2021/01/20 Ex.1

Marco Scarpelli

04 giugno, 2024

Dataset exploration

```
##    alcohol   region color
## 1 8.715131 Piemonte   red
## 2 8.387008 Piemonte   red
## 3 8.177352 Piemonte   red
## 4 7.749446 Piemonte   red
## 5 7.358778 Piemonte   red
## 6 7.647365 Piemonte   red
```

```
## [1] 150   3
```

Point a

Assumptions

The first assumption is Gaussianity for each combination of factors; the p-values are:

```
##          Ps
## 1 0.2369077
## 2 0.4612391
## 3 0.9702243
## 4 0.6525198
## 5 0.5279147
## 6 0.6612603
```

```
## [1] 0.2369077 0.4612391 0.9702243 0.6525198 0.5279147 0.6612603
```

```
## [1] 0.008333333
```

hence all combinations are Gaussian.

We also want the same covariance structure. We can try Bartlett's test, keeping in mind that it is very sensitive to departures from Gaussianity:

```
##
##  Bartlett test of homogeneity of variances
##
## data:  predicted_v and combined_factors
## Bartlett's K-squared = 0.7502, df = 5, p-value = 0.9801
```

the test succeeds (i.e. the covariance structure is the same).

Running ANOVA

We build the complete model:

```
##                     Df Sum Sq Mean Sq F value   Pr(>F)    
## factor_1             1  32.79   32.79  62.176 7.12e-13 ***
## factor_2             2   0.16    0.08   0.149    0.862    
## factor_1:factor_2    2   0.21    0.11   0.201    0.818    
## Residuals          144  75.94    0.53                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Point b

We can see that according to the summary, the interaction has low statistical significance. We can try to remove it and see whether the second factor still has low significance: ## Additive model

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## factor_1      1  32.79   32.79  62.864 5.26e-13 ***
## factor_2      2   0.16    0.08   0.151     0.86    
## Residuals   146  76.15    0.52                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

it does, so we can safely remove it too. We can check our assumptions on the single remaining group. ## Single-factor model ### Assumptions

Gaussianity; the p-values are:

```
## [1] 0.5949263 0.4291510
```

hence we accept; we also check the covariance structure:

```
## 
##  Bartlett test of homogeneity of variances
## 
## data:  predicted_v and factor_1
## Bartlett's K-squared = 0.074476, df = 1, p-value = 0.7849
```

we can now proceed with the new model.

Running ANOVA

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## factor_1      1  32.79   32.79   63.59 3.83e-13 ***
## Residuals   148  76.31    0.52                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We check the three models against each other:

```
## Analysis of Variance Table
## 
## Model 1: predicted_v ~ factor_1 + factor_2
## Model 2: predicted_v ~ factor_1 + factor_2 + factor_1:factor_2
## Model 3: predicted_v ~ factor_1
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    146 76.153                           
## 2    144 75.940  2   0.21245 0.2014 0.8178
## 3    148 76.310 -4  -0.36950 0.1752 0.9509
```

a high p-value means that the models are similar. This is expected, as we saw that neither the interaction, nor the second factor had any impact on the prediction.

Point c

Means:

```
## [1] 75 75
```

The confidence intervals are:

```
##           Lower    Center     Upper
## red   8.2854778 8.4493264 8.6131750
## white 7.3503926 7.5142412 7.6780898
## 1     0.4156921 0.5156073 0.6566446
```

Alternative formulation

Let us assume Bartlett's test failed: in this case, we determined that the covariance structure of the observations in the two groups is different; this means that we cannot create a satisfactory number ($\mathcal{S}_{\text{pooled}}$) that represents them jointly. We must resort to computing their individual variances:

```
##       red     white
## 0.5320155 0.4991990
```

Notice how the mean of these two number is exactly equal to $\mathcal{S}_{\text{pooled}}$ from the previous case: this is expected, as we have 75 observations in both groups and this means that the two groups contribute in equal part to the total variability.

After computing this, we can go on with the test manually:

```
##          Lower   Center    Upper
## red   8.282891 8.449326 8.615762
## white 7.353021 7.514241 7.675462
```

we find a very similar result. Notice how one of the factor has a slightly larger interval and one is slightly smaller, since they had similar variability (0.53 and 0.50), but not equal.