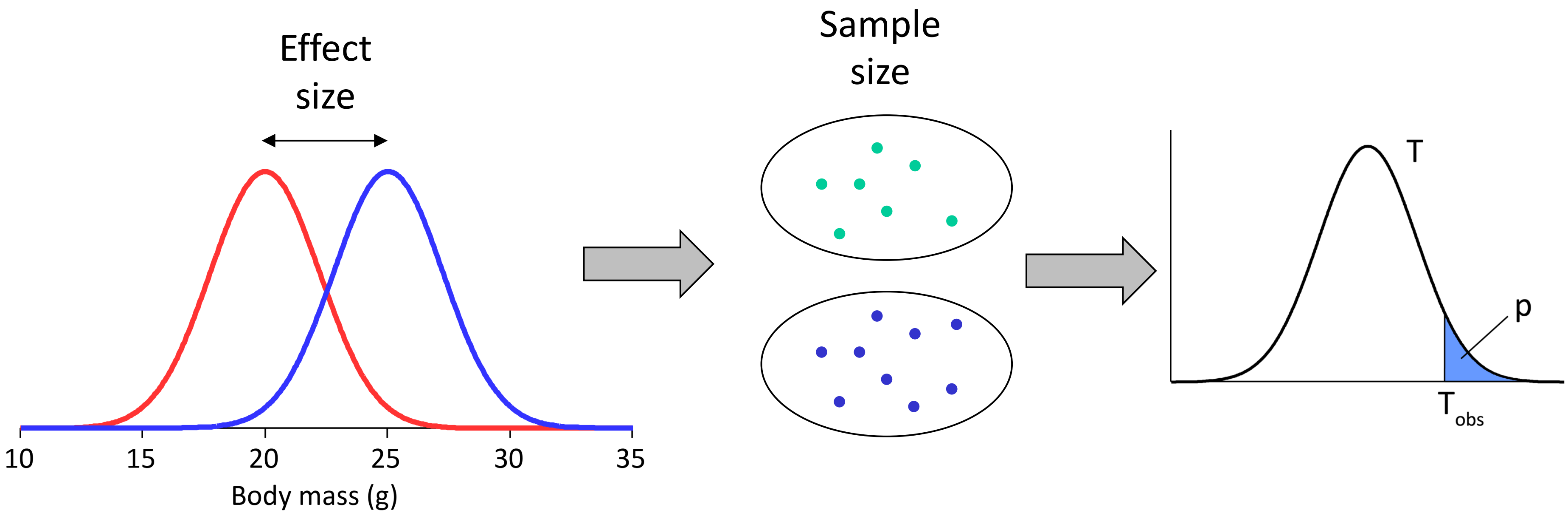
11. Statistical power

"If your experiment needs statistics, you ought to have done a better experiment"

Ernest Rutherford

Statistical power: what is it about?



Two populations (alternative hypothesis)

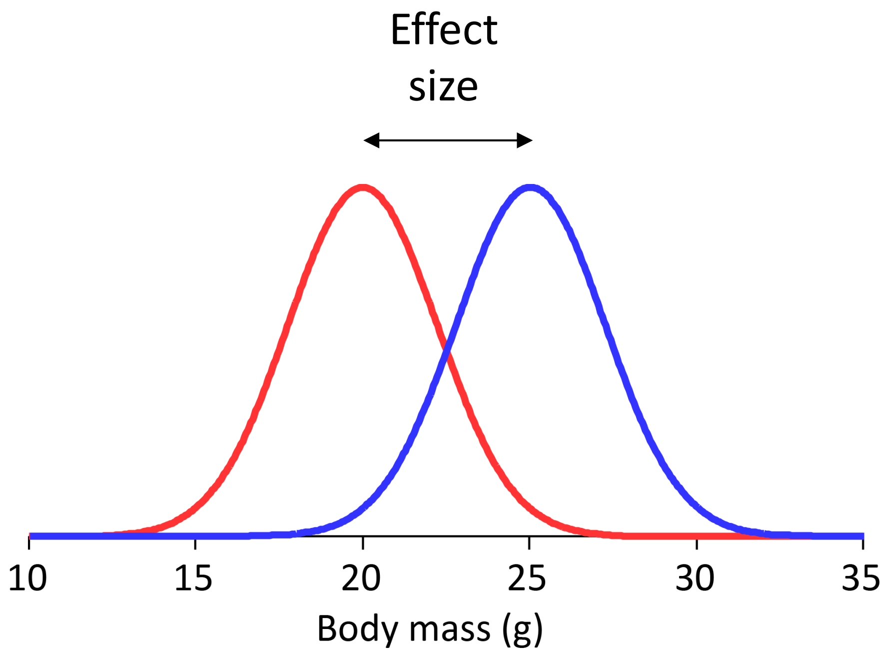
Two samples

Statistical significance

How does our ability to call a change "significant" depend on the effect size and the sample size?

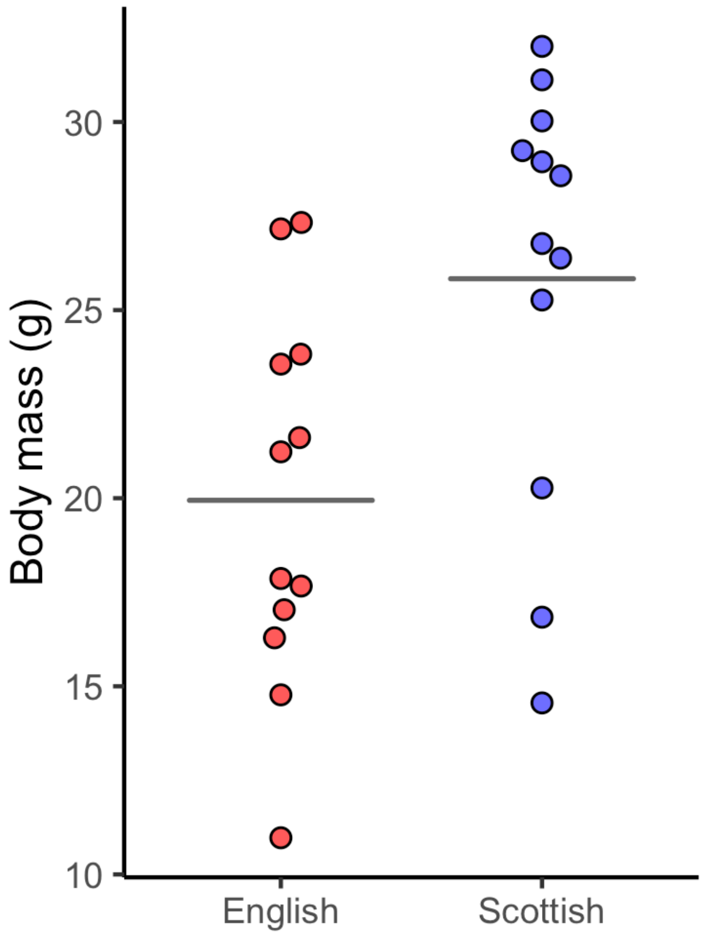
Effect size

Effect size describes the alternative hypothesis



Effect size for two sample means

$d = 1.1$



Horizontal bars represent sample means

$$d = \frac{M_1 - M_2}{SD}$$

Cohen's d

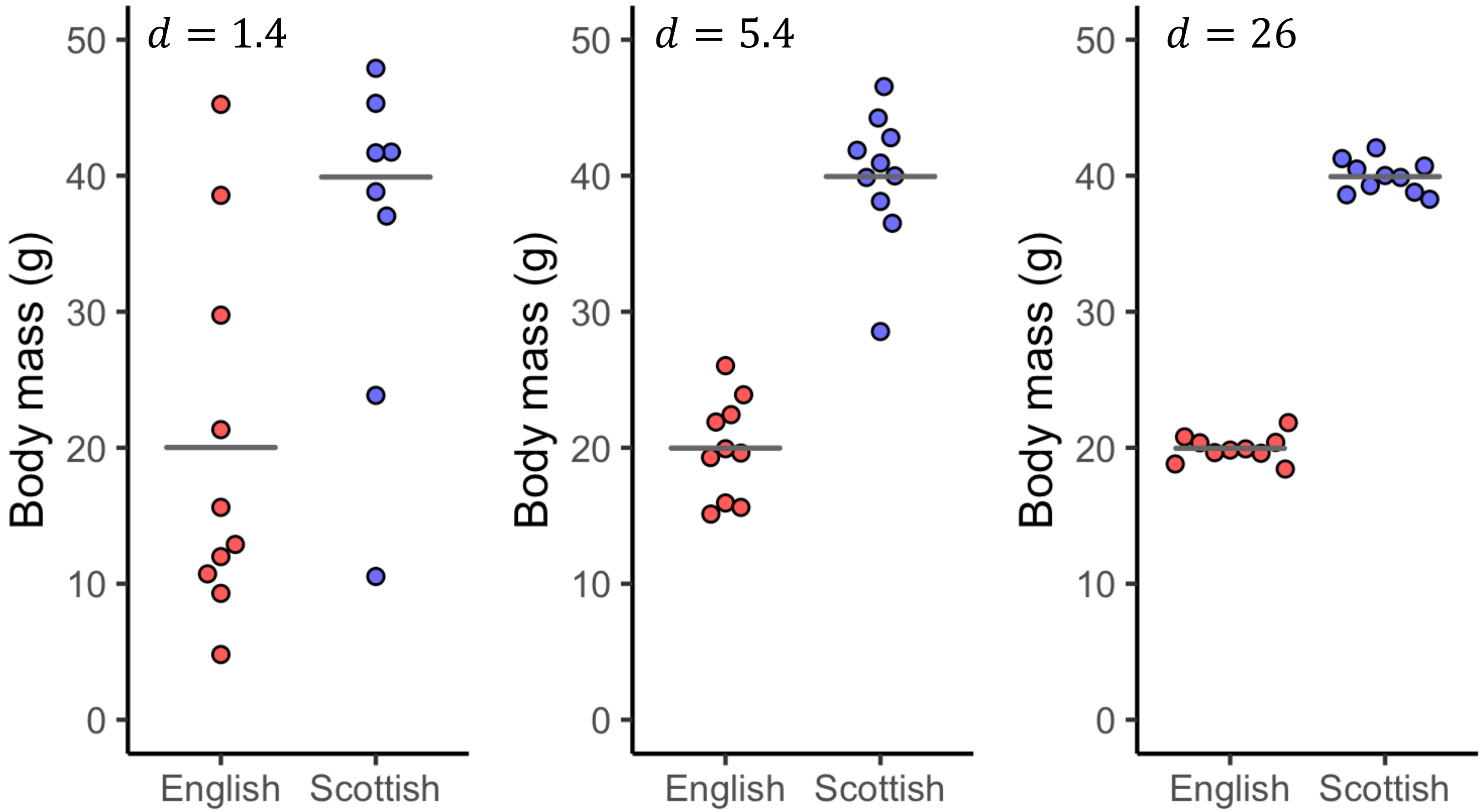
M – mean
SD – standard deviation

$$SD = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}}$$

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*

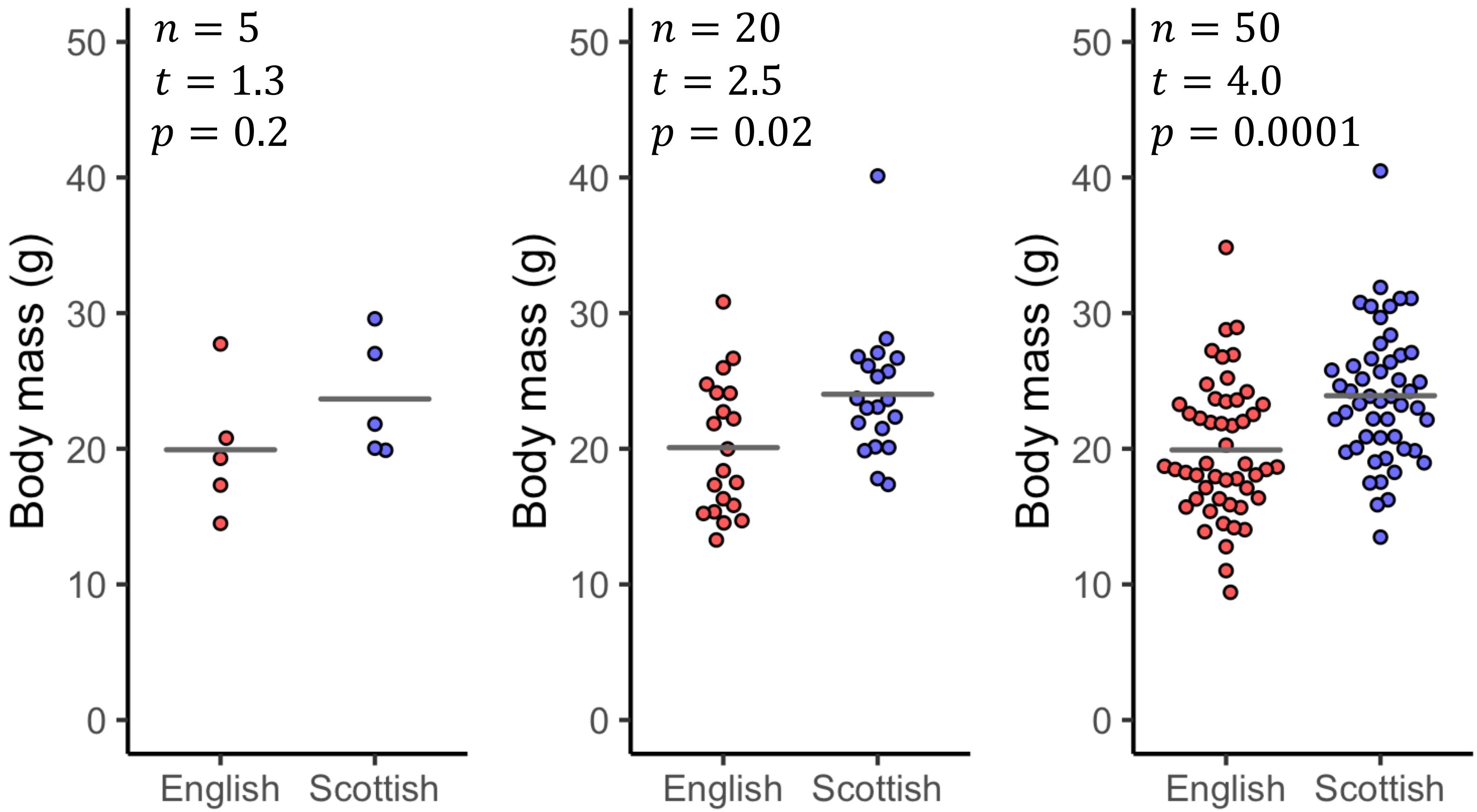
Effect size depends on the standard deviation

Difference between means = 20 g



Effect size does not depend on the sample size

Effect size = 0.8



Comparing two samples

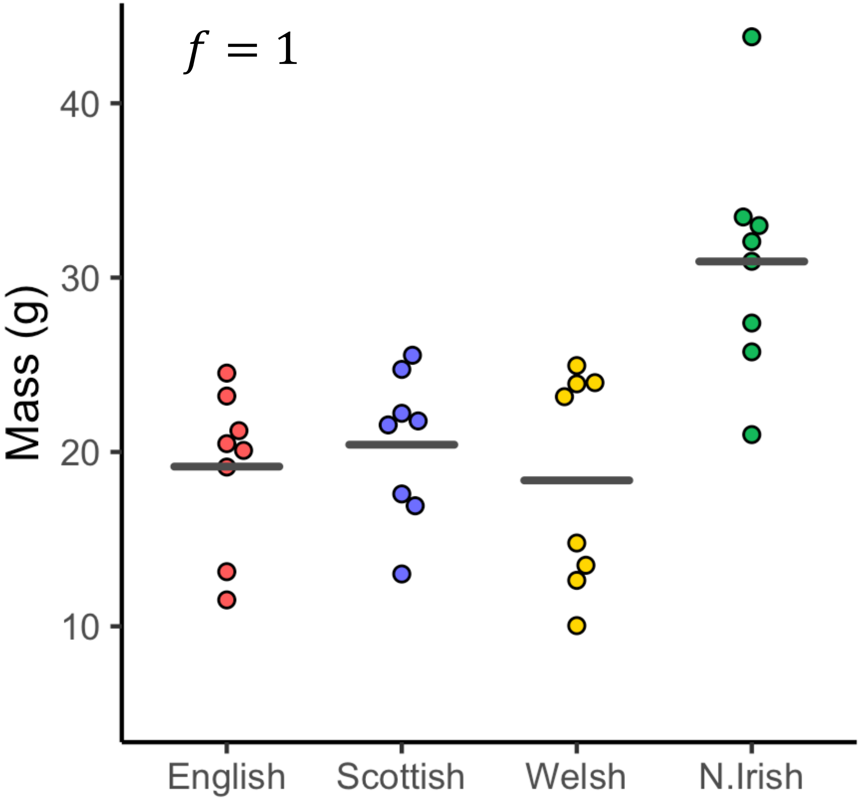
Statistic	Formula	Description
Difference	$\Delta M = M_1 - M_2$	Difference between sample means
Ratio	$r = \frac{M_1}{M_2}$	Often used as logarithm
Cohen's d	$d = \frac{M_1 - M_2}{SD}$	Effect size; takes spread in data into account
t-statistic	$t = \frac{M_1 - M_2}{SE}$	Directly relates to statistical significance; takes spread of data and sample size into account

M – mean
SD – standard deviation
SE – standard error

Effect size describes the alternative hypothesis

Effect size is not related to statistical significance

Effect size in ANOVA



Test statistic

$$F = \frac{MS_B}{MS_W}$$

H_0: $MS_B = MS_W$

H_1: $MS_B = MS_W + nMS_A$

Added variance

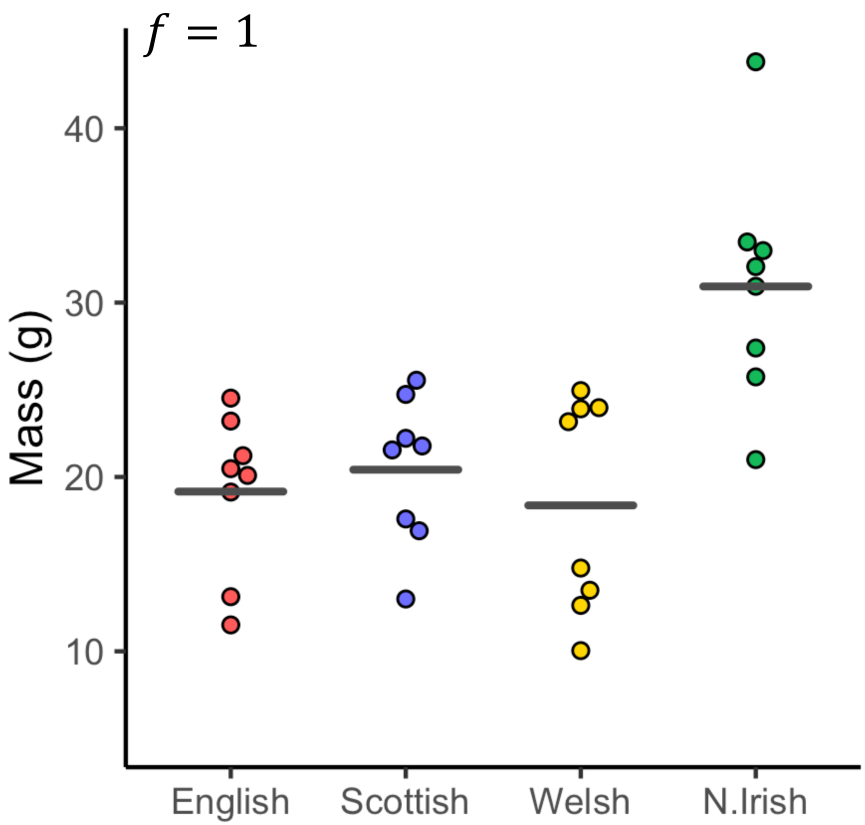
$$f^2 = \frac{MS_A}{MS_W}$$

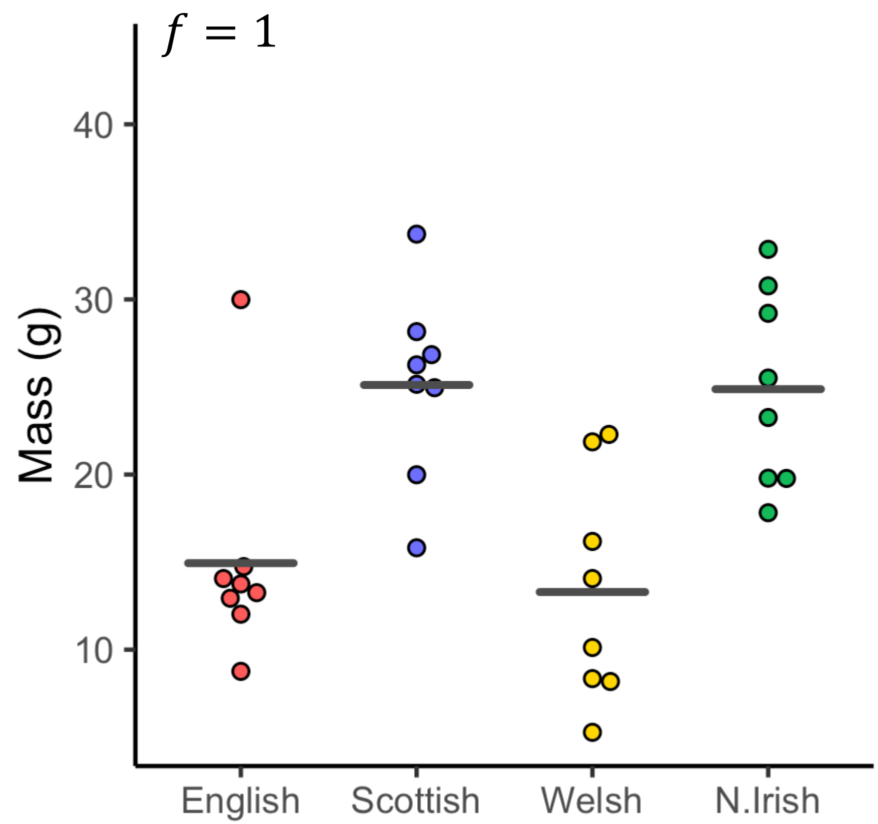
Cohen's f

$$f^2 = \frac{F-1}{n}$$

For the purpose of this calculation we only consider groups of equal sizes, n

Effect size in ANOVA





Effect size

Data	Statistical test	Effect size	Formula
Two sets, size n_1 and n_2	t-test	Cohen's d	$d = t\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$
k groups of n points each	ANOVA	Cohen's f	$f = \sqrt{\frac{F - 1}{n}}$
contingency table	chi-square	Cohen's w	$w = \sqrt{\frac{\chi^2}{N}}$
Paired data $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$	Significance of correlation	Pearson's r	$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - M_x}{SD_x}\right)\left(\frac{y_i - M_y}{SD_y}\right)$

Statistical power

t-test

Statistical testing

Statistical model

Null hypothesis

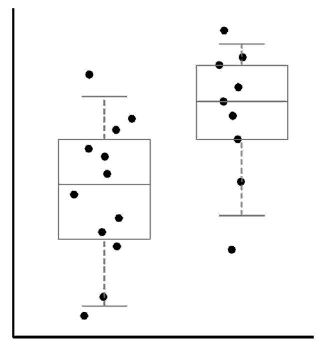
H_0: no effect

All other assumptions

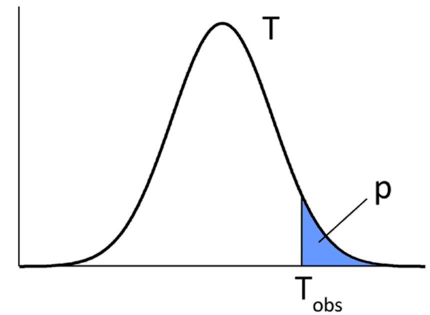
Significance level

$\alpha = 0.05$

Data



Statistical test against H_0



p-value: probability that the observed effect is random

$p < \alpha$

Reject H_0

(at your own risk)

Effect is real

$p \geq \alpha$

Accept H_0 (!!!)

This table

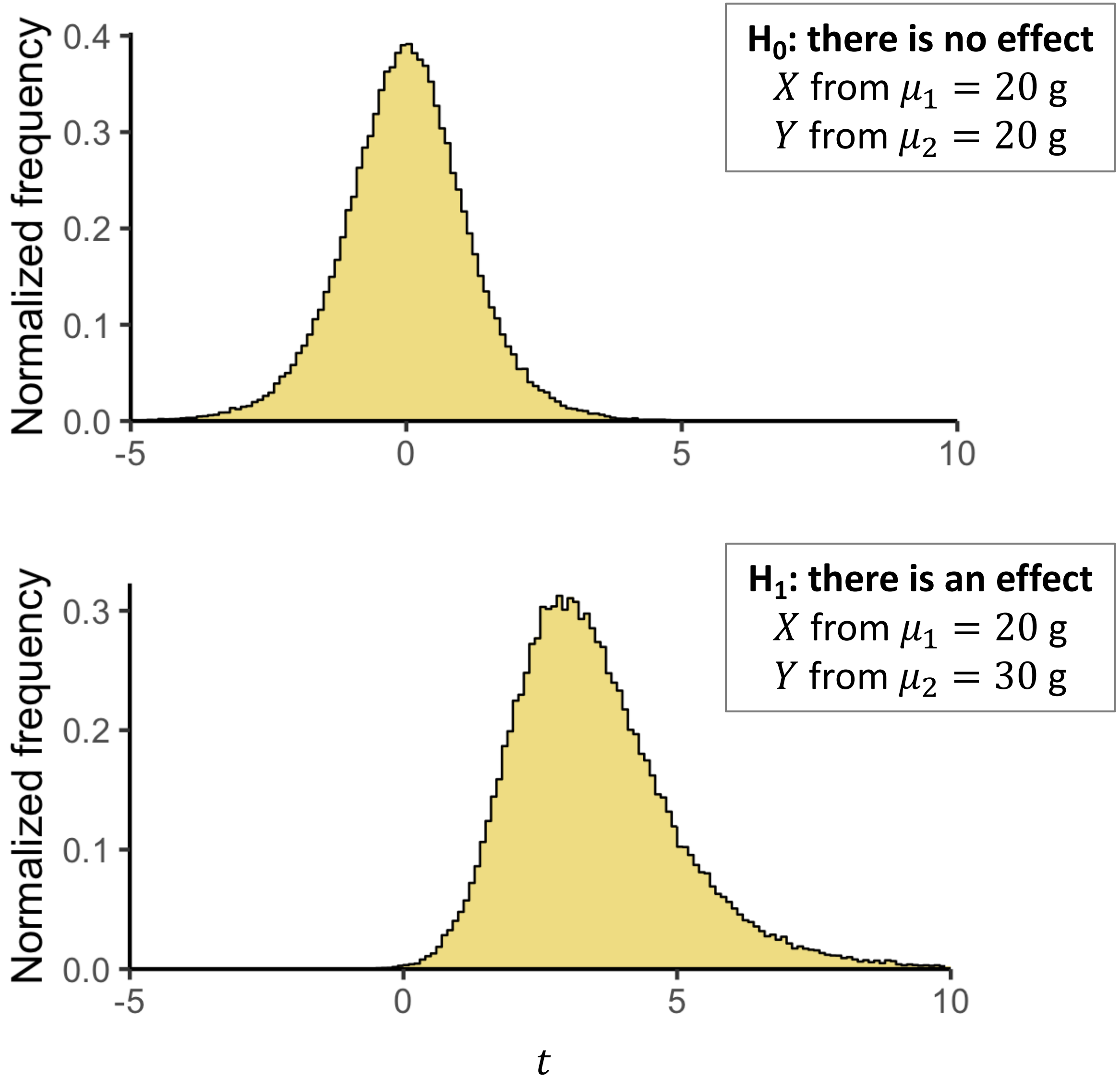
	H_0 is true	H_0 is false	
H_0 rejected	**type I error (α)** false positive	**correct decision** true positive	Positive
H_0 accepted	**correct decision** true negative	**type II error (β)** false negative	Negative
	No effect	Effect	

Gedankenexperiment

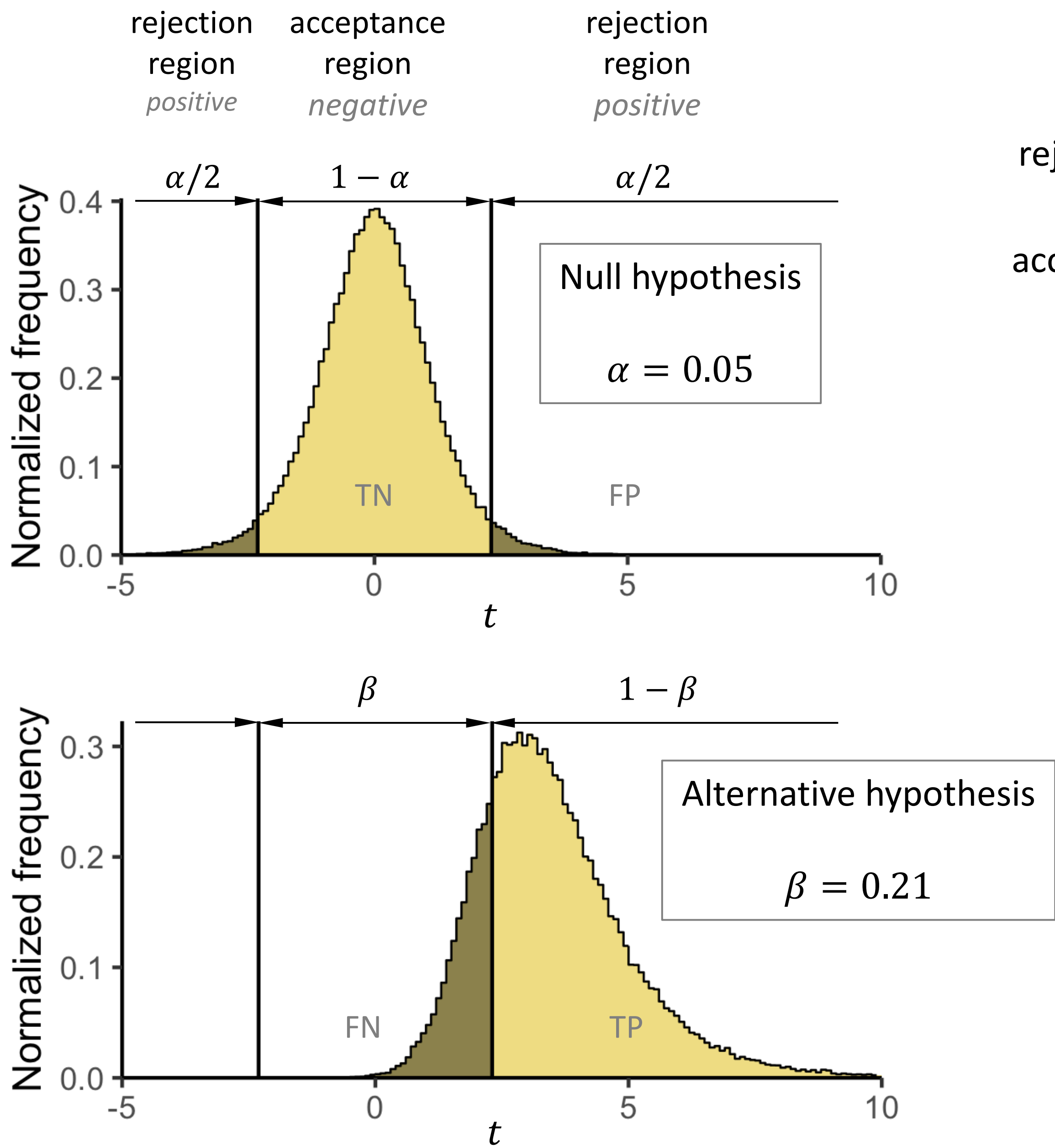
Draw 100,000 pairs of samples (X, Y) of size $n = 5$

Find $t = (M_1 - M_2)/SE$ for each pair

Build sampling distribution of t



One alternative hypothesis



	H_0 true	H_0 false	
reject	FP $\boldsymbol{\alpha}$	TP	positive
accept	TN	FN $\boldsymbol{\beta}$	negative
	no effect	effect	

Power of the test

$$P = 1 - \beta$$

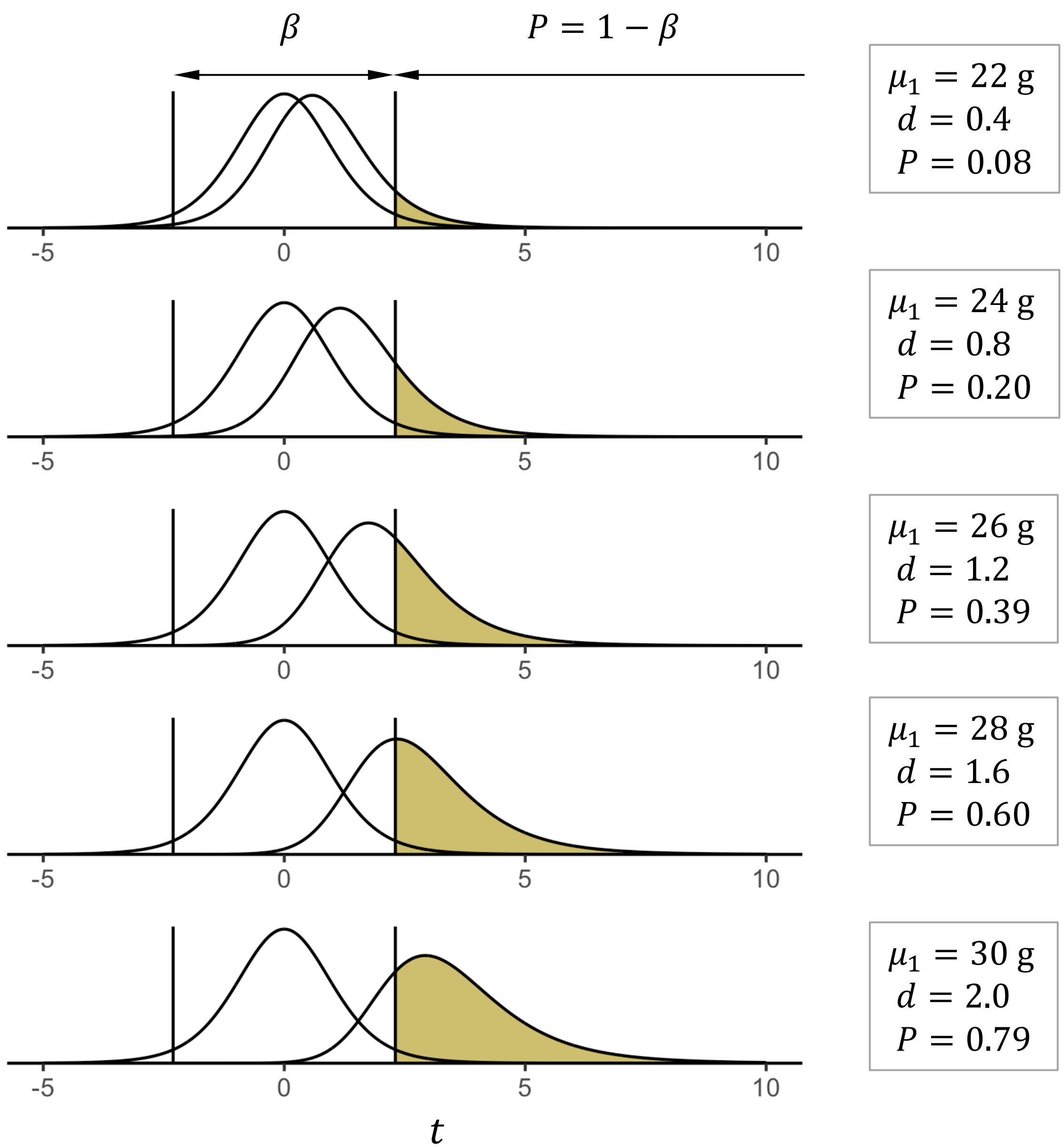
Probability that we correctly reject H_0

Statistical power

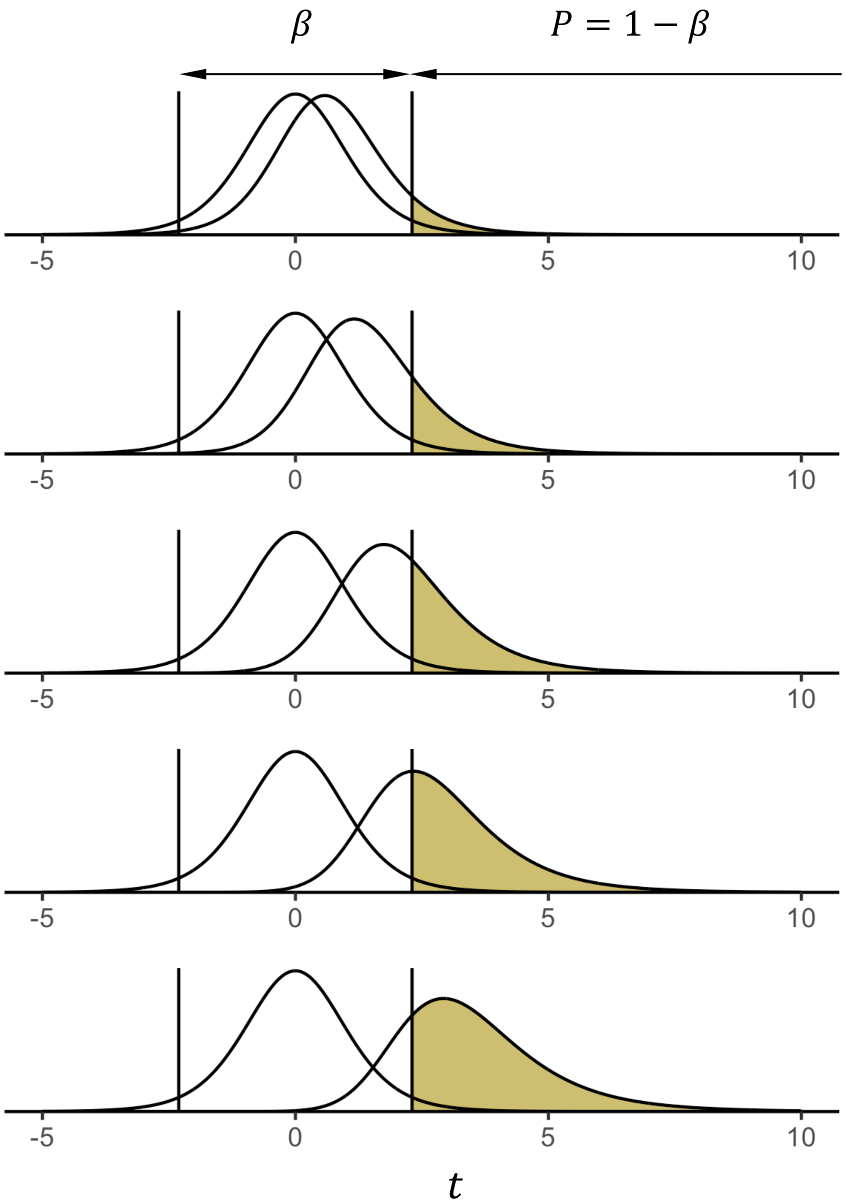
The probability of correctly rejecting the null hypothesis

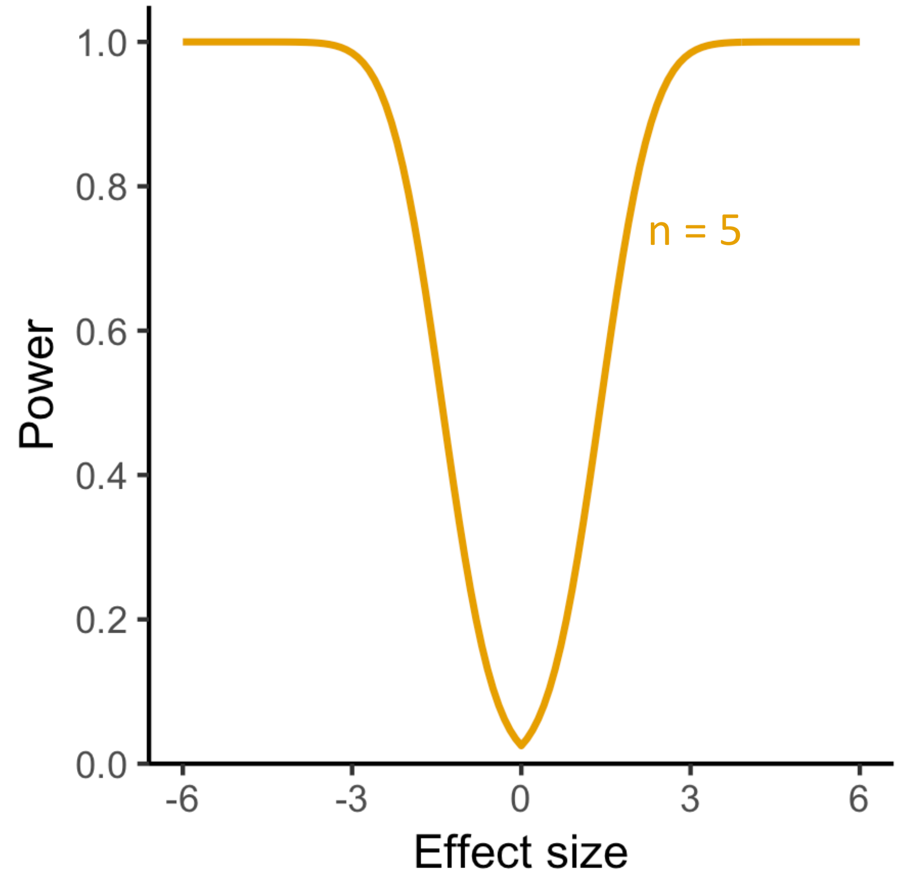
The probability of detecting an effect which is really there

Multiple alternative hypotheses



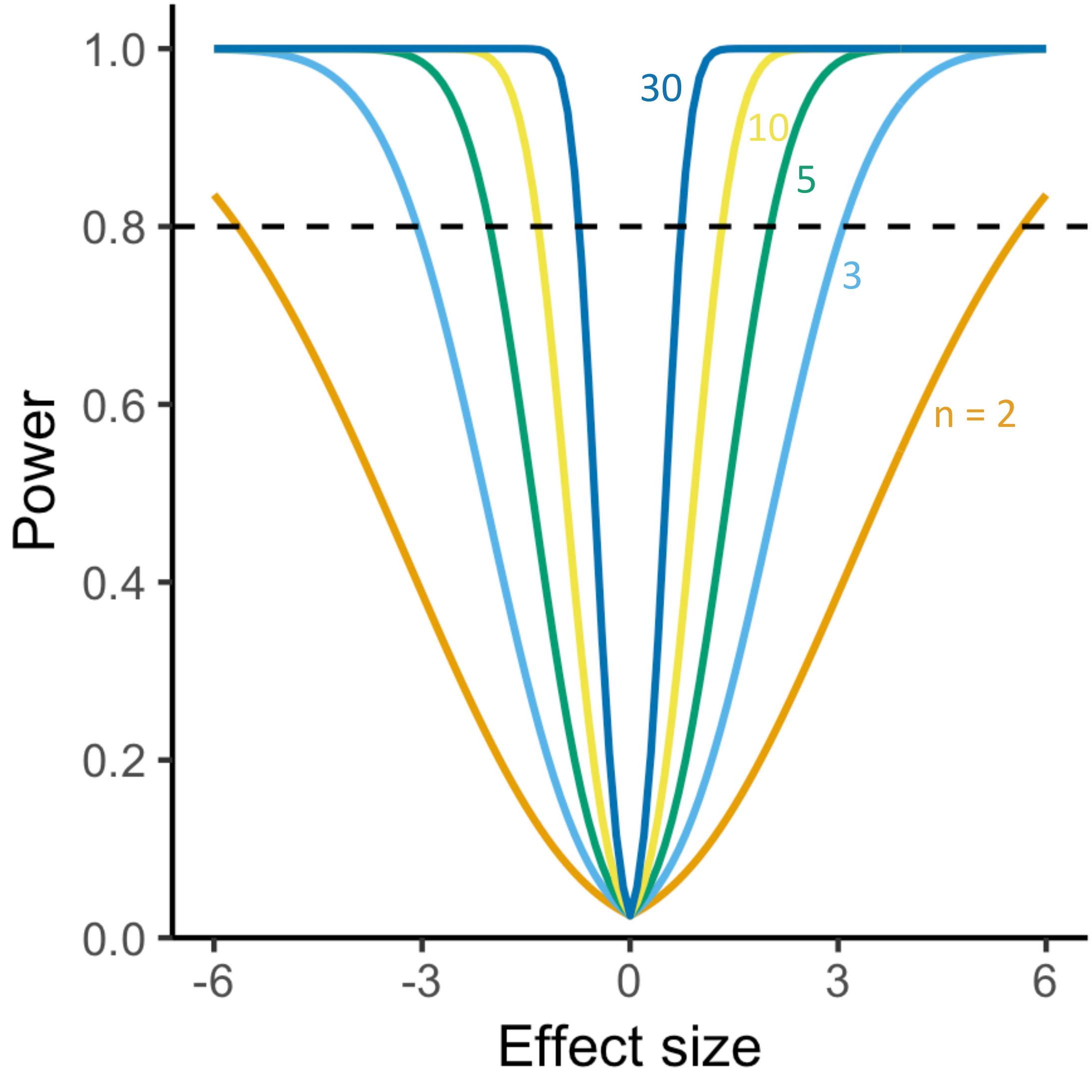
Power curve





Power: probability of detecting an effect when there is an effect

Power curves



$$d = \frac{M_1 - M_2}{SD}$$

How to do it in R?

```
> library(pwr)

# Find sample size required to detect the effect size d = 1, power = 0.8
> pwr.t.test(d=1, power=0.8, type="two.sample", alternative="two.sided")

     Two-sample t test power calculation

              n = 16.71472
              d = 1
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

# The same, but request power = 0.95
> pwr.t.test(d=1, power=0.95, type="two.sample", alternative="two.sided")

     Two-sample t test power calculation

              n = 26.9892
              d = 1
      sig.level = 0.05
          power = 0.95
    alternative = two.sided
```

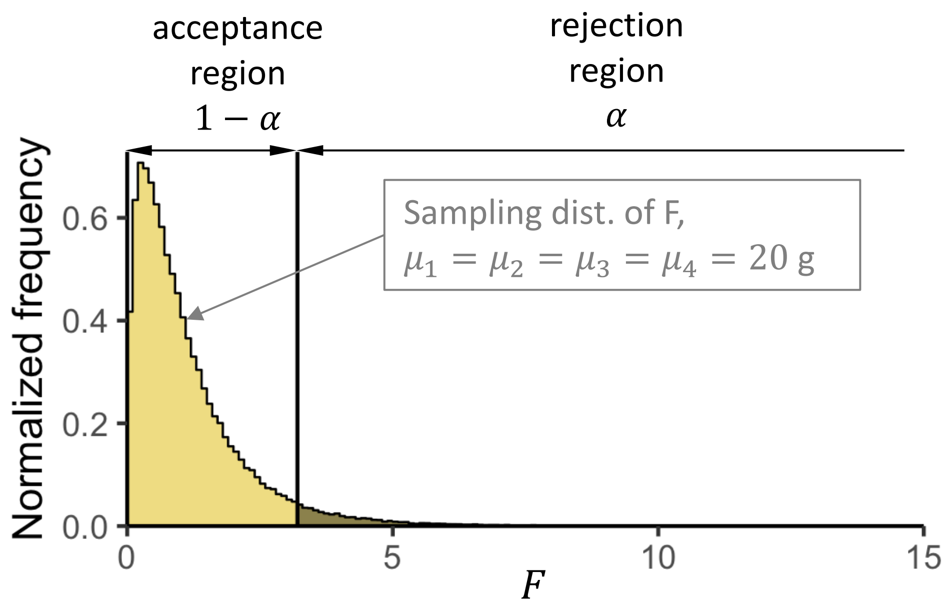
Statistical power

ANOVA

One alternative hypothesis

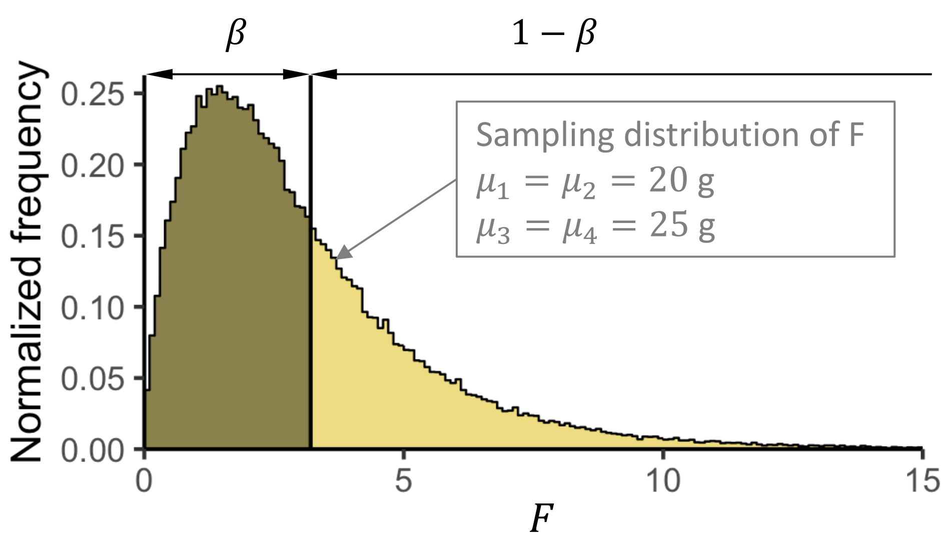
Null hypothesis

$\alpha = 0.05$

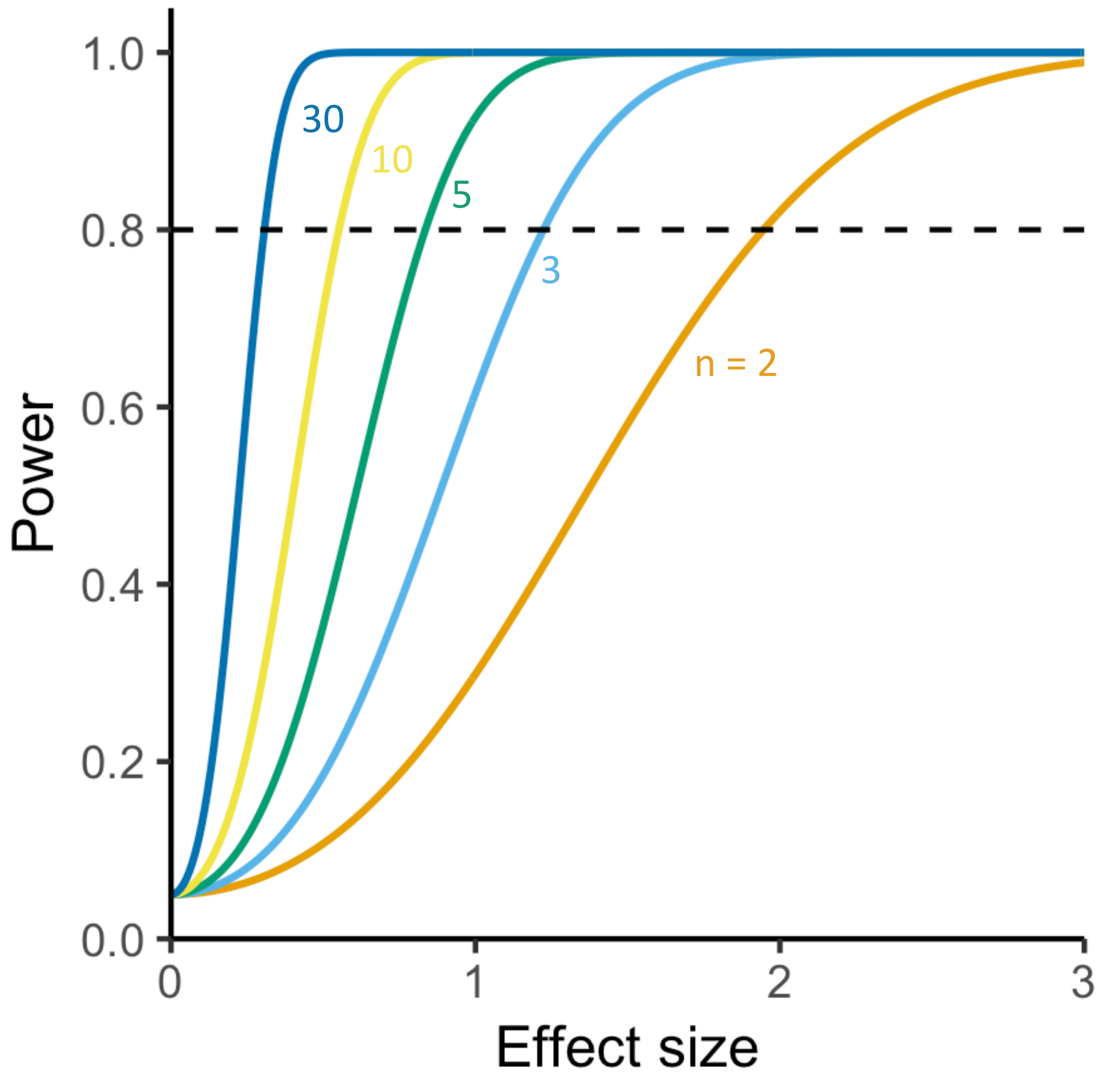


Alternative hypothesis

$\beta = 0.20$



Power curves



How to do it in R?

```
> library(pwr)

# Find sample size required to detect a “large” effect size f = 0.4
> pwr.anova.test(k=4, f=0.4, sig.level=0.05, power=0.8)

     Balanced one-way analysis of variance power calculation

              k = 4
              n = 18.04262
              f = 0.4
      sig.level = 0.05
          power = 0.8

NOTE: n is number in each group
```

Statistical power

Chi-square test

Significance of correlation

Power in chi-square test

```
> library(pwr)
# raw data
> d <- rbind(c(68,12), c(70,30))
# Chi-squared statistic
> chi2 <- chisq.test(d, correct = FALSE)$statistic
# Effect size
> w <- sqrt(chi2 / sum(d))
> w
[1] 0.1762268
# Power test
> pwr.chisq.test(w = w, df=(2-1)*(2-1), power=0.8)

     Chi squared power calculation

              w = 0.1762268
              N = 252.7333
             df = 1
      sig.level = 0.05
          power = 0.8

NOTE: N is the number of observations
```

	Dead	Alive	Total
Drug A	68	12	80
Drug B	70	30	100
Total	138	42	180

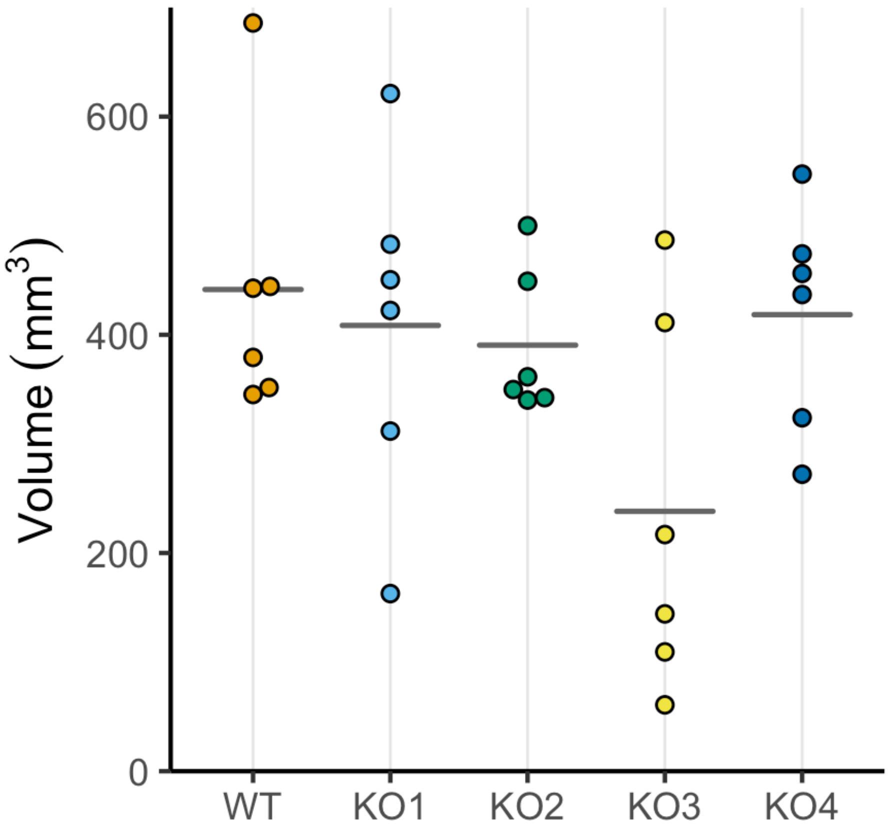
Power in correlation test

```
> library(pwr)
# Power test for correlation coefficient of 0.7
> pwr.r.test(r=0.7, power=0.8)

     approximate correlation power calculation (arctangh transformation)

              n = 12.81943
              r = 0.7
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

# Power test for correlation coefficient of 0.5
> pwr.r.test(r=0.5, power=0.8)

     approximate correlation power calculation (arctangh transformation)

              n = 28.24841
              r = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Worked example

Tumour growth in mice

Pilot experiment
WT and 4 KOs mice
Observe tumour growth
Measure volume after 10 days



Power analysis

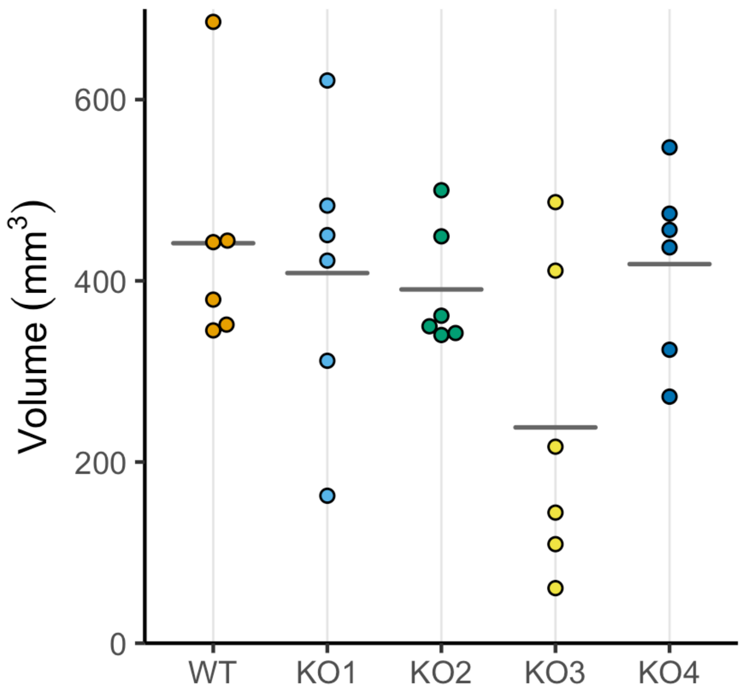
How many replicates do we need to...

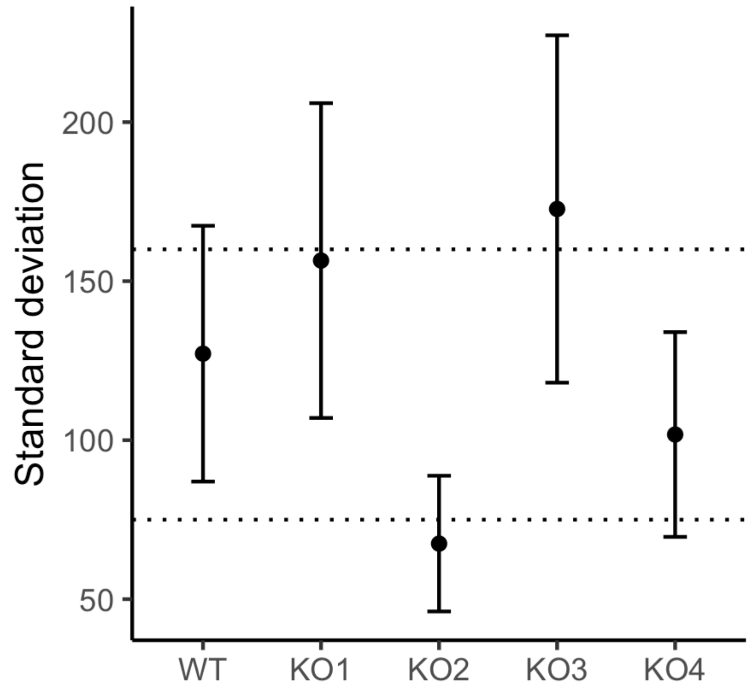
1) detect a 2-fold change between conditions? (power in t-test)

2) detect the observed effect in ANOVA? (power in ANOVA)

How many replicates to detect a 2-fold change between WT and a KO?

Estimate standard deviation





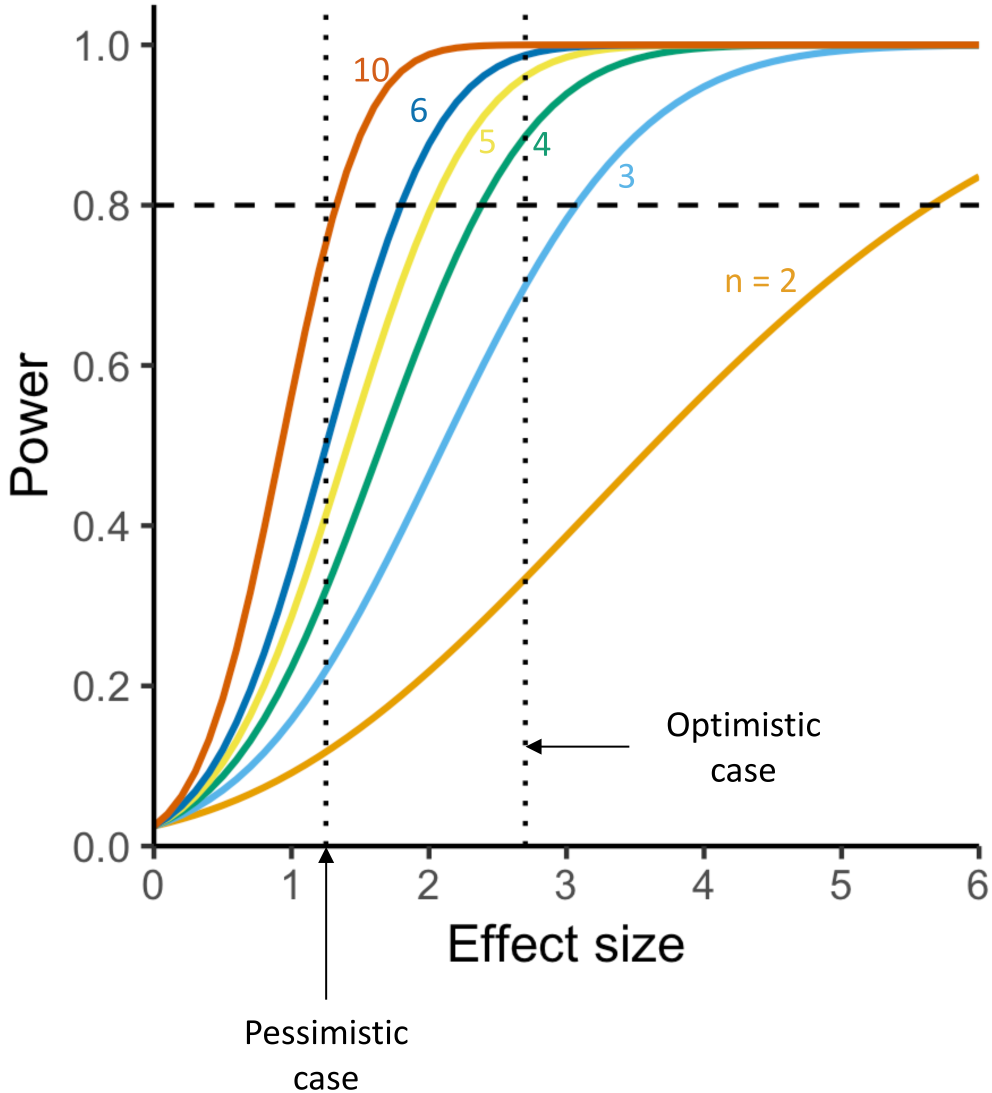
Pessimistic case
$SD = 160$

Optimistic case
$SD = 75$

Standard error of SD

$$SE_{SD} = \frac{SD}{\sqrt{2(n-1)}}$$

Two scenarios, $SD_1 = 75$ and $SD_2 = 160$



Cohen's d:

$$d_1 = \frac{\Delta M}{SD_1} = \frac{200}{75} = 2.7$$

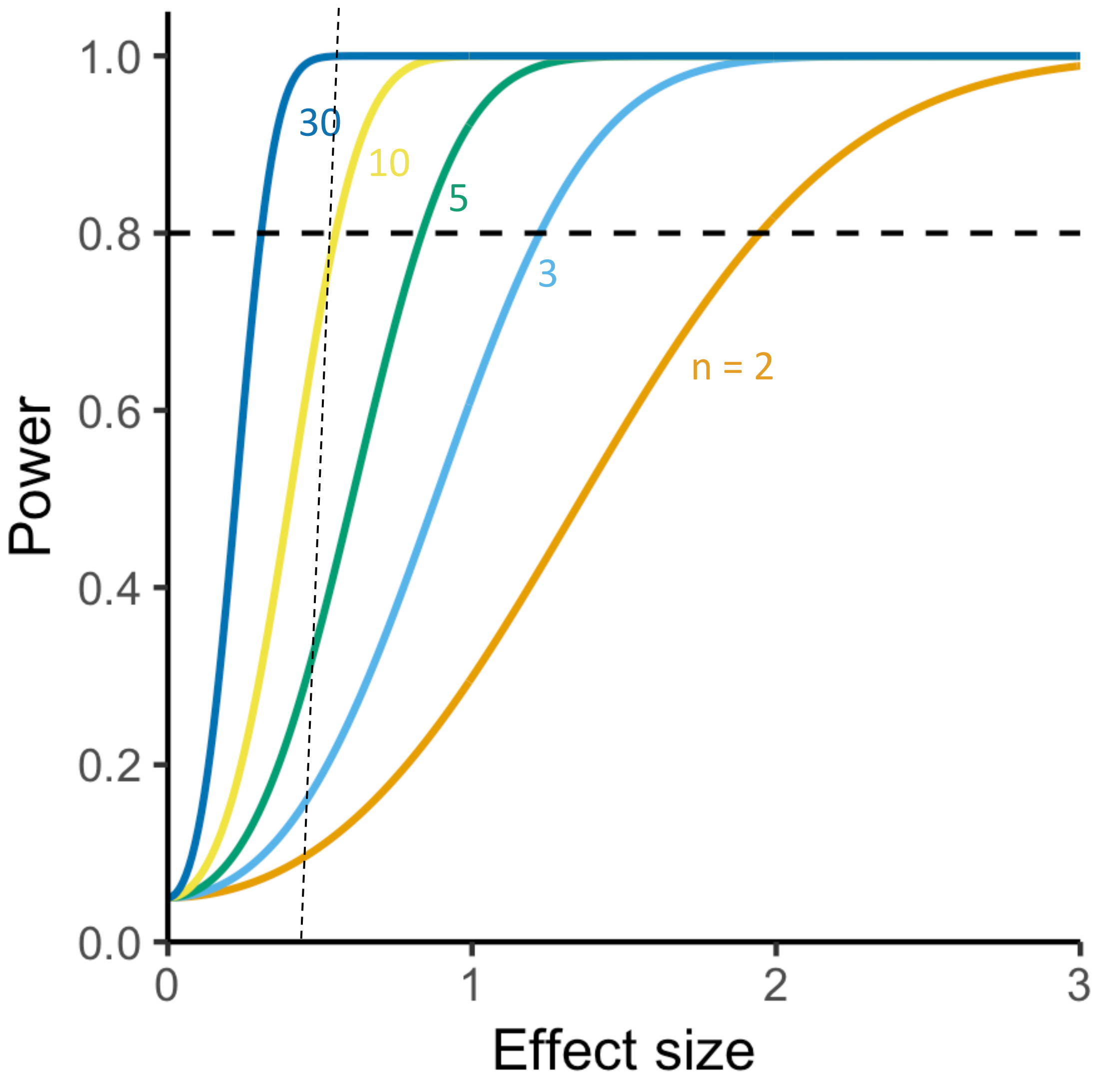
$$d_2 = \frac{\Delta M}{SD_2} = \frac{200}{160} = 1.25$$

R power calculations

```
> library(pwr)

# optimistic case
> pwr.t.test(d=2.7, power=0.8)

     Two-sample t test power calculation

              n = 3.435286
              d = 2.7
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group

# pessimistic case
> pwr.t.test(d=1.25, power=0.8)

     Two-sample t test power calculation

              n = 11.0942
              d = 1.25
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

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

How many replicates to detect the observed effect in ANOVA?

ANOVA power curves



From ANOVA on our data we have $F = 2.31$

Then, we find the observed effect size:

$$f = \sqrt{\frac{F-1}{n}} = 0.47$$

How many replicates do we need?

```
> library(pwr)
# Read data
> tumour <- read.table("http://tiny.cc/mouse_tumour", header=TRUE)
# Here n = 6 and k = 5
> tumour.lm <- lm(Volume ~ Group, data=tumour) # linear model
> tumour.av <- anova(tumour.lm) # perform ANOVA
> F <- tumour.av$`F value`[1] # extract statistic F
> f <- sqrt((F - 1)/6) # Effect size: Cohen's f
# What is the power of this experiment?
> pwr.anova.test(k=5, n=6, f=f)

              k = 5
              n = 6
              f = 0.4670469
      sig.level = 0.05
          power = 0.4293041
# How many replicates to get power of 0.8?
> pwr.anova.test(k=5, f=f, power=0.8)

              k = 5
              n = 11.93119
              f = 0.4670469
      sig.level = 0.05
          power = 0.8
```

Conclusions from our example

- Request power of 0.8

- To detect 2-fold change between WT and a KO in a pessimistic case we need 11 mice in each group

- To detect a change across all groups (ANOVA) we need 12 mice in each group

- We recommend an experiment with at least 12 mice in each group

Hand-outs available at
https://dag.compbio.dundee.ac.uk/training/Statistics_lectures.html