Sample Size

Tuan V. Nguyen
Professor and NHMRC Senior Research Fellow
Garvan Institute of Medical Research
University of New South Wales
Sydney, Australia

BMJ

LONDON, SATURDAY 29 JANUARY 1994

The scandal of poor medical research

We need less research, better research, and research done for the right reasons

Home Contents **Records**

jameslindlibrary.org

Key Passages Commentary Context

Bracken MB. Why animal studies are often poor predictors of human reactions to exposure.

Commentary on: Perel P, Roberts I, Sena E, Wheble P, Briscoe C, Sandercock S, Macleod M, Mignini LE, Jayaram P, Khan KS (2007). Comparison of treatment effects between animal

Research article **Open Access**

Methodological quality of systematic reviews of animal studies: a survey of reviews of basic research

Luciano E Mignini[1] and Khalid S Khan*[2]

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Practical difference vs statistical significance

Outcome	Group A	Group B
Improved	9	18
No improved	21	12
Total	30	30
% improved	30%	60%

Chi-square: 5.4; $P < 0.05$
"Statistically significant"

Outcome	Group A	Group B
Improved	6	12
No improved	14	8
Total	20	20
% improved	30%	60%

Chi-square: 3.3; $P > 0.05$
"Statistically insignificant"

What do you want?

Study design issues

- Setting
- Participants: inclusion / exclusion criteria
- Design: survey, factorial, etc
- Measurements: outcome, covariates
- Analysis

- **Sample size / power**

How many subjects do I need ?

- Tough question: dependent on many parameters
- Important question: money!
- Ethical question: too many or too few → unethical

Sample size issues

- How many patients / participants
  - Practical relevance vs statistical significance
  - Ethical issues
- Ethical issues
  - Unnecessarily large number of patients: unethical
  - Too small a sample: unethical

4 ingredients for sample size consideration

- How much error you are prepared to commit?
- Parameter (or outcome) of major interest
- Variability of the outcome of interest
- Effect size

Types of error in research

- **Type I error** (α): the probability of finding something in your sample but there's nothing going on in the population

"something" = effect

- **Type II error** (β): the probability of *not* finding something in your sample but there's something going on in the population
- **Power = 1 – β.**

Diagnosis and statistical reasoning

Medical diagnosis reasoning

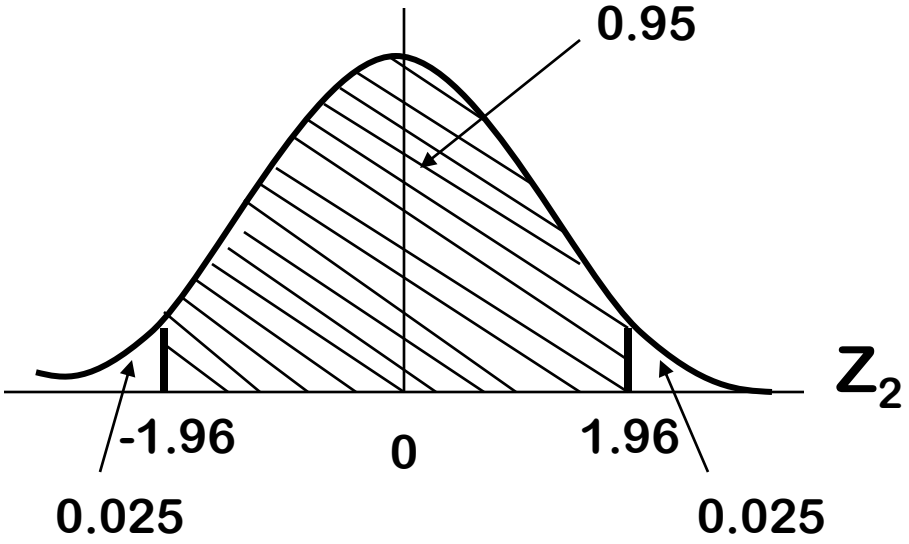
	Disease status	
	Present	Absent
Test result		
+ve	True +ve (sensitivity)	False +ve
-ve	False -ve	True -ve (Specificity)

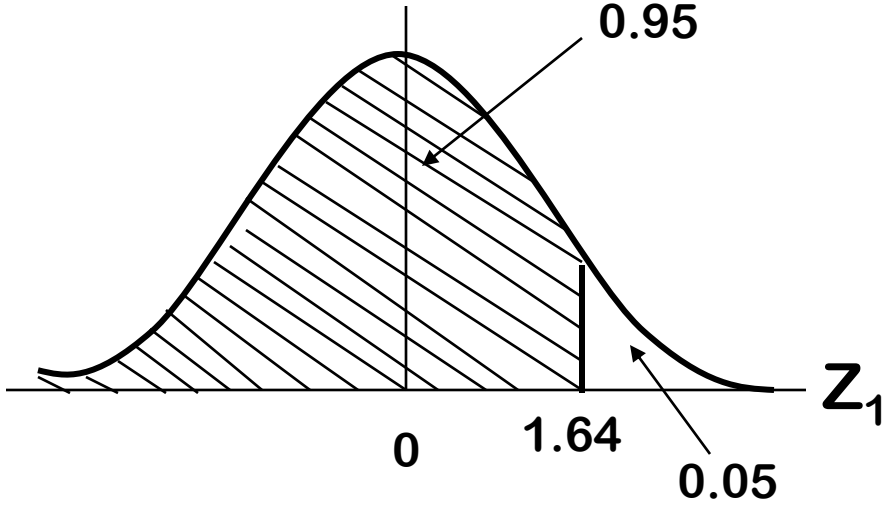
Statistical/research reasoning

	There is an effect	NO effect
Stat test		
Significant	No error $1-\beta$	Type I err. α
Not significant	Type II err. β	No error $1-\alpha$

"Significant": $P < 0.05$
"Not significant": $P > 0,05$

Constants associated with type I and II errors





Prob.	Z_1	Z_2
0.80	0.84	1.28
0.90	1.28	1.64
0.95	1.64	1.96
0.99	2.33	2.81

The Normal deviates (Gospel)

Alpha c	Z_{α} (One-sided)	$Z_{\alpha/2}$ (Two-sided)
0.20	0.84	1.28
0.10	1.28	1.64
0.05	1.64	1.96
0.01	2.33	2.81

Power	$Z_{1-\beta}$
0.80	0.84
0.90	1.28
0.95	1.64
0.99	2.33

Parameter of interest

- Type of measurement of outcome:
  - Continuous (blood pressure, bone density, glucose, etc)
  - Categorical outcome (dead/alive, fracture/no fracture)
- Examples:
  - Incidence of fracture, CVD, diabetes, death
  - Hedotic scale: 0-10
  - Nominal scale

Variability of the outcome of interest

- If the parameter is a continuous variable:
  - What is the **standard deviation** (SD) ?
- If the parameter is a categorical variable:
  - SD can be estimated from the proportion/probability.

Effect of sample size: a simulation

	True mean: 100 True SD: 15		True mean: 100 True SD: 35	
Sample size	Est. M	SD	Est. M	SD
10	98.0	11.0	108.9	32.2
50	100.4	13.6	95.3	41.4
100	101.3	14.4	99.1	35.5
200	99.9	15.2	100.3	33.2
500	99.8	15.3	98.9	33.8
1000	99.5	15.1	99.9	35.0
2000	99.7	15.0	99.9	34.7
10000	100.1	15.0	99.9	35.0
100000	100.0	15.0	100.0	35.0

Effect size

- Distinction between clinical relevance and statistical significance.
- A trivial effect can be statistically significant if the sample size is very large (meaningless study).
- A large effect may not be statistically significant if the sample size is inadequate
- QUESTION: what is "clinically relevant effect size"?
  - Context dependent
  - Educated guess (which more likely to be wrong)

Effect size

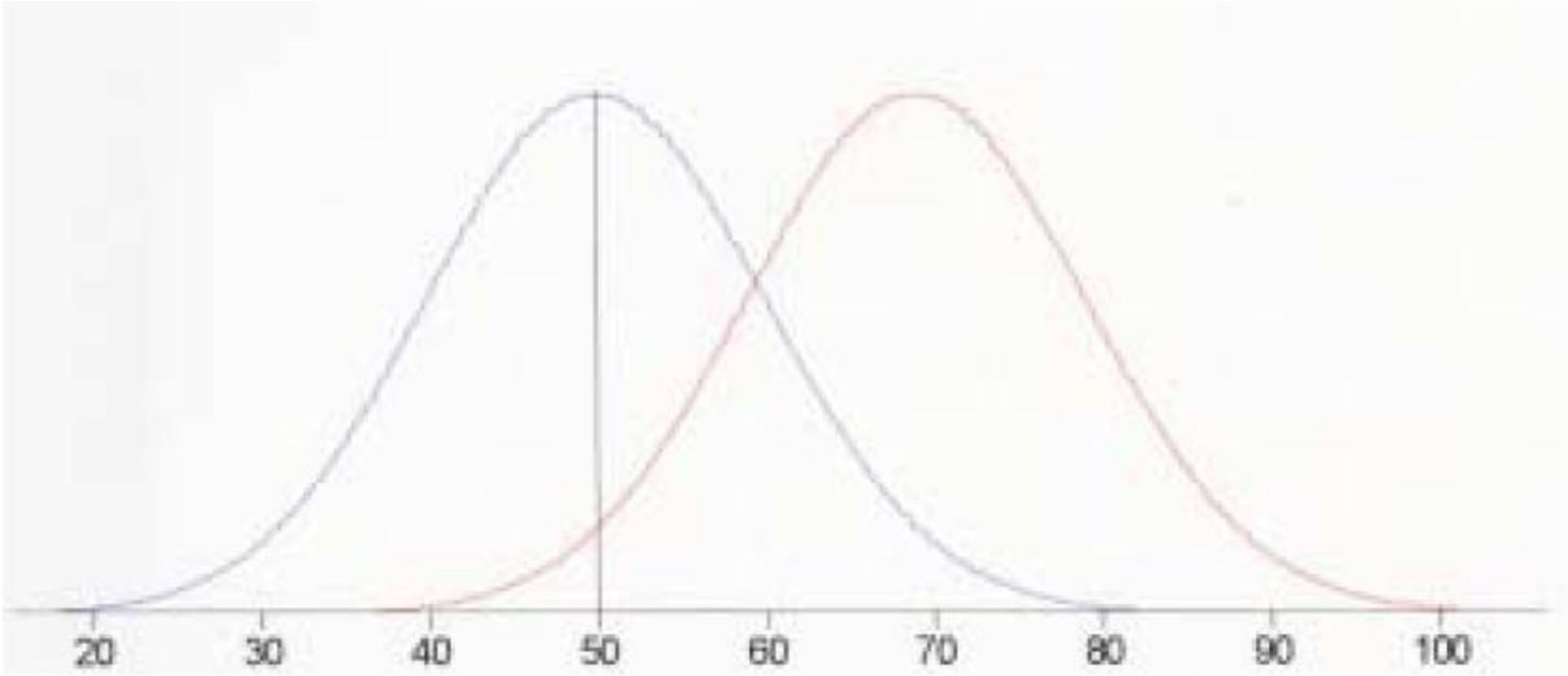
$$ES = \frac{\Delta}{S}$$

Δ = difference in *means* between 2 groups

S = Standard deviation

One way to see effect size

Effect size = 0.5



Interpretation of ES

Meaning	Effect size (ES)	Percent of overlap
Small	0.0	100
	0.1	92.3
	0.2	85.3
Medium	0.3	78.7
	0.4	72.6
	0.5	67.0
Large	0.6	61.8
	0.7	57.0
	0.8	52.6
	0.9	48.4
	1.0	44.6
	1.5	29.3
	2.0	18.9

General formula

$$N = \frac{2 \times C(\alpha, \beta)}{(ES)^2}$$

- **ES = effect size**
- **C(α, β) = constants defined by α and β**

α	β = 0.05	β = 0.10	β = 0.20
0.10	10.8	8.6	6.2
0.05	13.0	10.5	7.9
0.02	15.8	13.0	10.0
0.01	17.8	14.9	11.7

Case 1:
One sample
Outcome: continuous variable

Sample size for estimating a population mean

- We are interested in estimating the mean age at cancer diagnosis for a certain group of patients.
  - We would like to estimate the mean age within $\pm$ 2.5 years (95% CI of width 5 years).
  - Suppose that we estimate the population's standard deviation as 12 years.

Sample size for estimating a population mean

- Assume that we want 95% CI, so α = 0.05, C = 1.96
- Standard deviation: σ = 12
- Margin of error: d = 2.5
- The sample size needed :

$$N = \frac{(1.96)^2 \sigma^2}{d^2} = \frac{(1.96)^2 \times (12)^2}{(2.5)^2} = 89$$

Case 2:
One sample
Outcome: dichotomous variable

Sample size for estimating a proportion

- The prevalence of vertebral fracture is know to be ~20%.
  - We would like to estimate the prevalence within ± 3% accuracy
  - With 95% confidence interval

Sample size for estimating a proportion

- Let the prevalence be π, and the margin error be *d*. The sample size can be estimated as:

$$N = \frac{(1.96)^2 \pi(1-\pi)}{d^2} = \frac{(1.96)^2 \times (0.2)(0.8)}{(0.03)^2} = 683$$

Case 3:
One sample
Outcome: correlation between 2 variables

Sample size for estimating a correlation

- The correlation coefficient between fasting plasma glucose and BMD ranged between 0.08 đến 0.30
- We want to find sample size with α = 0.01 and power = 0.80 (eg β= 0.20)

Sample size for estimating a correlation

- **Let the correlation coefficient be *r***
- **For α = 0.01 and β= 0.20, we have C(α, β) = 11.7**

α	β = 0.05	β = 0.10	β = 0.20
0.01	17.8	14.9	11.7

- The sample size required:

$$N = 3 + \frac{C(\alpha,\beta)}{0.25\left[\log\left(\frac{1+r}{1-r}\right)\right]^2} = 3 + \frac{11.7}{0.25\left[\log\left(\frac{1+r}{1-r}\right)\right]^2}$$

Sample size for estimating a correlation

- The correlation coefficient between fasting plasma glucose and BMD ranged between 0.08 đến 0.30
- We want to find sample size with α = 0.01 and power = 0.80 (eg β= 0.20)

Corr coeff (r)	Sample size for power = 0.80 and	
	α = 0.01	α = 0.05
0.05	4527	3138
0.10	1128	783
0.15	499	347
0.20	279	194
0.25	177	123
0.30	121	85
0.35	88	62
0.40	66	47
0.45	51	36
0.50	41	29

Case 4:
Two samples
Outcome: continuous data

Sample size for two means

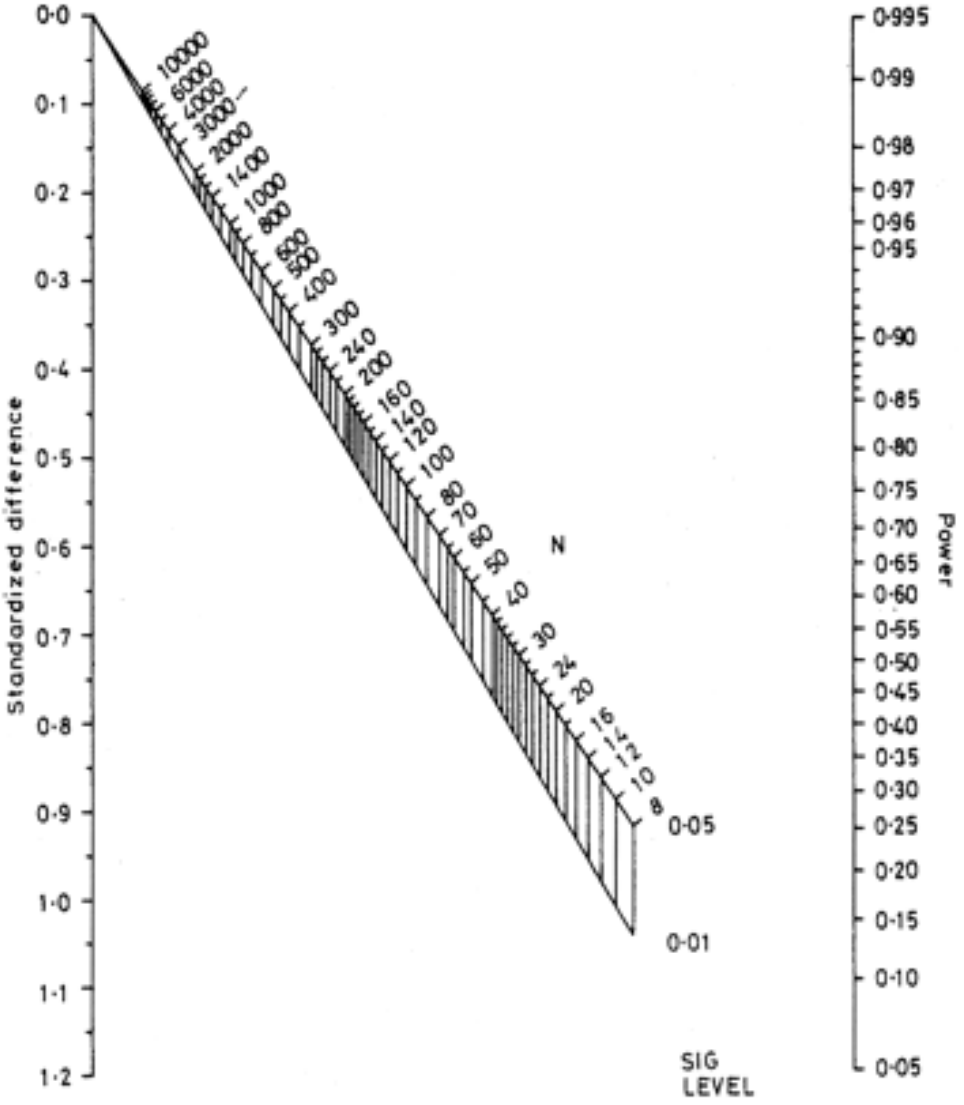
- An RCT is to be conducted on post-menopausal women
- Two groups: placebo and bisphosphonate
- Endpoint: bone mineral density (BMD)
- Bisphosphonate increases BMD by about 5%
- Baseline BMD = 0.80 g/cm^2
- Standard deviation = 0.12 g/cm^2

Sample size for two means: consideration

- **Difference: d = 0.80 x 0.05 = 0.04 g/cm^2**
- **Effect size: ES = 0.04 / 0.12 = 0.33**
- **Alpha = 0.05, beta = 0.90, C(α, β) = 10.5**
- **Sample size PER GROUP:**

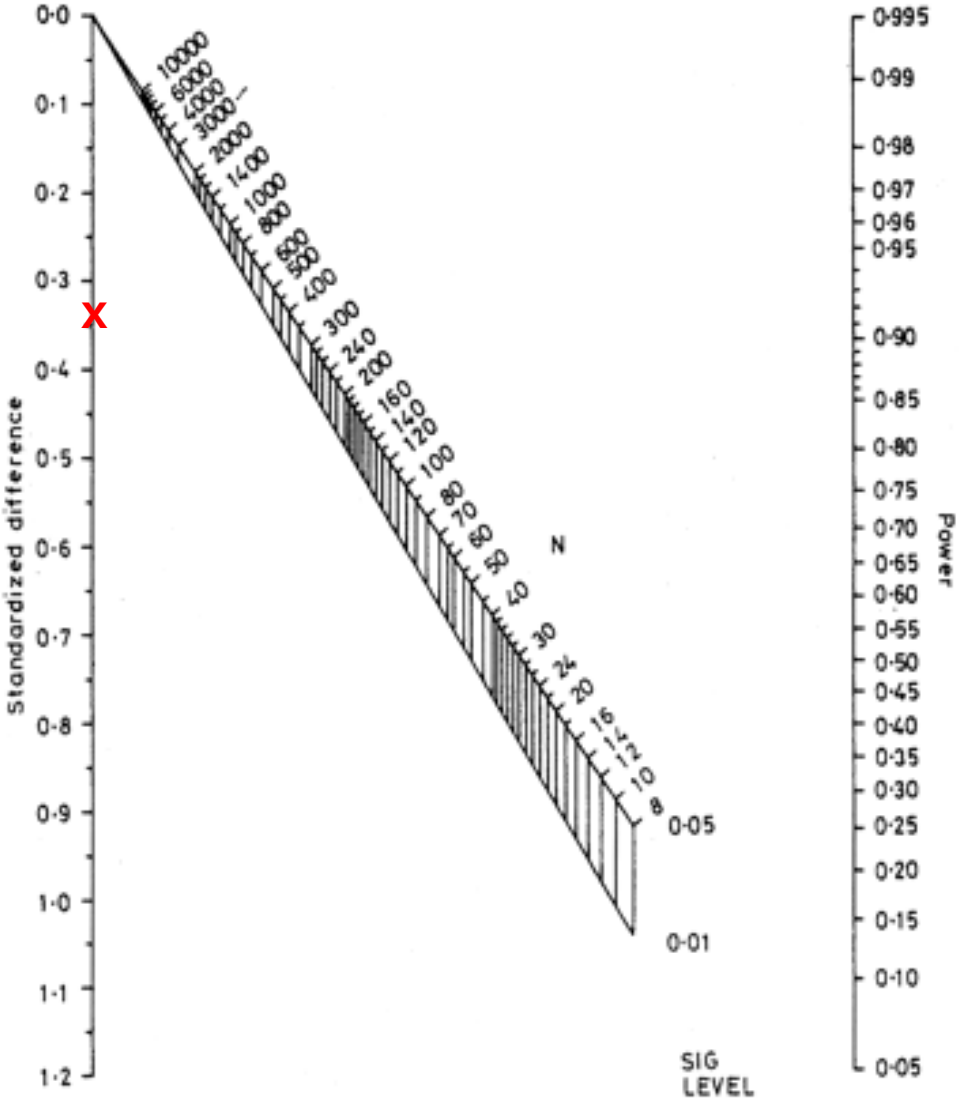
$$N = \frac{2 \times C(\alpha, \beta)}{(ES)^2} = \frac{2 \times 10.5}{(0.33)^2} = 189$$

Sample size nomogram



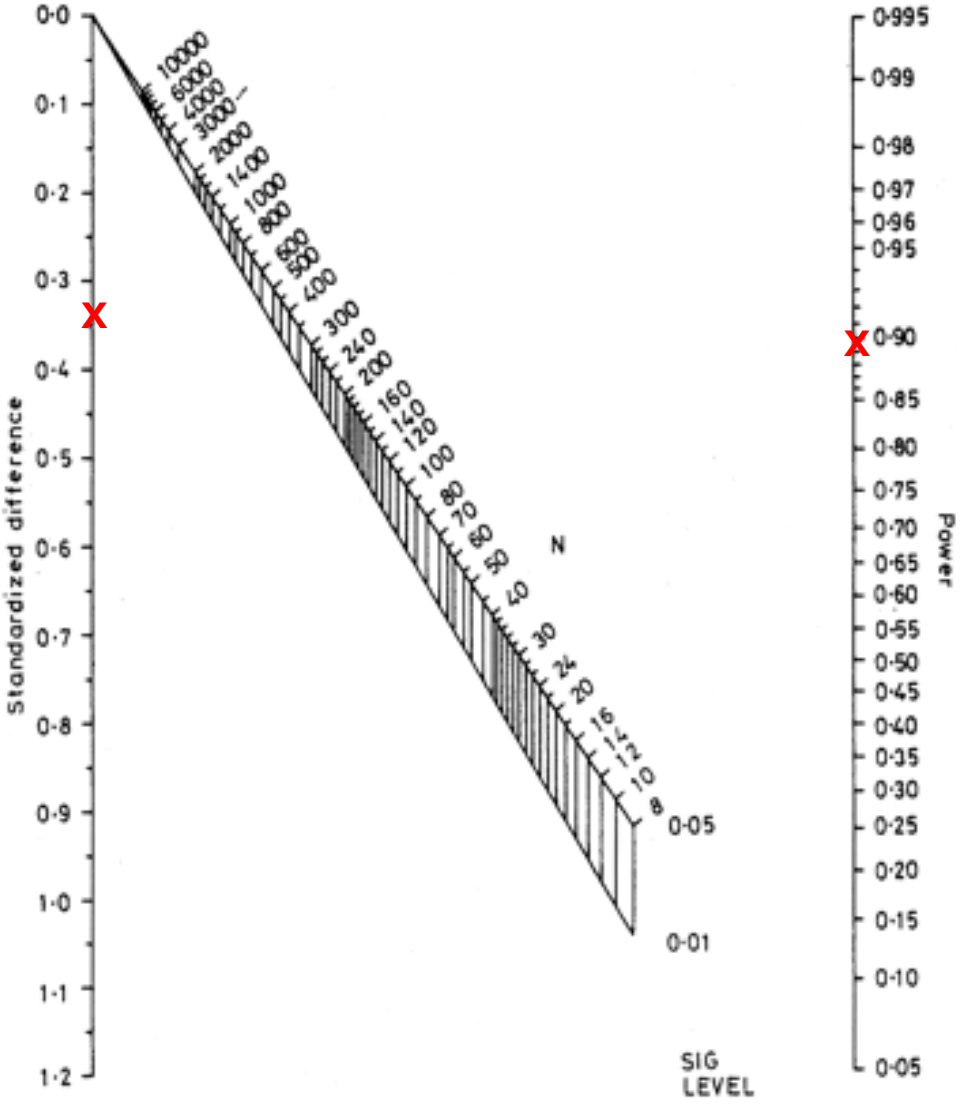
Sample size nomogram

ES = 0.33



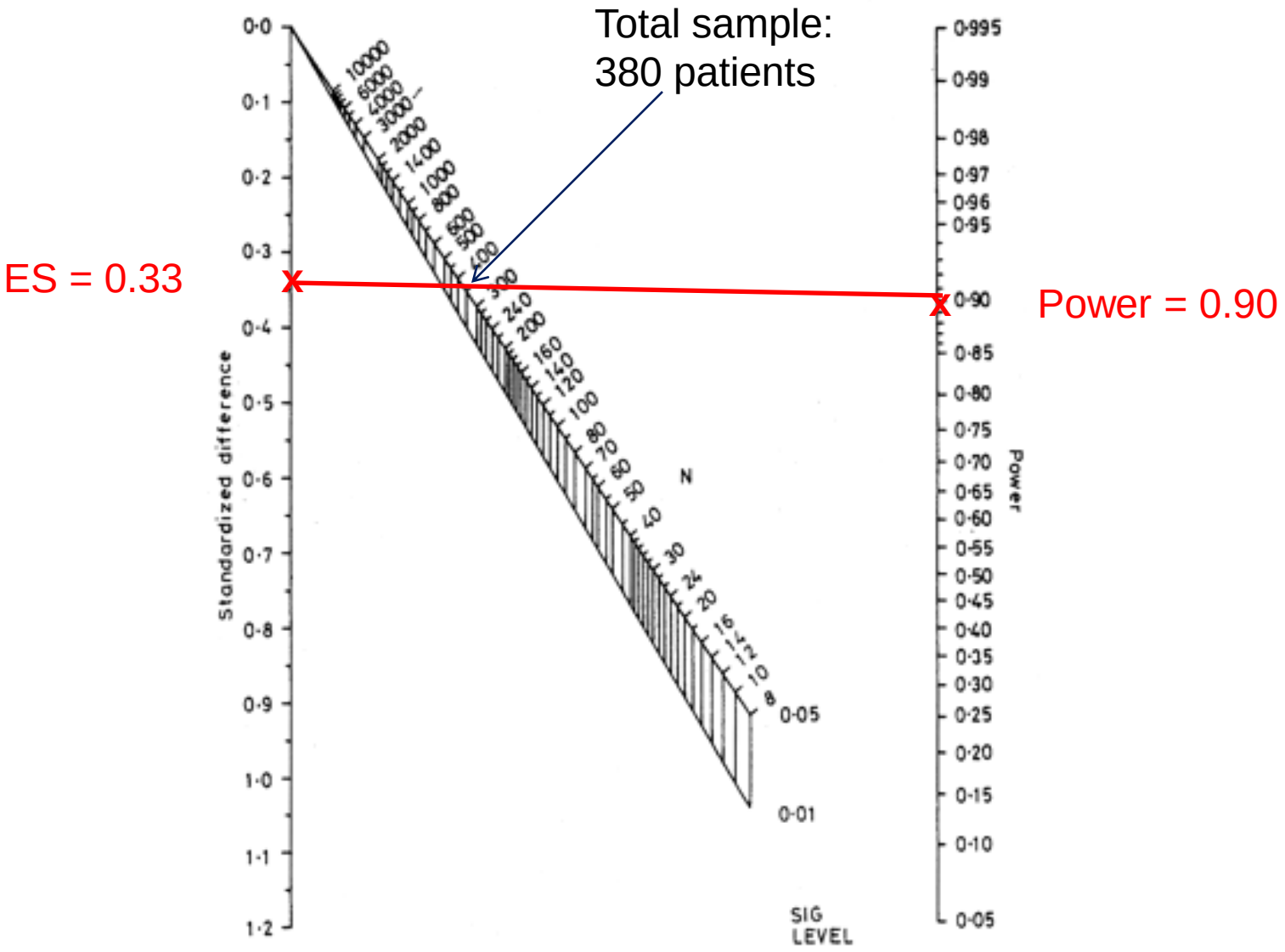
Sample size nomogram

ES = 0.33



Power = 0.90

Sample size nomogram



Simple approximation

- For alpha = 0.05, power = 0.80 (beta = 0.2)
- Sample size per group can be approximated by the following formula

$$N = \frac{16}{(ES)^2}$$

Case 5:
Two samples
Outcome: binary (dichotomous) data

Sample size for two proportions

- A clinical study has 2 groups of patients treated with
  - Standard drug
  - New drug
- Endpoint: death
- Expected percent of survival
  - Standard drug: 85%
  - New drug: 90%
- Power = 0.90

Sample size for two proportions

- **P_1 = probability of survival in group 1**
- **P_2 = probability of survival in group 2**
- **Difference (effect): $d = P_2 - P_1$**
- **Standard deviation of difference:**

$$SD = \sqrt{p_1(1-p_1) + p_2(1-p_2)}$$

- And effect size

$$ES = \frac{d}{SD}$$

- Sample size

$$N = \frac{C(\alpha, \beta)}{(ES)^2}$$

Sample size for two proportions

- $P_1 = 0.85$
- $P_2 = 0.90$
- $d = P_2 - P_1 = 0.90 - 0.85 = 0.05$
- $C(\alpha, \beta) = 10.5$

$$SD = \sqrt{0.85(1-0.85)+0.90(1-0.90)} = 0.467$$

$$ES = \frac{d}{SD} = \frac{0.05}{0.467} = 0.107$$

- And the sample size PER GROUP is:

$$N = \frac{C(\alpha,\beta)}{(ES)^2} = \frac{10.5}{(0.107)^2} = 917$$

Case 6:
Case-control study
Outcome: binary (dichotomous) data

Sample size for an *odds ratio:* example

- **Example**: The prevalence of vertebral fracture in a population is 25%. It is interested to estimate the effect of smoking on the fracture, with an odds ratio of 2, at the significance level of 5% (one-sided test) and power of 80%.
- The total sample size for the study can be estimated by:

$$N = \frac{4(1.64 + 0.85)^2}{(\ln 2)^2 \times 0.25 \times 0.75} = 275$$

Some comments

- Sample size consideration: a MUST in study design
- Mechanic of sample size computation: trivial
- Ingredients in sample size consideration: serious
- Do NOT try to remember formulae!
- Do remember concepts, idea, and logics
- Sample size has its own variability
- Sample size is only an approximation
- Non-response must be accounted for

Estimation of sample size using R

Function power.t.test

- Outcome: continuous variable
- Design: two groups
- Difference of interest: 0.5
- Common standard deviation: 1
- R code

```
power.t.test(delta=0.5, sd=1, power=0.8)
```

Function power.t.test

```
> power.t.test(delta=0.5, sd=1, power=0.8)

     Two-sample t test power calculation

              n = 63.76576
          delta = 0.5
             sd = 1
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

NOTE: n is number in *each* group

Function power.prop.test

- Outcome: binary variable
- Design: two groups
- Proportion group 1 = 0.75
- Proportion group 2 = 0.50
- R code

```
power.prop.test(p1=0.75, p2=0.50, power=0.8)
```

Function power.t.test

```
> power.prop.test(p1=0.75, p2=0.50, power=0.8)

Two-sample comparison of proportions power
calculation

              n = 57.67344
             p1 = 0.75
             p2 = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

Package `samplesize`

- Package `samplesize` is quite useful!
- n.indep.t.test.eq: independent t-test for equal groups
- n.indep.t.test.neq: independent t-test for unequal groups
- n.welch.test: sample size for t-test with Welch's approximation
- n.ipaired.t.test: sample size for paired t-test
- n.wilcox.ord: sample size for Wilcoxon-Mann-Whitney test for ordinal data

Two groups with equal sample size per group

```
library(samplesize)

n.indep.t.test.eq(mean.diff=0.5, sd.est=1,
power=0.80, alpha=0.95)


[1] "sample.size:" "101"
```

101 patients for 2 groups (~50 per group)

Two groups with unequal sample size per group

```
library(samplesize)

n.indep.t.test.neq(mean.diff=0.5, sd.est=1,
power=0.80, alpha=0.95, k=0.5)


"sample.size:"      "113"
"sample.size n.1:" "75.3"

"sample.size n.2:" "37.7"
```

Two groups with unequal variances

```
library(samplesize)

n.welch.test(mean.diff=0.5, sd.est=1,
sd.est2=1.2, power=0.80, alpha=0.95)


sample.size: 123

sample.size n1: 56

sample.size n2: 67
```

Sample size for Wilcoxon's test

```
library(samplesize)

n.wilcox.ord(t=0.5, p=c(0.66,0.15,0.19),
q=c(0.61, 0.23, 0.16), beta=0.20, alpha=0.05)


[1] 8340.925
```

Package `pwr`

- Package `pwr` is quite useful!
- pwr.2p.test: two proportions, equal sample size
- pwr.2p2n.test: two proportions, unequal sample size
- pwr.anova.test: sample size for one-way ANOVA
- pwr.chisq.test: sample size for Chi squared test
- pwr.p.test: sample size for 1 proportion
- pwr.r.test: sample size for a correlation coefficient
- pwr.t.test: sample size for t-test (one or two samples)
- pwr.t2n.test: sample size for t-test with unequal size

pwr package

- pwr package is based on the concept of effect size (ES)
- ES = 2*arcsin(sqrt(p1)) – 2*arcsin(sqrt(p2))

```
ES = ES.h(0.75, 0.50)
```

Sample size for 2 proportions

- Proportion in group 1: 0.75
- Proportion in group 2: 0.50
- Power = 0.80, alpha=0.05

```
ES = ES.h(0.75, 0.50)

pwr.2p.test(h=ES, power=0.80, sig.level=0.05)
```

Sample size for 2 proportions

```
> library(pwr)
> ES = ES.h(0.75, 0.50)
> pwr.2p.test(h=ES, power=0.80, sig.level=0.05)

Difference of proportion power calculation for binomial
distribution (arcsine transformation)

              h = 0.5235988
              n = 57.25842
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

NOTE: same sample sizes

```
Compared with:
power.prop.test(p1=0.75, p2=0.50, power=0.80)
```

Sample size for time to an event

- Package `gsDesign` with function `nSurvival`
- Rate of death in group 1: 0.30
- Rate of death in group 2: 0.20
- Drop-out rate: 0.10
- Duration of follow-up: 3 years
- Duration of recruitment: 1 year

```
library(gsDesign)

nSurvival(lambda1=0.3, lambda2=0.20, eta=0.1,
Ts=3, Tr=1, alpha=0.025, beta=0.20)
```

Sample size for time to an event

```
Fixed design, two-arm trial with time-to-event
outcome (Lachin and Foulkes, 1986).
Study duration (fixed):              Ts=3
Accrual duration (fixed):            Tr=1
Uniform accrual:                  entry="unif"
Control median:          log(2)/lambda1=2.3
Experimental median:     log(2)/lambda2=3.5
Censoring median:           log(2)/eta=6.9
Control failure rate:          lambda1=0.3
Experimental failure rate:     lambda2=0.2
Censoring rate:                   eta=0.1
Power:                    100*(1-beta)=80%
Type I error (1-sided):     100*alpha=2.5%
Equal randomization:              ratio=1
Sample size based on hazard ratio=0.667 (type="rr")
Sample size (computed):               n=466
Events required (computed): nEvents=192
```

Sample size for time to an event

- Relative reduction: 30% (RR = 0.70)
- Power = 0.80, alpha = 0.025
- How many patients are required, how many deaths are required

```
library(gsDesign)

nSurvival(lambda1=-log(0.5)/6, lambda2=-
0.7*log(0.5)/6, eta=-log(0.95)/12, Ts=30, Tr=36,
type="rr", entry="unif")
```

Sample size for time to an event

```
> nSurvival(lambda1=-log(0.5)/6, lambda2=-0.7*log(0.5)/6,
eta=-log(0.95)/12, Ts=30, Tr=36, type="rr", entry="unif")
Fixed design, two-arm trial with time-to-event
outcome (Lachin and Foulkes, 1986).
Study duration (fixed):              Ts=30
Accrual duration (fixed):            Tr=36
Uniform accrual:                  entry="unif"
Control median:          log(2)/lambda1=6
Experimental median: log(2)/lambda2=8.6
Censoring median:           log(2)/eta=162.2
Control failure rate:          lambda1=0.116
Experimental failure rate:  lambda2=0.081
Censoring rate:                   eta=0.004
Power:                    100*(1-beta)=90%
Type I error (1-sided):   100*alpha=2.5%
Sample size based on hazard ratio=0.7 (type="rr")
Sample size (computed):              n=680
Events required (computed): nEvents=329
```

Package `epiR`

- `epiR` is a very useful package for epidemiologic analyses
- Function for sample size: `epi.studysize`
- Computes the sample size, power, and minimum detectable difference for
  - cohort studies (using count data)
  - case control studies

  when comparing means and survival

epi.studysize

```
epi.studysize(treat, control, n, sigma, power, r
= 1, conf.level = 0.95, sided.test = 2, method =
"means")
```

`treat`: the expected value for the treatment group

`control`: the expected value for the control group

`sigma`: method = "means" this is the expected standard deviation of the variable of interest for both treatment and control groups.

When method = "case.control" this is the expected proportion of study subjects exposed to the risk factor of in- terest.

This argument is ignored when method = "proportions", method = "survival", or method = "cohort.count".

Two means

- Alpha = 5% =, power = 95%
- Treatment group = 5
- Control = 4.5
- Standard deviation = 1.4

```
epi.studysize(treat = 5, control = 4.5, n =
NA, sigma = 1.4, power = 0.95, r = 1,
conf.level = 0.95, sided.test = 1, method =
"means")
```

Two means

```
> epi.studysize(treat = 5, control = 4.5, n = NA, sigma
= 1.4, power = 0.95, r = 1, conf.level = 0.95,
sided.test = 1, method = "means")

$n.crude

[1] 340

$n.total

[1] 340

$n.treat

[1] 170

$n.control

[1] 170
```

Two proportions

- Prevalence of smoking = 0.30
- Control group = 0.32
- Alpha = 0.05 level, power = 0.90
- How many men need to be sampled?

```
epi.studysize(treat = 0.30, control = 0.32, n
= NA, sigma = NA, power = 0.90, r = 1,
conf.level = 0.95, sided.test = 1, method =
"proportions")
```

```
$n.total
```

```
[1] 4568
```

Survival data

- The 5-year survival probability of patients receiving a standard treatment = 0.30
- A new treatment will increase it to 0.45.
- Alpha = 0.05, Power = 0.90
- How many events are required?

```
epi.studysize(treat = 0.45, control = 0.30, n
= NA, sigma = NA, power = 0.90, r = 1,
conf.level = 0.95, sided.test = 2, method =
"survival")
```

Survival data

```
> epi.studysize(treat = 0.45, control = 0.30, n = NA,
sigma = NA, power = 0.90, r = 1, conf.level = 0.95,
sided.test = 2, method = "survival")

$n.crude

[1] 250

$n.total

[1] 250

$n.treat

[1] 125

$n.control

[1] 125
```

Sample size for hazard ratio

- What is the minimum detectable hazard in a study involving 500 subjects
- Power = 0.90 and alpha = 0.05 level

```
epi.studysize(treat = NA, control = NA, n =
500, sigma = NA, power = 0.90, r = 1,
conf.level = 0.95, sided.test = 2, method =
"survival")
```

```
$hazard

[1] 1.336334
```

Sample size for cohort study

- Follow-up period = 5 years
- Power = 0.90, alpha = 0.05
- Relative risk = 1.4
- Death rate in non-smokers is 413 per 100000 per year.
- Assuming equal numbers of smokers and non-smokers are sampled, how many should be sampled overall?

Sample size for cohort study

```
treat = 1.4 * (5 * 413)/100000

control = (5 * 413)/100000

epi.studysize(treat = treat, control = control, n = NA,
sigma = NA, power = 0.90, r = 1, conf.level = 0.95,
sided.test = 1, method = "cohort.count")


$n.total

[1] 12130

$n.treat

[1] 6065

$n.control

[1] 6065
```

Sample size for cohort study

- Follow-up period = 5 years
- Power = 0.90, alpha = 0.05
- Relative risk = 1.4
- Death rate in non-smokers is 413 per 100000 per year.
- Assuming number of non-smokers is 3 times the number of smokers, how many should be sampled overall?

Sample size for cohort study

```
treat = 1.4 * (5 * 413)/100000

control = (5 * 413)/100000

epi.studysize(treat = treat, control = control, n = NA,
sigma = NA, power = 0.90, r = 0.33, conf.level = 0.95,
sided.test = 1, method = "cohort.count")


$n.total

[1] 16061.08

$n.treat (smokers)

[1] 3985.08

$n.control (non-smokers)

[1] 12076
```

Sample size for case-control study

- Relationship between smoking and CHD
- Risk factor: smokers, 30% men are smokers
- OR = 2, power = 0.9, alpha=0.05
- How many subjects?

```
epi.studysize(treat = 2/100, control = 1/100,
n = NA, sigma = 0.30, power = 0.90, r = 1,
conf.level = 0.95, sided.test = 2, method =
"case.control")
```

Sample size for case-control study

```
> epi.studysize(treat = 2/100, control =
1/100, n = NA, sigma = 0.30, power = 0.90, r
= 1, conf.level = 0.95, sided.test = 2,
method = "case.control")

$n.total

[1] 376

$n.treat

[1] 188

$n.control

[1] 188
```