HW4 - SOLUTIONS

YOU WERE ASKED….ON MOVIES DATA SET

- “*What are the most promising alternative hypotheses about imdb scores to test? Name your three top candidates along with the evidence which backs them up: either in the form of R instruction(s) or plot*”

SLICING AND DICING MOVIES…..HOW MANY SLICES?

- High Budget Low Gross Comedies?
- UK Family movies?
- US Low Budget Dramas?
- High Budget Low Gross movies?
- R-rated Comedies?
- G-rated Family movies
- G-rated High Budget movies?

HOW MANY?

- Probably 100 slices? 1000?

USE PLOTS OR FUNCTIONS: TAPPLY, MEAN, SUBSETTING

- moviesSlice <- movies[Condition,]
- tapply(moviesSlice$imdb_score, attribute, mean)
- mean(moviesSlice$imdb_score)

WHAT DID I EMBED IN THE DATA?

- Low Budget History Movies > High Gross Action Movies   (mean imdb)

WHAT ELSE

- High Budget Family Movies < Low Budget Action Movies (mean imdb)

REAL JEWELS

8.46 mean imdb for **High gross UK History movies**

4.13 – mean imdb for **Low gross UK family movies**

P-VALUE HUNTING….

- Bad practice as we will discuss next week when talking about multiple hypotheses corrections

HOW MANY HYPOTHESES?

- SLICE 1 vs SLICE 2
- SLICE1 and SLICE2 better be disjoint
- Even then combinatorial explosion
- If 100 possible slices….could be up to 100000 comparisons

P-VALUE HUNTING WILL HAVE SEVERE CONSEQUENCES

For the significance level

5% will no longer do it!

Bonferroni correction will be needed