Forum 9
Regression: Categorical Predictor
Data Skills for Scientists

Inference

The 3 topics in the third module, 'Regression' are:

- One Numeric Predictor;
- Multiple Numeric Predictors; and
- Categorical Predictors.

In this forum we will cover the third topic, Categorical Predictors.

Key topics for Categorical Predictors.

- Analysis of Variance (ANOVA)
  - One-way ANOVA
  - ANOVA post-hoc testing
  - Two-way ANOVA
- Non-parametric alternatives to ANOVA, focusing on the one-way Kruskal-Wallis test.

One-way ANOVA

Multiple means

So far we have used the t-test to compare two population means. If we would like to compare any number of population means, we use **Analysis of Variance**.

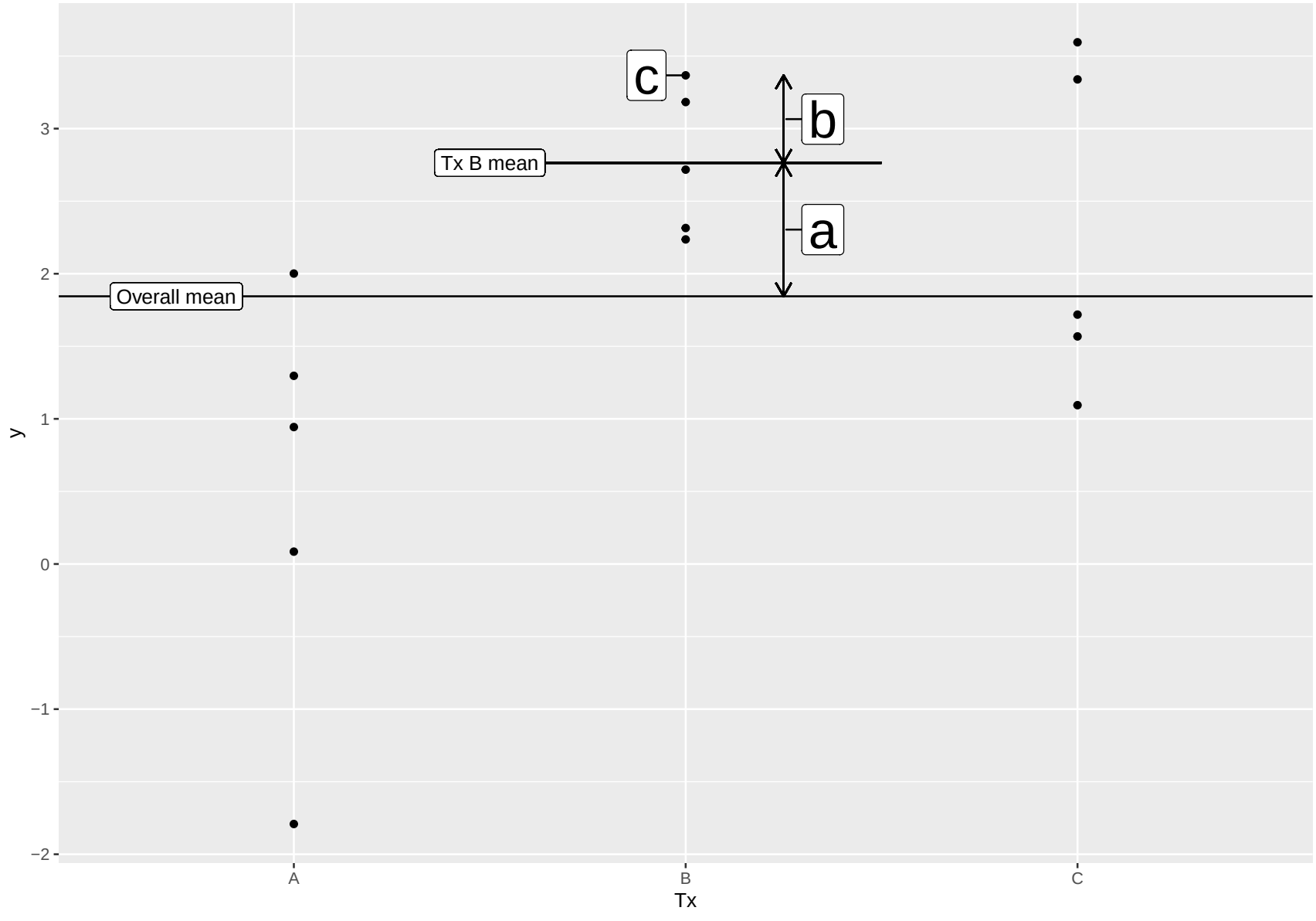
The idea of ANOVA

Analysis of variance (ANOVA) is the technique that is used to compare several means.

One-way ANOVA is used for situations in which there is only one way to classify the populations of interest.

Two-way ANOVA is used to analyse the effect of two factors.

The idea of ANOVA



ANOVA

We will split the total sum of squares (SST) of the observations compared to the overall mean into two parts:

- The sum of squares of the group sample means (SSM) compared to the overall mean.
- The sum of squares of the observations compared to the group sample means (SSE).

So that

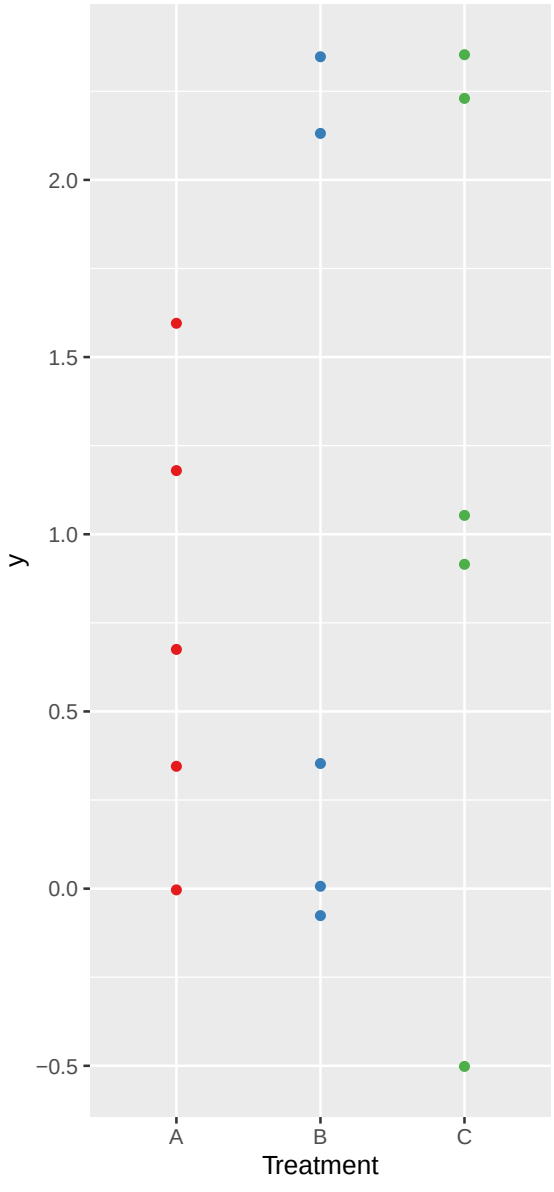
- $\text{SST} = \text{SSM} + \text{SSE}$.

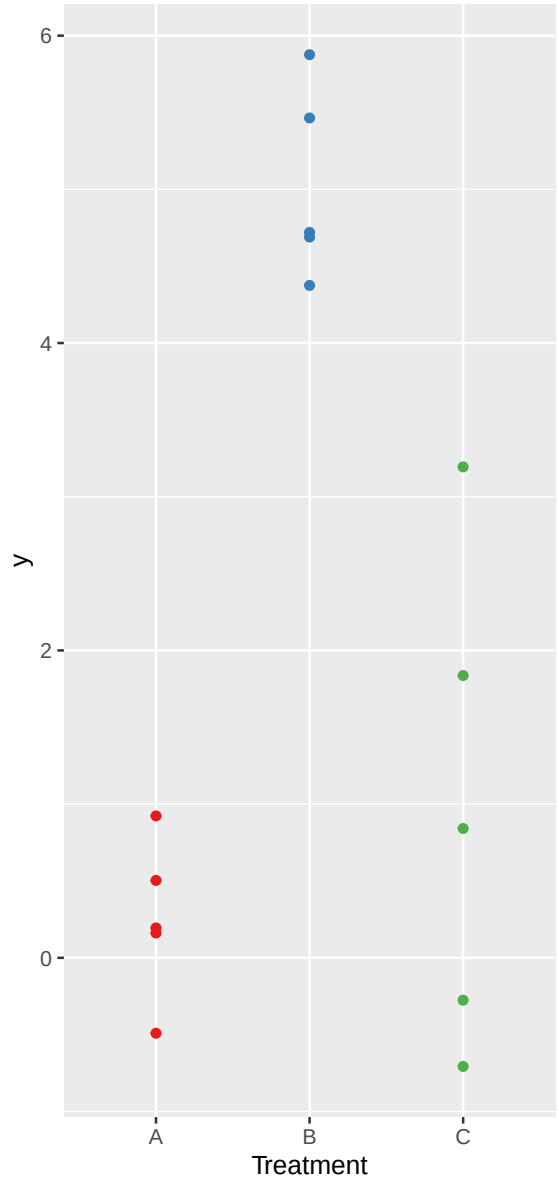
ANOVA maths

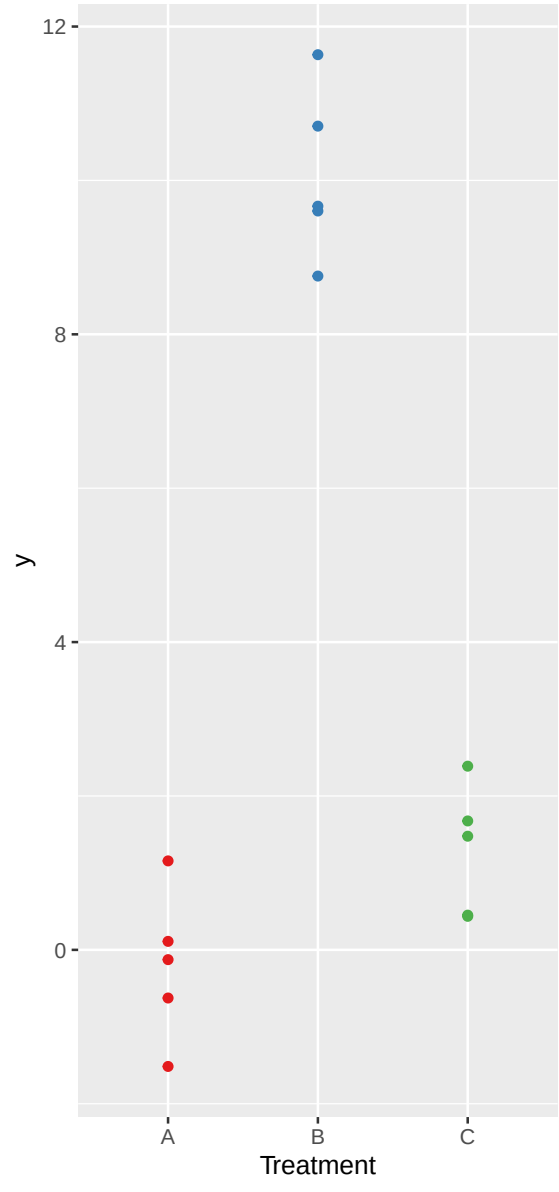
This can be expressed as

$$\sum_{k=1}^{K}\sum_{j=1}^{n_k}(y_{kj}-\bar{y}_{\bullet\bullet})^2 = \sum_{k=1}^{K} n_k(\bar{y}_{k\bullet}-\bar{y}_{\bullet\bullet})^2 + \sum_{k=1}^{K}\sum_{j=1}^{n_k}(y_{kj}-\bar{y}_{k\bullet})^2.$$

The idea of ANOVA







The idea of ANOVA

SSM	SSE	ratio
0.5132228	12.67270	0.0404983
66.0273086	12.70038	5.1988460
308.2892236	11.62813	26.5123636

Example

MPG

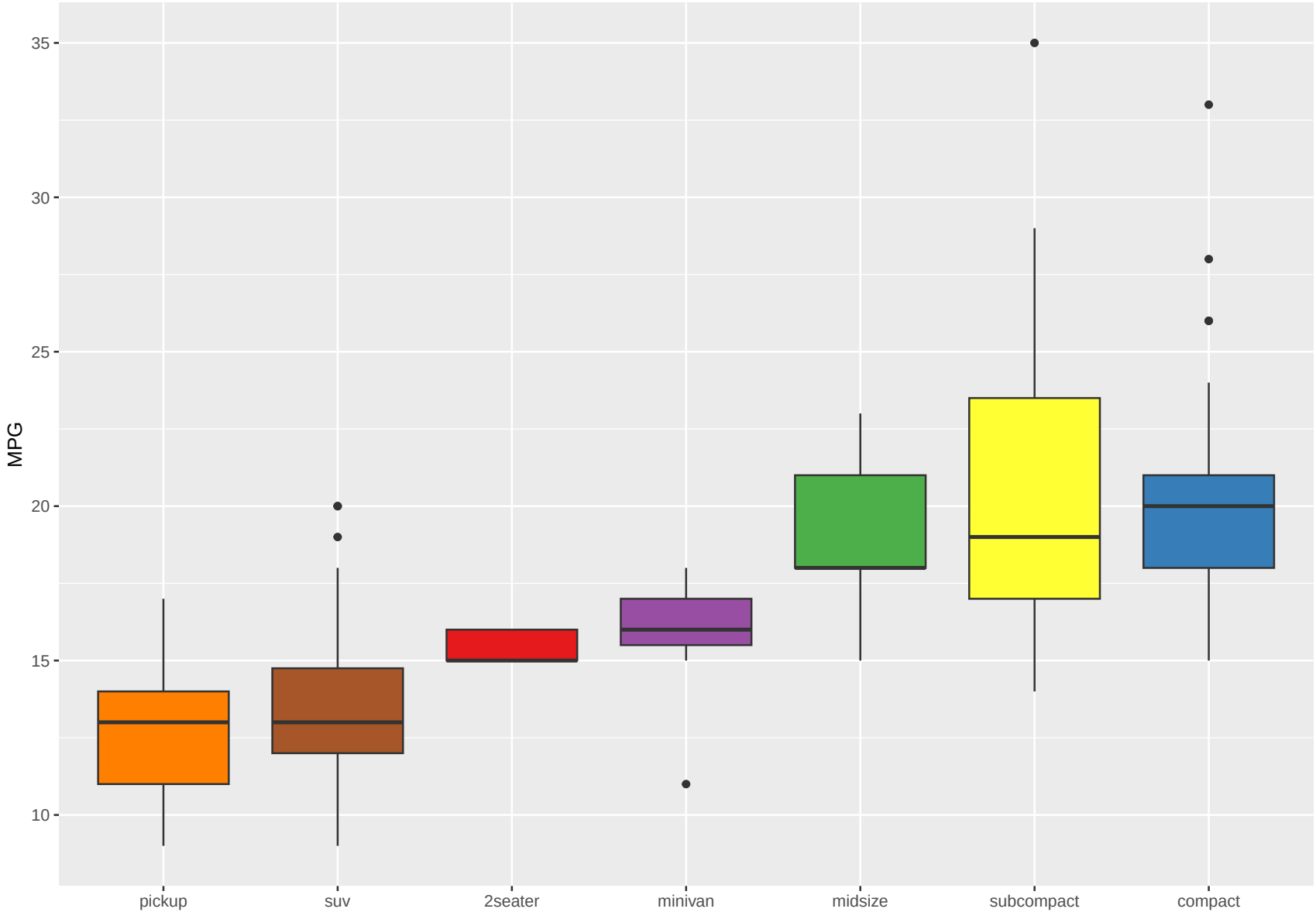
Which class of car has the best city mpg?

Response variable: petrol mileage (mpg) in the city.

Groups: vehicle classification

class	n
2seater	5
compact	47
midsize	41
minivan	11
pickup	33
subcompact	35
suv	62

Box-plots


MPG
35
30
25
20
15
10
pickup
suv
2seater
minivan
midsize
subcompact
compact

Example

Means

class	mean MPG
pickup	13.00
suv	13.50
2seater	15.40
minivan	15.82
midsize	18.76
compact	20.13
subcompact	20.37

Example

We could compare each *pair* of means to see if they are different, e.g.,

$$H_0 : \mu_{mid} = \mu_{pickup}$$

However, this gives rise to the problem of **multiple comparisons**

Problem of multiple comparisons

We have the problem of how to do many comparisons at the same time with some overall measure of confidence in all the conclusions.

Statistical methods for dealing with this problem usually have two steps:

- An **overall test** to find any differences among the parameters we want to compare.
- A detailed **follow-up analysis** to decide which groups differ and how large the differences are.

The one-way ANOVA model

The model is

$$x_{kj} = \mu_k + \epsilon_{kj},$$

where x_{kj} is the observation of the jth subject in the kth treatment group, μ_k is the population mean for the kth treatment group, and

$$\epsilon_{kj} \sim N(0, \sigma).$$

The ANOVA F-test

To test more than two population means, we use:

H_0 :$\mu_1 = \mu_2 = ... = \mu_K$,
H_a :at least one of the μ_k's is different from the rest.

The ANOVA F-statistic

The analysis of variance F-statistic for testing the equality of several means has this form:

$$F = \frac{\text{variation among the sample means}}{\text{variation among individuals in the same sample}}$$

The reference distribution

If the null hypothesis is true, then the test statistic F has a F-distribution, but unlike a t-distribution, has **two** degrees of freedom.

F-distributions

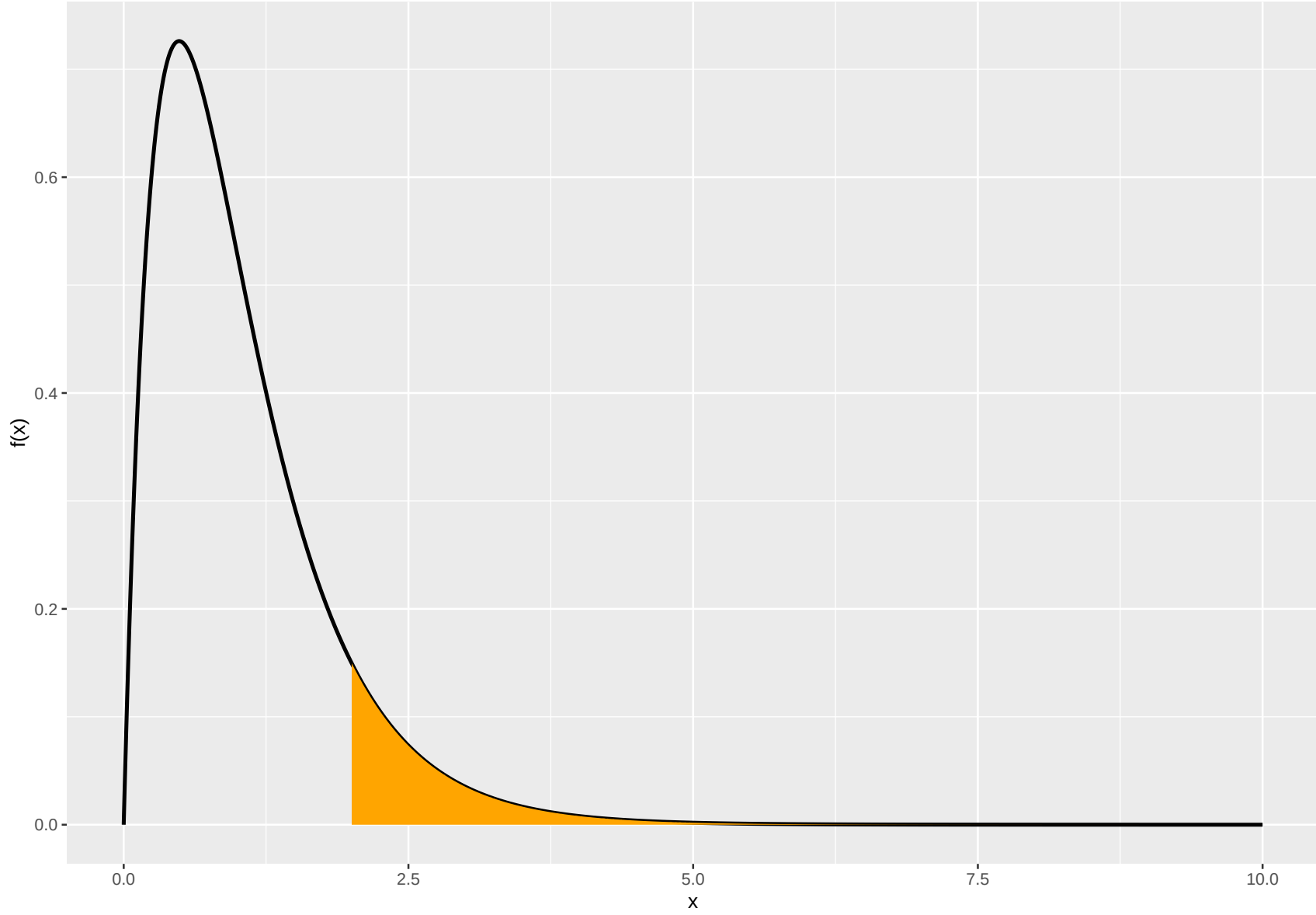
The F-distributions are a family of right-skewed distributions that take only values greater than 0. A specific F-distribution is determined by the degrees of freedom of the numerator and denominator of the F-statistic.

If we have K treatment groups and a total of N observations, then we use an

$$F_{K-1,N-K}$$

distribution.

F-distributions



Example: MPG data

- Write down the null and alternative hypotheses for the MPG example.
- Write down the distribution of the test statistic if the null hypothesis is true.

Example: MPG data

```
             Df Sum Sq Mean Sq F value Pr(>F)
class         6   2295   382.5    45.1 <2e-16 ***
Residuals   227   1925     8.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

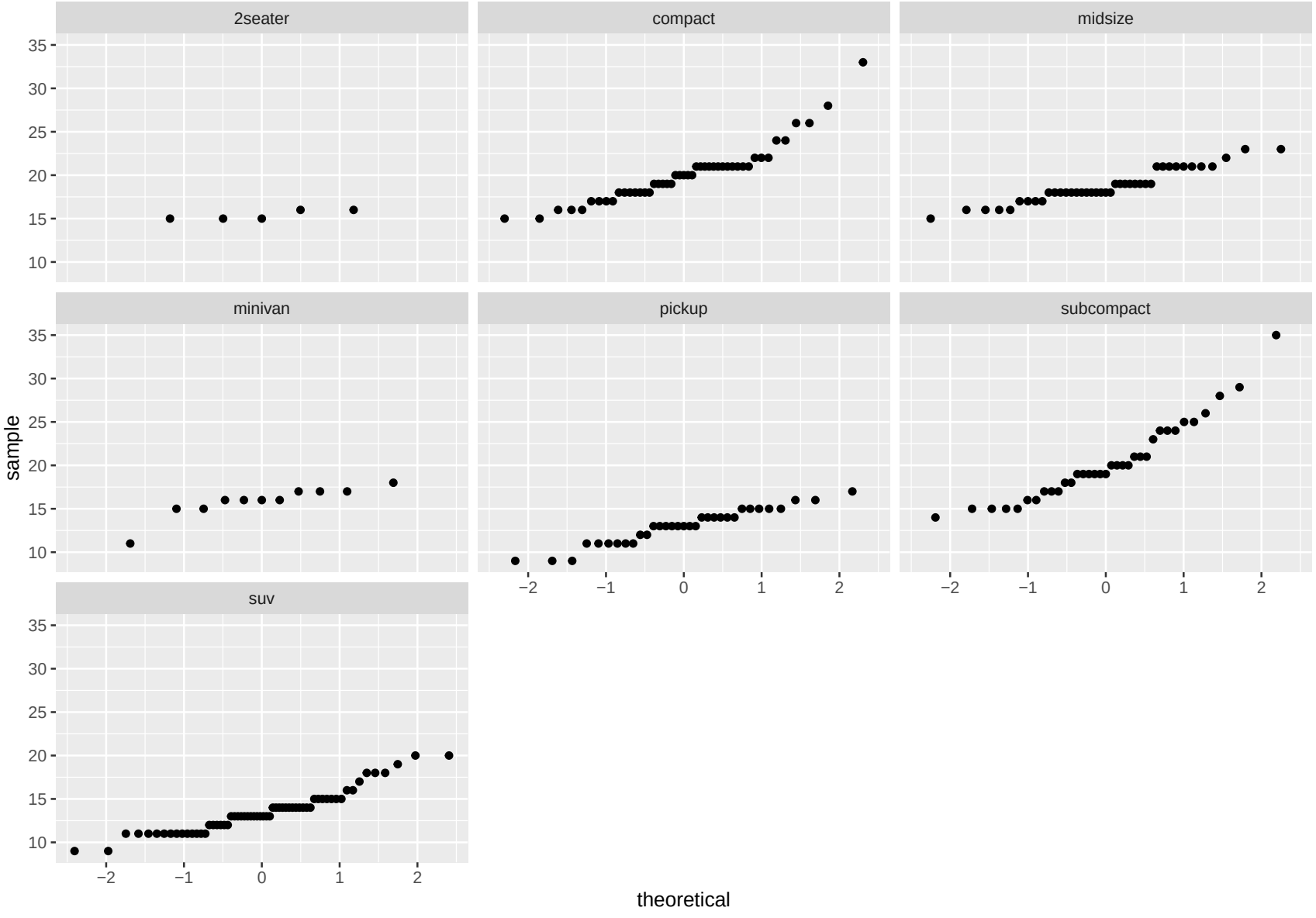
One-way ANOVA assumptions

- Normality of observations in each group.
- Constant variance of observations in each group.
- Independence of observations within each group.
- Independence of observations between each group.

Normality of observations in each group

- Look at normal quantile plot of observations in each group.
- Expect to see roughly linear points.

Example


2seater
compact
midsize
minivan
pickup
subcompact
suv
sample
theoretical
35
30
25
20
15
10
−2
−1
0
1
2

Constant variance of observations in each group

- Calculate the sample standard deviation for each group.
- Calculate

$$r = \frac{\max SD}{\min SD}.$$

- If $r < 2$, then the assumption is reasonable.

Example

class	n	mean	sd
2seater	77	15.40	0.55
compact	946	20.13	3.39
midsize	769	18.76	1.95
minivan	174	15.82	1.83
pickup	429	13.00	2.05
subcompact	713	20.37	4.60
suv	837	13.50	2.42

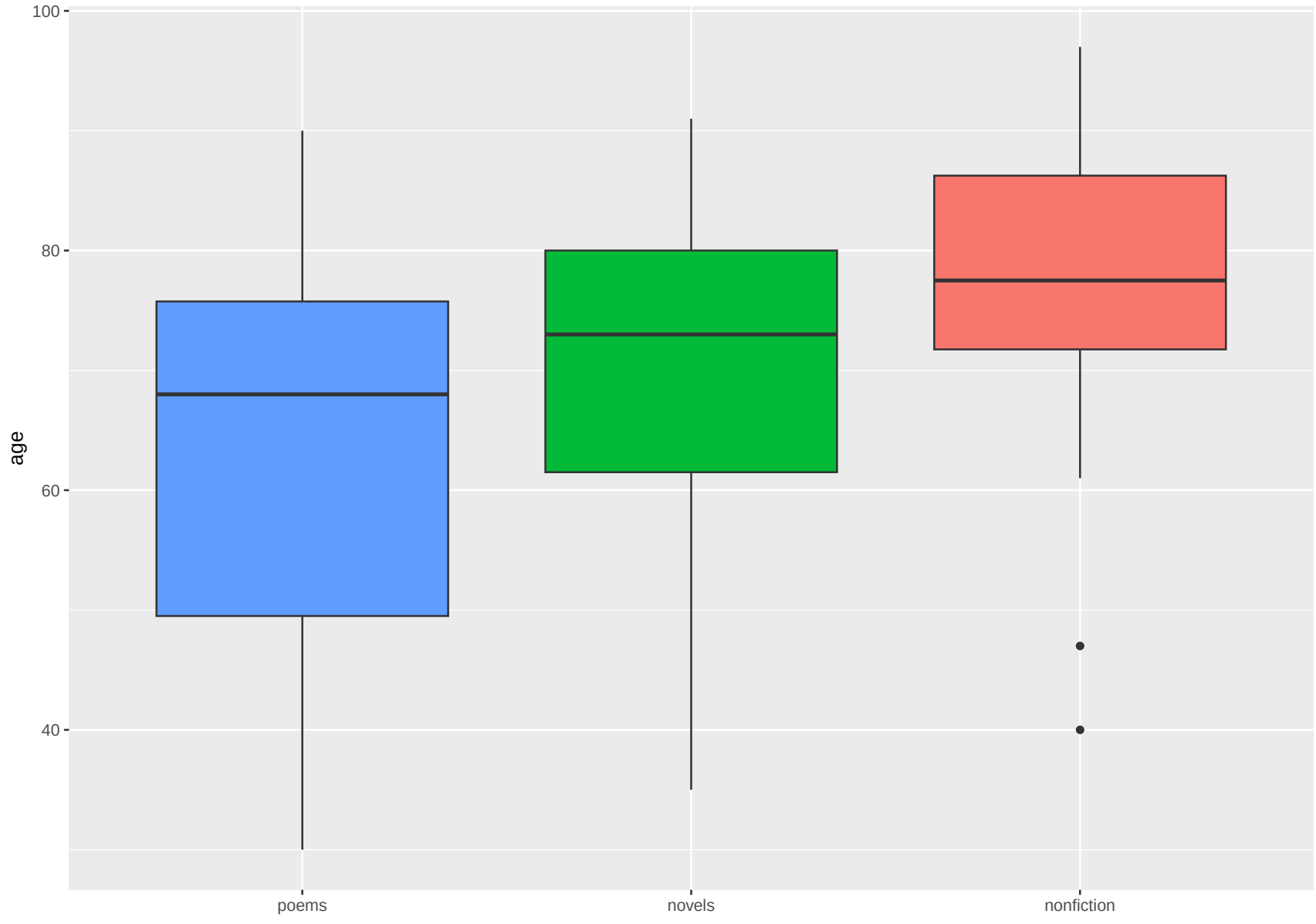
Independence

- Look at the experimental design.
- Looking for evidence that one observation does not give information about the other observations, e.g., random selection of subjects and random allocation to treatments.

Example

The following dataset examines whether there is a difference in lifespan for different female writers classified by the genre of their writing. For each female writer we have their age at death and the genre of writing they were classified as.

Boxplot



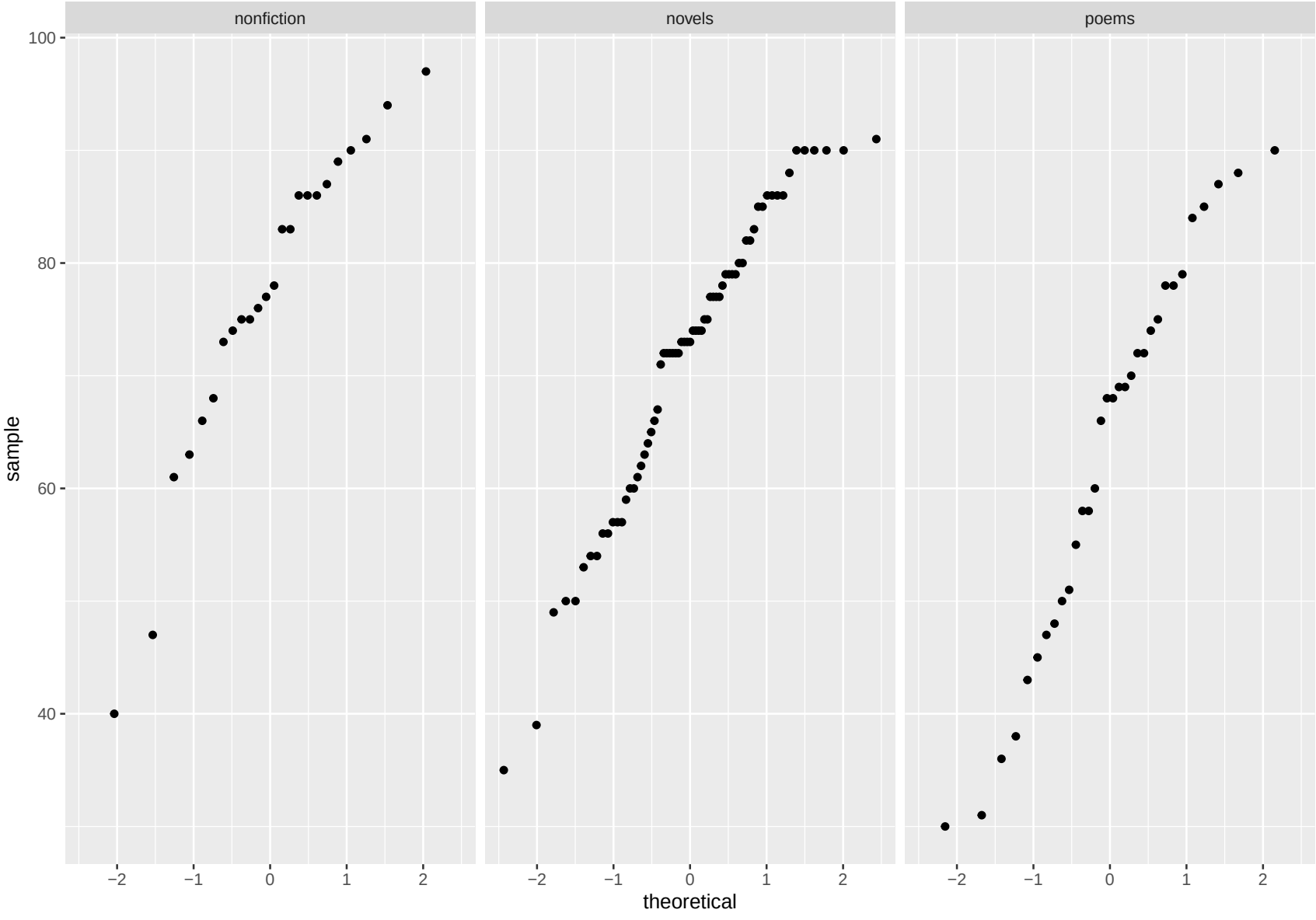
ANOVA output

```
             Df Sum Sq Mean Sq F value  Pr(>F)
type          2   2744  1372.1   6.563 0.00197 **
Residuals   120  25088   209.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary stats

type	n	mean	sd
nonfiction	1845	76.88	14.10
novels	4787	71.45	13.05
poems	2022	63.19	17.30

Normal QQ-plot



Multiple Comparisons

Multiple comparisons

You have calculated a P-value for your ANOVA test. Now what? If you have rejected the null hypothesis, you still need to determine the treatments effects.

- You can gain an insight by looking back at your plots, e.g. box-plots.
- There are several tests of statistical significance designed specifically for multiple tests. We will look at **multiple comparisons**.
- You can find the confidence interval for each mean μ_k shown to be significantly different from the others.

Multiple comparisons

Multiple comparisons should be used when there are no justified expectations. Those are **pairwise tests** of significance.

We compare gas mileage for types of vehicles. We have no prior knowledge to expect any type to perform differently from the rest. Pair-wise comparisons should be performed here, but only if an ANOVA test on all the types reached statistical significance first.

Multiple comparisons

Multiple comparison tests are variants on the two-sample t test.

- They use the pooled standard deviation S_p,
- the pooled degrees of freedom,
- and they compensate for the multiple comparisons.

How to do it

We compute the t-statistic for all pairs of means:

$$t_{ij} = \frac{\bar{x}_i - \bar{x}_j}{s_p\sqrt{1/n_i + 1/n_j}}.$$

We then use the value of t_{ij} to calculate a P-value. This P-value is calculated by adjusting for the multiple comparisons.

The pooled standard deviation

The pooled standard deviation is calculated by

$$S_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + ... + (n_k - 1)s_k^2}{n_1 + n_2 + ... + n_k - k}}.$$

The pooled degrees of freedom

The pooled degrees of freedom is

$$n_1 + n_2 + ... + n_k - k.$$

Example

class	mean	sd	n
2seater	15.40	0.55	5
compact	20.13	3.39	47
midsize	18.76	1.95	41
minivan	15.82	1.83	11
pickup	13.00	2.05	33
subcompact	20.37	4.60	35
suv	13.50	2.42	62

The Bonferroni correction

Using the Bonferroni correction, we calculate a new adjusted significance level by dividing the required significance level α by the number of comparisons made.

Example: Calculate the adjusted significance level for the MPG dataset if the significance level is 5%.

Simultaneous confidence intervals

We can also calculate simultaneous $C\%$ confidence intervals for all pairwise differences $(\mu_i - \mu_j)$ between population means:

$$(\bar{x}_i - \bar{x}_j) \pm t^*_{adj} s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

How to calculate t^*_{adj}

- Calculate the number of comparisons: m.
- Divide the significance level by m.
- Use this new significance level to calculate t^{**}.

Example

If there are 10 treatment levels and each level has 20 observations, calculate t^*_{adj}.

This is clearly extremely conservative! So we will use something that adjusts in a similar, but far less conservative way: A Tukey HSD test!

Example

	diff	lwr	upr	p adj
compact-2seater	4.728	0.652	8.803	0.012
midsize-2seater	3.356	-0.748	7.460	0.190
minivan-2seater	0.418	-4.255	5.091	1.000
pickup-2seater	-2.400	-6.558	1.758	0.605
subcompact-2seater	4.971	0.829	9.114	0.008
suv-2seater	-1.900	-5.928	2.128	0.800
midsize-compact	-1.372	-3.223	0.480	0.298
minivan-compact	-4.309	-7.211	-1.408	0.000
pickup-compact	-7.128	-9.095	-5.160	0.000
subcompact-compact	0.244	-1.691	2.178	1.000
suv-compact	-6.628	-8.303	-4.952	0.000
minivan-midsize	-2.938	-5.880	0.004	0.051
pickup-midsize	-5.756	-7.782	-3.730	0.000
subcompact-midsize	1.615	-0.379	3.609	0.199
suv-midsize	-5.256	-7.000	-3.512	0.000
pickup-minivan	-2.818	-5.835	0.198	0.084
subcompact-minivan	4.553	1.558	7.548	0.000
suv-minivan	-2.318	-5.153	0.516	0.190
subcompact-pickup	7.371	5.269	9.474	0.000
suv-pickup	0.500	-1.367	2.367	0.985
suv-subcompact	-6.871	-8.703	-5.040	0.000

Why is family wise error rate important?

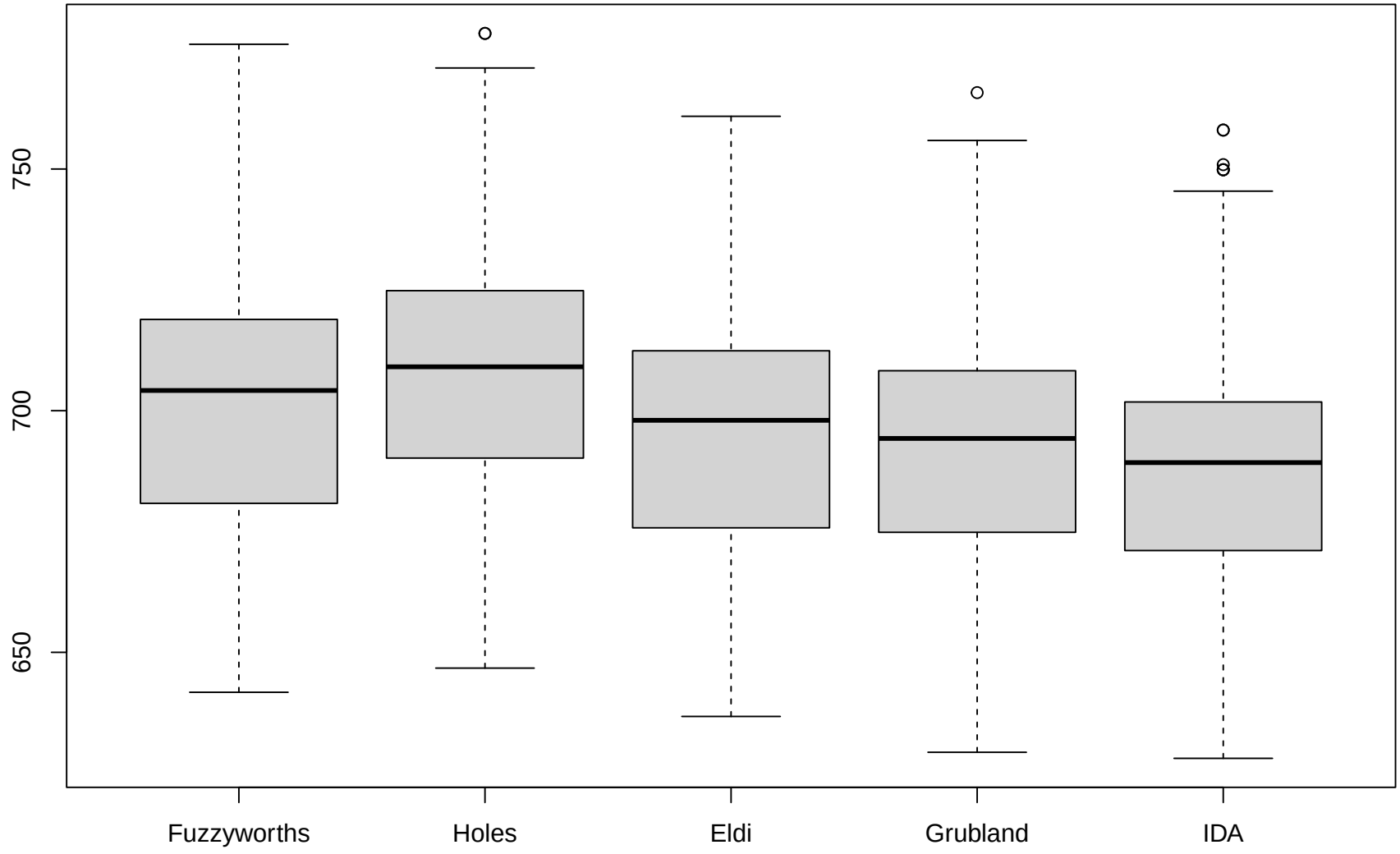
Consider the situation where you are a major supermarket... Lets call you Fuzzyworths.

You have four competitors within your area:

- Holes
- Eldi
- Grubland
- IDA

Every day, Fuzzyworths collects data on how many people go into each store.

Descriptive Statistics - Annual



Descriptive Statistics - Annual

Comparisons each week can lead to very different reulsts.

Comparing Fuzzyworths and Holes and each week:

- Significant 7.8% of the time

Comparing Fuzzyworths with the lowest group each week:

- Significant 19.9% of the time

Two-way ANOVA

Multiple means

We have already seen that, when we would like to compare more than two population means, we use **Analysis of Variance**.

We have only looked at how we do this with one categorical predictor in one-way ANOVA.

Now let's think about how we would analyse data that has **two** categorical variable.

Advantages of two-way ANOVA

More efficient: Running an ANOVA with two categorical predictors saves us from having to run two, separate, one-way ANOVAs (and increasing our family wise error rate).

Reduces residual variation: By including another predictor, we increase the amount of variability we can explain which reduces the amount of left over variability.

Investigate interactions: Sometimes the effect of one predictor can be different depending on whats happening with the other predictor. Two-way ANOVA can attempt to measure if this is happening or not.

Reducing residual variation in the real world

Response	Predictor 1	Predictor 2
Miles per gallon	Class of vehicle	
Age of lecturer	Gender	
Typing speed	Area of study	
Grade Point Average	Major	

Interactions in the real world

Response: GPA (Grade Point Average)

Predictor 1: Year at University (1^{st}, 2^{nd}, 3^{rd}, honours)

Predictor 2: Major (Psych, Bio, Maths)

Two-way ANOVA

For Two-Way ANOVA, we have **two** predictor variables so we will split the total sums of squares (SST) into four parts:

- The sum of squares of the **first predictor** variable group sample means (SSA) compared to the overall mean.
- The sum of squares of the **second predictor** variable group sample means (SSB) compared to the overall mean.
- The sum of squares of the **interaction** between the first and second predictor variable groups sample means (SSAB) compared to the overall mean.
- The sum of squares of the observations compared to the group sample means (SSE).

So that

$$\text{SST} = \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSE}.$$

The two-way ANOVA model

The model is

$$x_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \epsilon_{ijk},$$

where

- x_{ijk} is the observation of the i^{th} subject in the j^{th} treatment group of variable A and the k^{th} treatment group of variable B,
- μ is the grand mean,
- α_j is the adjustment to the mean for the j^{th} treatment group of A (main effect),
- β_k is the adjustment to the mean for the k^{th} treatment group of B (main effect),
- γ_{jk} is the adjustment to the mean for the combination of the j^{th} and k^{th} treatment group (interaction),
- $\epsilon_{ijk} \overset{\text{iid}}{\sim} N(0, \sigma)$.

Two-way ANOVA maths

This can be expressed as

$$\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk}-\bar{y}_{\bullet\bullet\bullet})^2 = NK\sum_{j=1}^{J}(\bar{y}_{\bullet j\bullet}-\bar{y}_{\bullet\bullet\bullet})^2+NJ\sum_{k=1}^{K}(\bar{y}_{\bullet\bullet k}-\bar{y}_{\bullet\bullet\bullet})^2+$$

$$N\sum_{j=1}^{J}\sum_{k=1}^{K}(\bar{y}_{\bullet jk}-\bar{y}_{\bullet j\bullet}-\bar{y}_{\bullet\bullet k}+\bar{y}_{\bullet\bullet\bullet})^2+\sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk}-\bar{y}_{\bullet jk})^2$$

Glue strength

Data:	Plastic	Wood
Thin	**52 64**	**72 60**
Moderate	**67 55**	**78 68**
Heavy	**86 72**	**43 51**

This data shows the force required (in newtons) to separate two components stuck together with either wood or plastic glue that has been applied in either a thin, moderate or thick layer.

Glue strength - the model

The model is

$$x_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \epsilon_{ijk},$$

where

- x_{ijk} is the observation of the i^{th} observation in the j^{th} Thickness group and the k^{th} Glue Type.
- μ is the grand mean,
  - e.g. here it's 64,
- α_j is the adjustment to the mean for the j^{th} Thickness group,
  - e.g. it's -2 for the Thin group.
- β_k is the adjustment to the mean for the k^{th} Glue Type,
  - e.g. it's $+2$ for the Plastic group.
- γ_{jk} is the adjustment to the mean for the combination of the j^{th} Thickness group and k^{th} Glue,
  - e.g. it's -6 for the Thin/Plastic group.
- ϵ_{ijk} is the residual error for each, individual observation.

Two-way ANOVA hypotheses

So, instead of just one hypothesis, we now have several. One for each of our main effects and one for the interaction.

Main effect

H_0 :$\alpha_1 = \alpha_2 = ... = \alpha_K$,
H_a :at least one of the α_k's is different from the rest.

Interaction

H_0 :$\gamma_{1,1} = \gamma_{1,2} = ... = \gamma_{J,K}$,
H_a :at least one of the $\gamma_{j,k}$'s is different from the rest.

Testing significance

The analysis of variance F-statistic for testing the equality of several means still has this form:

$$F = \frac{\text{variation among the sample means}}{\text{variation among individuals in the same sample}}$$

but now we have one for each of our main effects and the interaction.

The denominator for each of our F statistics stays the same. It is *always* MSE. The numerator will be the mean square of the specific main effect or interaction you are testing.

The reference distribution will be the same as we saw in one-way ANOVA but with a small change to the degrees of freedom...

Degrees of freedom

The degrees of freedom will change depending on which hypothesis you are testing:

Source	d.f.	S.S.	M.S.	t.s.	p
A	$K-1$	SSA	$SSA/(K-1)$	**MSA/MSE**	
B	$J-1$	SSB	$SSB/(J-1)$	**MSB/MSE**	
A × B	$(K-1)(J-1)$	$SSAB$	$SSAB$/df	**MSAB/MSE**	
Error	$KJ(c-1)$	SSE	SSE/df		
Total	$n-1$	SSY			

So the reference distributions will be:

A: $F_{K-1,KJ(c-1)}$, B: $F_{J-1,KJ(c-1)}$, AxB: $F_{(J-1)(K-1),KJ(c-1)}$

Degrees of freedom

What do you think would happen here if we only had one observation per combination of predictors?

Simple block design

Here we have data from five students across four exams

	Adam	Brenda	Cathy	Dave	Emily	Mean
Exam #1:	62	94	68	86	50	72
Exam #2:	87	95	93	97	63	87
Exam #3:	74	86	82	70	28	68
Exam #4:	77	89	73	79	47	73
Mean	75	91	79	83	47	75

As there is only one observation per cell (and there is no missing cells) this is called a *simple block design*.

With a simple block design, we can only test for main effects and *not for the interaction* as we do not have enough observations to cover the degrees of freedom we need.

Glue strength - hypotheses

Thickness

$H_0 : \alpha_{Thin} = \alpha_{Moderate} = \alpha_{Heavy}$,

H_a : {at least one of the mean thicknesses is different to at least one other.}

Glue Type

$H_0 : \beta_{Plastic} = \beta_{Wood}$,

$H_a : \beta_{plastic} \neq \beta_{Wood}$.

Interaction

$H_0 : \gamma_{Thin,Plastic} = \gamma_{Moderate,Plastic} = \cdots = \gamma_{Moderate,Wood} = \gamma_{Heavy,Wood}$,

H_a : { at least one of the cell means is different to at least one other.}

Two-way ANOVA - interpreting significance

Interaction

A significant interaction means that the relationship between one of your predictor variables and the response is different *depending on the level of the other predictor variable*.

If your interaction is significant, this overrides your main effects. You will have to perform follow-up, post-hoc testing of all of the groups in your data to investigate what these relationships are.

Two-way ANOVA - interpreting significance

Main Effects

If your main efects are significant then we can say that at least one of levels of our predictor variable is significantly different to at least one other.

If there are *more than two* levels in your predictor, you will need to follow up with post-hoc testing to find where these differences are.

Two-way ANOVA - interpreting significance

{Always test your interaction first.}

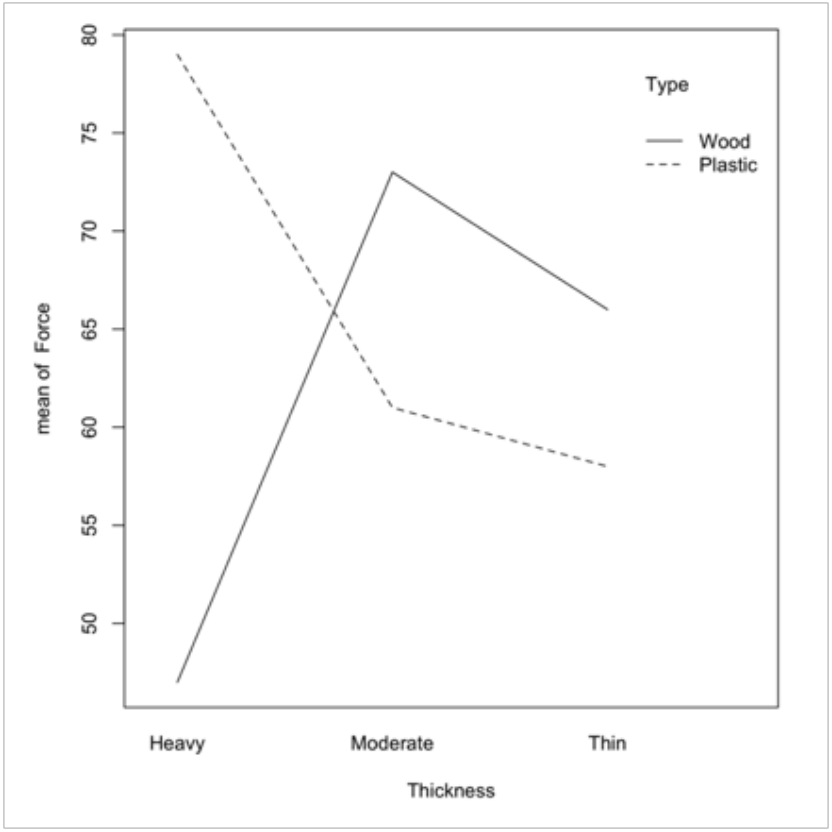
Glue strength - R output

```
                Df Sum Sq Mean Sq F value Pr(>F)
Thickness        2     56      28   0.424 0.6725
Type             1     48      48   0.727 0.4265
Thickness:Type   2   1184     592   8.970 0.0157 *
Residuals        6    396      66
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Here, the effect of Thickness on the mean force required to seperate the two components is *different depending on what glue you use*.

Glue strength - the picture

Given the picture below, how would we interpret the interaction?



Two-way ANOVA assumptions

The assumptions of two-way ANOVA are the same as for one-way ANOVA:

- Normality of observations in each group.
- Constant variance of observations in each group.
- Independence of observations within each group.
- Independence of observations between the groups.

And they can all be tested in exactly the same way as in one-way ANOVA.

Two-way ANOVA recap

- **Scenario** 2 categorical variables to investigate with a continuous response measure
- **Null and alternative hypotheses** There are up to three hypotheses; two for the main effects and one for the interaction. With:

Main effect

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_K,$$
$$H_a : \text{at least one of the } \alpha_k\text{'s is different from the rest.}$$

Interaction

$$H_0 : \gamma_{1,1} = \gamma_{1,2} = ... = \gamma_{J,K},$$
$$H_a : \text{at least one of the } \gamma_{j,k}\text{'s is different from the rest.}$$

Two-way ANOVA recap

- **Test statistic**

A: $F = \frac{MSA}{MSE}$, B: $F = \frac{MSB}{MSE}$, AxB: $F = \frac{MSAB}{MSE}$

- **Reference distribution**

A: $F_{K-1,KJ(c-1)}$, B: $F_{J-1,KJ(c-1)}$, AxB: $F_{(J-1)(K-1),KJ(c-1)}$

- P **value**

$P(F > f)$

Further reading/tools

- Textbook:
  - Chapter 12.1 Inference for one-way ANOVA
  - Chapter 12.2 Comparing the means
  - Chapter 13.1 The two-way ANOVA model
  - Chapter 13.2 Inference for two-way ANOVA
  - Chapter 15.3 The Kruskal-Wallis Test