Sample size calculation with GPower

Wilfried Cools (ICDS) & Sven Van Laere (BiSI)
https://icds-vubuz.github.io/

sample size: why bother ?

- what about my sample size ?
- more is better ≡ observation → information
  - but increasingly less so
  - but limited resources (time/money)
  - but ethical considerations
- the question → how many is good enough ?
- the workshop → understand how to do it !
  - simple but common situations
- not one simple formula for all !
  - GPower to the rescue

sample size calculation: a design issue

- linked to statistical inference
  - `testing` → power [probability to detect existing `effects`]
  - `estimation` → accuracy [size of `confidence intervals`]
- before data collection, during design of study
  - requires understanding: future data, analysis, inference (effect size, focus, ...)
  - conditional on assumptions & decisions
- not always possible nor meaningful !
  - easier for experiments (control), less for observational studies
  - easier for confirmatory studies, much less for exploratory studies
  - !! retrospective power analyses → OK for future study only
    Hoenig, J., & Heisey, D. (2001). The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. The American Statistician, 55, 19–24.

simple example

- evaluation of radiotherapy to reduce a tumor in mice
- comparing 2 groups, a treatment and a control (=condition)
- tumor induced, random assignment treatment or control (equal if no effect)
- after 20 days, measurement of size (=observations)
- analysis:
  - unpaired t-test to compare average for treatment with average for control
- goal:
  - if the average size in treatment is at least 20% less than control
    then we want to detect it (significance)
- the main issue:
  - how to calculate the required sample size ?

overview

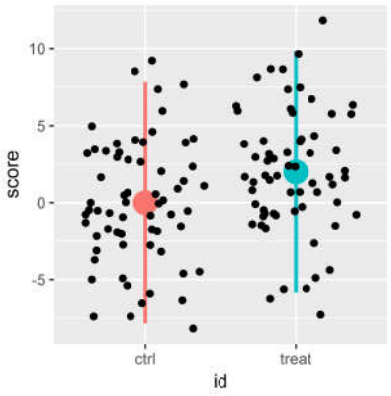
- PART I: building blocks in action for t-test
  - sizes: effect size, sample size
  - errors: type I (α), type II (β)
  - distributions: Ho, Ha
  - criterium: confidence (estimation), power (testing)
- PART II: moving beyong independent t-test
  - dependent groups
  - non-parametric distributions
  - multiple groups (ANOVA: omnibus, pairwise, focused)
  - proportions, correlations, ...

PART I: building blocks in action for t-test

reference example

- sample sizes easy and meaningful to calculate for well understood problems

- apriori specifications
  - intend to perform a statistical test
  - comparing 2 equally sized groups
  - to detect difference of at least 2
  - assuming an uncertainty of 4 SD on each mean
  - which results in an effect size of .5
  - evaluated on a Student t-distribution
  - allowing for a type I error prob. of .05 (α)
  - allowing for a type II error prob. of .2 (β)
- sample size conditional on specifications being true



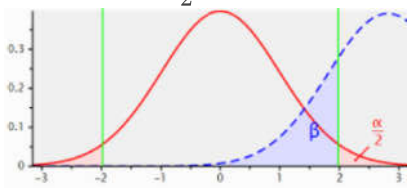
https://icds.shinyapps.io/shinyt/

a formula you could use

- for this particular case:
  - sample size (n → ?)
  - difference (d=signal → 2)
  - uncertainty (σ=noise → 4)
  - type I errors (α → .05, so $Z_{\alpha/2}$ → -1.96)
  - type II errors (β → .2, so Z_{β} → -0.84)

$$n = \frac{(Z_{\alpha/2}+Z_{\beta})^2 * 2 * \sigma^2}{d^2}$$

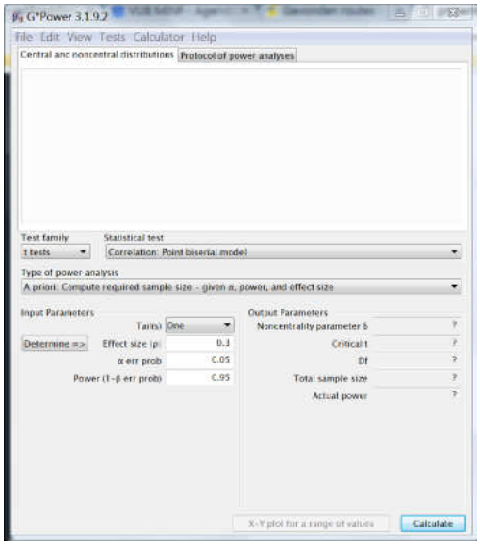
$$n = \frac{(-1.96-0.84)^2 * 2 * 4^2}{2^2} = 62.79$$



- sample size = 2 groups x 63 observations = 126
- note: formula's are tests and statistic specific but logic remains same
- this and other formula's implemented in various tools, our focus: GPower

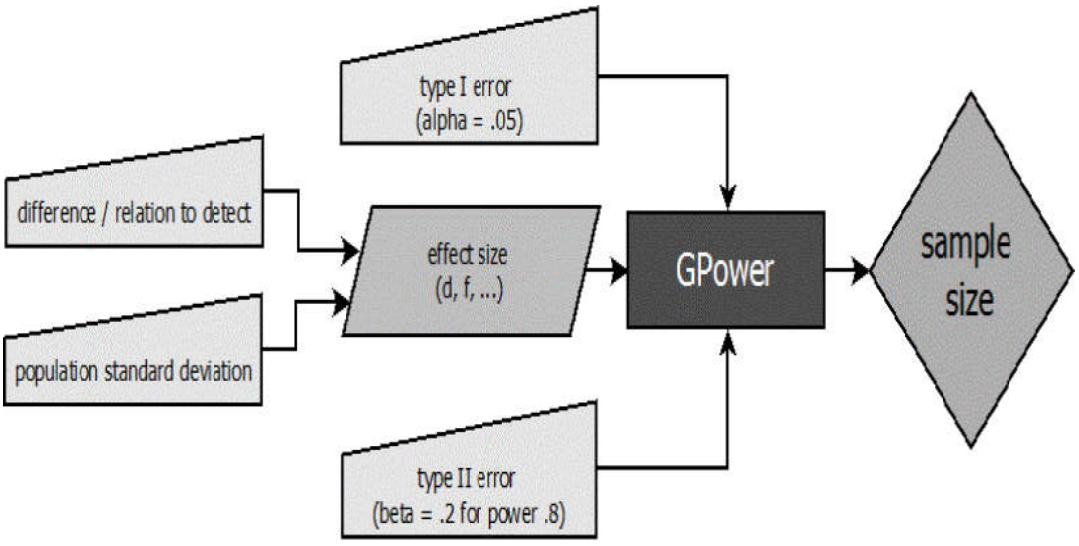
GPower: a useful tool

- popular and well established
- free @ http://www.gpower.hhu.de/
- implements wide variety of tests
- implements various visualisations
- documented fairly well
- note: not all tests are included !



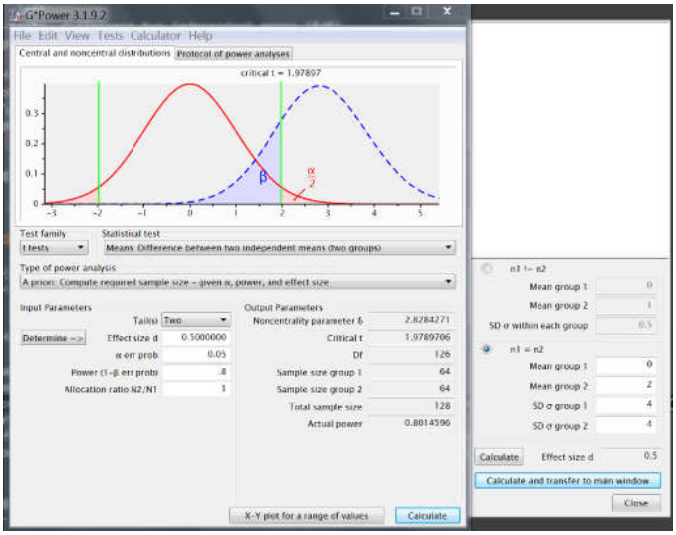
GPower: the building blocks in action

- For this example, calculate sample size based on
  - effect size (difference of interest, scaled on standard deviation)
  - type I and type II error



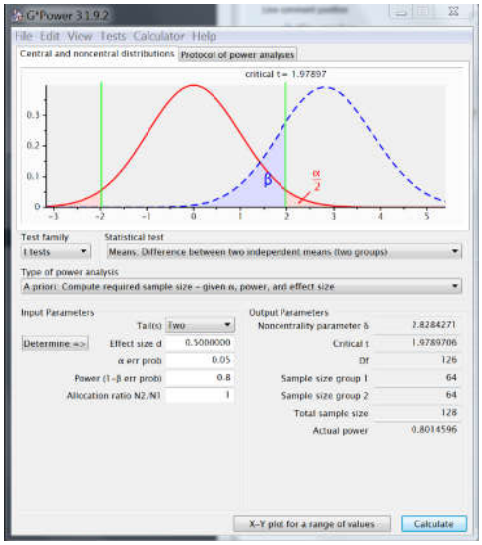
GPower input

- t-test : difference two indep. means
- apriori: calculate sample size
- 2 - tailed
- effect size = standardized difference
  - d = |difference| / SD_pooled
  - d = `|0-2|` / `4` = `.5`
- α = `.05` and $power = 1 - \beta$ = `.8`
- allocation ratio = `1`
- `~ reference example`



GPower output

- effect size: Cohen's d = 0.5 (`2/4`)
- sample size (n) = 64 x 2 = (`128`)
- degrees of freedom (df) = 126 (128 - 2)
- plot showing null and alternative distribution
  - in GPower central and non-central distribution
  - `Ho` & critical value → decision boundaries
    - critical t = 1.979, `qt(.975,126)`
  - `Ha`, shift with non-centrality parameter → truth
    - non centrality parameter (δ) = 2.8284
      `2/(4*sqrt(2))*sqrt(64)`
- power ≥ .80 (1-β) = 0.8015



reference example protocol

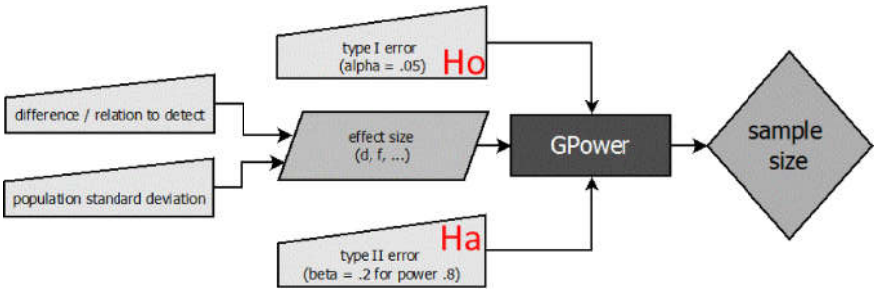
t tests - Means: Difference between two independent means (two groups)
Analysis: A priori: Compute required sample size

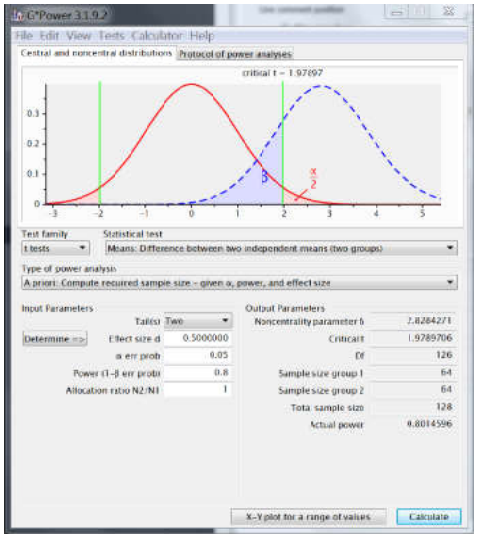
Input: Tail(s) = Two
Effect size `d` = `0.5000000`
`α` err prob = `0.05`
Power (`1-β` err prob) = `.8`
Allocation ratio N2/N1 = 1

Output: `Noncentrality` parameter δ = `2.8284271`
`Critical t` = `1.9789706`
Df = 126
Sample size group 1 = `64`
Sample size group 2 = `64`
Total `sample size` = `128`
Actual power = 0.8014596

sample size calculation for testing requires Ha

- null distribution `Ho ~ N(0,1)` cut off using α
  - used in the statistical test
- alternative distribution `Ha ~ N(ncp,1)`
  - to evaluate power (1-β) or sample size using cut off on `Ho` using α
- shift `Ho` → `Ha` ~ non-centrality parameter (`ncp`)
  - combines `effect size` & `sample size`

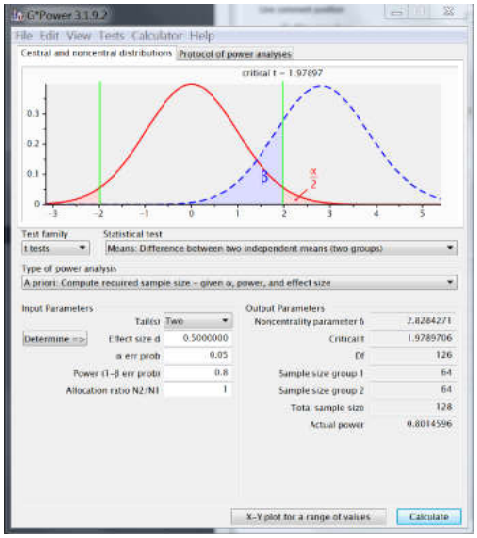




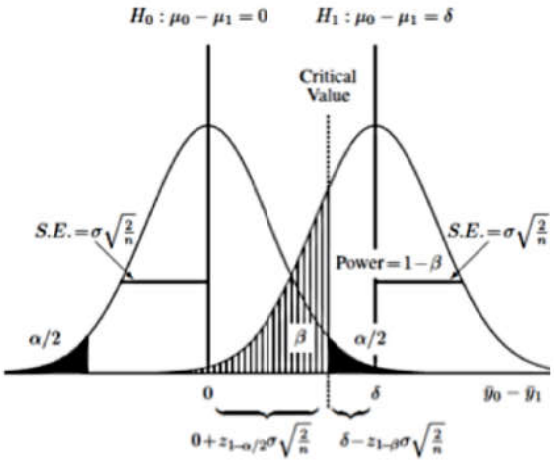
https://icds.shinyapps.io/shinyt/

Ho and Ha distributions

- Ho acts as $reference$ → eg., no difference
  - reject Ho if test returns `implausible` value
- Ha acts as $truth$ → eg., difference of .5 SD
  - particular violation of Ho expressed as `ncp`
    - assumed effect size (target or signal)
    - conditional on sample size (information)
  - non-central distribution
    - more often expressed as non-standardized effect size only !
- Ha is NOT interchangeable with Ho
  - absence of evidence $\neq$ evidence of absence
  - equivalence testing (Ha for 'no effect')



divide by n perspective on distributions

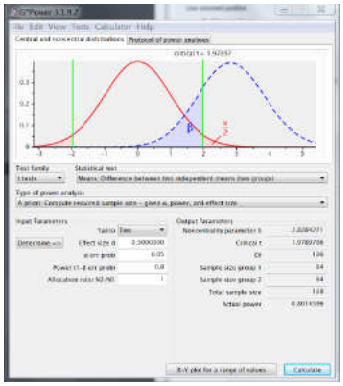


- remember:

$$n = \frac{(Z_{\alpha/2}+Z_{\beta})^2 * 2 * \sigma^2}{d^2}$$

$$n = \frac{(-1.96-0.84)^2 * 2 * 4^2}{2^2}$$

$$n = 62.79$$



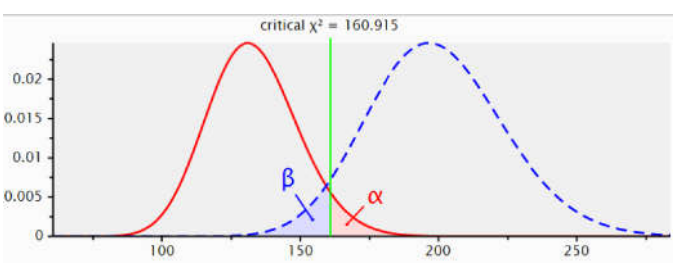
- non-centrality parameter, sample size translates Ha
- alternative: sample size changes standard deviation
- https://icds.shinyapps.io/shinyt/

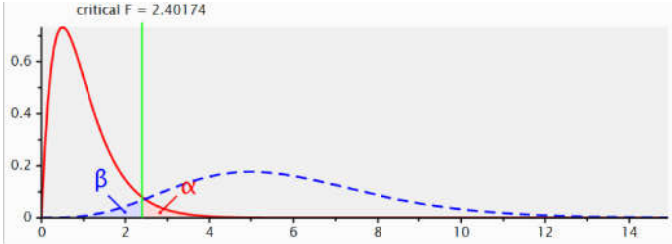
sample size calculation for estimation: no Ha

- sample size without type II error β, power, Ho or Ha
- focus on `estimation`, plausible values of effect, not testing or power
- ~ divide by n perspective but shifted to estimate
- precision analysis → set maximum width confidence interval
  - let E = maximum half width of confidence interval to accept
  - for confidence level $1 - \alpha$
  - $n = z^2_{\alpha/2} * \sigma^2 * 2 / E^2$ (for 2 groups)
- equivalence with statistical testing
  - if 0 (or other reference) outside confidence bounds → significant
- NOT GPower

GPower distributions or designs

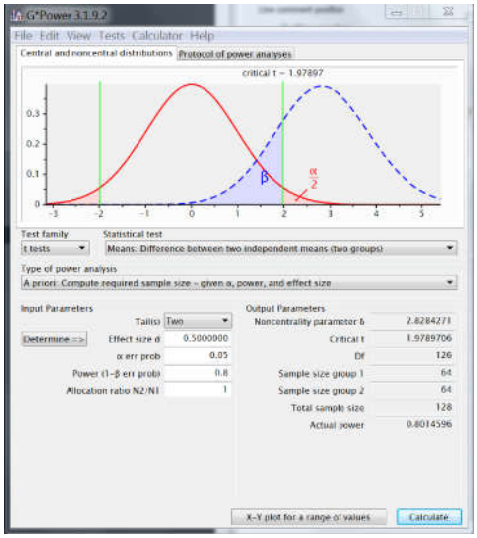
- distribution based test selection
  - Exact Tests (8)
  - t-tests (11) → `reference`
  - z-tests (2)
  - χ^2-tests (7)
  - F-tests (16)
- focus on the density functions
- design based test selection
  - correlation & regression (15)
  - means (19) → `reference`
  - proportions (8)
  - variances (2)
- focus on the type of parameters





type I/II error probability

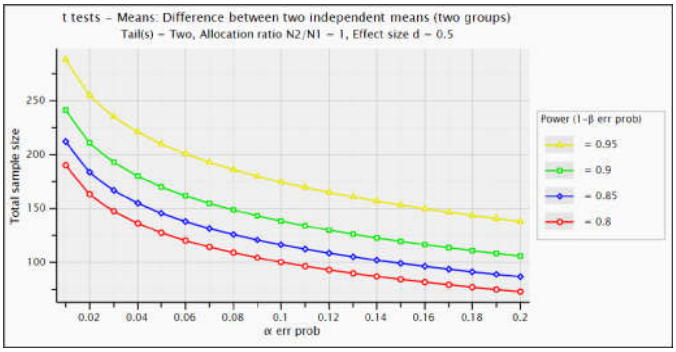
- decide whether to reject Ho **assuming** Ha
  - sample size calculation requires Ha assumption, test does not
- cut-off 'known' Ho
  - two tailed → both sides informative on Ho
  - one tailed → one side not informative on Ho
- two types of error
  - error: P(infer=Ha | truth=Ho) = α
  - error: P(infer=Ho | truth=Ha) = β
- confidence level = $1 - \alpha \rightarrow$ P(infer=Ho | truth=Ho)
- power = $1 - \beta \rightarrow$ P(infer=Ha | truth=Ha)



	infer=Ha	infer=Ho	sum
truth=Ho	α	1-α	1
truth=Ha	1-β	β	1

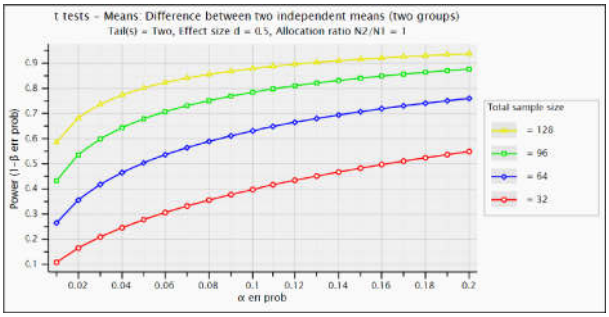
error exercise : create plot

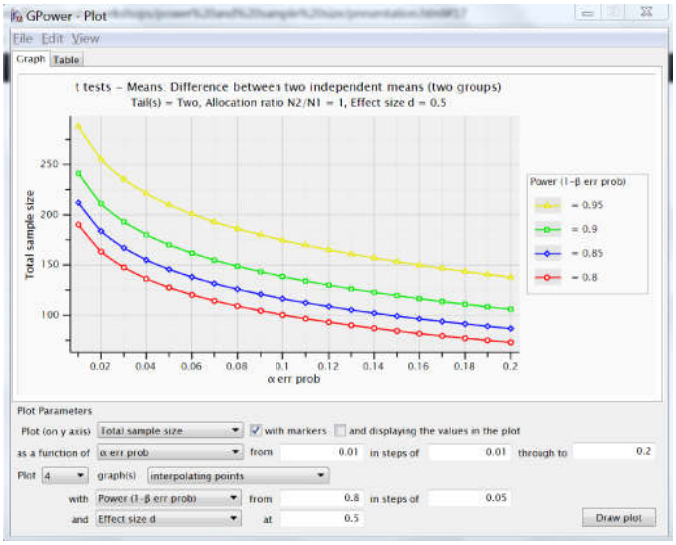
- create plot
  (X-Y plot for range of values)
- plot sample size by type I error
- set plot to 4 curves
  - for power .8 in steps of .05
- set α on x-axis
  - from .01 to .2 in steps of .01
- use effect size .5



error exercise : interpret plot

- where on the red curve (right) type II error = 4 * type I error ?
- when smaller effect size (.25), what changes ?
- switch power and sample size (32 in step of 32)
  what is relation type I and II error ?

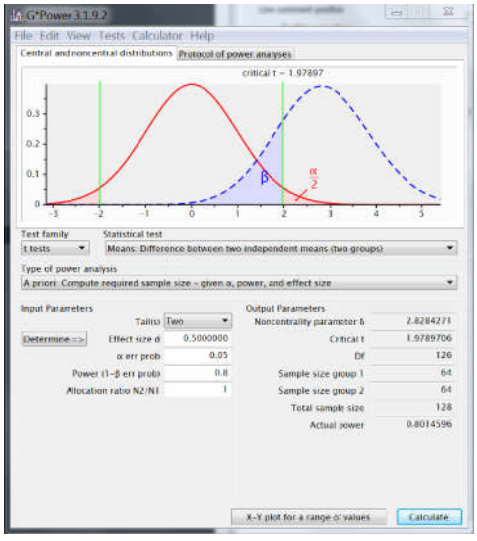




- where on the yellow curve (left) type II error = 4 * type I error ?

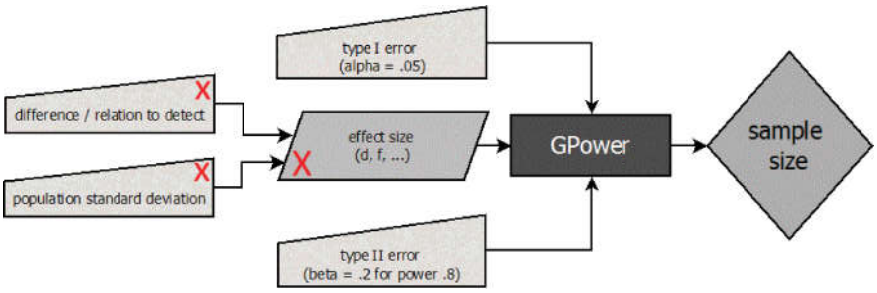
decide type I/II error probability

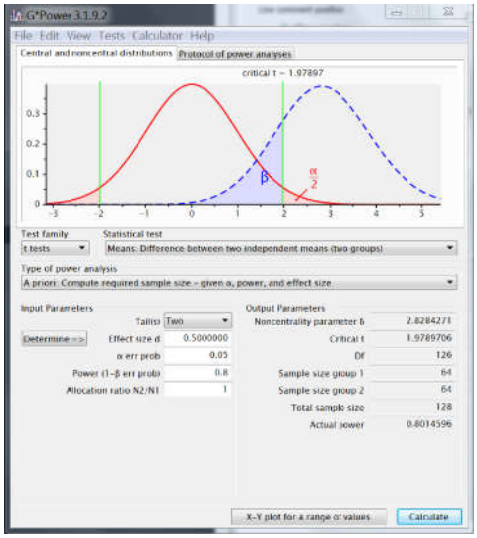
- α & β inversely related
  - if $\alpha = 0 \rightarrow$ never reject, no power
  - determine the balance
- which error you want to avoid most ?
  - cheap aids test ? $\rightarrow$ avoid type II
  - heavy cancer treatment ? $\rightarrow$ avoid type I
- rules of thumb ?
  - α = .05 ? $\rightarrow$ 1/20
  - β = .2 ? $\rightarrow$ power = 80%
  - α & β often selected in 1/4 ratio
    type I error is 4 times worse !!



effect size pushing Ha away from Ho

- bigger effect → more easy to detect it
- this is what you are looking for...
  - and how to `Determine` it





effect sizes, what it is and isn't

- degree with which a certain phenomenon holds (~ `Ho` is false)
  - part of non-centrality (as is sample size) → shift in GPower
  - signal to noise ratio
    - typically not just the signal, to provide scale
    - eg., difference on scale of pooled standard deviation
- 2 main families of effect sizes → test specific
  - differences d-family
  - association r-family
- transformations, eg., d = .5 → r = .243
  - $d = \frac{2r}{\sqrt{1-r^2}}$; $r = \frac{d}{\sqrt{d^2+4}}$; $d = ln(OR) * \frac{\sqrt{3}}{\pi}$

effect sizes of Cohen

Table 1
ES Indexes and Their Values for Small, Medium, and Large Effects

Test	ES index	Effect size: Small	Medium	Large
1. m_A vs. m_B for independent means	$d = \frac{m_A - m_B}{\sigma}$	.20	.50	.80
2. Significance of product–moment r	r	.10	.30	.50
3. r_A vs. r_B for independent rs	$q = z_A - z_B$ where z = Fisher's z	.10	.30	.50
4. P = .5 and the sign test	$g = P - .50$	.05	.15	.25
5. P_A vs. P_B for independent proportions	$h = \phi_A - \phi_B$ where ϕ = arcsine transformation	.20	.50	.80
6. Chi-square for goodness of fit and contingency	$w = \sqrt{\sum_{i=1}^{k} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}$	.10	.30	.50
7. One-way analysis of variance	$f = \frac{\sigma_m}{\sigma}$	.10	.25	.40
8. Multiple and multiple partial correlation	$f^2 = \frac{R^2}{1 - R^2}$	.02	.15	.35

- famous Cohen conventions
  - beware, just rules of thumb
  - Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed).

- Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155–159.

effect sizes of d and r family

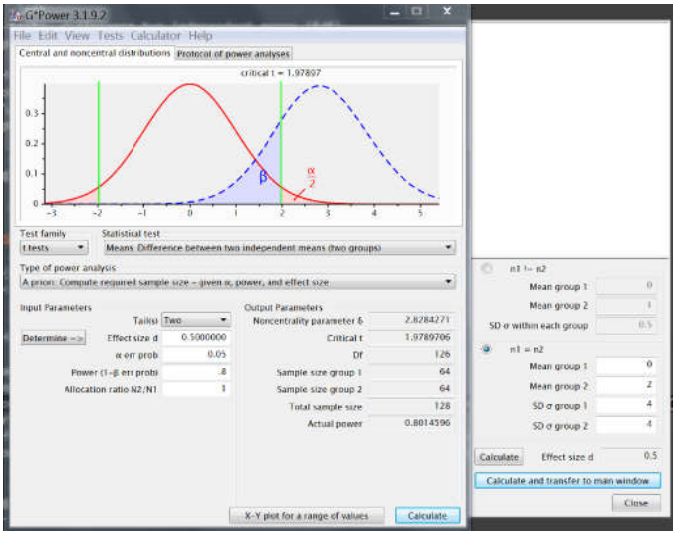
Measures of group differences (the *d* family)	Measures of association (the *r* family)	Measures of association (the *r* family)
(a) Groups compared on dichotomous outcomes	**(a) Correlation indexes**	**(b) Proportion of variance indexes**
RD The risk difference in probabilities: the difference between the probability of an event or outcome occurring in two groups	r The Pearson product moment correlation coefficient: used when both variables are measured on an interval or ratio (metric) scale	r^2 The coefficient of determination: used in bivariate regression analysis
RR The risk or rate ratio or relative risk: compares the probability of an event or outcome occurring in one group with the probability of it occurring in another	ρ (or r_s) Spearman's rho or the rank correlation coefficient: used when both variables are measured on an ordinal or ranked (non-metric) scale	R^2 R squared, or the (uncorrected) coefficient of multiple determination: commonly used in multiple regression analysis
OR The odds ratio: compares the odds of an event or outcome occurring in one group with the odds of it occurring in another	τ Kendall's tau: like rho, used when both variables are measured on an ordinal or ranked scale; tau-b is used for square-shaped tables; tau-c is used for rectangular tables	$_aR^2$ Adjusted R squared, or the coefficient of multiple determination adjusted for sample size and the number of predictor variables
(b) Groups compared on continuous outcomes	r_{pb} The point-biserial correlation coefficient: used when one variable (the predictor) is measured on a binary scale and the other variable is continuous	f Cohen's f: quantifies the dispersion of means in three or more groups; commonly used in ANOVA
d Cohen's d: the uncorrected standardized mean difference between two groups based on the pooled standard deviation	φ The phi coefficient: used when variables and effects can be arranged in a 2×2 contingency table	f^2 Cohen's f squared: an alternative to R^2 in multiple regression analysis and ΔR^2 in hierarchical regression analysis
Δ Glass's delta (or d): the uncorrected standardized mean difference between two groups based on the standard deviation of the control group	C Pearson's contingency coefficient: used when variables and effects can be arranged in a contingency table of any size	η^2 Eta squared or the (uncorrected) correlation ratio: commonly used in ANOVA
g Hedges' g: the corrected standardized mean difference between two groups based on the pooled, weighted standard deviation	V Cramér's V: like C, V is an adjusted version of phi that can be used for tables of any size	ε^2 Epsilon squared: an unbiased alternative to η^2
PS Probability of superiority: the probability that a random value from one group will be greater than a random value drawn from another	λ Goodman and Kruskal's lambda: used when both variables are measured on nominal (or categorical) scales	ω^2 Omega squared: an unbiased alternative to η^2
		R^2_C The squared canonical correlation coefficient: used for canonical correlation analysis

- Ellis, P. D. (2010). The essential guide to effect sizes: statistical power, meta-analysis, and the interpretation of research results.
- more than 70 different effect sizes... most of them related to eachother
- NOT p-value ~ partly effect size, partly sample size or power
  - do not simply compare p-values !

- most important effect size of
  - d: dichotomous - continuous
  - r: correlation - proportion variance

effect sizes in GPower (Determine)

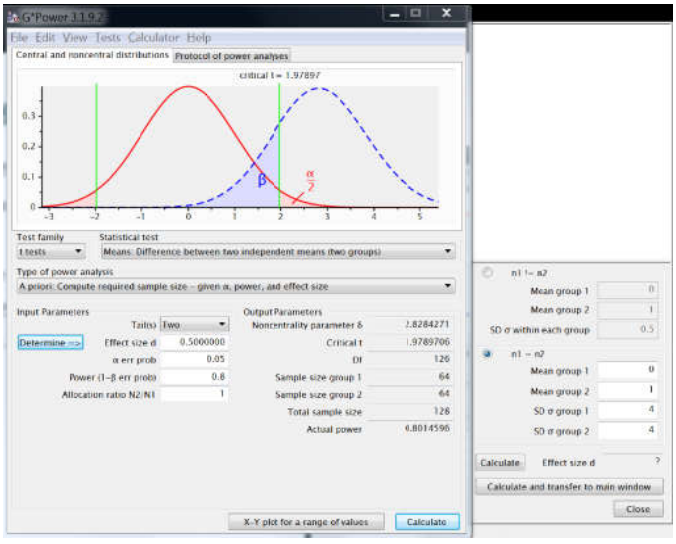
- often very difficult to specify
- GPower offers help with `Determine`
  - difference group means
    0-2 → signal ~ minimally relevant (or expected)
  - standard deviations (sd)
    4 each group → expected noise ~ natural diversity
  - written to Effect Size `d`
    .5 → difference in sd
- Reminder: effect size statistic depends on statistical test



effect size exercise : ingredients cohen d

For the `reference example`:

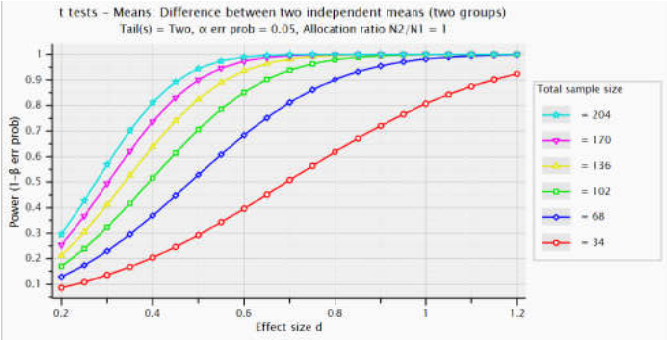
- change mean values from 0 and 2 to 4 and 6, what changes ?
- change sd values to 2 for each, what changes ?
  - effect size ?
  - total sample size ?
  - non-centrality ?
  - critical t ?
- change sd values to 6 for each, what changes ?



effect size exercise : plot

- plot power by effect size
- set plot to 6 curves
  - for sample sizes, 34 in steps of 34
- set effect sizes on x-axis
  - from .2 to 1.2 in steps of .05
- use α equal to .05

- create plot (X-Y plot for range of values)

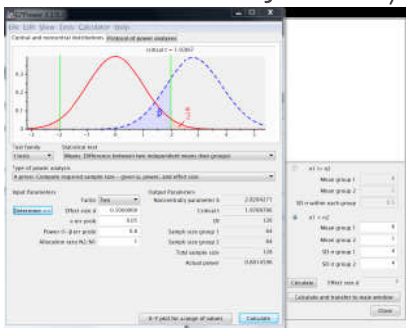


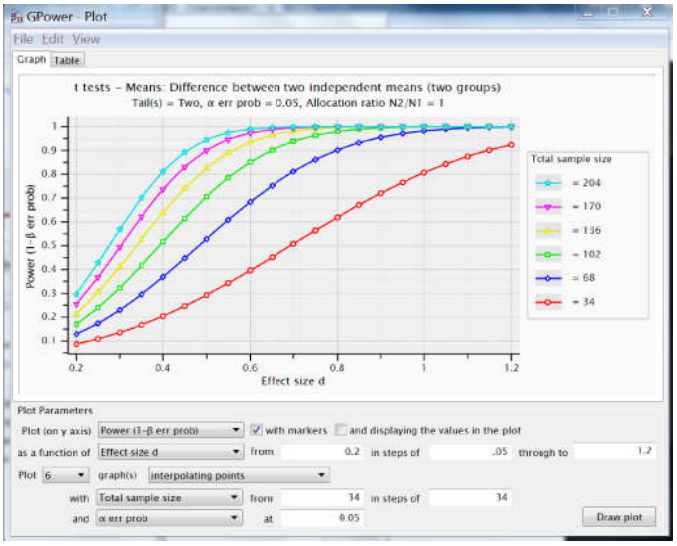
- determine (approximately) the three situations from previous slide on the plot
- how does power change when doubling the effect size, eg., from .5 to 1 ?

effect size exercise : imbalance

For the `reference example`:

- change allocation ratio from 1
  - to 2, .5, 3 and 4, what to conclude ?
    - ratio 2 and .5 ?
    - imbalance + 1 or * 2 ?
- ? no idea why n1 $\neq$ n2





effect sizes, how to determine them in theory

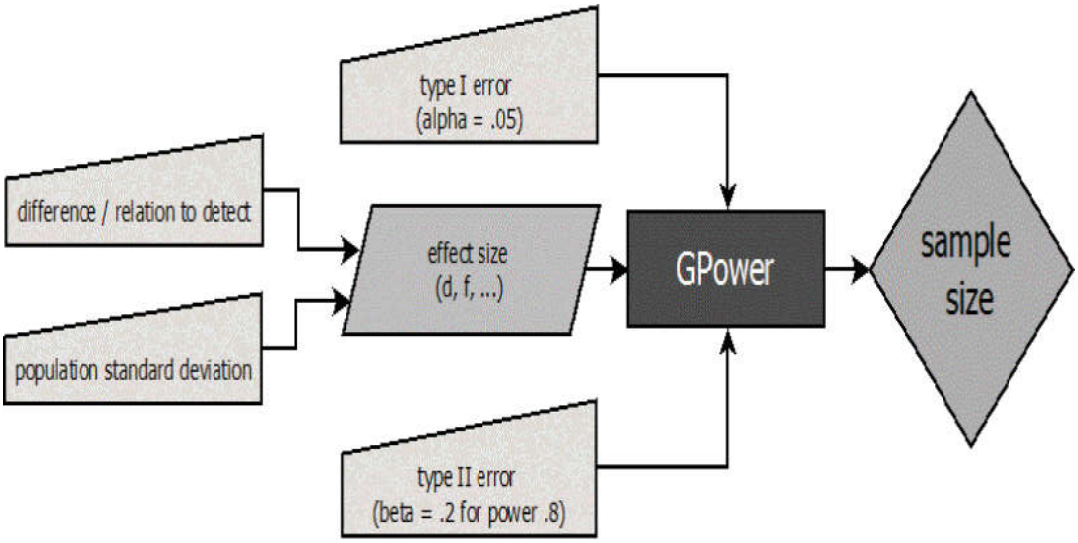
- choice of effect size matters → justify choice
- choice of effect size
  - NOT significant → meaningless, dependent on sample size
  - realistic (eg., previously observed effect) → replicate
  - important (eg., minimally relevant effect)
- use `Determine` to get started (check the manual)
  - for independent t-test → means and standard deviations
  - possible alternative is to use variance explained, eg., 1 versus 16
    - with one-way ANOVA (f=.25 instead of d=.5)
    - with linear regression (f^2=.0625 instead of d=.5)
    - https://www.psychometrica.de/effect_size.html#transform

effect sizes, how to determine them in practice

- experts / patients → use if possible → importance
- literature (earlier study / systematic review) → realistic
- pilot → guestimate dispersion estimate, but very small sample size
- internal pilot → stopping rule (sequential/conditional)
- turn to Cohen → use if everything else fails (rules of thumb)
- guestimate the input parameters, what can you do ?
  - sd from assumed range / 6 assuming normal distribution
  - sd for proportions (& percentages) at conservative .5
  - sd from control, assume treatment the same
  -

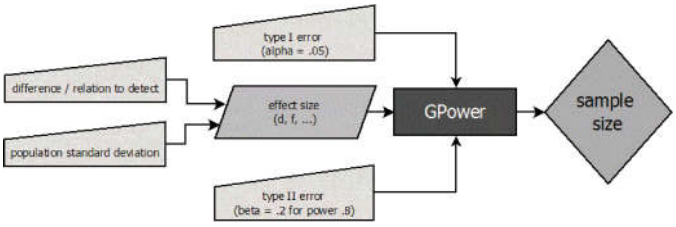
GPower: turning effect size into sample size

- For this example, calculate sample size based on
  - effect size (difference of interest, scaled on standard deviation)
  - type I and type II error



relation samples & effect size, errors I & II

- GPower → type of power analysis
  - Apriori: n ~ α, power, Δ
  - Post Hoc: power ~ α, n, Δ
  - Compromise: power, α ~ β / α, Δ, n
  - Criterion: α ~ power, Δ, n
  - Sensitivity: Δ ~ α, power, n



- building blocks:
  - sample size (n)
  - effect size (Δ)
  - alpha (α)
  - power ($1 - \beta$)
- each parameter fully determined conditional on others

type of power analysis exercise

- for given example, step through...
  - retrieve power given n, α and Δ
  - [1] for power .8, take half the sample size, how does Δ change ?
  - [2] set β/α ratio to 4, what is α & β ? what is the critical value ?
  - [3] keep β/α ratio to 4 for effect size .5, what is α & β ? critical value ?
- [1] .5 to .7115 = .2115, bigger effect size compensate for loss of sample size (sensitivity)
- [2] critical value 1.9990 with errors approx. .05 and .2 (compromise)
- [3] critical value 1.6994 with errors approx. double, .09 and .38

A practical example

- background
  - Nanobodies are antibody fragments of heavy chain only antibodies, of interest in cancer treatment because off their fast blood clearance, low nonspecific uptake and good specific binding.
  - They can direct the cytotoxicity power of silver-based nanoparticles towards acute myeloid leukemia (AML) cells to inhibit their proliferation and cell viability.
- setup
  - An anti-CD33 Nb (Nb 7) will be labeled with silver nanoparticles (NaNo) and in vivo biodistribution and tumor targeting will be evaluated. To this end, the Nb 7-NaNo will be radiolabeled with 99mTc and injected intravenously.
  - A subcutaneous (s.c.) tumor mouse model is included to evaluate the tumor targeting of Nb7-NaNo, an intravenous (i.v.) tumor mouse model is included to recapitulate the human situation of AML. Non-target control (ctrl) Nb-NaNo-99mTc is included to evaluate the specificity of Nb7-NaNo-99mTc.

A practical example: design

- focus: detect differences in the distribution of Nb7 and ctrl Nb labeled nanoparticles in the tumor, for both the s.c. and i.v. model ?
- question: what sample size → so, what now ?

- `t-test independent means`
- `one-tailed`
- `type I error alpha .05`
- `type II error beta .2, for power .8`
- `effect size (Cohen's d)`
  - `mean and pop. standard deviation for first group: .5% (1%)`
  - `mean and pop. standard deviation for second group: 1% (.25%)`
  - `resulting effect size 2.63 (big !!)`
- `sample size 6 (3 each group)`

A practical example: GPower protocol

```
t tests - Means: Difference between two independent means (two groups)
Analysis:    A priori: Compute required sample size

Input:
     Tail(s) =   One
     Effect size d    =   2.6261287
     α err prob    =   0.05
     Power (1-β err prob) =    .8
     Allocation ratio N2/N1  =   1

Output:
     Noncentrality parameter δ    =   3.2163377
     Critical t  =   2.1318468
     Df  =   4
     Sample size group 1 =   3
     Sample size group 2 =   3
     Total sample size   =   6
     Actual power    =   0.8366654
```

A practical example: possible issues

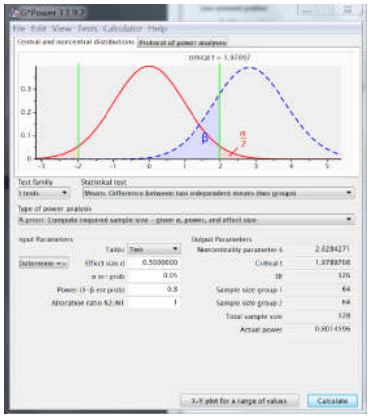
- assumed normality → non-parametric ?
- focus on 2 treatment groups only → ANOVA ?
- % used on linear scale → proportions ?
- what if you do not know the means and sd's ?

getting your hands dirty

- in R
  - qt → get quantile on Ho ($Z_{1-\alpha/2}$)
  - pt → get probability on Ha (non-central)

```
.n <- 64
.df <- 2*.n-2
.ncp <- 2 / (4 * sqrt(2)) * sqrt(.n)
.power <- 1 -
    pt(
        qt(.975,df=.df),
        df=.df, ncp=.ncp
    ) -
    pt( qt(.025,df=.df), df=.df, ncp=.ncp)
round(.power,4)
```

```
## [1] 0.8015
```



```
# calculator
m1=0;m2=2;s1=4;s2=4
alpha=.025;N=128
var=.5*s1^2+.5*s2^2
d=abs(m1-m2)/sqrt(2*var)
d=d*sqrt(N/2)
tc=tinv(1-alpha,N-1)
power=1-nctcdf(tc,N-1,d)
```

PART II: moving beyong independent t-test

GPower examples beyond the independent t-test

- so far, comparing two independent means
- selected topics
  - dependent instead of independent
  - non-parametric instead of assuming normality
  - more than 2 groups (compare jointly, pairwise, focused)
  - relations instead of groups (regression)
  - correlations
  - proportions, dependent and independent
- GPower manual describes 27 tests by example !!
  - equations effect size & non-centrality parameter

dependence between groups

- if 2 dependent groups (eg., before/after treatment) → account for correlations
- matched pairs (t-test / means, difference 2 dependent means)
- use original example
  - [1] use correlation .5 to compare (effect size, ncp, n)
  - [2] set original sample size (n=64*2) and compare (same error I & effect size)
  - [3] how many observations if no correlation exists (original setup) ?
  - [4] difference in sample size for correlation .875 ?
- [1] Δ looks same because $\sqrt{2 * (1 - \rho)}$, but ncp bigger, n much smaller (1 group)
- [2] -posthoc- for 64 subjects 2 measurements, ncp > 4 and power > .975
- [3] approx. independent means, here 65 but different effect size (dz)
- [4] double the effect size, almost 4 times smaller sample size (here 10 and 34)

non-parametric distribution

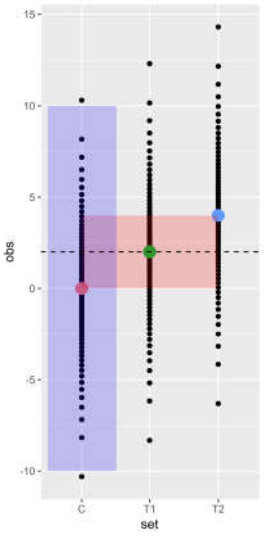
- expect non-normally distributed residuals, avoid normality assumption
- only considers ranks or uses permutations → price is efficiency
- two groups → Wilcoxon-Mann-Whitney (t-test / means, diff. 2 indep. means)
- use original example
  - [1] how about n ? compared to parametric → what is ARE in %loss?
  - [2] change parent distribution to 'min ARE' ? what now ?
- [1] a few more observations (3 more per group), less than 5 % loss (ARE)
- [2] several more observations, less efficient, more than 13 % loss (min ARE)
- maybe just ignore it and add 10 % observations ?!

more groups to compare, 4 cases

- simple example: assume one control, and two treatments
- if more than two groups, several options
  - test whether at least one differs → omnibus F-test
  - test whether two particular means differ → t-test
  - test whether all differ from eachother → pairwise comparisons
  - test whether control differs from each treatment → contrasts
- if multiple tests → inflation of type I error
  - correct, eg., using Bonferonni
  - take into account in your inference

F-test statistic

- multiple groups → no `d` and t-test statistic
- t-test extension → F-test statistic & effect size `f`
- `f` is ratio of variances $\sigma^2_{between}/\sigma^2_{within}$
- example: one control and two treatments
  - `reference example` + 1 group
  - within group observations normally distributed
  - means C=0, T1=2 and T2=4
  - sd for all groups (C,T1,T2) = 4



more groups: omnibus

- assume one control, and two treatments, test that at least one differs
- one-way Anova (F-test / Means, ANOVA - fixed effects, omnibus, one way)
- effect size f, with numerator/denominator df (derived from η^2)
- start from original example,
  - [1] what is the sample size ? ncp ? critical F ? does size matter ?
  - [2] set extra condition, mean 4, sd 4, 3 groups, what is the sample size ?
  - [3] change mean middle group away from 2, how effect / sample size change ?
  - [4] derive effect size with `variance` between .5 and within 3, or 1 and 6 ?
- [1] different effect size (f), distribution (ncp,critical-F), same sample size (size ~ imbalance)
- [2] n=63, with ncp 10.5 [3] effect size increases, difference with either 0 or 4!
- [4] same effect size, so, same sample size, 63, ncp 10.5 (1/7th explained)

more groups: pairwise

- assume one control, and two treatments
  - interested in all three pairwise comparisons → maybe Tukey
    - typically run aposteriori, after omnibus shows effect
  - use t-test with correction of α for multiple testing
- apply Bonferonni correction for original example
  - [1] resulting sample size for three tests ?
  - [2] what if biggest difference is ignored, sample size ?
  - [3] with original 64 sized groups, what is the power ?
- [1] divide α by 3 (86*2) → overall 86*3 = 258
- [2] or divide by 2 (78*2) (biggest difference implied) → overall 78*3 = 234
- [3] .6562 when /3 or .7118 when /2, power-loss

sample size calculation benefit from focus

- better to focus during the design on specific questions
  - only consider the main comparisons in focus (eg. primary endpoints)
    - only interested in comparing two treatments → t-test
  - only consider smallest of relevant effects, largest sample size
  - set up contrasts (next slide)
- sample size calculations (design)
  - not necessarily equivalent to statistics
  - requires justification to convince yourself and/or reviewers
- example:
  - statistics: group difference evolution 4 repeated measurements → mixed model
  - power: difference treatment and control last time point → t-test

more groups: contrasts

- assume one control and two treatments
  - set up 2 contrasts for T1 - C and T2 - C
  - set up 1 contrast for average(T1,T2) - C
- each contrast requires 1 degree of freedom
- each contrast combines a specific number of levels
- effect sizes for planned comparisons must be calculated !!
  - standard deviation of the effect using contrasts
  - $\sigma_m = \frac{|C|}{\sqrt{N \sum_i^k c_i^2 / n_i}}$
  - with $C = \sum \mu_i * c_i$

more groups: contrasts exercise

- one-way ANOVA (F-test / Means, ANOVA-fixed effects,special,main,interaction)
- obtain effect sizes for contrasts (assume equaly sized for convenience)
  - $\sigma_{contrast}$ T1-C: $\frac{(\ 1*0+1*2+0*4)}{\sqrt{(2*((-1)^2+1^2+0^2))}} = 1$; T2-C: $= 2$; (T1+T2)/2-C: $= 1.4142$
  - with σ = 4 → ratio of variances for effect sizes f .25, .5, .3536
- sample size for each contrast, each 1 df
  - [1] contrasts nrs. 1 or 2, number of groups = 2
  - [2] contrasts nrs. 1 AND 2, number of groups = 2
  - [3] contrasts nr. 3, number of groups = 3
- [1] total sample size 128 (again!!), 64 C and 64 T1, or 34 = 17 C and 17 T2
- [2] same with Bonferoni correction → 156 and 42; 78 C, 78 T1, 21 T2
- [3] total sample size 81 → 27 in each group

relations instead of groups

- differences between groups → relation observations & categorisation
- example → d = .5 → r = .243
- note: slope $\beta = r * \sigma_y / \sigma_x$
  for σ_x^2 = .25 (binary) and σ_y^2 = 17 (variance + bias^2) → β = .243*$\sqrt{17}/\sqrt{25}$ = 2
- regression coefficient (t-test / regression, fixed model single regression coef)
- use original example, regression style
  - [1] calculate sample size for variances explained 1 and residual 16, conclude ?
  - [2] what if σ_x (predictor values) or σ_y (effect and sd) increase ?
  - [3] what if also other predictors in the model ?
- [1] 128, same as for reference example, now with f^2 = .25^2.
- [2] sample size decreases with σ_x (opposite σ_y ~ effect size), for same slope
- [3] loss of degree of freedom, very little impact

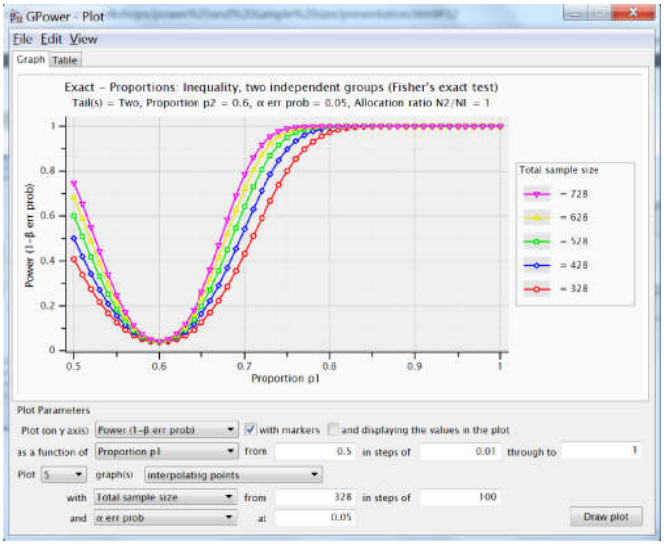
correlations

- when comparing two independent correlations
- z-tests / correlation & regressions: 2 indep. Pearson r's
- makes use of Fisher Z transformations → z = .5 * log($\frac{1+r}{1\ \ r}$) → q = z1-z2
- [1] assume correlation coefficients .7844 and .5 effect size & sample size ?
- [2] assume .9844 and .7, effect size & sample size ?
- [3] assume .1 and .3844 effect size & sample size ?
- [1] effect size q = 0.5074, sample size 64*2 = 128
- [2] effect size q = 1.5556, sample size 10*2 = 20, same difference, bigger effect
- [3] effect size q = -0.3048, sample size 172*2 = 344, negative and smaller effect
- note that dependent correlations are more difficult, see manual

proportions

- comparing two independent proportions → bounded between 0 and 1
- Fisher Exact Test (exact / proportions, difference 2 independent proportions)
- effect sizes in odds ratio, relative risk, difference proportion
  - [1] for odds ratio 2, p2 = .60, what is p1 ?
  - [2] sample size for equal sized, and type I and II .05 and .2 ?
  - [3] sample size when .95 and .8 (difference of .15) and .05 and .2 ?
- [1] odds ratio 2 * (.6/.4) = 3 (odds), 3/3+1 = .75
- [2] total sample size 328, [3] total sample size 164, either at .05 or .95
- treat as if unbounded, ok within .2 - .8, variance is p*(1-p) → maximally .25 !!
  - [4] use t-test for difference of .15
- [4] effect size .3, sample size 352 (> 328)

proportions exercise



- Fisher Exact Test
  - power over proportions .5 to 1
  - 5 curves, sample sizes 328, 428, 528...
  - type I error .05
  - [1] what happens ?
  - [2] repeat for one-tailed, what is different ?

- [1] power for proportion compared to reference .6, sample size determines impact
- [2] one-tailed, increases power, both sides (absolute value difference) tail to choose on the correct side

dependent proportions

- when comparing two dependent proportions
- McNemar test (exact / proportions, difference 2 dependent proportions)
  - include correlations implictely, discordant pairs → change
  - effect size as odds ratio → ratio of discordancy ?!
- assume odds ratio equal to 2, equal sized, type I and II errors .05 and .2, two-way
  - [1] what is the sample size for .25 proportion discordant, and [2] .5, and [3] 1
  - [4] for odds ratio 4 or .25, how the proportion p12 and p21 change ?
  - [5] repeat for third alpha option, and consider total sample size, what happens ?
- [1] total sample size 288 & [2] 144 & [3] impossible, but limits to 72
- [4] for proportion discordant, 1 to 4 or 4 to 1
- [5] sample size differs because side effects

dependence within groups (repeated)

- if repeated measures → account for correlations
- repeated measures (F-test / Means, repeated measures...)
- 3 main types
  - within: like dependent t-test for 2 or more measurements
  - between: use of multiple measurements per group
  - interaction: difference of change over groups
- correlation within subject (unit)
  - informative on group differences within subject
  - redundancy for between group differences

repeated measures within

- possible to have only 1 group (within subject comparison)
- use effect size f = .25 (1/16 explained versus unexplained)
  - [1] use zero correlation to compare with sample size independent t-test
  - [2] for one group use correlation .5, compare sample size dependent t-test
  - [3] double number of groups to 2
  - [4] double number of measurements to 4 (correlation 0 and .5), impact ?
- number of groups = 1, number of measurements = 2, sample size = [1] 65 and [2] 34
- [3] changed degrees of freedom (each group a set of measurements), thus F
- [4] impact bigger when correlation

repeated measures between

- use effect size f = .25 (1/16 for variance or 2/4 for means)
  - [1] use correlation 0 and .5 with 2 groups and 2 measurements, sample size ?
  - [2] for correlation .5, compare 2 or 4 measurements, sample size ?
  - [3] double number of groups to 2
- [1] sample size higher when higher correlation
- [2] sample size lower when more measurements (unless perfect correlation)
- [3] more groups require higher sample size

repeated measures within x between

- SPSS idiosyncrasies: https://www.youtube.com/watch?v=CEQUNYg80Y0
- use effect size f = .25 (1/16 for variance)
  - [1] use correlation 0, compare 2 groups 2 measurements with rep. between ?
  - [2] use correlation 0.5, compare 2 groups 2 measurements with rep. within ?
  - [3] use correlation .5, compare 2 groups and 4 measurements, sample size ?
  - [4] repeat with 4 groups and 4 measurements, sample size ?
- [1] same with indep and [2] same with dependent
- [3] more groups, higher sample size (identical to within)
- [4] difference between within and between

next...

- now only a few simple situations, but take us a long way...
  - use creative simple approach by focusing on what matters
- GPower allows for some more complex models, but more complex to specify
- GPower has limitations
  - GPower does not include all types of models
  - not for all types of models analytical solutions exist
    - use simulation (or bootstrap)

conclusion: keep it simple, keep it real

- sample size calculation is a design issue, not a statistical one
- building blocks: sample & effect sizes, type I & II errors, each conditional on rest
- simplify using a focus, if justifiable → then GPower can get you a long way
- otherwise simulation is always an option

The End, for now...

- in future: advanced sample size calculation
  - rest of GPower
  - models not in GPower, eg., survival analysis
  - models without closed formula, using simulation

about us ... 

- statistical consultancy at VUB / UZ
  - for PhD students and researchers
  - for master students' thesis
- collaboration on data analysis
- book us @ https://gf.vub.ac.be/stats.php

for fun: P(effect exists | test says so)

- power → P(test says there is effect | effect exists)
- $P(infer = Ha|truth = Ho) = \alpha$
- $P(infer = Ho|truth = Ha) = \beta$
- $P(infer = Ha|truth = Ha) = power$
- $P(truth = Ha|infer = Ha) = \frac{P(infer=Ha|truth=Ha)*P(truth=Ha)}{P(infer=Ha)}$
- $= \frac{P(infer=Ha|truth=Ha)*P(truth=Ha)}{P(infer=Ha|truth=Ha)*P(truth=Ha)+P(infer=Ha|truth=Ho)*P(truth=Ho)}$
- $P(truth = Ha|infer = Ha) = \frac{power*P(truth=Ha)}{power*P(truth=Ha)+\alpha*P(truth=Ho)}$
- what if probability my model is true is low ? eg., .01 ? → $P(truth = Ha) = .01$
- $P(truth = Ha|infer = Ha) = \frac{.8*.01}{.8*.01+.05*.99} = .14$
- probability that the effect exists if the test says so, in this case, is 14% chance!!