Introduction to Statistics and R

LSA Summer Institute 2009

Peter Graff

Statistics
Variables

- Variables
  A characteristic of persons, things, language, situations, actions, etc., which
  - Can be measured objectively for each instance of the population of interest, e.g.,
  - Can change between instances
  - Has multiple LEVELS = number of different values which variable can assume.
- There are two main roles variables can assume in experimental studies.

Statistics
Variables

- INDEPENDENT VARIABLE = IV
  Definition: the potential cause for an effect; a variable which the experimenter selects at – or adjusts to – different LEVELS
  - Levels of the i.v. determine the EXPERIMENTAL CONDITIONS (also called CELLS) of the design
  - Number of levels of i.v. used in experiment must be > 1, because there must be a CONTROL for comparison.

Statistics
Variables

- DEPENDENT VARIABLE = DV
  Definition: the potential effect
  - The dependent variable (‘d.v.’) is examined for the result.

Statistics

Scales: Nominal Data

- There are 3 major types of variable scales: NOMINAL, ORDINAL and INTERVAL/RATIO
- NOMINAL DATA
  In a fruit basket there are 5 bananas, 4 oranges and 2 cherries. All we can do is count the different types of fruit. It doesn't make sense to talk about an individual fruit as being more or less banana, it's either a banana or not.

Statistics
Scales: Ordinal Data

- ORDINAL DATA
  The different types of fruit in the fruit basket are rated for tastiness by a fruit enthusiast on a scale with values “not yummy”, “yummy” and “very yummy”. Bananas are rated “very yummy”, oranges are rated “yummy” and cherries “not yummy”. We can say that bananas are tastier than oranges which are tastier than cherries. However, we don’t know whether the difference in tastiness between bananas and oranges is the same as the difference in tastiness between oranges and cherries. The enthusiast might loathe cherries but only have a slight preference for bananas over oranges.

Statistics
Scales: Interval/Ratio Data

- INTERVAL DATA
  The fruit farmers' association measures the length of every banana in the basket.
  - Banana 1: 6 inches
  - Banana 2: 5.5 inches
  - Banana 3: 8 inches
  - Banana 4: 7 inches
  - Banana 5: 7.5 inches

  This time we can calculate the exact difference in length between banana X and banana Y and we know that the difference between 2 points on the scale is always the same.

Statistics

Scales

Scale	consistent coding				permitted operations	examples
	labels	order	Intervals	Zero point		
Category /nominal	Y	N	N	N	counting (freq)	Favourite fruit
Ordinal	Y	Y	N	N	counting (freq) Ranking	Syntactic judgments
Interval	Y	Y	Y	N	counting (freq) ranking + –	Shoe size
Ratio	Y	Y	Y	Y	counting (freq) ranking + – × ÷ √, etc	VOT

Statistics
Scales in R

- Interval/Ratio
  Anything numerical is automatically encoded as Interval/Ratio
- Ordered
  To turn a numerical vector into an ordered vector use as.ordered()
- Nominal (Factor)
  Anything alphabetical is automatically encoded as a Factor
  To turn a numerical vector into a factor vector use as.factor()

Statistics

Statistical Operators and the Formula in R

- Prediction “~”
- Crossing “+”
- Interaction “:”
- Crossing and Interaction “^” and “*”
- Grouping “|”

- Formula
  DV~IV+IV
  DV~IV*IV
  DV~IV:IV
  DV~(IV+IV)^2
  DV~IV+IV+Error(IV)
  DV~IV+IV+ (1|IV)

Statistics
Descriptive Statistics

- Interval/Ratio Data
  - mode()
  - median()
  - mean()
  - variance
    var()
  - standard deviation
    sd()
- Nominal Data
  - mode()
- Ordinal Data
  - mode()
  - median()

Statistics
Reasoning

- Abductive Reasoning
  - Generalizing from striking examples
- Observational Inductive Reasoning
  - Examine set of examples o1, o2, o3, o4, o5,...
  - Find an explanatory description of some of these
  - Test it on more observations: on+1 , on+2, on+3 , ….
  - If it fits, be encouraged
- Hypothetical Deductive Reasoning
  - Refine a conjecture into a Yes/No question
  - **NULL HYPOTHESIS**
    Usually: Any difference between samples is due to chance
  - **ALTERNATE/EXPERIMENTAL HYPOTHESIS**
    Usually: The difference between the samples is due to the experimental manipulation.

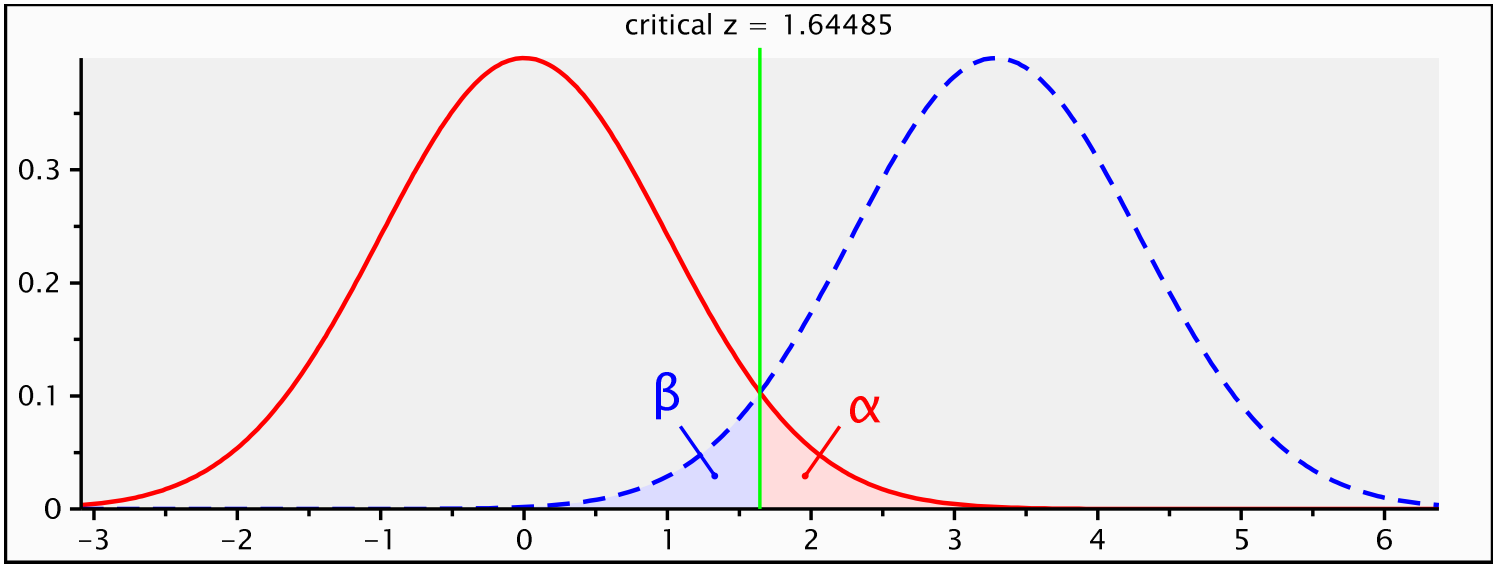
Statistics
Hypothesis Testing

- A hypothesis is the statement of a scientific question in terms of a proposed value of a parameter in a probability model. Hypothesis testing is a process of establishing proof by falsification. It has two essential components: a null hypothesis and an alternative hypothesis.
- The null hypothesis is a stated value of the parameter which defines the hypothesis we want to falsify. It is usually stated as a single value although it can be composite. We denote it as *H0.*
- The alternative hypothesis is the hypothesis whose veracity we wish to establish. It is usually defined by a value or set of values of the parameter that are different from the one specified in the null hypothesis. We denote it as *H1.*

Statistics

Error

- Type I Error
  Pr(Type I error)=Pr(rejecting H0|H0 true)=α
- Fisher's Criterion
  α = .5
- Type II Error
  Pr(Type II error)=Pr(not reject H0|H1 true)=β
- power = 1-β and should be .8 for important hypotheses



Statistics
Power Analysis: Say What?

- Power analysis has 4 components:
  - Effect Size: **d** OR **r**
  - Number of observations: **n**
  - Pr(Type I Error): **α**
  - Pr(Type II Error): **β**
- You choose or calculate 3 of the 4 and it gives you the 4th.

Statistics
Power Analysis: Effect Size

- How you calculate the effect size depends on the statistical test you're using:
- T-Test: $d = \frac{\bar{x}_1 - \bar{x}_2}{s}, \; s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2}},$
- ANOVA: $f^2 = \frac{R^2}{1 - R^2}$
- Chi-Squared: $\phi = \sqrt{\frac{\chi^2}{N}}$
- General: Pearson Correlation `lm(`*`formula`*`)`

Statistics
Power Analysis in R

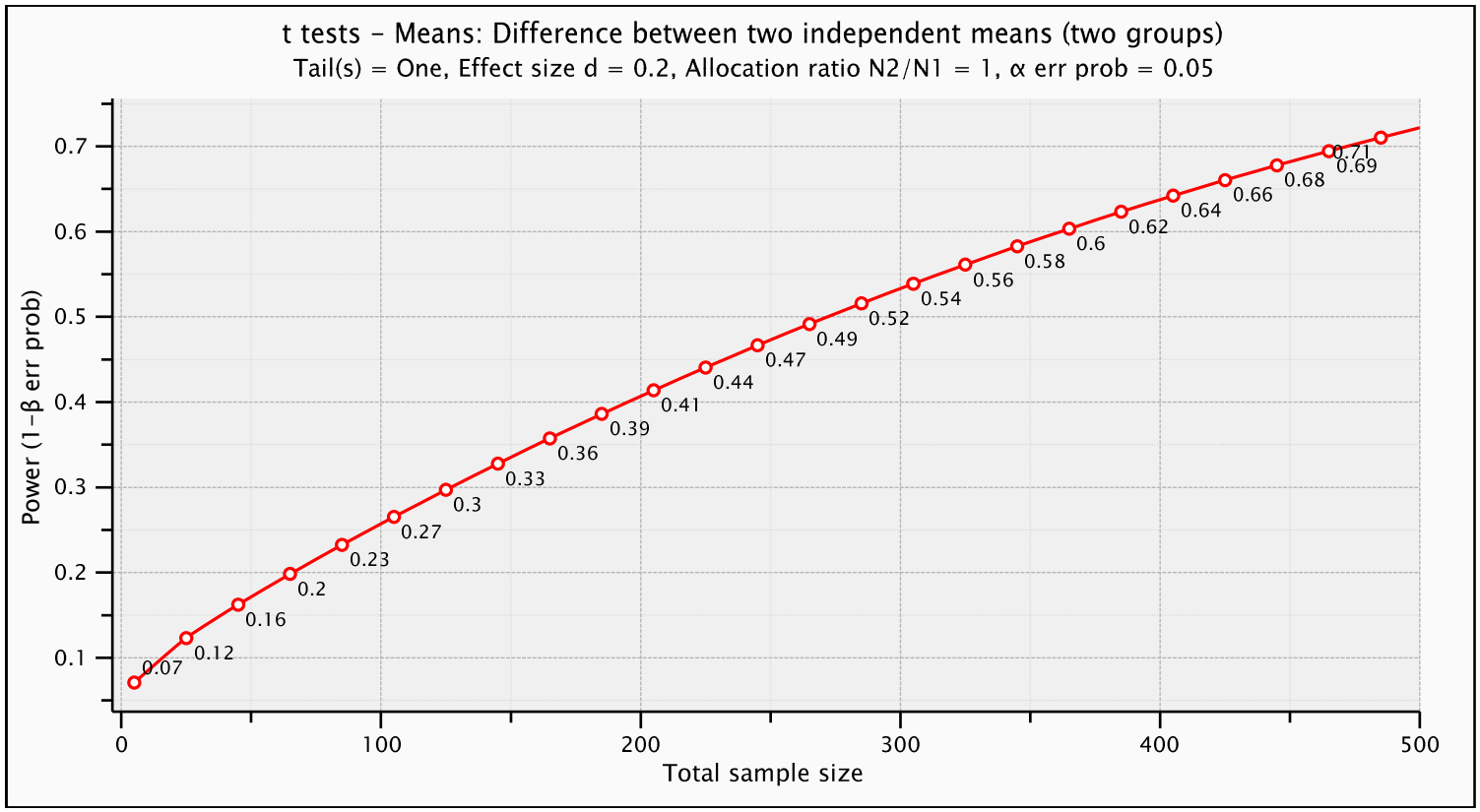
- In order to perform power analysis in R, you need to download the package `pwr`. The power calculation you use depends again on the test. One of the 4 arguments must be passed as NULL
  - `pwr.t.test(n = NULL, d = NULL, sig.level = 0.05, power = NULL, type = c("two.sample", "one.sample", "paired"), alternative = c("two.sided", "less","greater"))`
  - `pwr.anova.test(k = NULL, n = NULL, f = NULL, sig.level = 0.05, power = NULL)`
  - `pwr.chisq.test(w = NULL, N = NULL, df = NULL, sig.level = 0.05, power = NULL)`

Statistics
*G*Power*

- Yes, you can use R to do power analysis, but...
- There is a fantastic, free program out there that performs power analysis and makes beautiful figures at the same time. It's called G*Power and here's where you can download it:
  - **http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/**

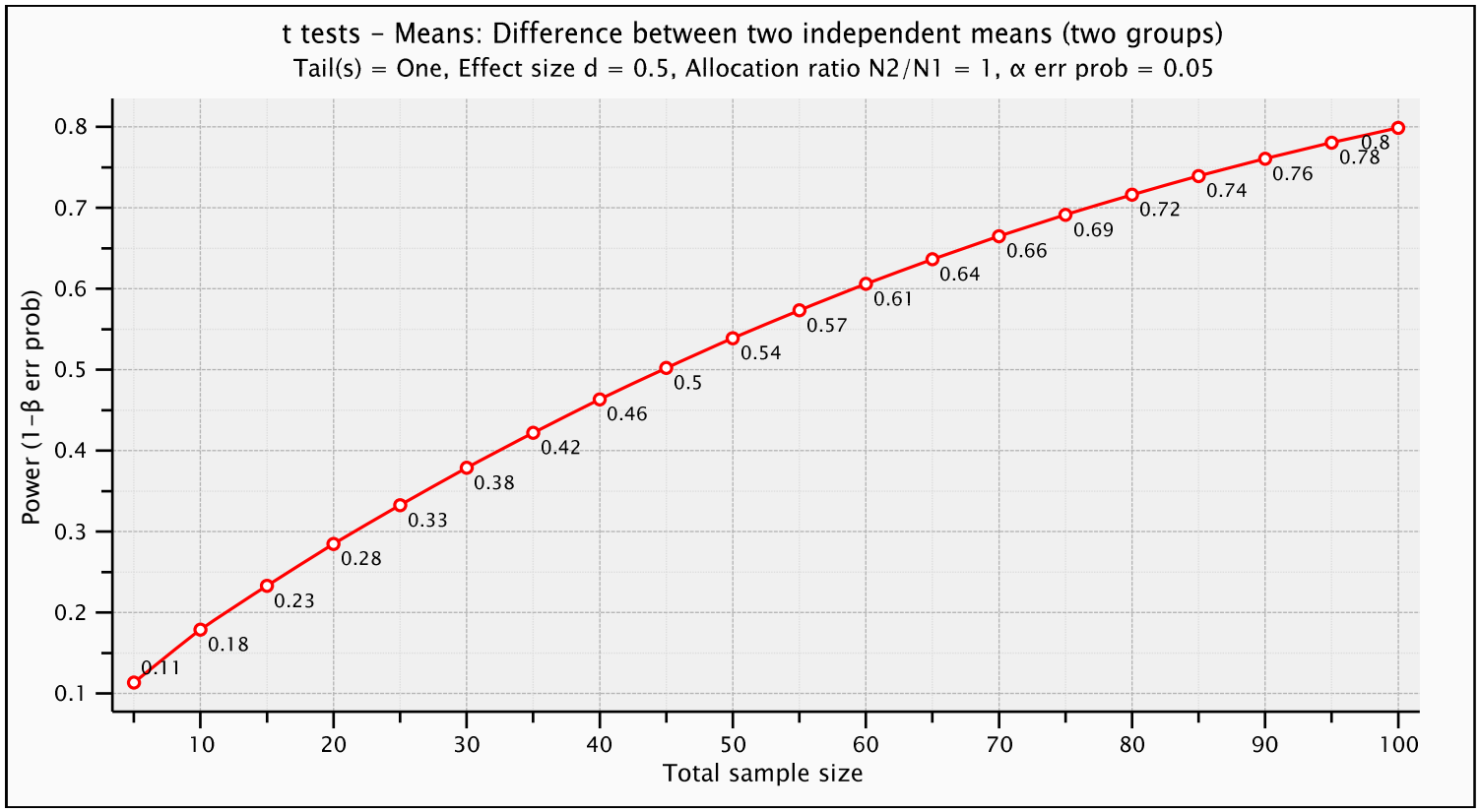
Statistics

Power vs. Sample Size (d=.2)



Statistics

Power vs. Sample Size (d=.5)



Statistics

Power vs. Sample Size (d=.8)