INTERFACULTY CENTER
DATA PROCESSING & STATISTICS
CORE FACILITY

Sample size calculation with GPower

Wilfried Cools (ICDS) & Sven Van Laere (BiSI)
https://www.icds.be/

sample size calculation

- the program
  - understand the reasoning
    - introduce building blocks
    - implement on t-test
  - explore more complex situations
    - simple but common
- not one simple formula for all → GPower to the rescue
  - a few exercises

sample size calculation: demarcation

- how many observations are sufficient ?
  - avoid too many: observations typically imply a cost
    - money / time → limited resources
    - risk / harm → ethical constraints
  - sufficient for what ? depends on
    - the aim of the study → statistical inference
- linked to statistical inference (using standard error)
  - `testing` → power [probability to detect `effect`]
  - `estimation` → accuracy [size of `confidence interval`]

sample size calculation: a difficult design issue

- 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
  - not possible for predictive models, because no standard error
  - NO retrospective power analyses → OK for future study only

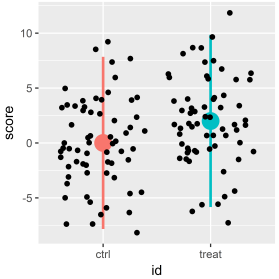
    Hoenig, J., & Heisey, D. (2001). The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. *The American Statistician, 55*, 19–24.
- alternative justifications:
  - common practice, feasibility → non-statistical (importance, low cost, ...)

simple example

- experimental - confirmatory
- evaluation of radiotherapy to reduce a tumor in mice
- comparing treatment group with control (=conditions)
- tumor induced, random assignment treatment or control (equal if no effect)
- after 20 days, measurement of tumor size (=observations)
- intended analysis: unpaired t-test to compare averages for treatment and control
- SAMPLE SIZE CALCULATION:
  - IF average tumor size for treatment at least 20% less than control (4 vs. 5mm)
  - THEN how many observations, sufficient to detect that difference (significance) ?

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



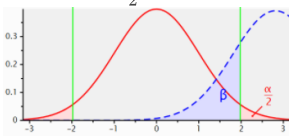
https://apps.icds.be/shinyt/

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)

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

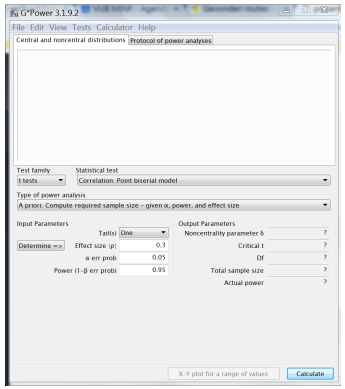
$$n = \frac{(-1.96-0.84)^2 * 2 * 4^2}{2^2} = 62.79$$



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

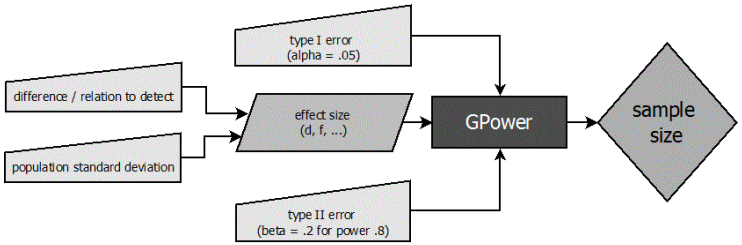
GPower: a useful tool

- popular and well established
- free @ http://www.gpower.hhu.de/
- implements wide variety of tests
- implements various visualizations
- documented fairly well
- note: not all tests are included !
- note: not without flaws !
- other tools exist (some paying)
- alternative: simulation (generate and analyze)



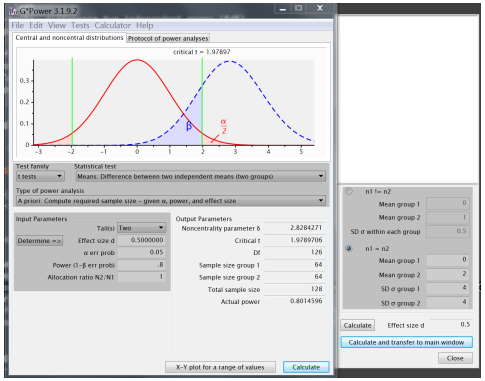
GPower: the building blocks in action

- **sizes**: effect size, sample size
- **errors**:
  - type I (α) defined on distribution Ho
  - type II (β) evaluated on distribution Ha
- calculate sample size based on effect size, and type I / II error



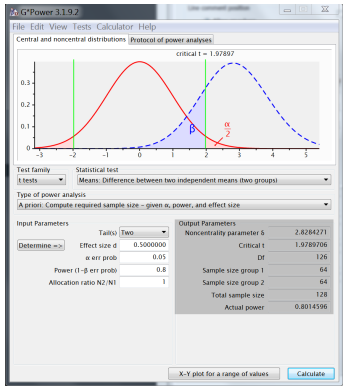
GPower input

- `~ reference example`
- t-test : difference two indep. means
- apriori: calculate sample size
- effect size = standardized difference [Determine]
  - Cohen's d
  - d = |difference| / SD_pooled
  - d = |`0-2`| / `4` = `.5`
- α = `.05`, 2 - tailed ($\alpha/2 \rightarrow$ .025 & .975)
- $power = 1 - \beta$ = `.8`
- allocation ratio = `1` (equally sized groups)



GPower output

- sample size (n) = 64 x 2 = (`128`)
- degrees of freedom (df) = 126 (128 - 2)
- plot showing null Ho and alternative Ha 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

- Protocol: summary for future reference or communication
- File/Edit save or print file (copy-paste)

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

building blocks

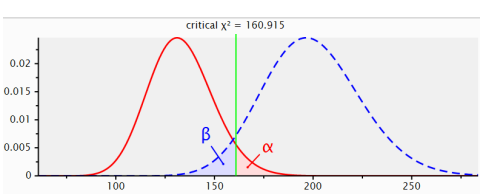
distributions: Ho & Ha, test dependent shape

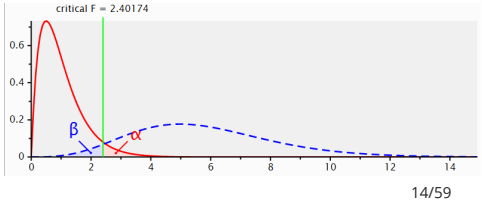
sizes: sample size and effect size in relation to distance between Ho & Ha

errors: type I error and type II error as cut-off on Ho & Ha

GPower distributions

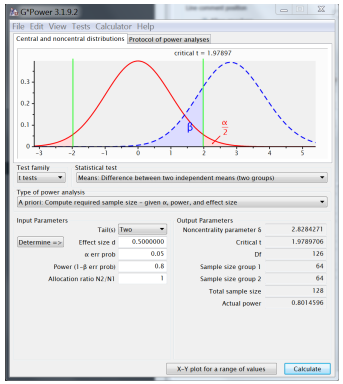
- test family - statistical tests [in window]
  - Exact Tests (8)
  - t-tests (11) → `reference`
  - z-tests (2)
  - χ^2-tests (7)
  - F-tests (16)
- tests [in menu]
  - correlation & regression (15)
  - means (19) → `reference`
  - proportions (8)
  - variances (2)
- focus on the type of parameters
- focus on the density functions





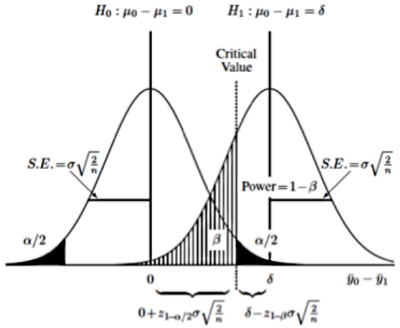
central Ho and non-central Ha distributions

- `Ho` acts as $benchmark$ → eg., no difference
  - `Ho ~ t(0,df)` $cutoff$ using α,
  - reject `Ho` if test returns `implausible` value
- `Ha` acts as $truth$ → eg., difference of .5 SD
  - `Ha ~ t(ncp,df)`
  - `ncp` as violation of `Ho` → shift (location/shape)
- `ncp` : non-centrality parameter combines
  - assumed `effect size` (target or signal)
  - conditional on `sample size` (information)
- `ncp` : determines overlap → power ↔ sample size



https://apps.icds.be/shinyt/

divide by n perspective on distributions

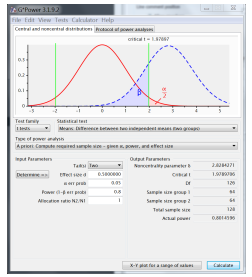


- remember:

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

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

$$n = 62.79$$



- non-centrality parameter, sample size translates `Ha`
- alternative: sample size changes standard deviation
- https://apps.icds.be/shinyt/

divide by n, for statistical estimation (no Ha)

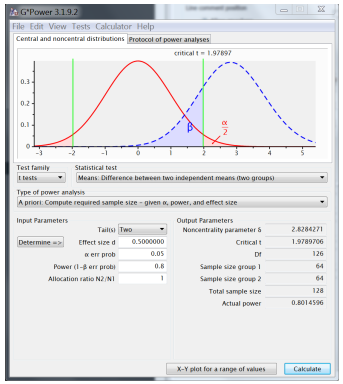
- focus on `estimation`, plausible values of effect (no testing)
- sample size without type II error β, power, Ho or Ha
- distribution on the estimate (not the null)
- precision analysis → set maximum width confidence interval
  - let E = maximum half width of confidence interval to accept
  - for confidence 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

Ho and Ha: a statistical note

- Ha is NOT interchangeable with Ho
  - α for cut-off at Ho → observe test statistics (Ha unknown)
  - fail to reject → remain in doubt
    - absence of evidence $\neq$ evidence of absence
    - p-value → P(statistic|Ho) != P(Ho|statistic)
    - η not significantly different from 0 → not from η * 2 either
  - equivalence testing
    - reject Ho that smaller than 0 - $|\delta|$
    - reject Ho that bigger than 0 + $|\delta|$
    - Ha for 'no effect'

type I/II error probability

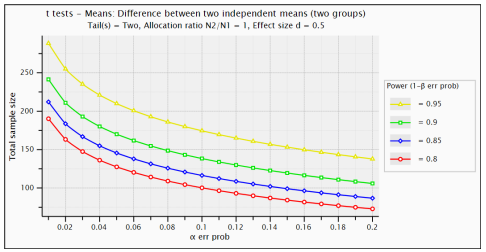
- inference, statistical testing
  - cut-off's →
    infer effect (+) vs. insufficient evidence (-)
  - distribution → true vs. false (density → AUC=1)
- type I error: incorrectly reject Ho (false positive):
  - cut-off at Ho, error prob. α controlled
  - one/two tailed → one/both sides informative ?
- type II error: incorrectly fail to reject Ho (false -):
  - cut-off at Ho, error prob. β depends on Ha
  - Ha assumed known in a power analyses
  - power = 1 - β = probability correct rejection (true +)



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

error exercise : create plot

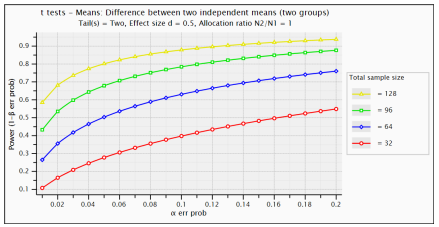
- `~ reference example`
- create plot
  (X-Y plot for range of values)
- plot sample size by type I error
- set plot to 4 curves
  - for power .8 in steps of .05
- set α on x-axis
  - from .01 to .2 in steps of .01
- use effect size .5

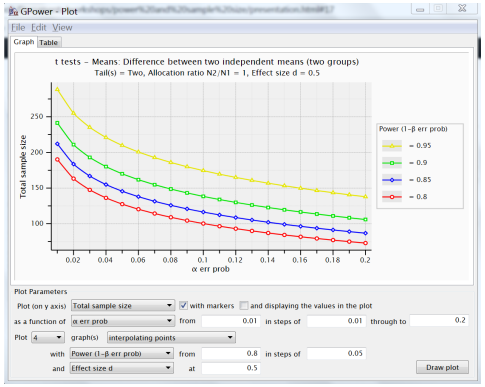


notice Table option

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 ?

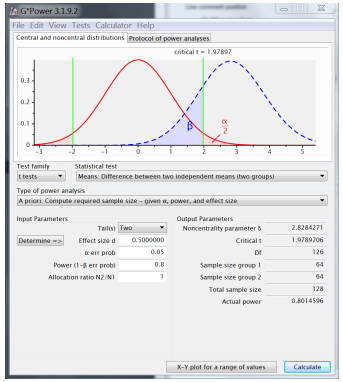




- where on the yellow curve (left) type II error = 4 * type I error ?

decide type I/II error probability

- frequent choices
  - α often in range .01 - .05 → 1/100 - 1/20
  - β often in range .1 to .2 → power = 80% to 90%
- α & β inversely related
  - if $\alpha = 0$ → never reject, no power
  - α & β often selected in 1/4 ratio
    type I error is 4 times worse !!
  - which error you want to avoid most ?
    - cheap aids test ? → avoid type II
    - heavy cancer treatment ? → avoid type I



control type I error

- multiple tests
  - inflates type I error α
  - family of tests: $1-(1-\alpha)^k$
    - correct, eg., Bonferroni (α/k)
  - interim analysis (analyze and proceed)
    - correct, eg., alpha spending
- suggested technique interim analysis: alpha spending
  - plan in advance
  - use O'Brien - Flemming bounds, more efficient than Bonferroni
  - NOT GPower
    - determine with simulation tool: https://apps.icds.be/aspending/
    - commercial packages eg., PASS and other software (eg., ldbounds in R)

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

- power → P(test says there is effect | effect exists)
- $P(infer = Ha | truth = Ho) = \alpha$
- $P(infer = Ho | truth = Ha) = \beta$
- $P(infer = Ha | truth = Ha) = power$
- $P(\underline{truth} = Ha | \underline{infer} = Ha) = \frac{P(infer=Ha|truth=Ha)*P(truth=Ha)}{P(infer=Ha)}$ → Bayes Theorem
- $P(truth = Ha | infer = Ha) = \frac{P(infer=Ha|truth=Ha)*P(truth=Ha)}{P(infer=Ha|truth=Ha)*P(truth=Ha)+P(infer=Ha|truth=Ho)*P(truth=Ho)}$
- $P(truth = Ha | infer = Ha) = \frac{power*P(truth=Ha)}{power*P(truth=Ha)+\alpha*P(truth=Ho)}$ → depends on prior probabilities
- IF very low probability model is true, eg., .01 ? → $P(truth = Ha) = .01$
- THEN probability effect exists if test says so is low, in this case only 14% !!
- $P(truth = Ha | infer = Ha) = \frac{.8*.01}{.8*.01+.05*.99} = .14$

effect sizes

- estimate/guestimate of magnitude or practical significance
- typically standardized: signal to noise ratio (noise provides scale)
  - eg., difference on scale of pooled standard deviation
- part of non-centrality (as is sample size) → shift in GPower
  - bigger effect → more easy to detect (pushing away `Ha`)
- 2 main families of effect sizes (test specific)
  - `d-family` (differences) and `r-family` (associations)
  - transform one into other, eg., d = .5 → r = .243

$$d = \frac{2r}{\sqrt{1-r^2}} \qquad r = \frac{d}{\sqrt{d^2+4}} \qquad d = ln(OR) * \frac{\sqrt{3}}{\pi}$$

- NOT p-value ~ partly effect size, but also partly sample size

effect sizes in literature

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

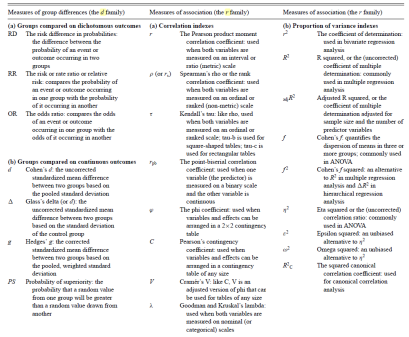
Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed).

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 but beware, just rules of thumb

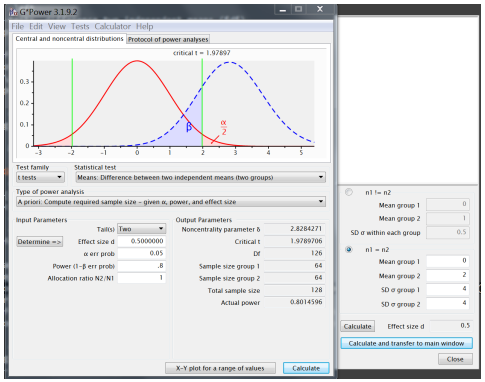
Ellis, P. D. (2010). The essential guide to effect sizes: statistical power, meta-analysis, and the interpretation of research results.



- more than 70 different effect sizes... most of them related to each other

effect sizes in GPower (Determine)

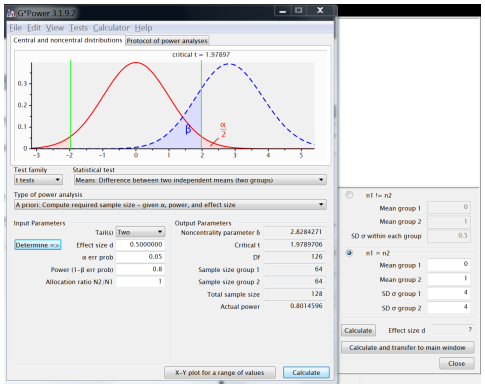
- often very difficult to specify
  - test specific, depends on various statistics
- GPower offers help with `Determine`
  - t-test → group means and sd's
  - one-way anova → variance explained & error
  - regression → again other parameters
  - . . .



effect size exercise : ingredients cohen d

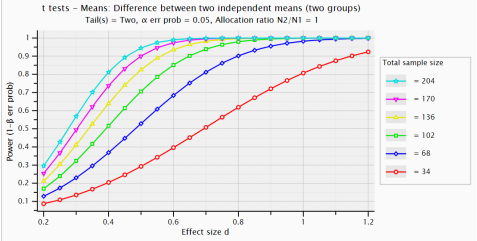
For the `reference example`:

- 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)

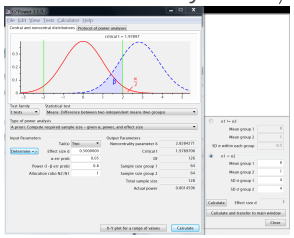


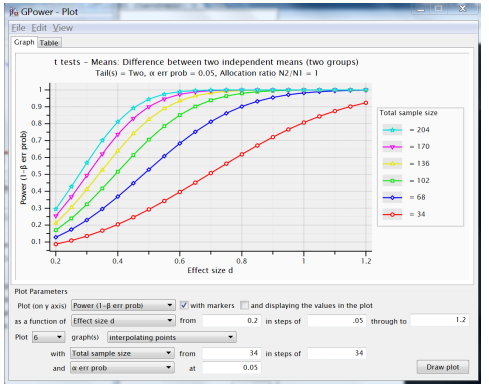
- determine (approximately) the three situations from previous slide on the plot
- how does power change when doubling the effect size, eg., from .5 to 1 ?

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 dependent on aim of the study
  - realistic (eg., previously observed effect) → replicate
  - important (eg., minimally relevant effect)
  - NOT significant → meaningless, dependent on sample size
- choice of effect size dependent on test of interest
  - 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

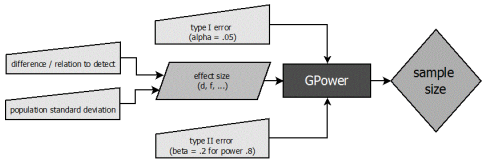
effect sizes, how to determine them in practice

- experts / patients → use if possible → importance
- literature (earlier study / systematic review) → beware of publication bias → realistic
- pilot → guestimate dispersion estimate (not effect size → small sample)
- internal pilot → conditional power (sequential)
- guestimate the input parameters, what can you do ?
  - sd from assumed range / 6 assuming normal distribution
  - sd for proportions at conservative .5
  - sd from control, assume treatment the same
  - ...
- turn to Cohen → use if everything else fails (rules of thumb)
  - eg., .2 - .5 - .8 for Cohen's d

relation sample & effect size, errors I & II

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

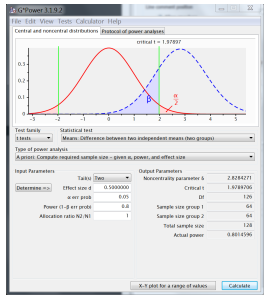
- GPower → type of power analysis
  - Apriori: n ~ α, power, Δ
  - Post Hoc: power ~ α, n, Δ
  - Compromise: power, α ~ β / α, Δ, n
  - Criterion: α ~ power, Δ, n
  - Sensitivity: Δ ~ α, power, n



type of power analysis exercise

- for given reference, step through consecutively ...
  - 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 .7, what is α & β ? critical value ?
- [1] .5 to .7115 = .2115, bigger effect compensates loss of sample size
- [2] α =.09 and β =.38, critical value 1.6994
- [3] α =.05 and β =.2, critical value 1.9990

getting your hands dirty



```
# 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)
```

- in R, assuming normality
  - 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
```


GPower beyond independent t-test

- so far, comparing two independent means
- selected topics with small exercises
  - dependent instead of independent
  - non-parametric instead of assuming normality
  - relations instead of groups (regression)
  - correlations
  - proportions, dependent and independent
  - more than 2 groups (compare jointly, pairwise, focused)
  - more than 1 predictor
  - repeated measures
- GPower manual 27 tests: effect size, non-centrality parameter and example !!

dependence between groups

- if 2 dependent groups (eg., before/after treatment) → account for correlations
- matched pairs (t-test / means, difference 2 dependent means)
- use `reference example`
  - [1] assume correlation .5 and compare (effect size, ncp, n)
  - [2] how many observations if no correlation exists (think then try) ? effect size ?
  - [3] difference sample size for corr = .875 (think: more or less, n/effect size) ?
  - [4] set original sample size (n=64*2) and effect size (dz=.5), power ?
- [1] Δ looks same, n much smaller = 34, BUT: 1 group and dz ~ $\sqrt{2 * (1 - \rho)}$
- [2] approx. independent means, here 65 (estimate the correlation), Δ=.3535 (not .5)
- [3] effect size * 2 → sample size from 34 to 10
- [4] power > .975: for 64 subjects 2 measurements, ncp > 4

non-parametric distribution

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

a relations perspective

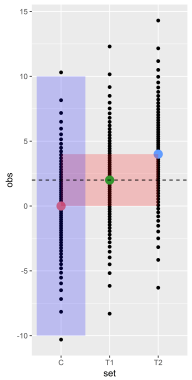
- differences between groups → relation observations & categorization
- example → d = .5 → r = .243 (note: slope $\beta = r * \sigma_y / \sigma_x$)
- regression coefficient (t-test / regression, one group size of slope)
- sample size for comparing the slope Ha with 0 (=Ho)
  - [1] determine slope (β, with SD = 4 and σ_x = .5) and σ_y, [2] calculate sample size
  - [3] determine σ_y for slope 6, σ_x = .5, and SD = 4
  - [4] what if σ_x (predictor values) or σ_y (effect and error) increase (think and try) ?
- [1] σ_x = $\sqrt{.25}$ = .5 (binary, 2 groups: 0 and 1) → slope = 2, σ_y = 4.12 = $\sqrt{4^2 + 1^2}$
- [2] 128, same as for reference example, now with effect size slope H1
- [3] σ_y = 5 = $\sqrt{4^2 + 3^2}$
- [4] sample size decreases with σ_x (opposite σ_y ~ effect size), for same slope

a variance ratio perspective

- between and within group variance → relation observations & categorization
- regression coefficient (t-test / regression, fixed model single regression coef)
- use `reference example`, regression style
  - variance within 4^2 and between 1^2, totaling σ_y^2 = 17
  - [1] calculate sample size, compare effect sizes ?
  - [2] what if also other predictors in the model ?
  - [3] what if 3 predictors extra reduce residual variance to 50% ?
- [1] 128, same as for reference example, now with f^2 = $.25^2$ = .0625.
- [2] loss of degree of freedom, very little impact BUT predictors explain variance
- [3] sample size much less (65) because less noise
- note: $f^2 = R^2/(1 - R^2)$

a variance ratio perspective on multiple groups

- multiple groups → not one effect size `d`
- 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



multiple groups: omnibus

- for 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 `reference example`, just 2 groups
  - [1] what is the sample size (ncp, critical F) ? does size matter ?
  - [2] set extra group, either mean 1, 2 or 4, what are sample sizes (think and try)?
  - [3] derive effect size with `variance` between 2.666667 and within 16 ?
- [1] different effect size (f), distribution, same sample size 128 (size ~ imbalance)
- [2] effect sizes f = .204, .236, .408; sample size n=237 (63x3), 177 (59*3), 63 (21*3) gradually easier to detect
- [3] same effect size (as 0-2-4), sample size 63, ncp 10.5
  (1/7th explained = 1 between / 6 within)

multiple 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 Bonferroni correction for original 3 group example
  - [1] resulting sample size for three tests ?
  - [2] what if biggest difference ignored (know that in between), sample size ?
  - [3] with original 64 sized groups, what is the power (both situations above) ?
- [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 (lower $\alpha \rightarrow \beta$)

multiple 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 !!
  - contrasts (linear combination)
  - standard deviation of contrasts

$$\sigma_{contrast} = \frac{|\sum \mu_i * c_i|}{\sqrt{N \sum_i^k c_i^2 / n_i}}$$

with group means μ_i, pre-specified coefficients c_i, sample sizes n_i and total sample size N

multiple groups: contrasts exercise

- one-way ANOVA (F-test / Means, ANOVA-fixed effects,special,main,interaction)
- obtain effect sizes for contrasts (assume equally 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
  - [2] contrasts nrs. 1 AND 2
  - [3] contrasts nr. 3
- [1] $d = 2f$ 128 (64 C - 64 T1) or 34 (17 C - 17 T2)
- [2] Bonferroni correction → /2 each: 155 and 41 → 177 (78 C, 78 T1, 21 T2)
- [3] total sample size 65 → 22 C, 22 T1, 22 T2

multiple factors

- multiway ANOVA (F-test / Means, ANOVA-fixed effects,special,main,interaction)
- multiple main effects and interaction effects
  - interaction: group specific difference between groups
    - degrees of freedom (#A-1)*(#B-1)
  - main effects: if no interaction (#X-1)
  - get effect sizes for two way anova https://icds.shinyapps.io/effectsizes/
- sample size for `reference example`, assume a second predictor is trivial
  - [1] what is partial η^2 ?
  - [2] sample size ?
- [1] .0588 (0-2, sd=4) → $f^2 = \eta^2/(1-\eta^2)$ & $d = 2f$ → d=.5
- [2] 128 again, with 2 groups (158 with 3 groups, df=2)

multiple factors: within group dependence

- if repeated measures → correlations
- repeated measures (F-test / Means, repeated measures...)
- 3 main types
  - within: similar to dependent t-test for multiple measurements
  - between: use of multiple measurements per group
  - interaction: change between over within
- correlation within subject (unit)
  - informative within subject (like paired t-test)
  - redundancy on information between subject
- note: Options 'as in SPSS' if based on effect sizes that include correlation

repeated measures within

- possible to have only 1 group (within subject comparison)
- use effect size f = .25 (1/16 explained versus unexplained)
  - [1] mimic dependent t-test, correlation .5
  - [2] mimic independent t-test
  - [3] double number of groups to 2, or 4 (cor = .5)
  - [4] double number of measurements to 4 (correlation 0 and .5), impact ?
- number of groups = 1, number of measurements = 2, sample size = [1] 34 and [2] 65
- [3] changed degrees of freedom, sample size could change little bit
- [4] impact nr measurements depend on correlation 0: (65x2)-45x4-30x8 / .5: (34x2)-24x4-16x8

repeated measures between

- use effect size f = .25 (1/16 for variance or 2/4 for means)
  - [1] mimic independent t-test
  - [2] use correlation 0 and .5 with 2 groups and 2 measurements, sample size ?
  - [3] for correlation .5, compare 2, 4, 8 measures, sample size (think and try) ?
  - [4] double number of groups to 4, 8 for 4 measures and corr .5
- [1] 128 (2 groups of 64, each 2 measurements, 256, but second uninformative)
- [2] more if higher corr., sample sizes up 66x2=132~128 for 0, 98x2=196 for .5
- [3] sample size lower when more measurements, unless correlation is 1 (82x2=164)
- [4] more groups require higher sample size (82-116-152) but effect size ignored

repeated measures interaction within x between

- issue effect sizes and correlation: https://www.youtube.com/watch?v=CEQUNYg80Y0
- get effect sizes for two way anova https://icds.shinyapps.io/effectsizes/
  - [1] if change 0-1-2 vs. 2-3-4 or 4-3-2, with cor=.5, what are effect sizes ?
  - [2] compare effect sizes with correlation 0 and .5 ?
  - [3] sample size to detect the interaction with cor .5 or cor 0 ?
- [1] no interaction, or no main effect, f=.3535
- [2] f=.25, higher correlation higher effect size
- [3] beware: corr. 0, 82x2, each 3; corr. .5, 44x2, each 3

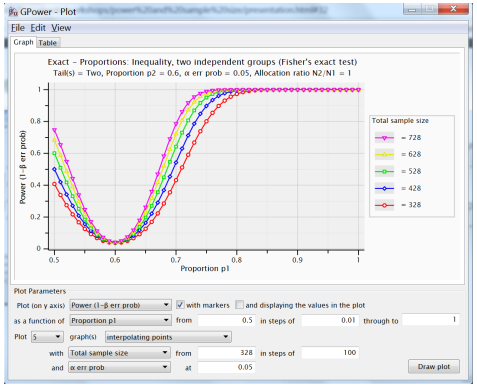
correlations

- 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] generate plot: explain curve minimum, relation sample size ?
  - [2] repeat for one-tailed, difference ?

- [1] power for proportion compared to reference .6, sample size determines impact
- [2] one-tailed, increases power (both sides !?)

dependent proportions

- when comparing two dependent proportions
- McNemar test (exact / proportions, difference 2 dependent proportions)
  - information from changes → discordant pairs
  - effect size as odds ratio → ratio of discordance ?!
- 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, .5, and 1
  - [2] odds ratio 4-.25, prop discordant .25, how about p12, p21 and sample sizes ?
  - [3] repeat for third alpha option, and consider total sample size, what happens ?
- [1] sample size 288-144-72 (limits), less with more info
- [2] 1 to 4 or 4 to 1 → same sample size 80 (.25)
- [3] sample size differs because side effects

not included

- various statistical tests difficult to specify in gpower
  - various statistics that are difficult to guestimate
  - manual for more complex tests not always very elaborate
- various statistical tests not included in gpower
  - eg., survival analysis
  - many tools online, not all with high quality
- various statistical tests no closed form solution
  - simulation may be the only tool
    - iterate many times: generate and analyze → proportion of rejections
      - generate: simulated outcome ← model and uncertainties
      - analyze: simulated outcome → model

simulation example t-test

```
gr <- rep(c('T','C'),64)
y <- ifelse(gr=='C',0,2)
dta <- data.frame(y=y,X=gr)
cutoff <- qt(.025,nrow(dta))

sim1 <- function(){
    dta$y <- dta$y+rnorm(length(dta$X),0,4)     # generate (with uncertainty)
    res <- t.test(data=dta,y~X)                 # analyze
    c(res$estimate %*% c(-1,1),res$statistic,res$p.value)   # keep results
}
sims <- replicate(10000,sim1())     # large number of iterations
dimnames(sims)[[1]] <- c('diff','t.stat','p.val')

mean(sims['p.val',] < .05)  # p-values
[1] 0.8029
mean(sims['t.stat',] < cutoff)  # t-statistics
[1] 0.8029
mean(sims['diff',] > sd(sims['diff',])*cutoff*(-1)) # estimated differences
[1] 0.8024
```

focus / simplify

- complex statistical models
  - simulate BUT program and model well understood
  - focus on essential elements → simplify the aim
- sample size calculations (design) for simpler research aim
  - not necessarily equivalent to final statistical testing / estimation
  - requires justification to convince yourself and/or reviewers
    - successful already if simple aim is satisfied
    - ignored part is not too costly
- example:
  - statistics: group difference evolution 4 repeated measurements → mixed model
  - focus: difference treatment and control last time point → t-test
  - argument: first 3 measurements cheap, difference at end interesting

conclusion:

- sample size calculation is a design issue, not a statistical one
- building blocks: sample & effect sizes, type I & II errors, each conditional on rest
- effect sizes express the amount of signal compared to the background noise
- bigger effects require less information to detect them (smaller sample size)
- complex models → complex sample size calculations, maybe only simulation
- GPower deals with not too complex models
  - more complex complex models imply more complex specification
  - simplify using a focus, if justifiable → then GPower can get you a long way

about us ...

- statistical consultancy at VUB / UZ
  - for PhD students and researchers
  - for master students' thesis
- collaboration on data analysis
- website @ www.icds.be
- book us @ www.icds.be/consulting/