# 662 Anova

**Assumptions**: As a linear regression model, you'd generally want HILE Gauss

## ANOVA 1

ANOVA models test hypotheses about the mean of more than two groups

same as a linear regression model in which:

the predictor variables are categorical

### Cell Means Model

$Y_{ij} = \mu_i + \epsilon_{ij}$, where: $Y_{ij}$ is the $j^{th}$ observation in the $i^{th}$ group

Generally, the primary hypothesis of interest is:

$H_0 : \mu_i = \mu_k$ For all group means where $i \neq k$

The alternative hypothesis of interest is:

$H_a : \mu_i \neq \mu_k$ $(i \neq k)$ for at least one $i, k$

### Factor Effects Model

$Y_{ij} = \mu_i + \alpha_i + \epsilon_{ij}$, where: $\mu = \frac{1}{N} \sum_{i=1}^{K} n_i \mu_i$ and $\alpha_i = \mu_i - \mu$

Has constraint $\sum_{i=1}^{K} n_i \alpha_i = 0$

$\alpha_i$ is the $i^{th}$ main effect or factor effect

The equivalent null hypothesis to the cell means model is that $H_0 : \alpha_i = 0$ for $i = 1, 2, ..., K$

### Reference Group Model

Choose one group as reference, for example group 1

$Y_{1j} = \mu_1 + \epsilon_{1j}$

$Y_{ij} = \mu_1 + (\mu_i - \mu_1) + \epsilon_{ij}$

```
$=\mu_1+\beta_i+\epsilon_{ij}$ for $i=2,3,...,K$
```

The equivalent null hypothesis is $H_0 : \beta_2 = \beta_3 = ... = \beta_K = 0$

See Lecture 15, Slides 24-27 for sample code (though it's less useful without the datasets)

## ANOVA 2

Multiple comparisons

==> want to check where the inequalities lie if we reject the overall null hypothesis

==> conducting multiple independent tests

increases the probability of making at least 1 type I error

Methods to adjust for this:

-Tukey (preferred in balanced designs, where you have similar numbers in each cell

==> Tukey CIs are narrower than Scheffe, so easier to reject)

-Scheffe (Use for contrasts (linear combinations where the coefficients sum to 0))

-Bonferroni (Use for factor level means ($\mu_1$, $\mu_2$,...), contrasts (linear combinations where the coefficients sum to 0), linear combinations)

```
#Tukey in R (Lecture 16, Slide 21)
group <- as.factor(group)
fit <-aov(dose~group,data=dat1)
TukeyHSD(fit,"group")
```

```
#Bonferroni in R (Lecture 16, Slide 25)
pairwise.t.test(dose, group,p.adj="bonf")
#Compare adjusted p-values
#to your regular alpha or compare the non-adjusted p-values
#to your alpha divided by (K choose 2)
```

## ANOVA 3

### Assumption Diagnostics

Again, HILE Gauss ==> Homogeneity of variance, Normality (and existence) of residual errors, Independence of Errors, and Linearity

With categorical covariates, we tend to assume the linearity assumption holds

Independence is typically a result of the sampling scheme and experimental design

### Testing Homogeneity of Variance

(Lecture 17, Slides 3-7)

Can plot raw data vs. group means AND plot standard deviation of data within each group vs. group means

```
#Use the modified Levene test (Brown-Forsythe test),
#which applies ANOVA to absolute deviation from group medians
#Robust to normality and does not require equal sample sizes
#REJECTION INDICATES LACK OF HOMOGENEITY
#In R Lecture 17 (Slide 7):
Levene <- function(y, group){
group <- as.factor(group) # precautionary
medians <- tapply(y, group, median)
resp <- abs(y - medians[group])
anova(lm(resp ~ group))[1, 4:5]
}
Levene(age, group)
#Can also change anova(lm(resp ~ group))[1, 4:5]
#to anova(lm(resp ~ group)) to get the ANOVA table
#instead of just the F-stat and p-value
```

### Testing Normality

(Lecture 17, Slides 8-12)

```
#QQ-plot
group<-as.factor(group)
av<-aov(age~group)
qq<-qqnorm(av$residuals)
```

```
qq
qqline(av$residuals,col="red") #Reference line

#Kolmogorov-Smirnov GOF test
#REJECTION suggests that residuals do not follow normal distribution
lillie.test(av$residuals)

#Pearson's Product-Moment Correlation on the theoretical and
#sample quartiles of the qqnorm plot
#See Lecture 17, Slide 9 for alpha=0.05 critical values
#Assumption of normality is question if observed correlation is
#less than or equal to the critical value listed on the above slide
#Lecture 17, Slide 11
cor.test(qq$x,qq$y)
```

## Remedial measures for normality include

1) Appealing to the CLT if your sample size is large enough
2) Transformations (especially Box-Cox)
3) Nonparametrics, especially Kruskal-Willis

```
#Box-Cox (in the MASS library)
boxcox(x,
    lambda = {if (optimize) c(-2, 2) else seq(-2, 2, by = 0.5)},
    optimize = FALSE, objective.name = "PPCC",
    eps = .Machine$double.eps, include.x = TRUE, ...)
#x is your "lm" object
#You can change the range/increments of powers to test for the transformation
#Pick the lambda value that maximizes the log-likelihood
#You can set optimize=TRUE for R to pick the value that
#maximizes log-likelihood in the range of c(-lower,upper),
#but it is better to pick a lambda that is more easily interpretable
#Use the value of lambda to transform your Y's
#(See Lecture 17, Slide 13 for how to apply transformation)
```

## Kruskal-Wallis

$Y_{ij} = \mu_i + \epsilon_{ij}$

$\epsilon_{ij}$ are independent and identically distributed with mean zero, but are not necessarily normal

The primary hypothesis of interest is:

$H_0 : \mu_i = \mu_k$ For all group means where $i \neq k$

The alternative hypothesis of interest is:

$H_a : \mu_i \neq \mu_k$ $(i \neq k)$ for at least one $i, k$

Test statistic is approximately $\chi^2_{K-1}$

```
#Kruskal-Wallis (Lecture 17, Slide 26)
kruskal.test(change,dose)
```

If K=2, then Kruskal-Wallis is equivalent to Wilcoxon rank sum test

For multiple comparison of means, use Wilcoxon rank sum tests with Bonferroni correction

```
pairwise.wilcox.test(change,dose,p.adjust.method="bonf")
```