Analysis of Variance

CSU, Chico Math 314

2018-10-24

outline

outline

recap, Two Sample t-test

The two sample t-test, compares the means of two groups in the hypothesis

$$\begin{aligned} H_0 : &\quad \mu_1 = \mu_2 \\ H_A : &\quad \mu_1 \neq \mu_2. \end{aligned}$$

The test brought us a new statistic with a new standard error.

recap, Two Sample t-test

The two sample t-test formulas

$$t_{df} = \frac{(\bar{X}_a - \bar{X}_b) - (\mu_a - \mu_b)}{s^2_{\bar{X}_a - \bar{X}_b}} \qquad \text{and} \qquad (\bar{X}_a - \bar{X}_b) \pm t^*_{df} s^2_{\bar{X}_a - \bar{X}_b}$$

outline

Analysis of Variance, motivation

If there were three or more groups, the two sample t-test would not work. We could force the test on the data by comparing two groups at a time, but this has dangerous implications[1]. We thus require a new statistical method, **analysis of variance**.

[1] This is a phrase that I will reference when we come to the lecture Decision Making is Hard

Analysis of Variance, definition

ANOVA

Analysis of variance uses a single hypothesis to check whether the *means* across two or more groups are equal, and uses a new test statistic called F to evaluate this hypothesis.

ANOVA, hypotheses

The ANOVA hypotheses for k groups are

$$\begin{aligned} H_0: &\quad \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k \\ H_A: &\quad \text{at least one mean is different.} \end{aligned}$$

ANOVA, hypothesis examples

With ANOVA you can compare the means by groups for many different data sets.

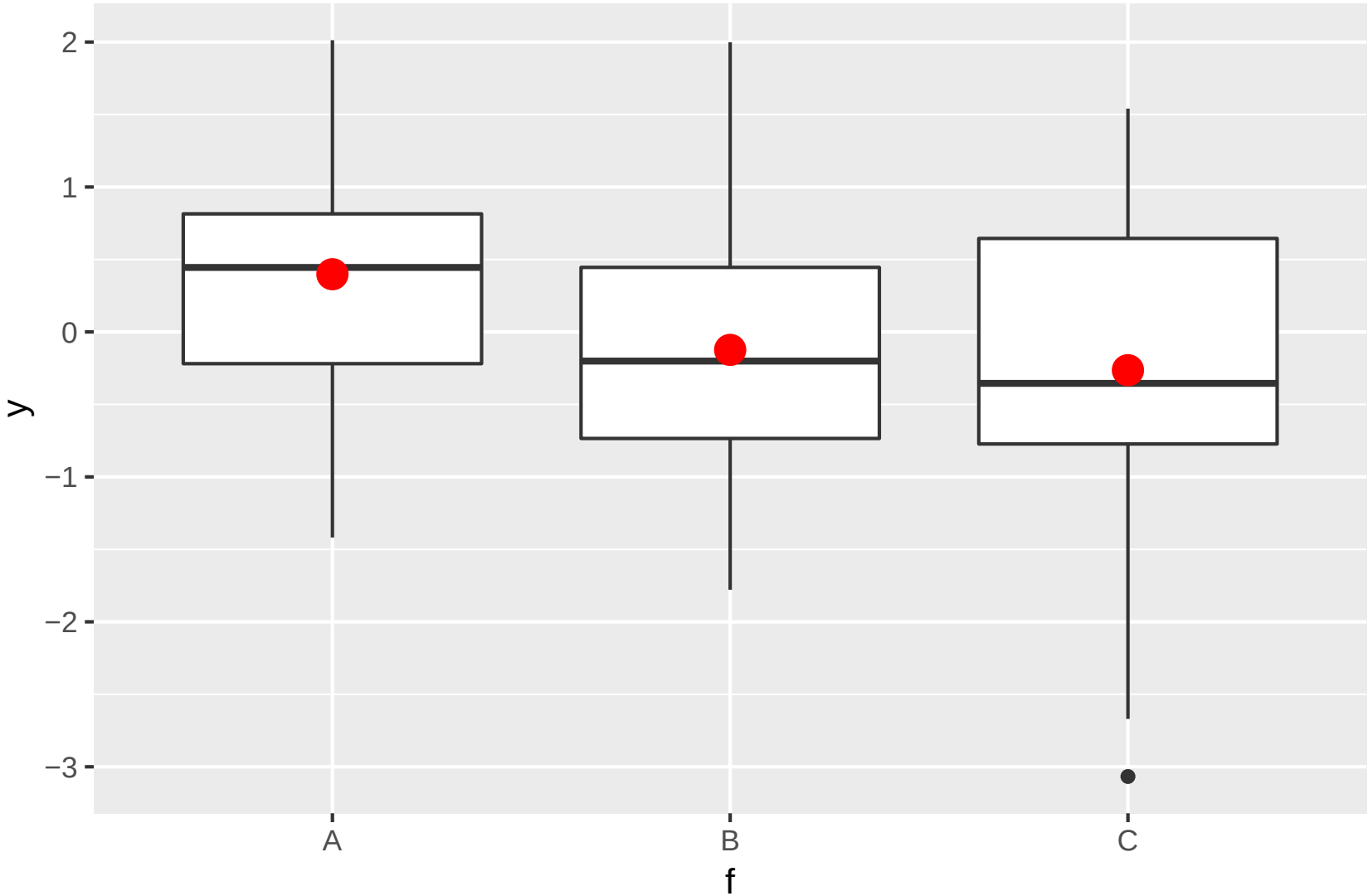
- mean batting average by position,
- mean movie budget by genre, (or year),
- mean CO2 intake by treatment, or location, (or concentration)
- mean rem sleep by conservation status,
- mean birth/body weight by family,
- ... numerical variable by levels of a categorical variable ...
- ...

ANOVA, intuition

Analysis of variance tells us about means by groups, despite its name. Large variation amongst the groups relative to small variation within the groups indicates different population means.

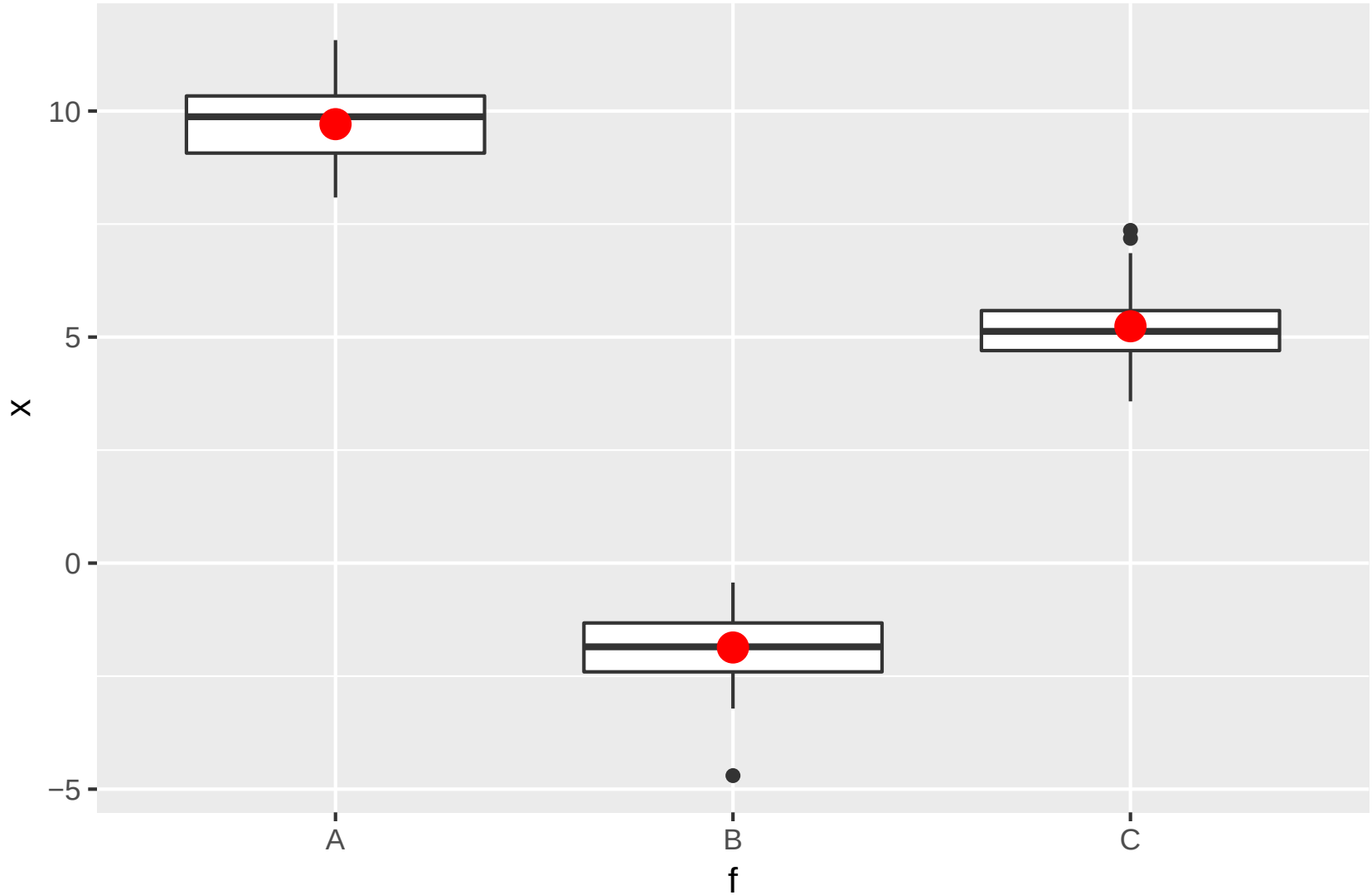
ANOVA, intution via picture

What do you think of these means – think of variation within and amongst groups?



ANOVA, intution via picture

What do you think of these means – think of variation within and amongst groups?



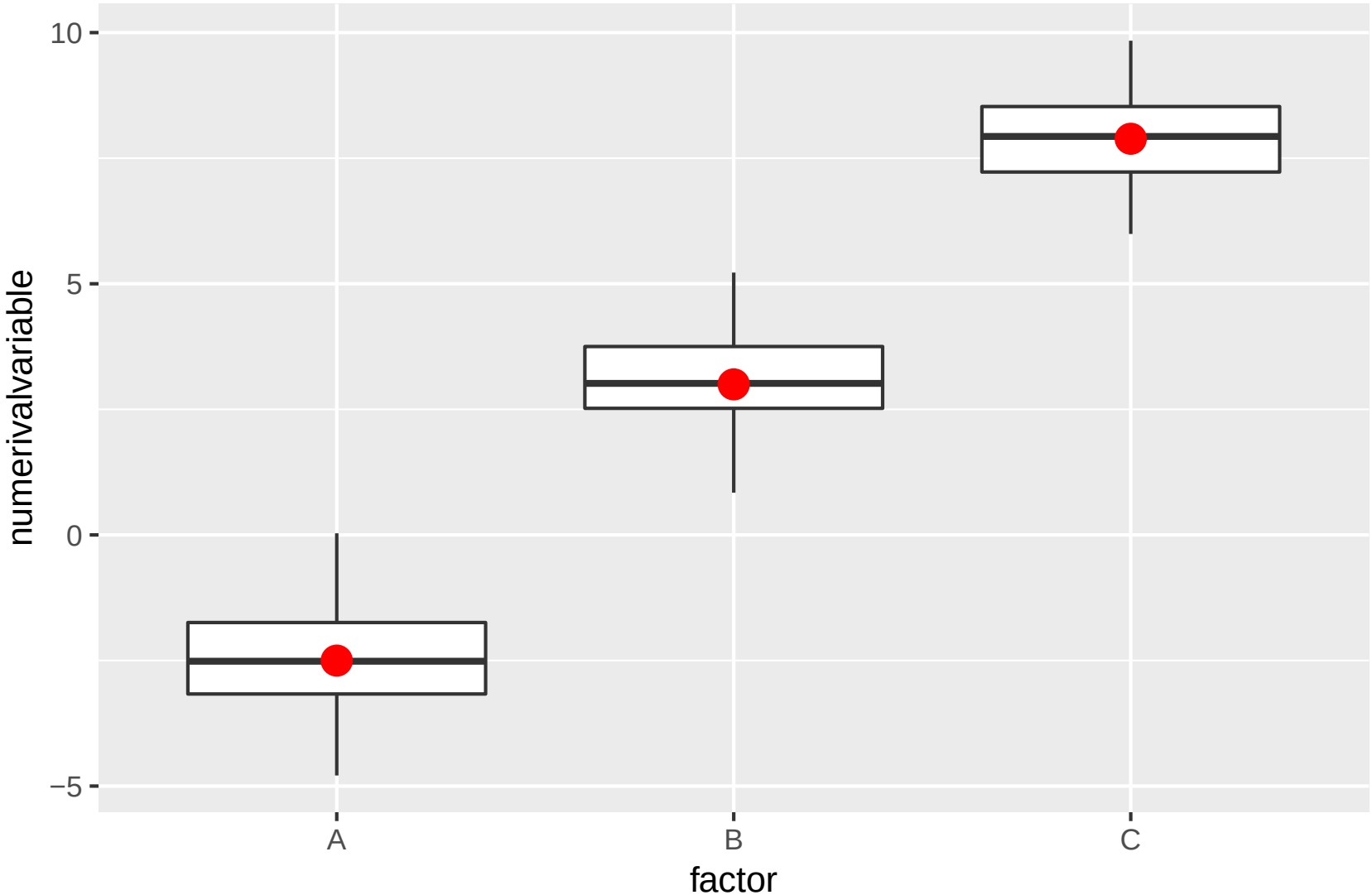
ANOVA, assumptions

The ANOVA has three basic assumptions,

- independent observations (amongst and within groups)
- data within each group are nearly normal, and
- the variability of each group is about equal.

ANOVA, picture

The last two assumptions can help us visualize just what is going on. It starts with box plots by groups – draw more pictures on board.



ANOVA, test statistic

ANOVA calculates one fraction based on two numbers, variation amongst (between) groups and variation within groups. These two numbers are generally referred to as mean square values.

mean squared amongst groups

MSG is a strictly positive measure of the variation across all groups, and has $df_G = k - 1$ where k represents the number of groups.

mean squared error

MSE is a strictly positive measure of the variation within groups, and has $df_E = n - k$ where k represents the number of groups and n is the sample size.

ANOVA, test statistic

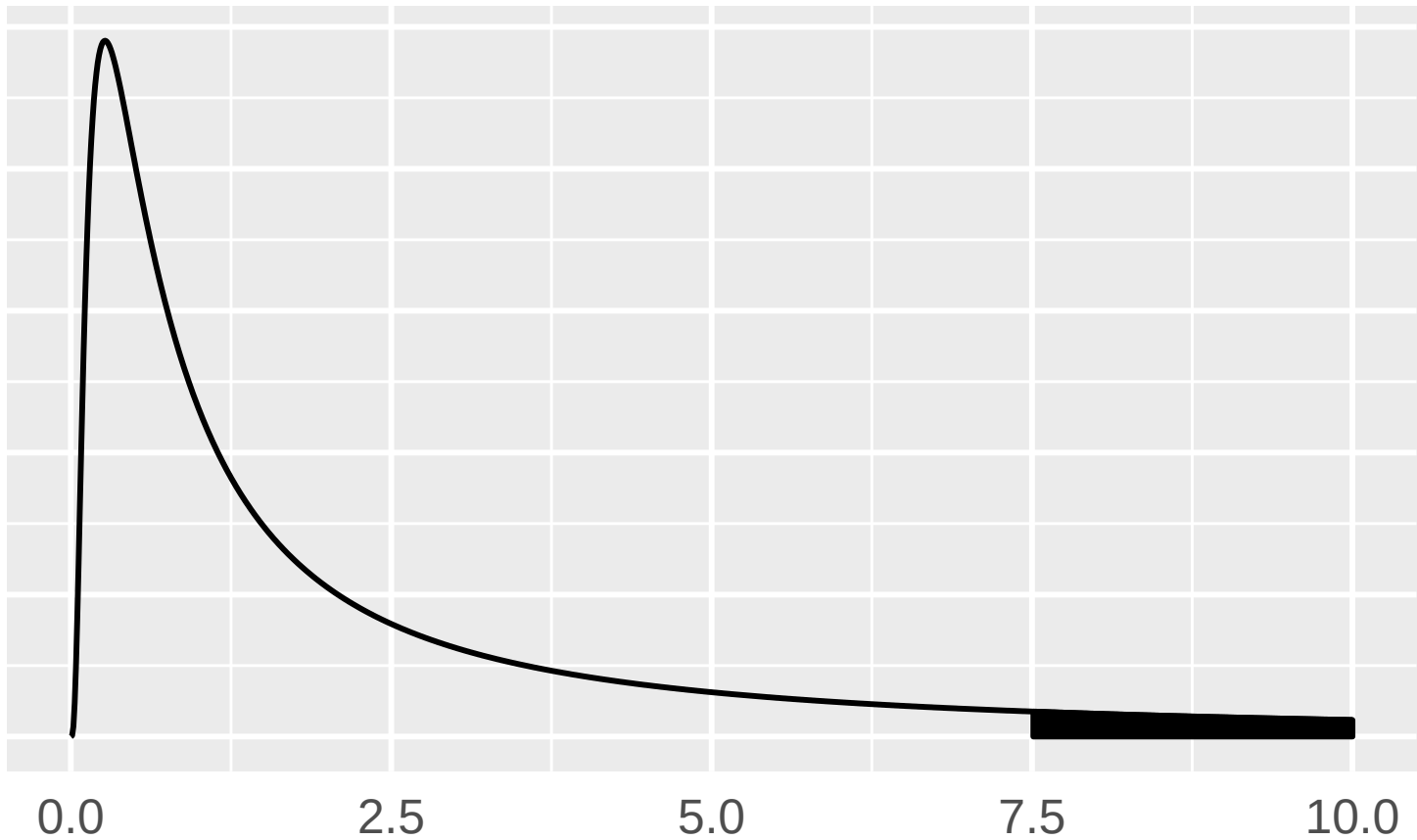
The test statistic of ANOVA follows an F-distribution,

$$F = \frac{MSG}{MSE},$$

and has degrees of freedom associated with the numerator and denominator, df_G and df_E, respectively.

F-distribution

The F-distribution is a probability density function over non-negative numbers. P-values are strictly calculated from the right tail – hence large F statistics indicate evidence against the null hypothesis.



outline

ANOVA, example I

Are baseball players paid on average differently by position?

$$H_0: \quad \mu_{catcher} = \mu_{dh} = \mu_{first} = \ldots = \mu_{third}$$
$$H_A: \quad \text{at least one mean salary is different.}$$

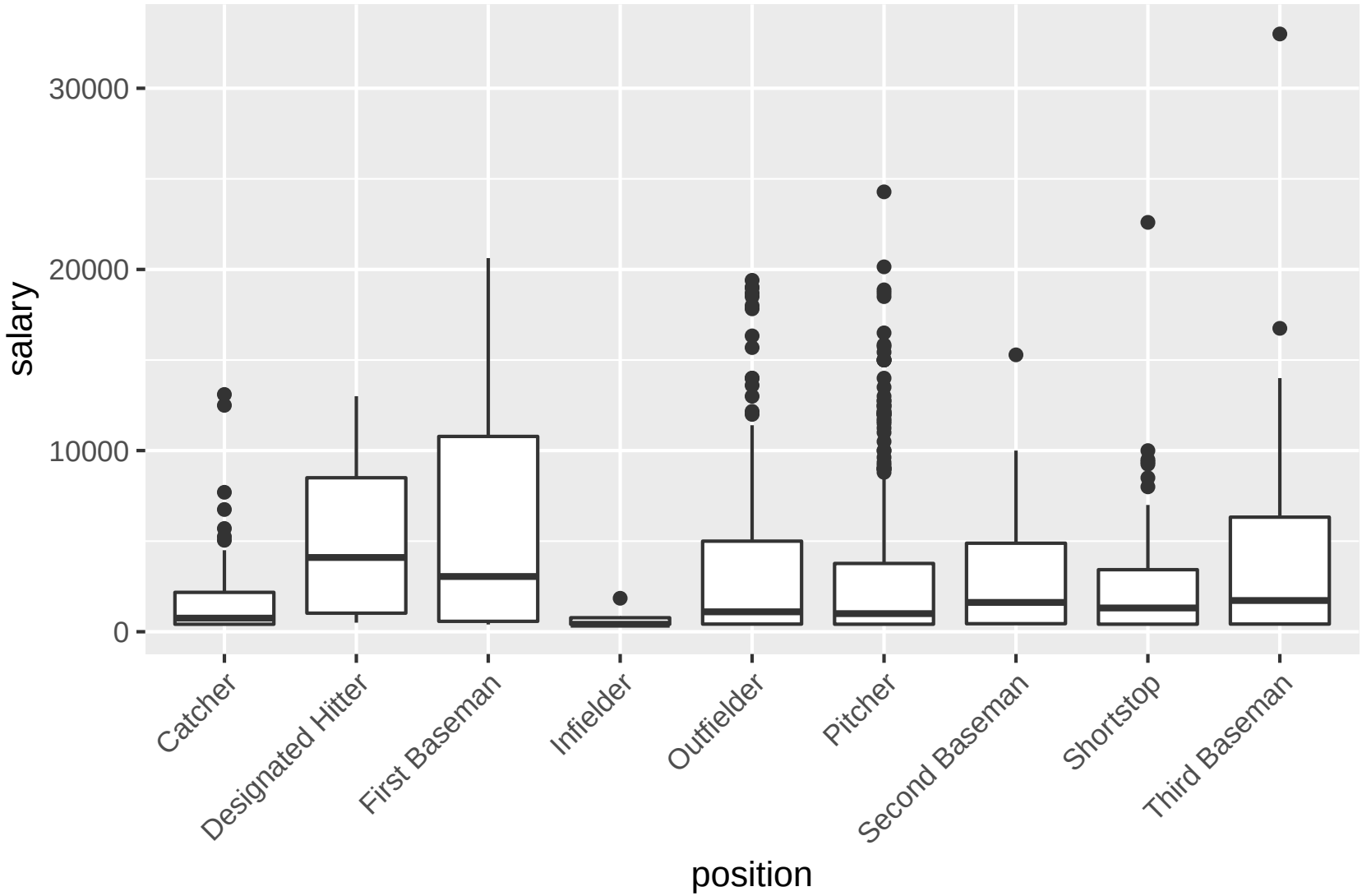
with $\alpha = 0.05$.

ANOVA, example I

Plot the data!

```
suppressMessages(library(ggplot2))
mlb <- read.csv("https://roualdes.us/data/mlb.csv")
p <- ggplot(data=mlb, aes(x=position, y=salary)) +
    geom_boxplot() +
    theme(axis.text.x=element_text(angle=45, hjust=1))
```

ANOVA, example



ANOVA, in R

The R code to run ANOVA

```
model <- lm(salary~position, data=mlb)
anova(model)
```

ANOVA, in R

The tilde $\sim$ is meant to be read as, predict the left hand side with the right hand side. Hence, we read the following as, predict salary by different levels of position.

```
model <- lm(salary~position, data=mlb)
```

We thus need to ensure that the variable `position` is a categorical/factor variable.

```
is.factor(mlb[,"position"])

## [1] TRUE
```

ANOVA, example output

The F-statistic and p-value are the most important pieces of information to extract from an ANOVA table. Print output tables from R with

```
anova(model)

## Analysis of Variance Table
##
## Response: salary
##             Df     Sum Sq  Mean Sq
## position     8 6.0975e+08 76219146
## Residuals  819 1.5881e+10 19390502
##            F value    Pr(>F)
## position    3.9307 0.0001422 ***
## Residuals
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   ' ' 0.1 ' ' 1
```

ANOVA, example conclusion

Because the p-value $= 10^{-4} < \alpha = 0.05$ we reject the null in favor of the alternative. There is sufficient evidence to claim that the mean salaray of baseball players varies by position.

outline

ANOVA

Let's consider the variables `Family` and average brain weight `SB` from the dataset `ape::carnivora`.

```
suppressMessages({library(ape)
  library(dplyr)
  library(ggplot2)
  data(carnivora)})
## Step 1?
```

ANOVA

We'll filter down to just five families.

```
data(carnivora)
carnivs <- filter(carnivora,
                  !(Family %in% c("Ailuridae",
                                  "Procyonidae",
                                  "Viverridae")))
```

ANOVA

Then run ANOVA.

```
mod <- lm(SB~Family, data=carnivs)
anova(mod)
```

ANOVA, details

ANOVA is all about breaking up the response variable, in general denoted by Y, into component pieces made up of the explanatory variable(s). The slide above breaks up the response variable brain weight into an intercept term, and then group means (of sorts) for the remaining $k - 1$ levels of Family.

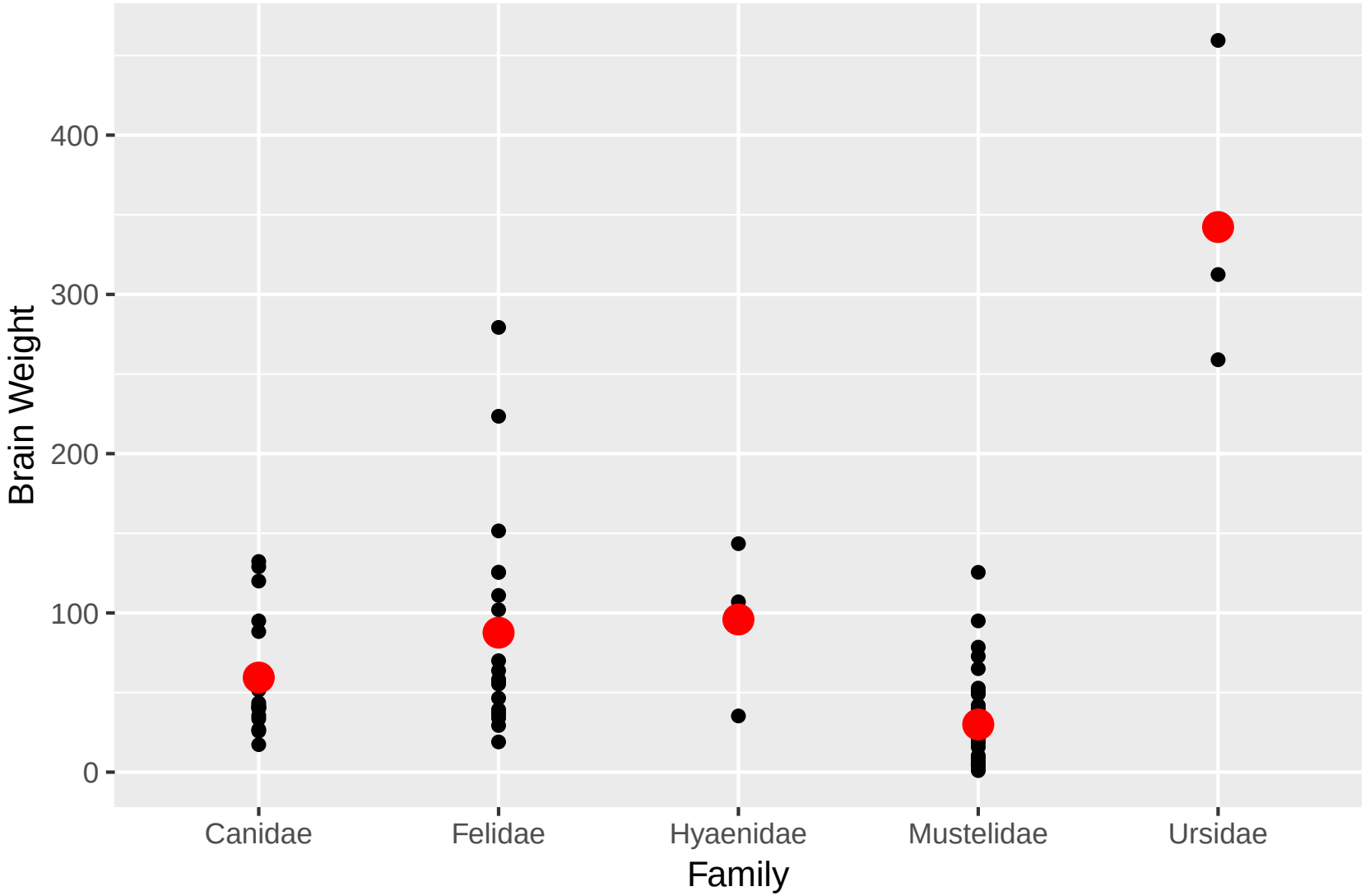
ANOVA, details graphed

We think this, but ...

```
p <- ggplot(carnivs, aes(Family, SB)) +
    geom_point() +
    stat_summary(fun.y=mean, colour="red",
                 geom="point", size=4) +
    labs(y="Brain Weight")
```

ANOVA, details graphed

We think this, but . . .



ANOVA, model

The general ANOVA[2] model is

$$Y_i = \alpha + \beta_1 X_{i,1} + \ldots + \beta_{k-1} X_{i,k-1} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

for $i = 1, \ldots, n$. Details

- α is called the intercept
- $X_{i,j}$ equals 1 or 0 dependent on the ith observation
- $k - 1$ levels of the categorical explanatory variable X_j get their own estimated value β_j
  - β_j is difference between the respective group mean and the intercept

[2] This is the most common model used to fit ANOVA, but it is hardly the most intuitive.

ANOVA, details in R

Calling the function `coefficients` on the linear model returns the intercept and $k-1$ values β_j. Each group's mean is actually the sum, $\hat{\alpha} + \hat{\beta}_j$ relative to the specific group, j, of interest.

```
coefficients(mod)
```

```
##       (Intercept)    FamilyFelidae
##          59.45556          28.14444
##  FamilyHyaenidae FamilyMustelidae
##          36.44444         -29.47056
##    FamilyUrsidae
##         282.86944
```

ANOVA, details in R

Compare coefficients to the group means

```
carnivs %>%
    group_by(Family) %>%
    summarise(m=mean(SB))
```

```
## # A tibble: 5 x 2
##   Family         m
##   <fct>      <dbl>
## 1 Canidae     59.5
## 2 Felidae     87.6
## 3 Hyaenidae   95.9
## 4 Mustelidae  30.0
## 5 Ursidae    342.
```

ANOVA, details in math

The model predicts the mean value of Y_i, $\mu_{Y|X}$, dependent on the group that observation Y_i came from. For instance, suppose observation i is a member of Felidae,

$$\begin{aligned}\hat{Y}_i = E(Y_i|X) &= E(Y_i|Felidae) \\ &= \hat{\alpha} + \hat{\beta}_1(X_{i,1} == 1) + \ldots + \hat{\beta}_{k-1}(X_{i,k-1} == 0) \\ &= \hat{\alpha} + \hat{\beta}_1.\end{aligned}$$

ANOVA, fitted values

The **predicted/fitted** values of the model are the group means; the values the model would predict for each group

$$\hat{y}_{ij} = \hat{\alpha} + \hat{\beta}_j.$$

```
yhat <- fitted(mod) # retreive fitted values
```

ANOVA, details in math

The model says that each observation is the sum of the intercept, the appropriate β_j, and some left over piece, called the **residuals**, denoted by ϵ_i,

$$y_i = \alpha + \beta_1 X_{i,1} + \ldots + \beta_{k-1} X_{i,k-1} + \epsilon_i.$$

The **residuals** make up the difference between the observation and the sum of the intercept and each β_j.

ANOVA, residuals

residuals/errors

The theoretical (or random) residuals/errors are often denoted by ϵ_i in statistical models. Once observed, after fitting the model, we define the residuals as $e_i = y_i - \hat{y}_i$.

```
residuals(mod) # retrieve residuals
```

Note

The difference in the notation between ϵ_i and e_i is similar to that with random variables: X denotes a random variable, not yet observed, and x is the no longer random value that X took on. Here, e_i denotes the observed value of the random variable ϵ_i.

outline

Take away

- ANOVA breaks up the mean of the response variable into components
  - components are specified by a categorical explanatory variable
- Normality assumption is only necessary for small sample sizes
  - interest lies in means $\Rightarrow$ Central Limit Theorem
- equal variance assumption qualitatively checked with plots
  - there exists more formal ways to check this ...
- you are best bet to ensure independent observations
- multiple comparisons is a silent killer