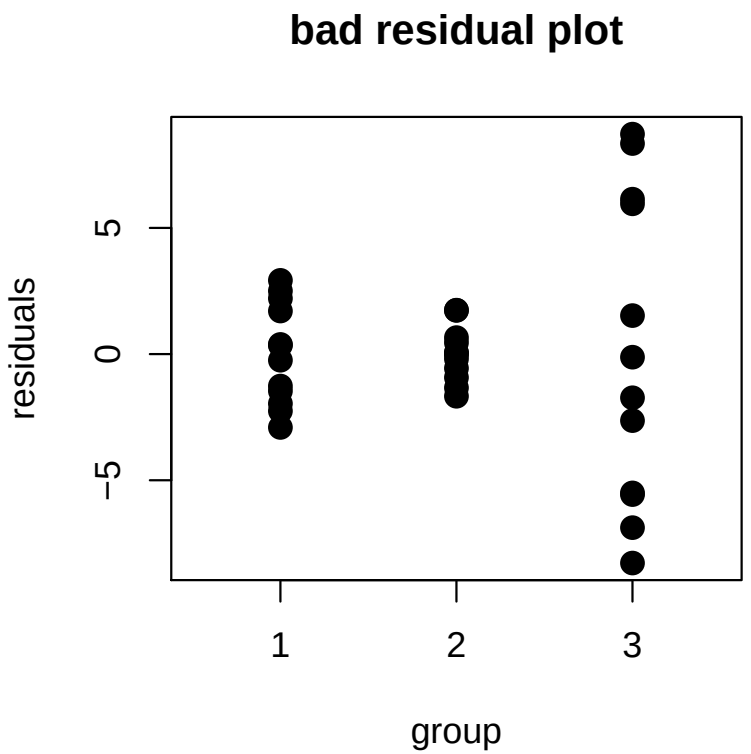
$$\hat{\epsilon}_{ij} = y_{ij} - \bar{y}_{i\cdot}$$

To assess the normality assumption. Make a normal quantile plot of the $\hat{\epsilon}_{ij}$. A straight-line pattern indicates approximate normality.

To assess the equal variances assumption. Plot the residuals by group. Visually investigate whether all of the groups have similar scatter around zero.

Residual by group plot





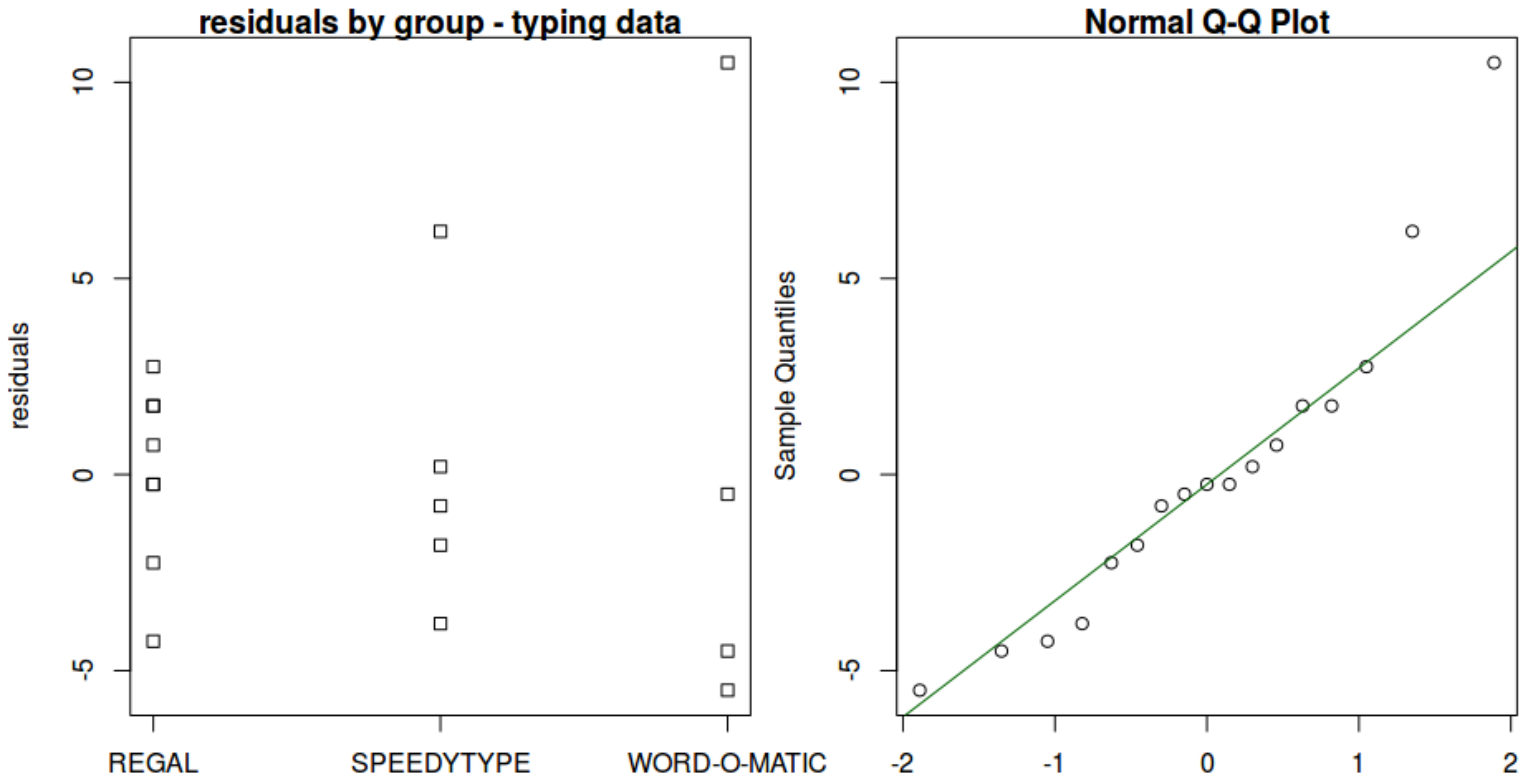
Transformations

If the modeling assumptions appear to be violated (i.e., the residuals appear to have different variances or are highly non-normal), consider transforming the variable before conducting ANOVA test.

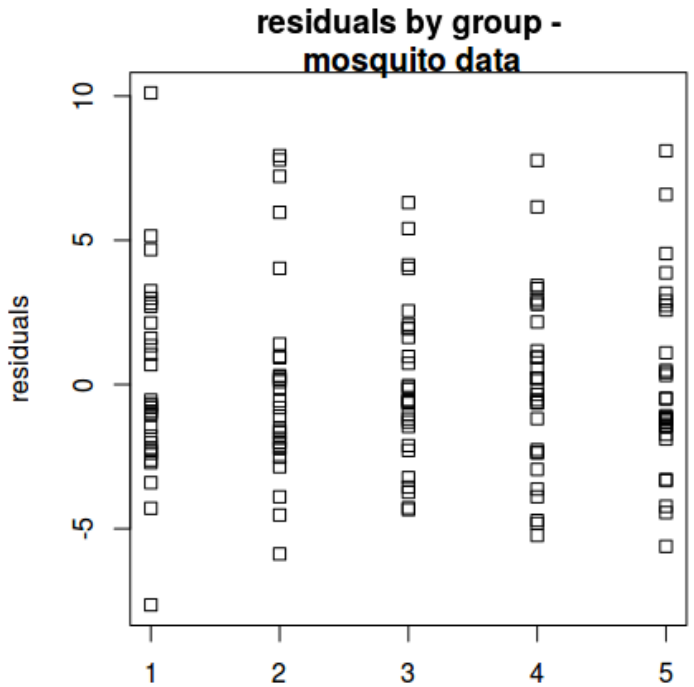
Some popular transformations:

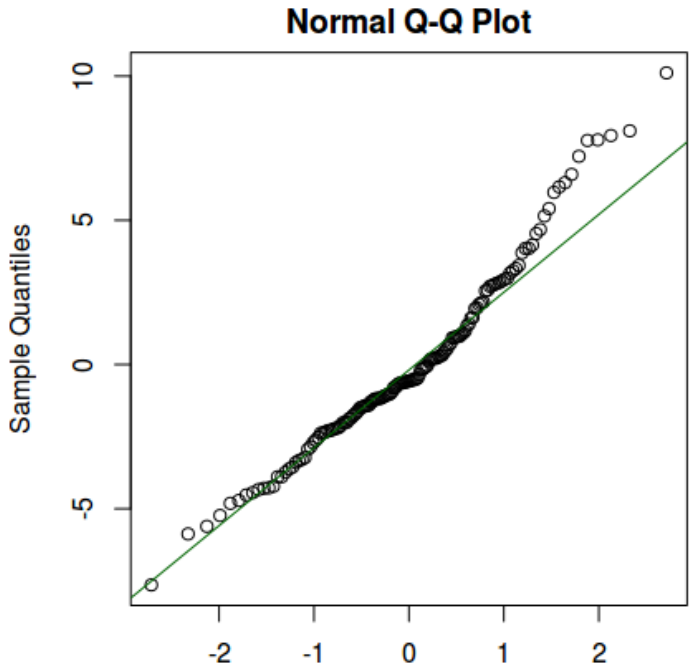
- For counts - $\sqrt{y}$
- For proportions/percentages - $arcsin(\sqrt{y})$

Typing speeds (residual analysis)



Mosquito (residual analysis)





Multiple comparisons

Example (mosquitos)

Five different mosquito repellents were tested on 150 subjects. The number of mosquito bites in a pre-specified time period was measured. Is there a significant difference among the mean number of bites across the five repellents? Use $= 0.05$.

```
> anova(anova.results)
Analysis of Variance Table

Response: Mosquito.bite.rate
                 Df  Sum Sq Mean Sq F value Pr(>F)
Treatment.group   4  113.06  28.265  2.7946 0.0284 *
Residuals       145 1466.55  10.114
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Error rates

Individual error rate. The probability of making a Type I error for a single hypothesis test.

Notation: α_I.

Experimentwise error rate. The probability of making 1 or more type I errors in a set of hypothesis tests.

Notation: α_E.

Error rates

Individual error rate. The probability of making a Type I error for a single hypothesis test.

Notation: $_I$.

Experimentwise error rate. The probability of making 1 or more type I errors in a set of hypothesis tests.

Notation: $_E$.

If you perform one test using $= 0.05$, you control the probability of type 1 error for one test; that is, the $_I$ is controlled to be 0.05.

Inflation of experimentwise error rate

Hypothesis tests. If we perform more than one test, each with individual error rate $\quad_I$, then $\quad_E$ will be $\geq \quad_I$.

Confidence intervals. If we create more than one confidence interval, each with confidence level of $(1 - \quad) * 100\%$ the "experimentwise confidence level" will be $\leq$ to $(1 - \quad_I) * 100\%$.

"Multiple comparison procedures" are statistical methods that allow us to perform several comparisons while controlling the experimentwise error rate at a pre-specified value $\quad_E$.

Bonferroni adjustment

The **Bonferroni adjustment** controls the experimentwise error rate to be no greater than by

- Using $_I = /m$ for each of m hypothesis tests.
- Using $(1 - _I) = (1 - /m)$ as the confidence level for each of m confidence intervals.

This procedure is very conservative. For moderate values of m, $_I$ becomes very small. This makes it harder to reject H_0 for each individual comparison; i.e., power is low.

Example: exam scores

A university offers four sections of Chem 101, each of which is taught by a different professor. An analysis of variance on their Exam 1 scores revealed the the scores differed significantly across sections. Professor Q has a reputation for being an easy grader, so the goal is to test the hypotheses

$$H_0 : \mu_4 - \mu_1 = 0; \quad H_a : \mu_4 - \mu_1 > 0$$
$$H_0 : \mu_4 - \mu_2 = 0; \quad H_a : \mu_4 - \mu_2 > 0$$
$$H_0 : \mu_4 - \mu_3 = 0; \quad H_a : \mu_4 - \mu_3 > 0$$

to evaluate if any of the sections had significantly lower scores than Professor Q's section. (Professor Q's section is labeled as $i = 4$.) What $_I$ should be used in order to use a Bonferroni adjustment to control the experimentwise error rate at $_E = 0.05$?

Tukey's honest significant difference (HSD)

Used for inference on "all pairwise comparisons," the differences $\mu_i - \mu_j$ for every combination of i and j.

- Tukey's HSD will simultaneously test $H_0 : \mu_i - \mu_j = 0$ for all $i \neq j$ or create confidence intervals for $\mu_i - \mu_j$ for all $i \neq j$.
- The confidence intervals have simultaneous confidence level of $(1 - \quad) \times 100\%$.
- The p-values are adjusted for multiple comparisons. Compare each given p-value to the desired $\quad_E$ to determine whether to reject H_0.

Example (mosquitos)

Five different mosquito repellents were tested on 150 subjects. The number of mosquito bites in a pre-specified time period was measured. Is there a significant difference among the mean number of bites across the five repellents? Use $= 0.05$.

```
> anova(anova.results)
Analysis of Variance Table

Response: Mosquito.bite.rate
                 Df  Sum Sq Mean Sq F value Pr(>F)
Treatment.group   4  113.06  28.265  2.7946 0.0284 *
Residuals       145 1466.55  10.114
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Example (mosquitos)

Which treatments (repellents) have significant differences in average number of bites? Use $E = 0.10$.

```
  Tukey multiple comparisons of means
    99% family-wise confidence level

Fit: aov(formula = lm_mosquito)

$Treatment.group
          diff       lwr       upr     p adj
2-1  0.2323333 -2.490866 2.9555322 0.9985833
3-1 -0.2346667 -2.957866 2.4885322 0.9985266
4-1 -1.5673333 -4.290532 1.1558655 0.3173749
5-1 -1.9003333 -4.623532 0.8228655 0.1462845
3-2 -0.4670000 -3.190199 2.2561988 0.9793812
4-2 -1.7996667 -4.522866 0.9235322 0.1887192
5-2 -2.1326667 -4.855866 0.5905322 0.0761721
4-3 -1.3326667 -4.055866 1.3905322 0.4852531
5-3 -1.6656667 -4.388866 1.0575322 0.2577218
5-4 -0.3330000 -3.056199 2.3901988 0.9942560
```

Example (mosquitos)

Find a confidence interval for the difference in average number of bites between group 4 and group 2. Use a 99% experimentwise confidence level.

```
  Tukey multiple comparisons of means
    99% family-wise confidence level

Fit: aov(formula = lm_mosquito)

$Treatment.group
          diff       lwr       upr     p adj
2-1  0.2323333 -2.490866 2.9555322 0.9985833
3-1 -0.2346667 -2.957866 2.4885322 0.9985266
4-1 -1.5673333 -4.290532 1.1558655 0.3173749
5-1 -1.9003333 -4.623532 0.8228655 0.1462845
3-2 -0.4670000 -3.190199 2.2561988 0.9793812
4-2 -1.7996667 -4.522866 0.9235322 0.1887192
5-2 -2.1326667 -4.855866 0.5905322 0.0761721
4-3 -1.3326667 -4.055866 1.3905322 0.4852531
5-3 -1.6656667 -4.388866 1.0575322 0.2577218
5-4 -0.3330000 -3.056199 2.3901988 0.9942560
```

Example (mosquitos)

Find a confidence interval for the difference in average number of bites between group 3 and group 5. Use a 99% experimentwise confidence level.

```
  Tukey multiple comparisons of means
    99% family-wise confidence level

Fit: aov(formula = lm_mosquito)

$Treatment.group
          diff       lwr       upr     p adj
2-1  0.2323333 -2.490866 2.9555322 0.9985833
3-1 -0.2346667 -2.957866 2.4885322 0.9985266
4-1 -1.5673333 -4.290532 1.1558655 0.3173749
5-1 -1.9003333 -4.623532 0.8228655 0.1462845
3-2 -0.4670000 -3.190199 2.2561988 0.9793812
4-2 -1.7996667 -4.522866 0.9235322 0.1887192
5-2 -2.1326667 -4.855866 0.5905322 0.0761721
4-3 -1.3326667 -4.055866 1.3905322 0.4852531
5-3 -1.6656667 -4.388866 1.0575322 0.2577218
5-4 -0.3330000 -3.056199 2.3901988 0.9942560
```