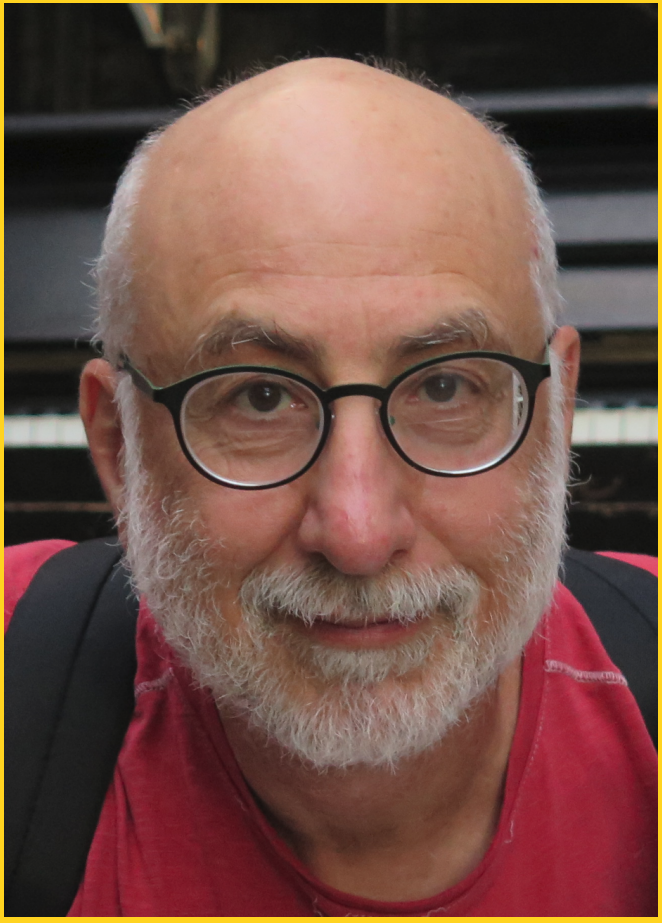
P-values and related concepts, part 2

Matt Kramer
NEA Statistics Group

matt.kramer@ars.usda.gov
301.504.8121



Recap of Part 1

- *p*-values were developed ad-hoc
- It is the probability of getting results at least as extreme as the ones you observed, *given that the null hypothesis is correct*
- Not a measure of the *strength* of an effect or relationship
- Gives the signal/noise ratio as a single value
- Confidence intervals are more informative

Outline for Part 2

- Power
- Effect size
- Variance decomposition
- Exploratory and confirmatory research and multiple comparisons
- Multiple dependent variables
- Conclusions

p-values and power

If there are true differences among treatments or among slopes of regression lines, a more powerful design for the same sample size will produce lower *p*-values.

- A power analysis will tell you (a) how large a difference among treatments is detectable, given (b) sample size and (c) estimates of the random effects parameters (e.g. residual variance, block-to-block variance).
- If you know (a) and (c), a power analysis can find (b) sample size needed. This is a common use of power analysis.

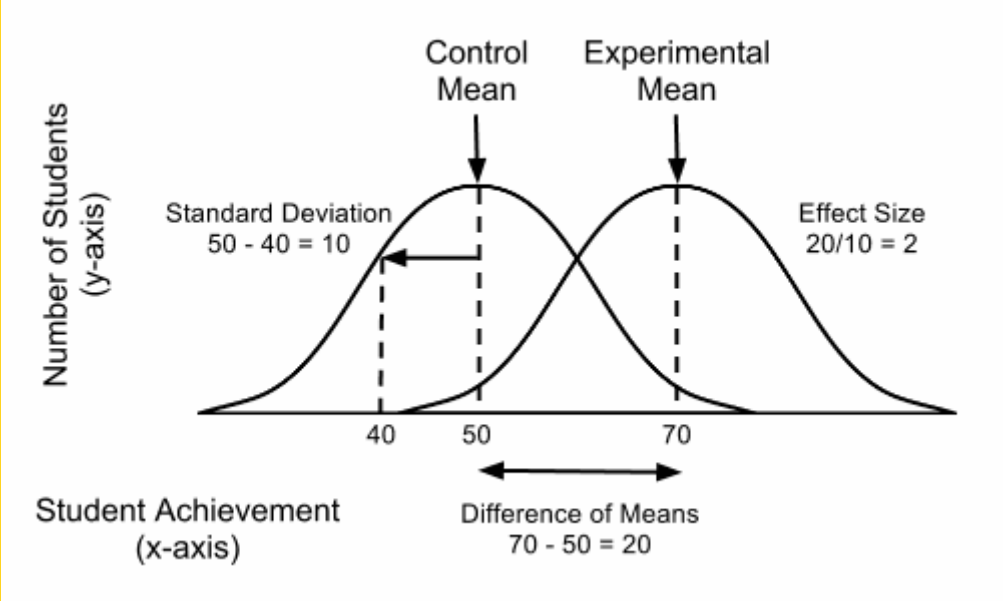
How to do a power analysis

- For *t*-tests, the simplest way is explained in Festing, MFW. 2018. On determining sample size in experiments involving laboratory animals. Laboratory Animals: https://doi.org/10.1177/0023677217738268.
- For all fixed effects models (ANOVA, regression), there is readily available software (SAS, R, etc.) as well as websites to do the calculation, see:

  http://psych.wisc.edu/henriques/power.html
  http://www.datavis.ca/online/power/
- For models with random effects, it is more difficult; you either need to set up an exemplary data set (see *Stroup, W.W. 2012. Generalized Linear Mixed Models: Modern Concepts, Methods and Applications, CRC Press* for details and how to do subsequent calculations) or use simulation (setting up many example data sets and seeing how many give you $p < 0.05$).

Effect Size

- One idea closely linked with *p*-values (and power) is effect size, how the large the difference is between treatments, relative to other sources of variation.



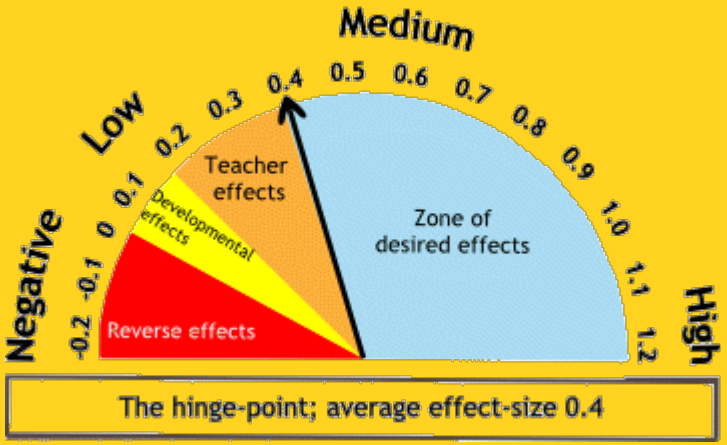
- If the effect size is large, it will be easy to separate the treatments; *p*-values (both a priori and a posteriori) will be small (i.e. the treatment effect is significant).
- If the effect size is small, it will be difficult to find a treatment effect without a large sample size.

Calculating and reporting effect size

- Effect size for the difference between two means is calculated as

  Cohen's $d = (\overline{X}_1 - \overline{X}_2)/S_p$,

  where S_p is the pooled standard deviation.



  Effect sizes: small: $d = 0.2$; medium: $d = 0.5$; large: $d = 0.8$

- Some journals are now recommending reporting effect size.

Variance decomposition

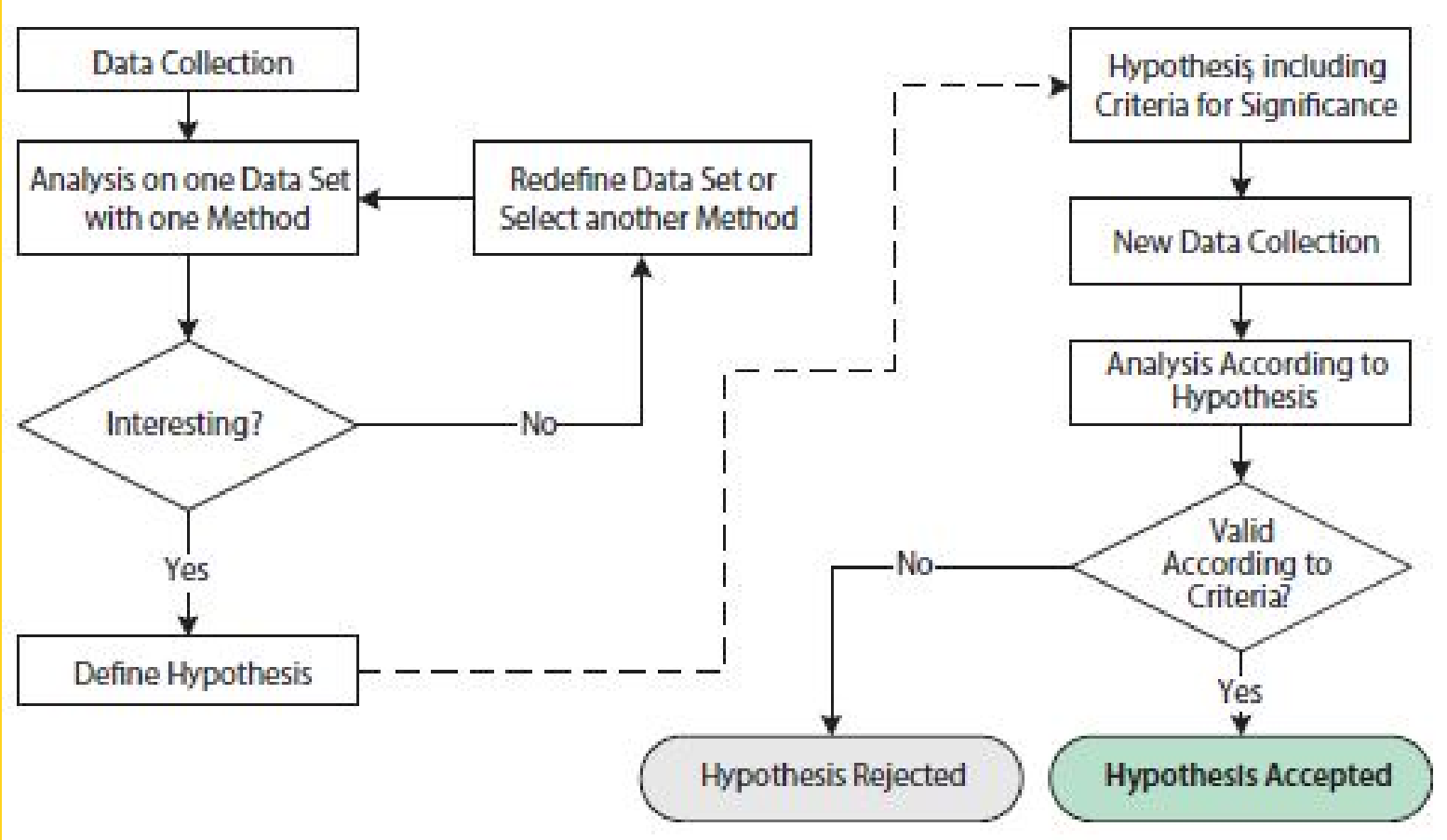
- A potentially superior way to impart how much of the variability in the data set is attributable to the treatment effect is to perform a variance decomposition.
- This is especially useful when there are many potential independent variables affecting samples, e.g., fixed and random effects and their interactions.
- If the model also includes quantitative variables, they can be binned into levels of qualitative factors.

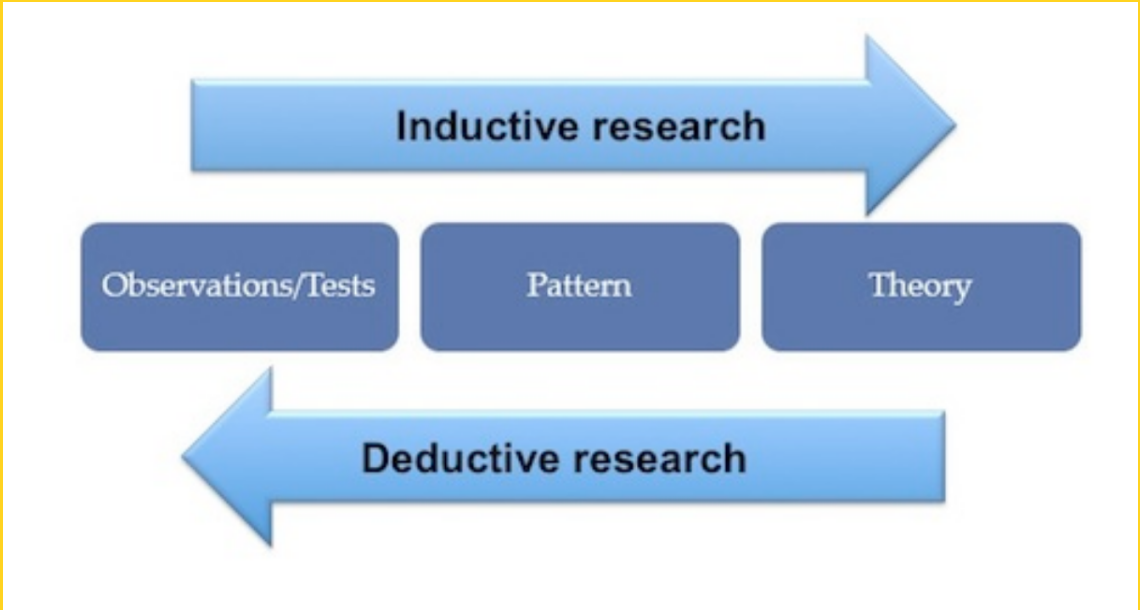
Example variance decomposition table

Source of variation	variance	percent
Factor 1	140.5	8
Factor 2	158.0	9
Covariate	210.7	12
Block-to-block	316.1	18
Residual	930.1	53
Total	1755.4	100

Exploratory or Confirmatory Research and Multiple Comparisons

- JW Tukey: “Finding the question is often more important than finding the answer.”
- Whether a project is in an exploratory phase (e.g., which bio-engineered cultivars produce reasonable yields in non-irrigated mid-west farms?) or a confirmatory phase (e.g., do three new cultivars outperform the traditional one?) should dictate the kinds of statistics used.
- In particular, the balance between type I and type II error changes.


Data Collection
Analysis on one Data Set with one Method
Redefine Data Set or Select another Method
Interesting?
No
Yes
Define Hypothesis
Hypothesis including Criteria for Significance
New Data Collection
Analysis According to Hypothesis
Valid According to Criteria?
No
Yes
Hypothesis Rejected
Hypothesis Accepted


Inductive research
Observations/Tests
Pattern
Theory
Deductive research

Exploratory Research

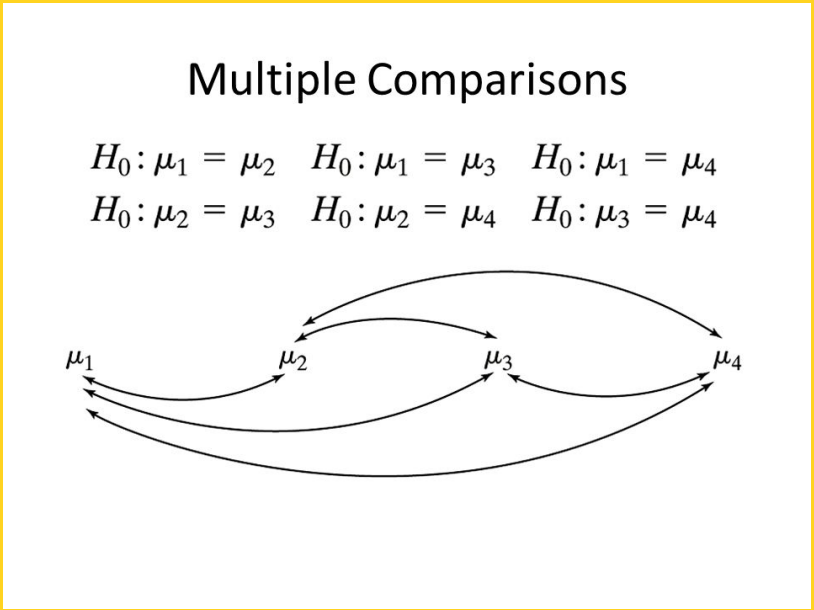
- Controlling type II error is more important
- More concern about saying that a finding is not important when in truth it is (and conversely, less concern saying that a finding is important when it isn’t)
- In a paper, explain that the research is exploratory in nature and use that to justify multiple comparisons methodologies that do not control type I error rates (e.g. Duncan’s—and this is the only circumstance this method should be used).

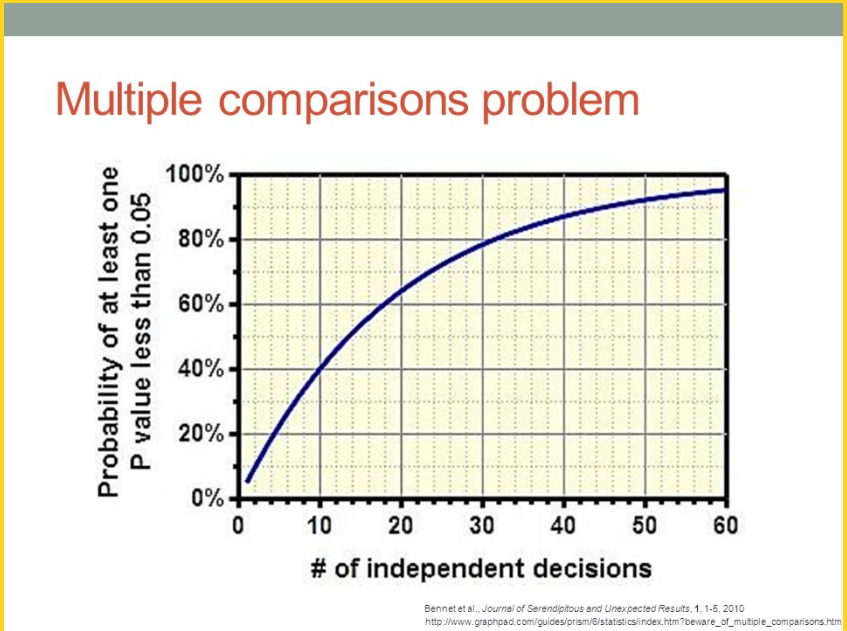
Confirmatory Research

- Controlling type I error rate is more important
- More concern about establishing ‘proof’ that something is true
- Most published research is assumed to be confirmatory, so multiple comparisons should be made with good type I error control

Multiple Comparisons for Confirmatory Research

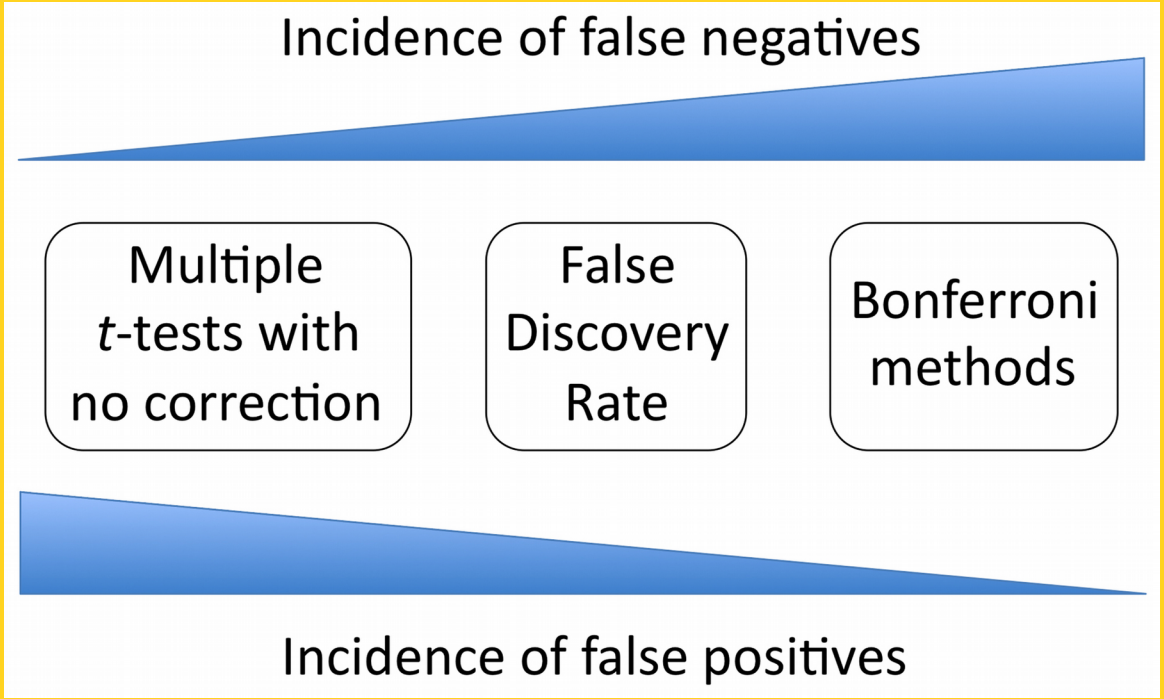
- Your factor has more than two levels and you want to know which levels differ. These are not a priori comparisons.
- You can not make these tests at $\alpha = 0.05$ because you are testing subsets of the data that were previously tested. This increases the probability of falsely rejecting hypotheses of no difference.
- To control for this, one must adjust α; there are many acceptable techniques.



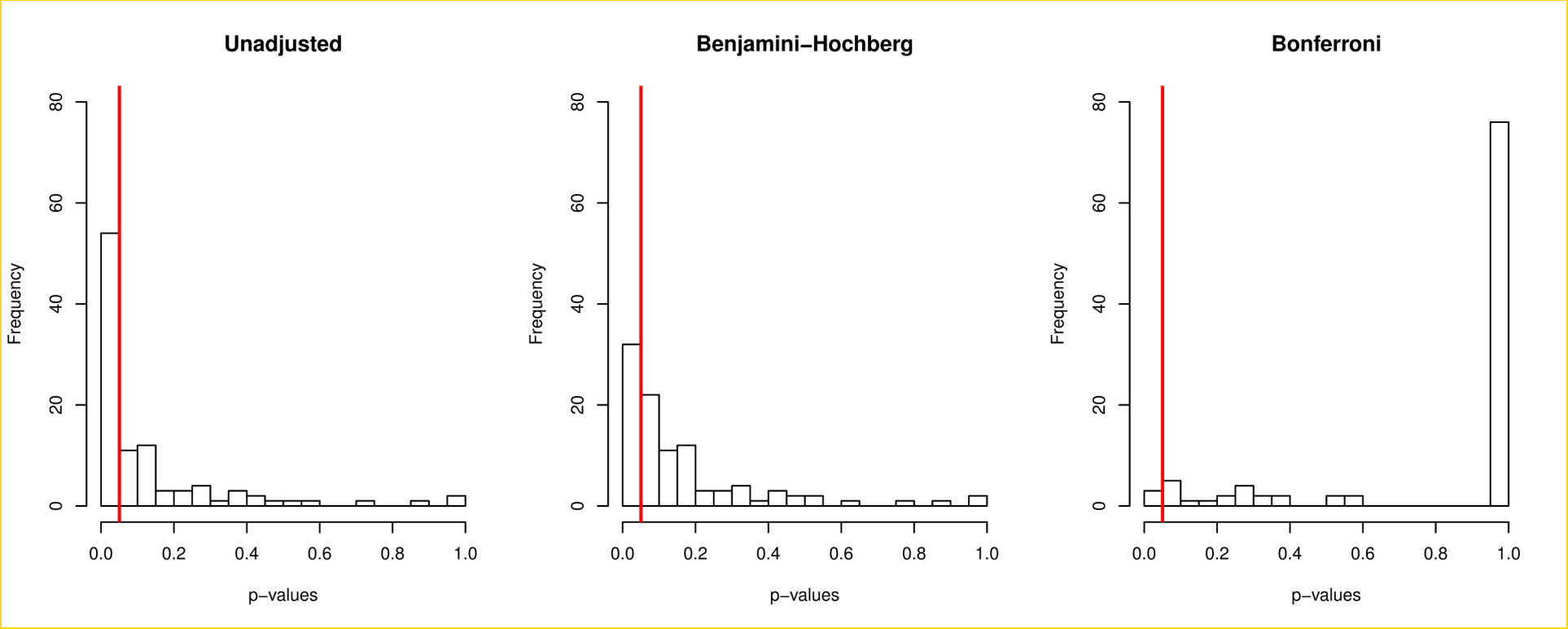


False discovery rate

- Another approach; instead of controlling for the probability of at least one error (family-wise error rate), one controls for the expected proportion of errors (false discoveries).
- This is a useful idea when hundreds or thousands of tests are made, as is common in genomics; controlling for family-wise error rate would be prohibitively conservative.

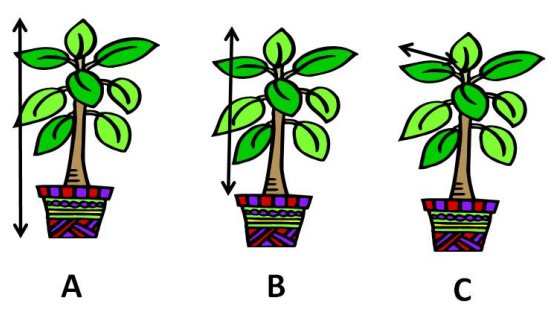


Example using Benjamini-Hochberg for an experiment with 100 *p*-values



Multiple dependent variables

- You did an experiment where several characteristics were measured on each sample.



- For example, on each plant you measured: approximate number of mature leaves on last 30 cm of terminal (or longest) branch, leaf area of the youngest five mature leaves on the terminal shoot, stem diameter at base, plant height, plant width (average of largest and smallest measure of ellipsoid created from plant shadow at noon), number of flowers, average size of five flowers (maximum diameter), flower quality (score of 0 – 5), and health (score of 0 – 5).
- There is some overlying treatment structure (e.g. varieties, growth conditions, origin) and design structure (e.g. location, blocking, greenhouse).

These variables have different properties (distributions).

- Quantitative (but not necessarily normally distributed)
  - Leaf area
  - Stem diameter
  - Height
  - Width
  - Flower size
- Categorical (or ordered categorical)
  - Flower quality
  - Plant health
- Count
  - Number mature leaves
  - Number of flowers

Some of these variables are positively correlated with others, e.g. plant height, plant width, stem diameter, health, number of flowers; others possibly negatively correlated, e.g. number of flowers and flower quality.

They may be on different scales, e.g. height (cm), leaf area (cm^2), or unknown or poorly defined scales (flower quality).

Multiple dependent variables

- DO NOT analyze each dependent variable separately without taking into account that you are in a repeated measures scenario.
- DO NOT test each dependent variable at $\alpha = 0.05$.
- The different characteristics are correlated through the individual plants; they are not independent. For them to be independent, each characteristic would have to be measured on a different set of plants.

Approaches

- Under a fairly restrictive set of conditions, MANOVA may be useful (e.g., no random effects, all variables are approx. normally distributed, the same independent variables explain all the DVs). This will correctly adjust testing and *p*-values for the multiple dependent variables.
- Dimension reduction methods to reduce the number of dependent variables. Note that several of the variables are related to “size”. This might be captured in the first dimension of a PCA. Rescale variables to the same scale if possible (e.g., take the square root of area measures). Transformations may also be useful.

Approaches (2)

- For ordered or categorical data, perhaps the number of categories can be reduced, e.g., a plant can be diseased or not diseased; flowers can be of high or low quality.
- If each variable is modeled independently, there must be experiment-wise error control. The *p*-values need to be adjusted using an acceptable method, e.g., multiple comparisons or FDR.

Conclusions

There are better, though not as universally acceptable, ways to demonstrate the importance of research results than *p*-values, e.g., effect size, variance decomposition.

In the face of repeated measures, multiple comparisons, etc., unadjusted *p*-values can incorrectly inflate the number of “significant” differences. The balance between type I and type II error can be used to decide what kind of statistical adjustment (if any) should be made to *p*-values.