ANOVA
Analysis of variance

*Torture numbers, and they'll confess to anything.*

Gregg Easterbrook

Cláudia Mendes
Ewa Chrostek
Null hypothesis and alternative hypothesis

Does the chicks diet influences their weight?

$H_0$: no effect of the diet on the weight of chicks

$H_1$: the weight of the chicks depends on the food they eat
Collected Data

<table>
<thead>
<tr>
<th>Horsebean</th>
<th>Linseed</th>
<th>Soybean</th>
<th>Sunflower</th>
<th>Meat Meal</th>
<th>Casein</th>
</tr>
</thead>
<tbody>
<tr>
<td>179</td>
<td>309</td>
<td>243</td>
<td>423</td>
<td>325</td>
<td>368</td>
</tr>
<tr>
<td>160</td>
<td>229</td>
<td>230</td>
<td>340</td>
<td>257</td>
<td>390</td>
</tr>
<tr>
<td>136</td>
<td>181</td>
<td>248</td>
<td>392</td>
<td>303</td>
<td>379</td>
</tr>
<tr>
<td>227</td>
<td>141</td>
<td>327</td>
<td>339</td>
<td>315</td>
<td>260</td>
</tr>
<tr>
<td>217</td>
<td>260</td>
<td>329</td>
<td>341</td>
<td>380</td>
<td>404</td>
</tr>
<tr>
<td>168</td>
<td>203</td>
<td>250</td>
<td>226</td>
<td>153</td>
<td>318</td>
</tr>
<tr>
<td>108</td>
<td>148</td>
<td>193</td>
<td>320</td>
<td>263</td>
<td>352</td>
</tr>
<tr>
<td>124</td>
<td>169</td>
<td>271</td>
<td>295</td>
<td>242</td>
<td>359</td>
</tr>
<tr>
<td>143</td>
<td>213</td>
<td>316</td>
<td>334</td>
<td>206</td>
<td>216</td>
</tr>
<tr>
<td>140</td>
<td>257</td>
<td>267</td>
<td>322</td>
<td>344</td>
<td>222</td>
</tr>
<tr>
<td>244</td>
<td>199</td>
<td>297</td>
<td>258</td>
<td>283</td>
<td></td>
</tr>
<tr>
<td>271</td>
<td>171</td>
<td>318</td>
<td></td>
<td></td>
<td>332</td>
</tr>
<tr>
<td>158</td>
<td>248</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Groups: 6 different feeds

Number of items: Weight of 71 chickens
(10-14 chickens in each group)
ANOVA test for chicks:

ANOVA – testing whether or not the means of several groups are equal (by the analysis of their variance). It generalizes the Student t-test for more than two groups.

We will employ one-way ANOVA – testing the influence of only one factor (food).

Our aim is to check if the 7 different feeds will affect the mean weights of the chickens.

Group = chicks fed with the same food
WARNING: There is variance and variance, and they are not the same!

The population or populations we are sampling from have the true value of sigma (but we do not know this one).

In ANOVA we know that the single samples values differ from the mean value for the group.

And that the means of all groups differ from the global mean (mean of all means).

This differences will be called variances.
Variance within the groups
Variance between the groups
ANOVA Assumptions

1. Random sampling
2. Independence of items
3. Normality of residuals
4. Homogeneity of variances
The difference between the sample and the estimated value (mean for the group in this case)

Deviations for each individual within the group (caused for example, by genetic and environmental effects).

We can under or overestimate the weight of this individual chicken, so the residual error can be a positive or a negative value.

<table>
<thead>
<tr>
<th>Horse-bean</th>
<th>Linseed</th>
<th>Soybean</th>
<th>Sunflower</th>
<th>Meat Meal</th>
<th>Case-in</th>
</tr>
</thead>
<tbody>
<tr>
<td>179</td>
<td>309</td>
<td>243</td>
<td>423</td>
<td>325</td>
<td>368</td>
</tr>
<tr>
<td>160</td>
<td>229</td>
<td>230</td>
<td>340</td>
<td>257</td>
<td>390</td>
</tr>
<tr>
<td>136</td>
<td>181</td>
<td>248</td>
<td>392</td>
<td>303</td>
<td>379</td>
</tr>
<tr>
<td>227</td>
<td>141</td>
<td>327</td>
<td>339</td>
<td>315</td>
<td>260</td>
</tr>
<tr>
<td>217</td>
<td>260</td>
<td>329</td>
<td>341</td>
<td>380</td>
<td>404</td>
</tr>
<tr>
<td>168</td>
<td>203</td>
<td>250</td>
<td>226</td>
<td>153</td>
<td>318</td>
</tr>
<tr>
<td>108</td>
<td>148</td>
<td>193</td>
<td>320</td>
<td>263</td>
<td>352</td>
</tr>
<tr>
<td>124</td>
<td>169</td>
<td>271</td>
<td>295</td>
<td>242</td>
<td>359</td>
</tr>
<tr>
<td>143</td>
<td>213</td>
<td>316</td>
<td>334</td>
<td>206</td>
<td>216</td>
</tr>
<tr>
<td>140</td>
<td>257</td>
<td>267</td>
<td>322</td>
<td>344</td>
<td>222</td>
</tr>
<tr>
<td>244</td>
<td>199</td>
<td>297</td>
<td>258</td>
<td>283</td>
<td></td>
</tr>
<tr>
<td>271</td>
<td>171</td>
<td>318</td>
<td></td>
<td></td>
<td>332</td>
</tr>
<tr>
<td>158</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>248</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1. Random sampling

2. Independence of items – errors are identically and independently distributed

3. Normality of residuals - error terms are normally distributed

4. Homogeneity of variances
ANOVA Assumptions

1. Random sampling
2. Independence of items
3. Normality of residuals
4. Homogeneity of variances
Does the variance between groups differ from the variance within groups?

Variance within groups = WSS

\[ \frac{WSS}{df_{\text{WSS}}} \]

Variance between groups = BSS

\[ \frac{BSS}{df_{\text{BSS}}} \]

Both variances are good estimators of the population variance \( \sigma^2 \) under the assumption that \( H_0 \) is true.
Does the variance between groups differ from the variance within groups?

\[ F = \frac{BSS / df_{BSS}}{WSS / df_{WSS}} \]

- \( F \approx 1 \) We accept the null hypothesis \((H_0)\)

- \( F \gg 1 \) We reject the null hypothesis \((H_0)\) and accept the alternative hypothesis \((H_1)\)

**Fisher-Snedecor’s F or F-statistics** in honor of R.A. Fisher and George W. Snedecor.

**F-distribution** is the distribution of the ratio of two estimates of variance.

It is used to compute probability values in the analysis of variance (ANOVA).
F-distribution under the assumption that $H_0$ is true

$$\frac{BSS}{df_{BSS}} = \frac{WSS}{df_{WSS}}$$

$$F \approx 1$$

One-tailed test

$$\frac{BSS}{df_{BSS}} > \frac{WSS}{df_{WSS}}$$

$$F > 4.426348$$

R code:

```r
qf(q, df_wss, df_Bss)
```

$df_{Bss} = \text{number of groups} - 1 = 5$

$df_{Wss} = \text{number of items} - \text{number of groups} = 65$
Data Set

### ANOVA assumptions:
- Random sampling and
- Independence of items

### Number of items:
- Weight of 71 chickens

<table>
<thead>
<tr>
<th>Horsebean</th>
<th>Linseed</th>
<th>Soybean</th>
<th>Sunflower</th>
<th>Meat Meal</th>
<th>Casein</th>
</tr>
</thead>
<tbody>
<tr>
<td>179</td>
<td>309</td>
<td>243</td>
<td>423</td>
<td>325</td>
<td>368</td>
</tr>
<tr>
<td>160</td>
<td>229</td>
<td>230</td>
<td>340</td>
<td>257</td>
<td>390</td>
</tr>
<tr>
<td>136</td>
<td>181</td>
<td>248</td>
<td>392</td>
<td>303</td>
<td>379</td>
</tr>
<tr>
<td>227</td>
<td>141</td>
<td>327</td>
<td>339</td>
<td>315</td>
<td>260</td>
</tr>
<tr>
<td>217</td>
<td>260</td>
<td>329</td>
<td>341</td>
<td>380</td>
<td>404</td>
</tr>
<tr>
<td>168</td>
<td>203</td>
<td>250</td>
<td>226</td>
<td>153</td>
<td>318</td>
</tr>
<tr>
<td>108</td>
<td>148</td>
<td>193</td>
<td>320</td>
<td>263</td>
<td>352</td>
</tr>
<tr>
<td>124</td>
<td>169</td>
<td>271</td>
<td>295</td>
<td>242</td>
<td>359</td>
</tr>
<tr>
<td>143</td>
<td>213</td>
<td>316</td>
<td>334</td>
<td>206</td>
<td>216</td>
</tr>
<tr>
<td>140</td>
<td>257</td>
<td>267</td>
<td>322</td>
<td>344</td>
<td>222</td>
</tr>
<tr>
<td>244</td>
<td>199</td>
<td>297</td>
<td>258</td>
<td>283</td>
<td></td>
</tr>
<tr>
<td>271</td>
<td>171</td>
<td>318</td>
<td></td>
<td></td>
<td>332</td>
</tr>
</tbody>
</table>

### Groups:
- 6 different feeds
Let’s go to R :

data<-datasets::chickwts
weight<-data$weight
feed<-factor(data$feed)
Data<-data.frame(feed,weight)

**ANOVA assumptions:**
• Homogeneity of variances (Bartlett test or Levene test)

bartlett.test(weight,feed)

levene.test<-function (x,groups)
{
  medians<-sapply(split(x,groups),median,na.rm=T)
  resid.x<-abs(x-median(groups))
  anova(lm(resid.x~groups))
}
levene.test(weight,feed)
Let's go to R :) 

**ANOVA:**

```r
anova<-aov(weight~feed,data=Data)
summary(anova)
```

**ANOVA assumptions:**

- Normality of residuals (Shapiro-Wilk normality test)

```r
res<-residuals(anova)
shapiro.test(res)

qqnorm(anova$res,ylab="Residuals",pch=16,col="blue")
qqline(anova$res,col="red")
```
Thank you for your attention!
ANOVA vs t-test

Probability of type I error

\[ 1 - (1 - \alpha)^N \]

\[ 1 - (1 - \alpha)^6 \approx 0.26 \]

\[ 1 - (1 - \alpha)^{50} \approx 0.92 \]

\[ 1 - (1 - \alpha)^{100} \approx 0.99 \]
Oh no! The table is here!!!

<table>
<thead>
<tr>
<th>a groups (a = 7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>41</td>
</tr>
<tr>
<td>44</td>
</tr>
<tr>
<td>48</td>
</tr>
<tr>
<td>43</td>
</tr>
<tr>
<td>42</td>
</tr>
</tbody>
</table>

\[
\sum Y = n_1 + n_2 + n_3 + n_4 + n_5 \\
\bar{Y} = \frac{n_1 + n_2 + n_3 + n_4 + n_5}{5} \\
\sum Y^2 = n_1^2 + n_2^2 + n_3^2 + n_4^2 + n_5^2 \\
\sum j^2 = (n_1 - \bar{Y})^2 + (n_2 - \bar{Y})^2 + (n_3 - \bar{Y})^2 + (n_4 - \bar{Y})^2 + (n_5 - \bar{Y})^2
\]

Modified from 'Biometry', Sokal R., Rohlf F., W.H. Freeman and Company

\[ a(n-1) = a(n - a) \text{ DEGREES OF FREEDOM} \]

\[ \sum j^2 = \frac{(n_1 - \bar{Y})^2 + (n_2 - \bar{Y})^2 + (n_3 - \bar{Y})^2 + (n_4 - \bar{Y})^2 + (n_5 - \bar{Y})^2}{a(n-1)} \text{ VARIANCE WITHIN GROUPS} \]

\[ \frac{n}{a-1} \sum_{i=1}^{a} (\bar{Y}_i - \bar{Y})^2 \text{ VARIANCE BETWEEN GROUPS} \]