# One-Way ANOVA (MD3) 

Paul Gribble

Winter, 2017

## Review from last class

- sample vs population
- estimating population parameters based on sample
- null hypothesis $H_{0}$
- probability of $H_{0}$
- meaning of "significance"
- t-test: what precisely are we testing?


## General Linear Model (GLM)

- we will develop logic \& rationale for ANOVA (and computational formulas) based on GLM
- any phenomenon is affected by multiple factors
- observed value on dependent variable (DV) $=$
- sum of effects of known factors +
- sum of effects of unknown factors
- similar to the idea of "accounting for variance" due to various factors


## General Linear Model (GLM)

- let's develop a model that expresses DV as a sum of known and unknown factors
- $\mathrm{DV}=\mathrm{C}+\mathrm{F}+\mathrm{R}$
- $\mathrm{C}=$ constant factors (known)
- $\mathrm{F}=$ factors systematically varied (known)
- $\mathrm{R}=$ randomly varying factors (unknown)
- notation looks like this:

$$
Y_{i}=\beta_{0}+\beta_{1} X_{1_{i}}+\beta_{2} X_{2_{i}}+\cdots+\beta_{n} X_{n_{i}}+\epsilon_{i}
$$

## Single-Group Example

- a little artificial (who ever does experiments using just one group?)
- but it will help us develop the ideas
- imagine we collect scores on some DV for a group of subjects
- we want to compare the group mean to some known population mean
- e.g. IQ scores where by definition, $\mu=100$ and $\sigma=15$


## Single-Group Example

- We know that:

$$
\begin{aligned}
& H_{0}: \bar{Y}=\mu \\
& \mathrm{H}_{1}: \bar{Y} \neq \mu
\end{aligned}
$$

- let's reformulate in terms of a GLM of the effects on DV:

$$
\begin{aligned}
& H_{0}: Y_{i}=\mu+\epsilon_{i} \text { where } \mu=100 \\
& H_{1}: Y_{i}=\hat{\mu}+\epsilon_{i} \text { where } \hat{\mu}=\bar{Y}
\end{aligned}
$$

- we call $H_{0}$ the restricted model - no parameters need to be estimated
- we call $H_{1}$ the full model - we need to estimate one parameter (can you see what it is?)


## Computing Model Error

- how well do these two models fit our data?
- let's use the sum of squared deviations of our model from the data, as a measure of goodness of fit

$$
\begin{aligned}
& H_{0}: \sum_{i=1}^{N}\left(e_{i}^{2}\right)=\sum_{i=1}^{N}\left(Y_{i}-100\right)^{2} \\
& H_{1}: \sum_{i=1}^{N}\left(e_{i}^{2}\right)=\sum_{i=1}^{N}\left(Y_{i}-\hat{\mu}\right)^{2}=\sum_{i=1}^{N}\left(Y_{i}-\bar{Y}\right)^{2}
\end{aligned}
$$

- remember: SSE about the sample mean is lower than SSE about any other number
- so the error for $H_{0}$ will be greater than for $H_{1}$
- so the relevant question then is, how much greater must $H_{0}$ error be, for us to reject $H_{0}$ ?


## Computing Model Error

- consider the proportional increase in error (PIE)
- $\left(E_{R}-E_{F}\right) / E_{F}$
- PIE gives error increase for $H_{0}$ compared to $H_{1}$ as a $\%$ of $H_{1}$ error
- but we want a model that is both
- adequate (low error)
- simple (few parameters to estimate)
- question: why do we want a simpler model?
- philosophical reason
- statistical reason


## Computing Model Error

- how big is increase in error with $H_{0}$ (restricted model), per unit of simplicity?
- let's design a test statistic that takes into account simplicity
- simplicity will be related to the number of parameters we have to estimate
- degrees of freedom $d f$ :
- \# independent observations in the dataset minus \# independent parameters that need to be estimated
- so higher $d f=$ a simpler model


## Computing Model Error

- let's normalize model errors (PIE) by model $d f$

$$
\frac{\left(E_{R}-E_{F}\right) /\left(d f_{R}-d f_{F}\right)}{\left(E_{F} / d f_{F}\right)}
$$

- guess what: this is the equation for the F statistic!

$$
F=\frac{\left(E_{R}-E_{F}\right) /\left(d f_{R}-d f_{F}\right)}{\left(E_{F} / d f_{F}\right)}
$$

- so if we can compute $F_{o b s}$, then we can look up in a table (or compute in R using pf() ) probabilities of obtaining that $F_{\text {obs }}$


## Two-Group Example

- let's look at a more realistic situation
- 2 groups, 10 subjects in each group
- test mean of group 1 vs mean of group 2
- do we accept $H_{0}$ or $H_{1}$ ?
- we will formulate this question as before in terms of 2 linear models
- full vs restricted model
- is the error for the restricted model significantly higher than for the full model?
- is the decrease in error for the full model large enough to justify the need to estimate a greater \# parameters?


## Hypotheses \& Models

$H_{0}: \mu_{1}=\mu_{2}=\mu$

- restricted model: $Y_{i j}=\mu+\epsilon_{i j}$
$H_{1}: \mu_{1} \neq \mu_{2}$
- full model: $Y_{i j}=\mu_{j}+\epsilon_{i j}$
symbols
- the subscript ${ }_{j}$ represents group (group 1 or group 2)
- ${ }^{\prime}$ represents individuals within each group (1 to 10 ) restricted model
- each score $Y_{i j}$ is the result of a single population mean plus random error $\epsilon_{i j}$
full model
- each score $Y_{i j}$ is the result of a different group mean plus random error $\epsilon_{i j}$


## Deciding between full and restricted model

- how do we decide between these two competing accounts of the data?
key question
- will a restricted model with fewer parameters be a significantly less adequate representation of the data than a full model with a parameter for each group?
- we have a trade-off between simplicity (fewer parameters) and adequacy (ability to accurately represent the data)


## Error for the restricted model

- let's determine how to compute errors for each model, and how to esimate parameters
error for restricted model
- sum of squared deviations of each observation from the estimate of the population mean (given by the grand mean of all of the data)

$$
\begin{aligned}
E_{R} & =\sum_{j} \sum_{i}\left(Y_{i j}-\hat{\mu}\right)^{2} \\
\hat{\mu} & =\left(\frac{1}{N}\right) \sum_{j} \sum_{i}\left(Y_{i j}\right)
\end{aligned}
$$

## Error for the full model

error for the full model

- now we have 2 parameters to be estimated (a mean for each group)

$$
\begin{aligned}
E_{F} & =\sum_{j=1}^{2} \sum_{i}\left(Y_{i j}-\hat{\mu_{j}}\right)^{2} \\
E_{F} & =\sum_{i}\left(Y_{i 1}-\hat{\mu}_{1}\right)^{2}+\sum_{i}\left(Y_{i 2}-\hat{\mu_{2}}\right)^{2} \\
\hat{\mu}_{j} & =\left(\frac{1}{n_{j}}\right) \sum_{i}\left(Y_{i j}\right), \quad j \in\{1,2\}
\end{aligned}
$$

## Deciding between full and restricted model

- now we formulate our measure of proportional increase in error (PIE) as before:

$$
F=\frac{\left(E_{R}-E_{F}\right) /\left(d f_{R}-d f_{F}\right)}{E_{F} / d f_{F}}
$$

- this is the F statistic!
- df-normalized proportional increase in error for restricted model $\left(H_{0}\right)$ relative to the full model $\left(H_{1}\right)$


## Model Comparison approach vs traditional approach to ANOVA

- how does our approach compare to the traditional terminology for ANOVA? (e.g. in the Keppel book and others)
- traditional formulation of ANOVA asks the same question in a different way
- is the variability between groups greater than expected on the basis of the within-group variability observed, and random sampling of group members?
- MD Ch 3: proof that computational formulae are same
- see MD Chapter 3 for description of the general case of one-way designs with more than 2 groups ( N groups)


## Assumptions of the F test

1. the scores on the dependent variable $Y$ are normally distributed in the population (and normally distributed within each group)
2. the population variances of scores on $Y$ are equal for all groups
3. scores are independent of one another

## Violations of Assumptions

- how close is close enough to normally distributed?
- ANOVA is generally robust to violations of the normality assumption
- even when data are non-normal, the actual Type-I error rate is close to the nominal value $\alpha$
- what about violations of the homogeneity of variance assumption?
- ANOVA is generally robust to moderate violations of homogeneity of variance as long as sample sizes for each group are equal and not too small ( $>5$ )
- independence?
- ANOVA is not robust to violations of the independence assumption


## Testing assumptions in $R$

In R you can test for:

- normality
- homogeneity of variance


## Some example data

| Group 1 | Group 2 | Group 3 |
| ---: | ---: | ---: |
| 4 | 7 | 6 |
| 5 | 4 | 9 |
| 2 | 6 | 8 |
| 1 | 3 | 5 |
| 3 | 5 | 7 |
| mean=3 | mean=5 | mean=7 |

## Some example data: Restricted model

## 1 Parameter to Estimate



## Some example data: Full model

## 3 Parameters to Estimate



## Next Class

- testing differences between specific pairs of means
- controlling Type-I error rate
- statistical power calculations


## R code

- one-way single factor ANOVA using R, using the aov() function
- tests for homogeneity of variance
- var.test() (2 groups)
- bartlett.test() (> 2 groups)
- test for normality using shapiro.test()

