Day 4
Analysis of Qualitative data
Anne Segonds-Pichon
v2019-03
Outline of this section

• Comparing 2 groups (one factor)
• Fisher’s exact or Chi² tests
  • Example: Dancing cats

• More than one factor
• Logistic regression
  • Example: Methylation data
Qualitative data

• = not numerical

• = values taken = usually names (also nominal)
  • e.g. causes of death in hospital

• Values can be numbers but not numerical
  • e.g. group number = numerical label but not unit of measurement

• Qualitative variable with intrinsic order in their categories = ordinal

• Particular case: qualitative variable with 2 categories: binary or dichotomous
  • e.g. alive/dead or presence/absence
Comparing 2 groups
One factor
Example: *cats.dat*

- Cats trained to line dance
- 2 different rewards: food or affection
- **Question**: Is there a difference between the rewards?

- **Is there a significant relationship between the 2 variables?**
  - does the reward significantly affect the likelihood of dancing?

To answer this type of question:
- **Contingency table**
- **Fisher’s exact** or **Chi² tests**

<table>
<thead>
<tr>
<th></th>
<th>Food</th>
<th>Affection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dance</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>No dance</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

But first: **how many cats** do we need?
Exercise 14: Power calculation cats.dat

- Preliminary results from a pilot study: **25%** line-danced after having received affection as a reward vs. **70%** after having received food.
- **How many cats** do we need?

![](power.prop.test(p1= 0.25, p2= 0.7, sig.level= 0.05, power= 0.8)

Two-sample comparison of proportions power calculation

n = 18.10585
p1 = 0.25
p2 = 0.7
sig.level = 0.05
power = 0.8
alternative = two.sided

NOTE: n is number in *each* group

- Providing the effect size observed in the experiment is similar to the one observed in the pilot study, we will need 2 samples of **18 to 19 cats** to reach significance (p<0.05) with a Fisher’s exact test.
Plot cats data (From raw data)

cats<-read.delim("cats.dat")
head(cats)

plot(cats$Training, cats$Dance, xlab = "Training", ylab = "Dance")

table(cats)
contingency.table <- table(cats)
contingency.table100 <- prop.table(contingency.table, 1)
contingency.table100 <- round(contingency.table100 * 100)
contingency.table100

```
   Training Dance
  Affection as Reward 114 48  
    Food as Reward  10 28  
```

```
   Training Dance
  Affection as Reward 0.7037037 0.2962963  
    Food as Reward 0.2631579 0.7368421  
```

```
   Training Dance
  Affection as Reward 70 30  
    Food as Reward  26 74  
```

barplot(t(contingency.table100), legend.text=TRUE)
barplot(t(contingency.table100),
    col=c("chartreuse3","lemonchiffon2"),
    cex.axis=1.2,
    cex.names=1.5,
    cex.lab=1.5,
    ylab = "Percentages",
    las=1)

legend("topright",
    title="Dancing",
    inset=.05,
    c("Yes","No"),
    horiz=TRUE,
    pch=15,
    col=c("chartreuse3","lemonchiffon2"))
Chi-square and Fisher’s tests

• Chi$^2$ test very easy to calculate by hand but Fisher’s very hard
• Many software will not perform a Fisher’s test on tables > 2x2
• **Fisher’s test more accurate** than Chi$^2$ test on **small samples**
• **Chi$^2$ test more accurate** than Fisher’s test on **large samples**
• Chi$^2$ test assumptions:
  • 2x2 table: no expected count <5
  • Bigger tables: all expected > 1 and no more than 20% < 5
• **Yates’s continuity correction**
  • All statistical tests work well when their assumptions are met
  • When not: probability Type 1 error increases
  • **Solution**: corrections that increase p-values
  • Corrections are dangerous: no magic
  • Probably best to avoid them
Chi-square test

- In a chi-square test, the observed frequencies for two or more groups are compared with expected frequencies by chance.

\[ \chi^2 = \sum \frac{(\text{Observed Frequency} - \text{Expected Frequency})^2}{\text{Expected Frequency}} \]

- With observed frequency = collected data

- Example with ‘cats’
Chi-square test

- Formula for Expected frequency = \( \frac{\text{(row total)} \times \text{(column total)}}{\text{grand total}} \)

Example: expected frequency of cats line dancing after having received food as a reward:

Expected = \( \frac{(38 \times 76)}{200} = 14.44 \)

Alternatively:
Probability of line dancing: \( \frac{76}{200} \)
Probability of receiving food: \( \frac{38}{200} \)

\( \left( \frac{76}{200} \right) \times \left( \frac{38}{200} \right) = 0.072 \)

Expected: 7.2% of 200 = 14.44

\( \chi^2 = \frac{(114-100.4)^2}{100.4} + \frac{(48-61.6)^2}{61.6} + \frac{(10-23.6)^2}{23.6} + \frac{(28-14.4)^2}{14.4} \)

= 25.35

Is 25.35 big enough for the test to be significant?
Chi-square and Fisher’s Exact tests

Randomness:

Answer: Training significantly affects the likelihood of cats line dancing (p=4.8e-07).
Sometimes data to test are just a few values: 23 342 221 29395

These numbers come from an experiment/query

In the entire genome (n=29395), 221 genes are associated with a function (e.g. antigen binding). From an experiment, we identified 342 genes out of which 23 are showing such an association.

**Question**: Is the enrichment significant?

We need to create a matrix with the 4 values:

```r
enrichment <- matrix(c(23, 342, 221, 29395), nrow=2, ncol=2, byrow = TRUE)
```

Chi-square and Fisher’s Exact tests

From a matrix

Chi-square and Fisher’s Exact tests

Enrichment is significant (p < 2.2e-16).
Comparisons with more than one factor
Real Example: DNA Methylation

Genome

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Position</th>
<th>5-Methylcytosine</th>
<th>Uracil</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRR1179533.2884_0021:5:1:1361:11782</td>
<td>-9</td>
<td>54434159</td>
<td>u</td>
</tr>
<tr>
<td>SRR1179533.2884_0021:5:1:1361:11782</td>
<td>-9</td>
<td>54434156</td>
<td>u</td>
</tr>
<tr>
<td>SRR1179533.2884_0021:5:1:1361:11782</td>
<td>-9</td>
<td>54434153</td>
<td>u</td>
</tr>
<tr>
<td>SRR1179533.2817_0021:5:1:1355:11753</td>
<td>+5</td>
<td>77694920</td>
<td>m</td>
</tr>
<tr>
<td>SRR1179533.2817_0021:5:1:1355:11753</td>
<td>+5</td>
<td>77694916</td>
<td>m</td>
</tr>
<tr>
<td>SRR1179533.2884_0021:5:1:1355:11782</td>
<td>-5</td>
<td>24348106</td>
<td>u</td>
</tr>
<tr>
<td>SRR1179533.2884_0021:5:1:1355:11782</td>
<td>-5</td>
<td>24348146</td>
<td>u</td>
</tr>
<tr>
<td>SRR1179533.2803_0021:5:1:1353:10123</td>
<td>-9</td>
<td>108500170</td>
<td>u</td>
</tr>
</tbody>
</table>
Chi-square and Fisher’s Exact tests
More than one factor

- Still comparing 2 groups: one factor/condition of interest (e.g. treatment vs control)
  - but each group has 3 biological samples/3 experiments
- Example: dna_methylation_format2.csv

<table>
<thead>
<tr>
<th>Gene</th>
<th>Sample</th>
<th>Group</th>
<th>Unmeth</th>
<th>Meth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pno1</td>
<td>AL-1174</td>
<td>AL</td>
<td>166</td>
<td>443</td>
</tr>
<tr>
<td>Pno1</td>
<td>AL-1180</td>
<td>AL</td>
<td>116</td>
<td>276</td>
</tr>
<tr>
<td>Pno1</td>
<td>AL-1220</td>
<td>AL</td>
<td>108</td>
<td>305</td>
</tr>
<tr>
<td>Pno1</td>
<td>DR-1191</td>
<td>DR</td>
<td>188</td>
<td>332</td>
</tr>
<tr>
<td>Pno1</td>
<td>DR-1185</td>
<td>DR</td>
<td>204</td>
<td>320</td>
</tr>
<tr>
<td>Pno1</td>
<td>DR-1197</td>
<td>DR</td>
<td>248</td>
<td>342</td>
</tr>
<tr>
<td>Pnpla5</td>
<td>AL-1174</td>
<td>AL</td>
<td>287</td>
<td>254</td>
</tr>
<tr>
<td>Pnpla5</td>
<td>AL-1180</td>
<td>AL</td>
<td>121</td>
<td>97</td>
</tr>
<tr>
<td>Pnpla5</td>
<td>AL-1220</td>
<td>AL</td>
<td>163</td>
<td>197</td>
</tr>
<tr>
<td>Pnpla5</td>
<td>DR-1181</td>
<td>DR</td>
<td>144</td>
<td>224</td>
</tr>
<tr>
<td>Pnpla5</td>
<td>DR-1185</td>
<td>DR</td>
<td>116</td>
<td>240</td>
</tr>
<tr>
<td>Pnpla5</td>
<td>DR-1197</td>
<td>DR</td>
<td>173</td>
<td>334</td>
</tr>
</tbody>
</table>

Values are independent observations of individual DNA bases.
A DNA base is either methylated (Meth) or unmethylated (Unmeth).

- Equivalent to a 2-way ANOVA but with a binary outcome:
  - Binary logistic regression
Logistic regression
dna_methylation_format2.csv

• For each gene: Is there a difference in methylation between the 2 conditions (AL and DR) accounting for the variability between samples?

  • **Question 1**: is the difference between samples significant? (experimental variability)

  • **Question 2**: is the difference between conditions significant accounting for the variation between samples?
Exercise 15: dna_methylation_format2.csv

• Load dna_methylation_format2.csv
• Use `head()` to get to know the structure of the data
• Calculate the percentage of methylation for each sample
• Plot the percentage for each gene separately and within each gene for each sample.
  • No clues but here is how the graph should look -ish.
## Load the file ##

```r
dna.methylation <- read.csv("dna_methylation_format2.csv")
head(dna.methylation)
```

<table>
<thead>
<tr>
<th>Gene</th>
<th>Sample</th>
<th>Group</th>
<th>Unmeth</th>
<th>Meth</th>
<th>MethPercent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pno1</td>
<td>AL-1174</td>
<td>AL</td>
<td>166</td>
<td>443</td>
<td></td>
</tr>
<tr>
<td>Pno1</td>
<td>AL-1180</td>
<td>AL</td>
<td>116</td>
<td>276</td>
<td></td>
</tr>
<tr>
<td>Pno1</td>
<td>AL-1220</td>
<td>AL</td>
<td>108</td>
<td>305</td>
<td></td>
</tr>
<tr>
<td>Pno1</td>
<td>DR-1181</td>
<td>DR</td>
<td>188</td>
<td>332</td>
<td></td>
</tr>
<tr>
<td>Pno1</td>
<td>DR-1185</td>
<td>DR</td>
<td>204</td>
<td>320</td>
<td></td>
</tr>
<tr>
<td>Pno1</td>
<td>DR-1197</td>
<td>DR</td>
<td>248</td>
<td>342</td>
<td></td>
</tr>
</tbody>
</table>

## calculate percentage ##

```r
dna.methylation$MethPercent <- (dna.methylation$Meth/(dna.methylation$Meth+dna.methylation$Unmeth)*100)
head(dna.methylation)
```

<table>
<thead>
<tr>
<th>Gene</th>
<th>Sample</th>
<th>Group</th>
<th>Unmeth</th>
<th>Meth</th>
<th>MethPercent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pno1</td>
<td>AL-1174</td>
<td>AL</td>
<td>166</td>
<td>443</td>
<td>72.7420</td>
</tr>
<tr>
<td>Pno1</td>
<td>AL-1180</td>
<td>AL</td>
<td>116</td>
<td>276</td>
<td>70.40816</td>
</tr>
<tr>
<td>Pno1</td>
<td>AL-1220</td>
<td>AL</td>
<td>108</td>
<td>305</td>
<td>73.84988</td>
</tr>
<tr>
<td>Pno1</td>
<td>DR-1181</td>
<td>DR</td>
<td>188</td>
<td>332</td>
<td>62.84625</td>
</tr>
<tr>
<td>Pno1</td>
<td>DR-1185</td>
<td>DR</td>
<td>204</td>
<td>320</td>
<td>61.06870</td>
</tr>
<tr>
<td>Pno1</td>
<td>DR-1197</td>
<td>DR</td>
<td>248</td>
<td>342</td>
<td>57.98610</td>
</tr>
</tbody>
</table>

## create 2 files: one for each gene ##

```r
dna.methylation.Pno1 <- dna.methylation[dna.methylation$Gene == "Pno1",]
dna.methylation.Pnpla5 <- dna.methylation[dna.methylation$Gene == "Pnpla5",]
```

<table>
<thead>
<tr>
<th>Gene</th>
<th>Sample</th>
<th>Group</th>
<th>Unmeth</th>
<th>Meth</th>
<th>MethPercent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pnpla5</td>
<td>AL-1174</td>
<td>AL</td>
<td>287</td>
<td>254</td>
<td>46.95009</td>
</tr>
<tr>
<td>Pnpla5</td>
<td>AL-1180</td>
<td>AL</td>
<td>121</td>
<td>97</td>
<td>44.49541</td>
</tr>
<tr>
<td>Pnpla5</td>
<td>AL-1220</td>
<td>AL</td>
<td>163</td>
<td>187</td>
<td>53.42857</td>
</tr>
<tr>
<td>Pnpla5</td>
<td>DR-1181</td>
<td>DR</td>
<td>144</td>
<td>224</td>
<td>60.86957</td>
</tr>
<tr>
<td>Pnpla5</td>
<td>DR-1185</td>
<td>DR</td>
<td>116</td>
<td>240</td>
<td>67.41573</td>
</tr>
<tr>
<td>Pnpla5</td>
<td>DR-1197</td>
<td>DR</td>
<td>173</td>
<td>334</td>
<td>65.87771</td>
</tr>
</tbody>
</table>
Exercise 15: dna_methylation_format2.csv - Answers

```r
## plot percentage methylation for each sample and for each gene ##
par(mfrow=c(1,2))

## dna.methylation.Pno1<-as.data.frame(dna.methylation.Pno1) ##
barplot(dna.methylation.Pno1$MethPercent,
    names.arg = dna.methylation.Pno1$Sample,
    col=rep(c("lightblue","lightpink"), each=3),
    main = "Pno1",
    ylim=c(0,100),
    cex.names=0.6,
    ylab = "Percentage Methylation",
    las=1)

## dna.methylation.Pnpla5<-as.data.frame(dna.methylation.Pnpla5) ##
barplot(dna.methylation.Pnpla5$MethPercent,
    names.arg = dna.methylation.Pnpla5$Sample,
    col=rep(c("lightblue","lightpink"), each=3),
    main = "Pnpla5",
    ylim=c(0,100),
    cex.names=0.6,
    ylab = "Percentage Methylation",
    las=1)
```

Ta-dah!
Exercise 16: dna_methylation_format2.csv

- **Question 1**: is the difference between samples significant? (experimental variability)
  - Work a Chi^2 test for each group/gene so that you can workout the experimental variability.

- Restructure the file in the long format (1 file per gene) for **Question 2**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Sample</th>
<th>Group</th>
<th>Unmeth</th>
<th>Meth</th>
<th>MethPercent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pno1</td>
<td>AL-1174</td>
<td>AL</td>
<td>166</td>
<td>443</td>
<td>72.74220</td>
</tr>
<tr>
<td>Pno1</td>
<td>AL-1180</td>
<td>AL</td>
<td>116</td>
<td>276</td>
<td>70.40816</td>
</tr>
<tr>
<td>Pno1</td>
<td>AL-1220</td>
<td>AL</td>
<td>108</td>
<td>305</td>
<td>73.84988</td>
</tr>
<tr>
<td>Pno1</td>
<td>DR-1181</td>
<td>DR</td>
<td>188</td>
<td>332</td>
<td>63.84615</td>
</tr>
<tr>
<td>Pno1</td>
<td>DR-1185</td>
<td>DR</td>
<td>204</td>
<td>320</td>
<td>61.06870</td>
</tr>
<tr>
<td>Pno1</td>
<td>DR-1197</td>
<td>DR</td>
<td>248</td>
<td>342</td>
<td>57.96610</td>
</tr>
</tbody>
</table>
Exercise 16: dna_methylation_format2.csv - Answers

• Question 1: is the difference between samples significant? (experimental variability)
  • Chi² test

```r
dna.methylation.Pnpla5.AL <- dna.methylation.Pnpla5[dna.methylation.Pnpla5$Group == "AL",]
```

<table>
<thead>
<tr>
<th>Gene</th>
<th>Sample</th>
<th>Group</th>
<th>Unmeth</th>
<th>Meth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pnpla5</td>
<td>AL-1174</td>
<td>AL</td>
<td>287</td>
<td>254</td>
</tr>
<tr>
<td>Pnpla5</td>
<td>AL-1180</td>
<td>AL</td>
<td>121</td>
<td>97</td>
</tr>
<tr>
<td>Pnpla5</td>
<td>AL-1220</td>
<td>AL</td>
<td>163</td>
<td>187</td>
</tr>
<tr>
<td>Pnpla5</td>
<td>DR-1181</td>
<td>DR</td>
<td>144</td>
<td>224</td>
</tr>
<tr>
<td>Pnpla5</td>
<td>DR-1185</td>
<td>DR</td>
<td>116</td>
<td>240</td>
</tr>
<tr>
<td>Pnpla5</td>
<td>DR-1197</td>
<td>DR</td>
<td>173</td>
<td>334</td>
</tr>
</tbody>
</table>

```r
chisq.test(dna.methylation.Pnpla5.AL[,4:5], correct=F)
```

3x2 contingency table

<table>
<thead>
<tr>
<th>Unmeth</th>
<th>Meth</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL-1174</td>
<td>287</td>
</tr>
<tr>
<td>AL-1180</td>
<td>121</td>
</tr>
<tr>
<td>AL-1220</td>
<td>163</td>
</tr>
</tbody>
</table>
Exercise 16: dna_methylation_format2.csv - Answers

• Question 1: is the difference between samples significant? (experimental variability)

```r
# test difference between samples for each gene/condition
   chisq.test(dna.methylation.Pn101.AL[,4:5], correct=F)

   chisq.test(dna.methylation.Pn101.DR[,4:5], correct=F)

3 dna.methylation.Pnpla5.AL <- dna.methylation.Pnpla5[dna.methylation.Pnpla5$Group == "AL",]
   chisq.test(dna.methylation.Pnpla5.AL[,4:5], correct=F)

4 dna.methylation.Pnpla5.DR <- dna.methylation.Pnpla5[dna.methylation.Pnpla5$Group == "DR",]
   chisq.test(dna.methylation.Pnpla5.DR[,4:5], correct=F)
```

1. Pearson's Chi-squared test
   data: dna.methylation.Pnpla5.DR[, 3:4]
   X-squared = 3.8193, df = 2, p-value = 0.1481

2. Pearson's Chi-squared test
   data: dna.methylation.Pn101.DR[, 3:4]
   X-squared = 4.0289, df = 2, p-value = 0.1334

3. Pearson's Chi-squared test
   data: dna.methylation.Pnpla5.AL[, 3:4]
   X-squared = 5.3236, df = 2, p-value = 0.06982

4. Pearson's Chi-squared test
   data: dna.methylation.Pn101.AL[, 3:4]
   X-squared = 1.2487, df = 2, p-value = 0.5356
**Exercise 16: dna_methylation_format2.csv - Answers**

- **Question 2**: is the difference between conditions significant accounting for the variation between samples?
  - Data need to be in the long format (1 file per gene)
    - Restructure `dna.methylation.Pno1` and `dna.methylation.Pnpla5`
    - Rename the columns “Methylation” and “Counts”
    - Extra credits if you remove the first and the last columns 😊
Exercise 16: dna_methylation_format2.csv - Answers

```r
## prepare file (long format)##
dna.methylation.Pno1.long <- melt(dna.methylation.Pno1[,2:5], ID=c("Sample", "Group"))
dna.methylation.Pnpla5.long <- melt(dna.methylation.Pnpla5[,2:5], ID=c("Sample", "Group"))


head(dna.methylation.Pno1.long)
```

<table>
<thead>
<tr>
<th>Gene</th>
<th>Sample</th>
<th>Group</th>
<th>Unmeth</th>
<th>Meth</th>
<th>MethPercent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pnol</td>
<td>AL-1174</td>
<td>AL</td>
<td>166</td>
<td>443</td>
<td>72.74220</td>
</tr>
<tr>
<td>Pnol</td>
<td>AL-1180</td>
<td>AL</td>
<td>116</td>
<td>276</td>
<td>70.40816</td>
</tr>
<tr>
<td>Pnol</td>
<td>AL-1220</td>
<td>AL</td>
<td>108</td>
<td>305</td>
<td>73.84988</td>
</tr>
<tr>
<td>Pnol</td>
<td>DR-1181</td>
<td>DR</td>
<td>188</td>
<td>332</td>
<td>63.84615</td>
</tr>
<tr>
<td>Pnol</td>
<td>DR-1185</td>
<td>DR</td>
<td>204</td>
<td>320</td>
<td>61.06870</td>
</tr>
<tr>
<td>Pnol</td>
<td>DR-1197</td>
<td>DR</td>
<td>248</td>
<td>342</td>
<td>57.96610</td>
</tr>
</tbody>
</table>

```

```
```
• **Question 2**: is the difference between conditions significant accounting for the variation between samples?

• **Logistic regression**: Equivalent to a 2-way ANOVA or a linear model but with a binary outcome

• **Family: Linear Model** (General Linear Model or Generalized Linear Model)
  
  • $y = f(x)$
  • One factor: $y = \text{Constant} + \text{Coefficient} \ast x$
  • Two factors: $y = \text{Constant} + \text{Coefficient}_1 \ast x_1 + \text{Coefficient}_2 \ast x_2$
  • $y = f(x_1, x_2 \ldots x_n)$

• In R: **Outcome ~ factor1 + factor2**

• With the interaction: **Outcome ~ factor1 + factor2 + factor1*factor2**
Question 2: is the difference between conditions significant accounting for the variation between samples?

Logistic regression: Equivalent to a 2-way ANOVA but with a binary outcome

In R:
- Continuous outcome:
  \[ \text{lm or aov}(\text{Outcome} \sim \text{factor1} + \text{factor2} + \text{factor1} \times \text{factor2}) = \text{glm}(\text{Outcome} \sim \text{factor1} + \text{factor2} + \text{factor1} \times \text{factor2}) \]

- Binary outcome:
  \[ \text{lm or aov(\text{does not work}) = glm(Outcome \sim \text{factor1} + \text{factor2} + \text{factor1} \times \text{factor2}, \text{family = binomial()})} \]

Data need to be in the long format (1 file per gene)
Logistic regression

**dnas_methylation_format2.csv**

- **Question 2:** is the difference between conditions significant accounting for the variation between samples?

```r
dna.model.Pno1 <- glm(Methylation ~ Group+Sample, data=dna.methylation.Pno1.long, family = binomial(), weights=Counts)
```

<table>
<thead>
<tr>
<th>Group</th>
<th>Methylation</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL</td>
<td>Unmeth</td>
<td>2</td>
</tr>
<tr>
<td>AL</td>
<td>Meth</td>
<td>3</td>
</tr>
<tr>
<td>AL</td>
<td>Unmeth</td>
<td>1</td>
</tr>
<tr>
<td>DR</td>
<td>Meth</td>
<td>4</td>
</tr>
</tbody>
</table>

```r

<table>
<thead>
<tr>
<th>Sample</th>
<th>Group</th>
<th>Methylation</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL-1174</td>
<td>AL</td>
<td>Unmeth</td>
<td>166</td>
</tr>
<tr>
<td>AL-1180</td>
<td>AL</td>
<td>Unmeth</td>
<td>116</td>
</tr>
<tr>
<td>AL-1220</td>
<td>AL</td>
<td>Unmeth</td>
<td>108</td>
</tr>
<tr>
<td>DR-1181</td>
<td>DR</td>
<td>Unmeth</td>
<td>188</td>
</tr>
<tr>
<td>DR-1185</td>
<td>DR</td>
<td>Unmeth</td>
<td>204</td>
</tr>
<tr>
<td>DR-1197</td>
<td>DR</td>
<td>Unmeth</td>
<td>248</td>
</tr>
</tbody>
</table>
```
Logistic regression

dna_methylation_format2.csv

- **Question 2**: is the difference between conditions significant accounting for the variation between samples?

dna.model.Pno1 <- glm(Methylation ~ Group+Sample, data=dna.methylation.Pno1.long, family = binomial(), weights=Counts)
supply(dna.model.Pno1)

- **Deviance Residuals:**
  - Min 1Q Median 3Q Max
  -20.774 -19.574 -1.605 16.909 19.313

- **Coefficients:**
  - (Intercept): Estimate 0.9816, Std. Error 0.0910, z value 10.787, p < 2e-16 ***
  - Group1: Estimate -0.6602, Std. Error 0.1234, z value -5.348, p = 8.87e-08 ***

- **Signif. codes:** 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

- **Null deviance:** 3899.3 on 11 degrees of freedom
- **Residual deviance:** 3848.2 on 6 degrees of freedom
- **AIC:** 3860.2

- Number of Fisher Scoring iterations: 4
• **Question 2**: is the difference between conditions significant accounting for the variation between samples?

dna.model.Pnpla5<-glm(Methylation ~ Group+Sample, data=dna.methylation.Pnpla5.long, family = binomial(), weights=Counts)

summary(dna.model.Pnpla5)
Real Example: DNA Methylation

Measures per gene
Chi-Square p<0.001 (FDR)
Our email addresses:

anne.segonds-pichon@babraham.ac.uk