Analysis of Variance :: CHEAT SHEET

Basics

Are the means amongst the k levels equal?

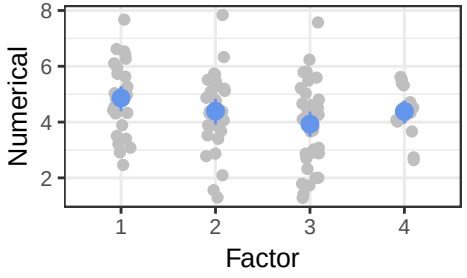
H_0 : $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_K$

H_1 : at least one mean is different

$\alpha = 0.01$

Step 1, plot the data!

```
ggplot(df, aes(Factor, Numerical)) +
    geom_jitter(width=0.1, colour="grey75") +
    stat_summary(fun.data="mean_cl_boot",
                 colour="cornflowerblue")
```




```
fit <- lm(Numerical ~ Factor, data=df)
print(anova(fit), signif=FALSE)
```

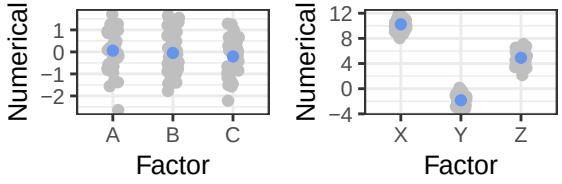
```
## Analysis of Variance Table
##
## Response: Numerical
##            Df Sum Sq Mean Sq F value Pr(>F)
## Factor      3     14    4.67    2.71  0.048
## Residuals 109    187    1.72
```

Assumptions

1. independent observations y_{nk} for $n = 1, \ldots, N$,
2. within each group $k = 1, \ldots, K$, data are normal, and
3. the variability of each group is equal.

Intuition

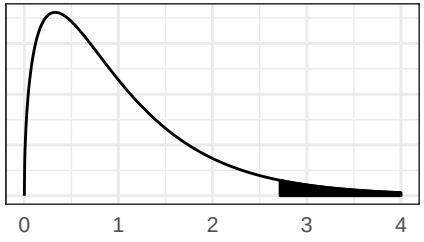
Generally, means will be different when the average square error amongst the groups is significantly larger than average square error within the groups.



$$F_{K-1,N-K} = \frac{\text{average square error amongst the groups}}{\text{average square error within the group}}$$

F distribution

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



Details

ANOVA decomposes observation n into three pieces: the first group's mean α, the offset β_k for group k, and an error term ϵ_{nk}. Each x_{nk} is a numerically encoded, binary variable that indicates when observation n belongs to group k.

$$y_{nk} = \alpha + \beta_1 x_{n1} + \ldots + \beta_{K-1} x_{n,K-1} + \epsilon_{nk}, \quad \epsilon_{nk}\ \mathcal{N}(0, \sigma^2)$$

Coefficients

The coefficients β_k are group offsets relative to the first group: group k's mean is equal to $\hat{\alpha} + \hat{\beta}_k$.

```
##             [,1]
## alpha    4.8607
## beta_1 -0.4670
## beta_2 -0.9337
## beta_3 -0.4795
```

Group Means

The expected value of y_{nk}, $E(y_{nk}|x_{nk})$, is found by averaging over the observations within group k. If observation $n = 5$ belongs to group $k = 3$ then

$$\begin{aligned} \hat{y}_{5,3} = E(y_{5,3}|x_{5,3} == 1) &= \hat{\alpha} + \hat{\beta}_3 \\ &= 4.8607 + -0.4795 \\ &= 3.927 \end{aligned}$$

Notice $\hat{y}_{5,3}$ is exactly equal to group 3's mean.

```
df %>%
    group_by(Factor) %>%
    summarise(group_mean=mean(Numerical))
```

```
## # A tibble: 4 x 2
##   Factor group_mean
##   <fctr>      <dbl>
## 1      1      4.861
## 2      2      4.394
## 3      3      3.927
## 4      4      4.381
```