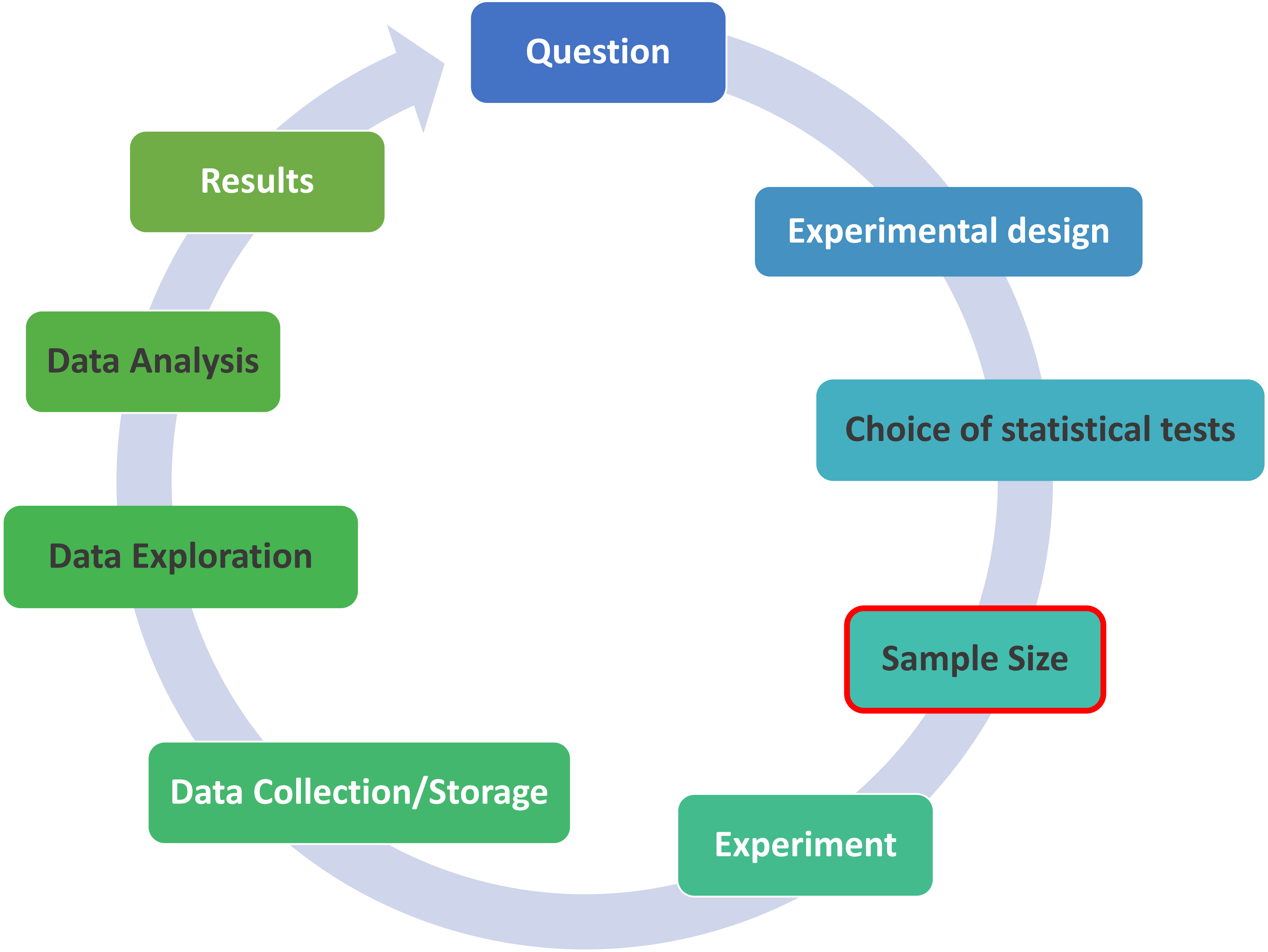
Power Analysis

Anne Segonds-Pichon

v2020-12

Babraham Institute

Babraham Bioinformatics

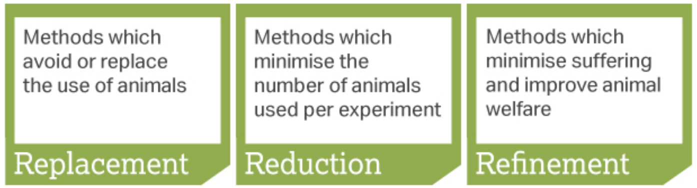

Question
Experimental design
Choice of statistical tests
Sample Size
Experiment
Data Collection/Storage
Data Exploration
Data Analysis
Results

Sample Size: Power Analysis

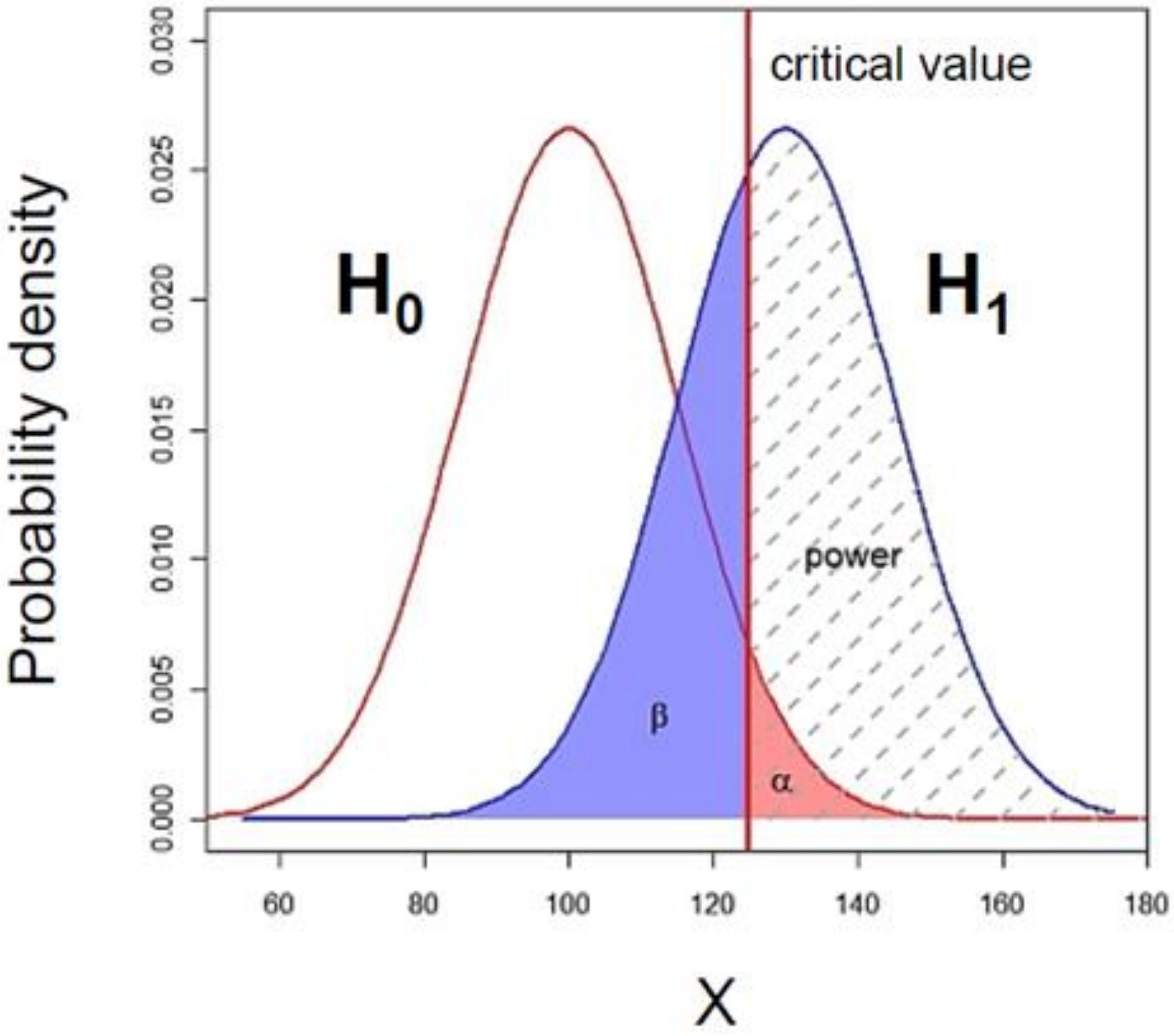
- **Definition of power**: probability that a statistical test will reject a false null hypothesis (H_0).
    - **Translation**: the probability of detecting an effect, given that the effect is really there.

- **In a nutshell**: the bigger the experiment (big sample size), the bigger the power (more likely to pick up a difference).

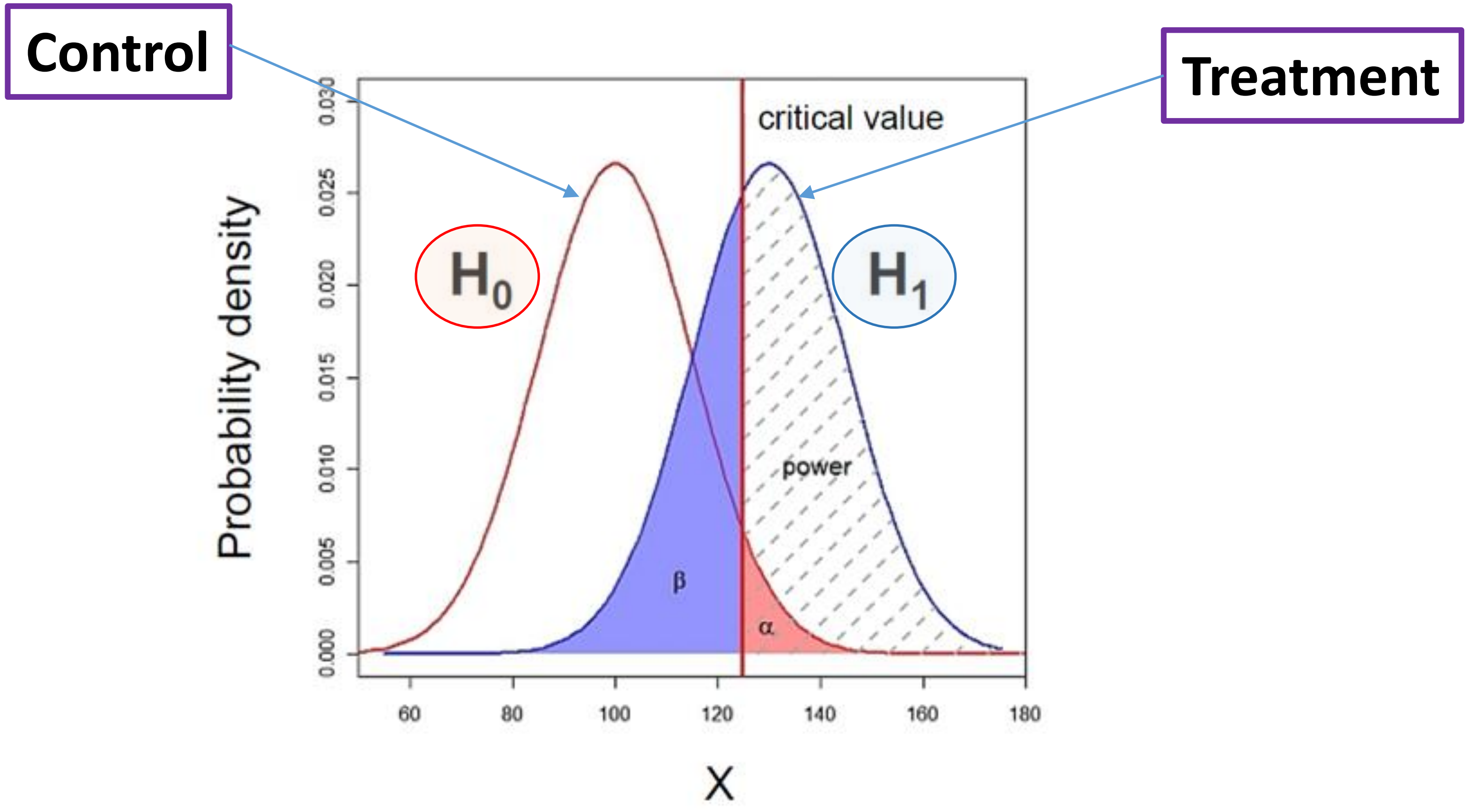
- Main output of a **power analysis**:
    - Estimation of an appropriate **sample size**
        - **Too big**: waste of resources,
        - **Too small**: may miss the effect ($p>0.05$)+ waste of resources,
        - **Grants**: justification of sample size,
        - **Publications:** reviewers ask for power calculation evidence,
        - **Home office**: the 3 Rs: Replacement, **Reduction** and Refinement.



What does Power look like?

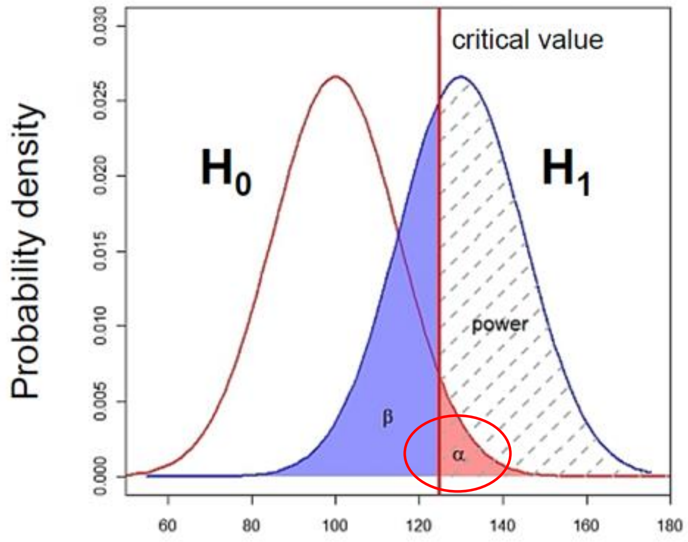


What does Power look like? Null and alternative hypotheses



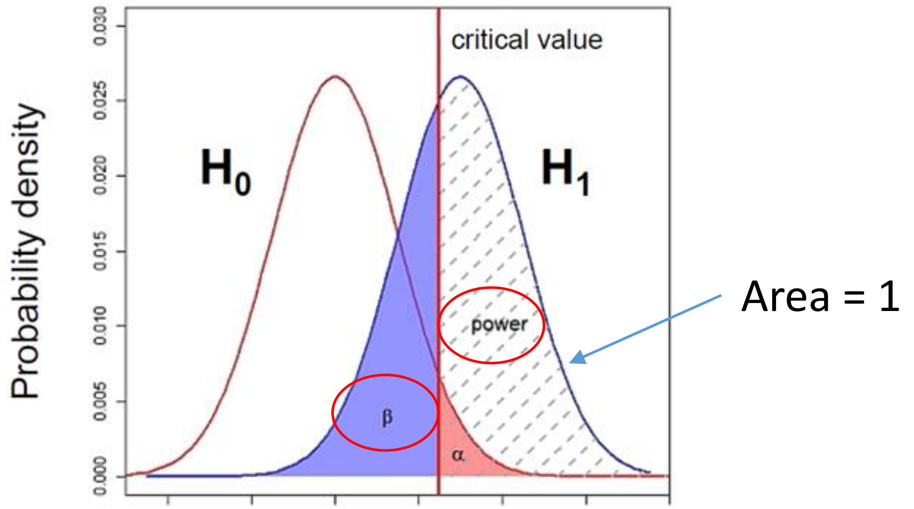
- Probability that the observed result occurs if $\mathbf{H_0}$ is true
  - H_0 : **Null hypothesis** = absence of effect
  - H_1: **Alternative hypothesis** = presence of an effect

What does Power look like? Type I error (α)



- **Type I error** is the failure to reject a <u>true</u> H_0
  - **α :** probability of claiming an effect which is not there.
- **p-value**: probability that the observed statistic occurred by chance alone
  - probability that a difference as big as the one observed could be found even if there is no effect.
- **Statistical significance**: comparison between **α** and the **p-value**
  - p-value < 0.05: there is a difference ☺ (reject H_0)
  - p-value > 0.05: there is no difference ☹ (fail to reject H_0)

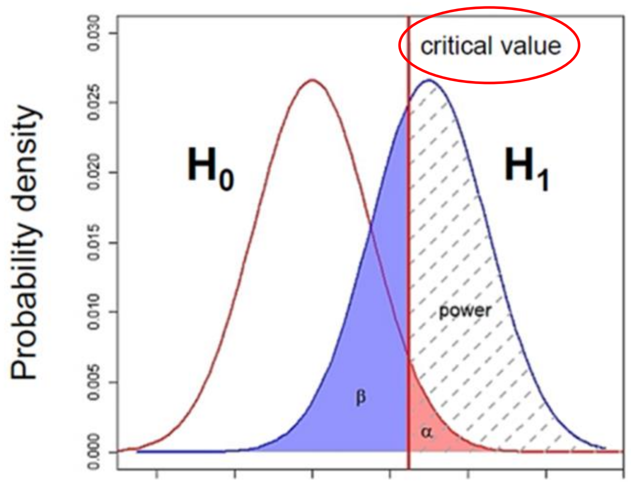
What does Power look like? Type II error (β) and Power



- **Type II error** (**β**) is the failure to reject a <u>false</u> H_0
  - **β :** Probability of missing an effect which is really there.
  - **Power**: probability of detecting an effect which is really there.

  - Direct relationship between **Power** and **type II error**:
    - **Power** = 1 – **β**

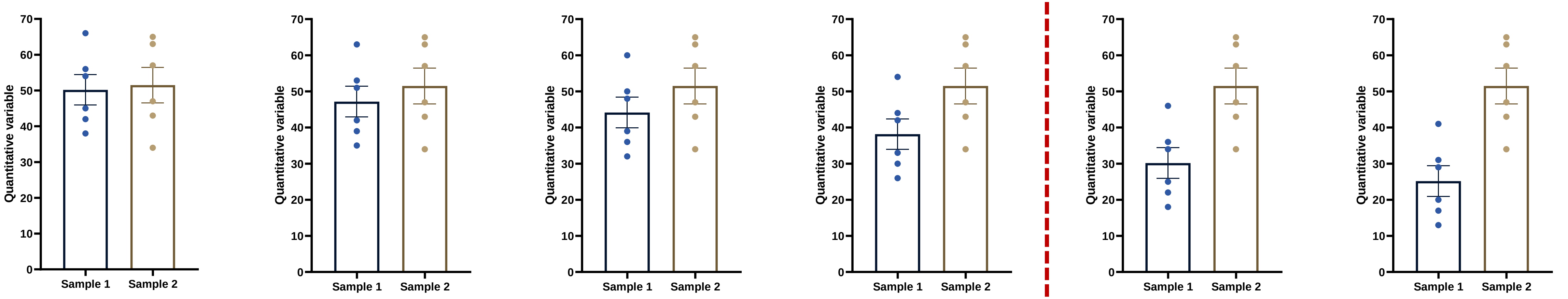
What does Power look like? Power = 80%



- **General convention: 80%** but could be more
  - if **Power** = 0.8 then β = 1- **Power** = 0.2 (20%)
- Hence a true difference will be missed 20% of the time
- Jacob Cohen (1962):
  - For most researchers: Type I errors are four times more serious than Type II errors so: **0.05 * 4 = 0.2**
    - Compromise: 2 groups comparisons:
      - 90% = +30% sample size
      - 95% = +60% sample size

The critical value

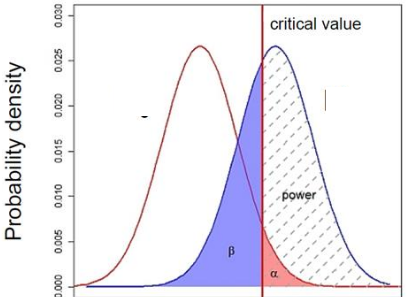
Critical value



Small difference → Big difference

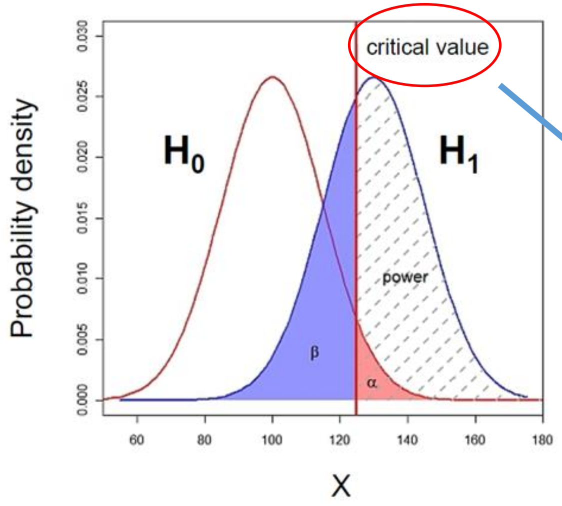
Not significant: $p>0.05$

Significant: $p<0.05$



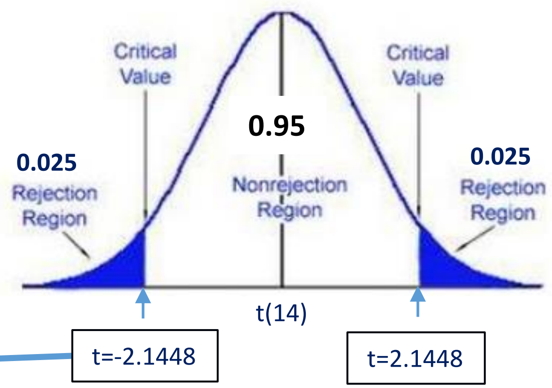
Critical value = size of difference + sample size + significance

What does Power look like? Example with the *t*-test



Example: 2-tailed t-test with n=15 (df=14)

df	0.20	0.10	0.05	0.02	0.01	0.001
1	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192
2	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240
4	1.5332	2.1318	2.7764	3.7469	4.6041	8.6103
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.9588
7	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079
8	1.3968	1.8595	2.3060	2.8965	3.3554	5.0413
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.5869
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.4370
12	1.3562	1.7823	2.1788	2.6810	3.0545	4.3178
13	1.3502	1.7709	2.1604	2.6503	3.0123	4.2208
14	1.3450	1.7613	2.1448	2.6245	2.9768	4.1405
15	1.3406	1.7531	2.1314	2.6025	2.9467	4.0728



- In **hypothesis testing**:
  - **test statistic** is compared to the **critical value** to determine significance
  - Example of test statistic: t-value

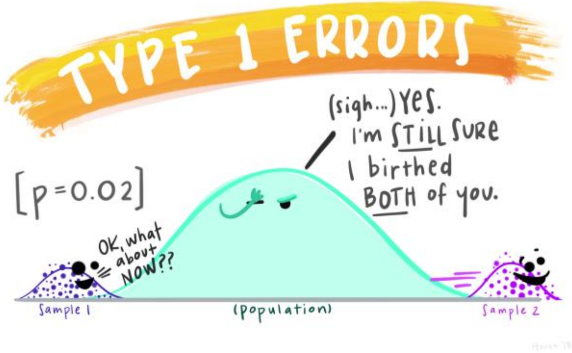
- If **test statistic** > **critical value**: statistical significance and rejection of the null **hypothesis**
  - Example: t-value > critical t-value

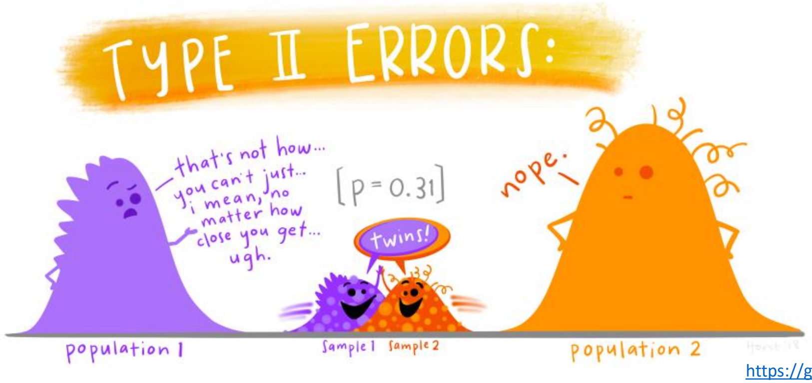
To recapitulate:

- The null hypothesis (H_0): H_0 = no effect
- The aim of a statistical test is to reject or not H_0.

Statistical decision	True state of H_0	
	H_0 True (no effect)	**H_0 False (effect)**
Reject H_0	**Type I error α** **False Positive**	Correct **True Positive**
Do not reject H_0	Correct **True Negative**	**Type II error β** **False Negative**

- **High specificity** = low **False Positives** = low **Type I error**
- **High sensitivity** = low **False Negatives** = low **Type II error**





https://github.com/allisonhorst/stats-illustrations#other-stats-artwork

Sample Size: Power Analysis

The power analysis depends on the relationship between 6 variables:

- the **difference** of biological interest
- the **variability** in the data (**standard deviation**)

Effect size

- the significance level (5%)
- the desired power of the experiment (80%)
- the **sample size**
- the alternative hypothesis (ie one or two-sided test)

The difference of biological interest

- This is to be determined scientifically, not statistically.
  - minimum meaningful effect of biological relevance
  - the larger the effect size, the smaller the experiment will need to be to detect it.
- **How to determine it?**
  - Previous research, pilot study …

The Standard Deviation (SD)

- Variability of the data
- **How to determine it?**
  - Data from previous research on WT or baseline …

The effect size: what is it?

- The **effect size**: Absolute difference + variability

- How to determine it?
  - Substantive knowledge
  - Previous research
  - Conventions

- **Jacob Cohen**
  - Defined small, medium and large effects for different tests

Test	Relevant effect size	Effect Size Threshold		
		Small	Medium	Large
t-test for means	d	0.2	0.5	0.8
F-test for ANOVA	f	0.1	0.25	0.4
t-test for correlation	r	0.1	0.3	0.5
Chi-square	w	0.1	0.3	0.5
2 proportions	h	0.2	0.5	0.8

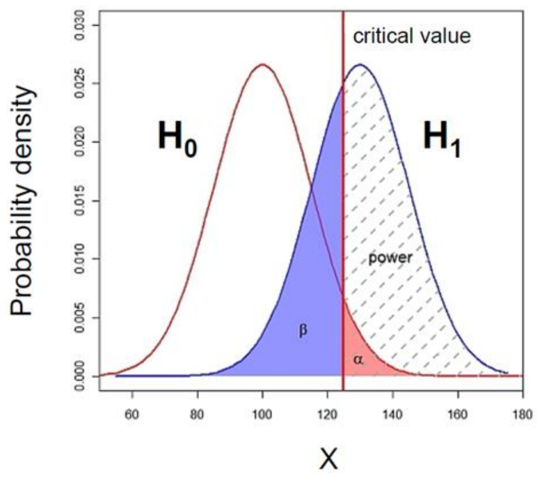
The effect size: how is it calculated?
The absolute difference

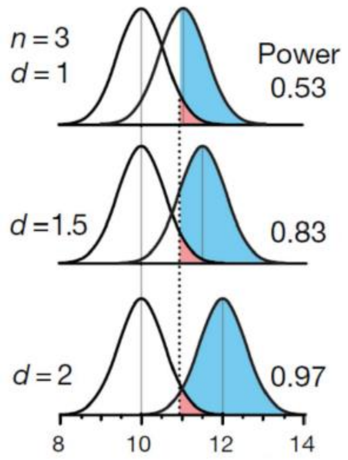
- It depends on the type of difference and the data
  - Easy example: comparison between 2 means

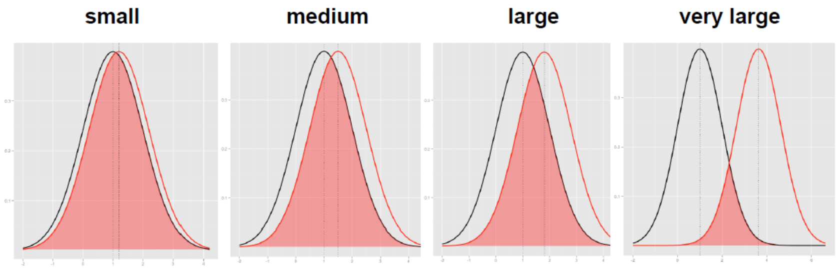
Absolute difference

$$\text{Effect Size} = \frac{[\text{Mean of experimental group}] - [\text{Mean of control group}]}{\text{Standard Deviation}}$$

- The bigger the effect (the absolute difference), the bigger the power
= the bigger the probability of picking up the difference





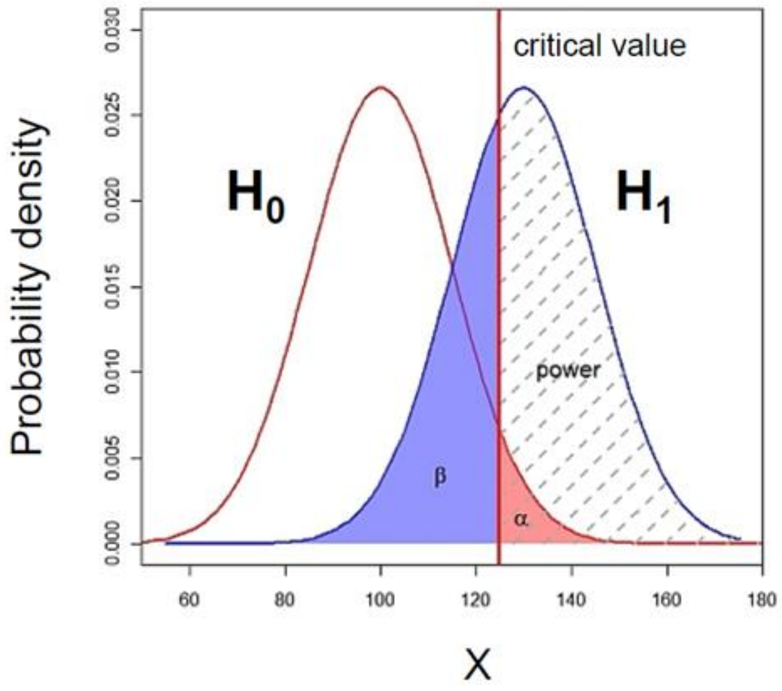


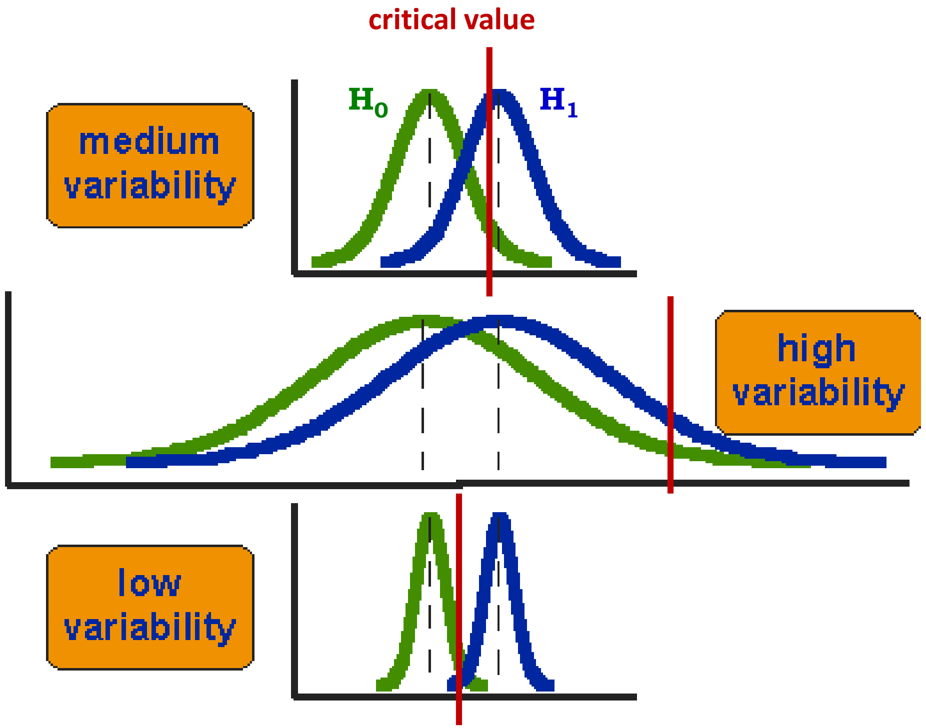
http://rpsychologist.com/d3/cohend/

The effect size: how is it calculated?
The standard deviation

- The bigger the variability of the data, the smaller the power

$$\text{Effect Size} = \frac{[\text{Mean of experimental group}] - [\text{Mean of control group}]}{\boxed{\text{Standard Deviation}}}$$





Power Analysis

The power analysis depends on the relationship between 6 variables:

- the **difference** of biological interest
- the **standard deviation**
- **the significance level (5%) (p< 0.05) α**
- **the desired power of the experiment (80%) β**
- the **sample size**
- the alternative hypothesis (ie one or two-sided test)

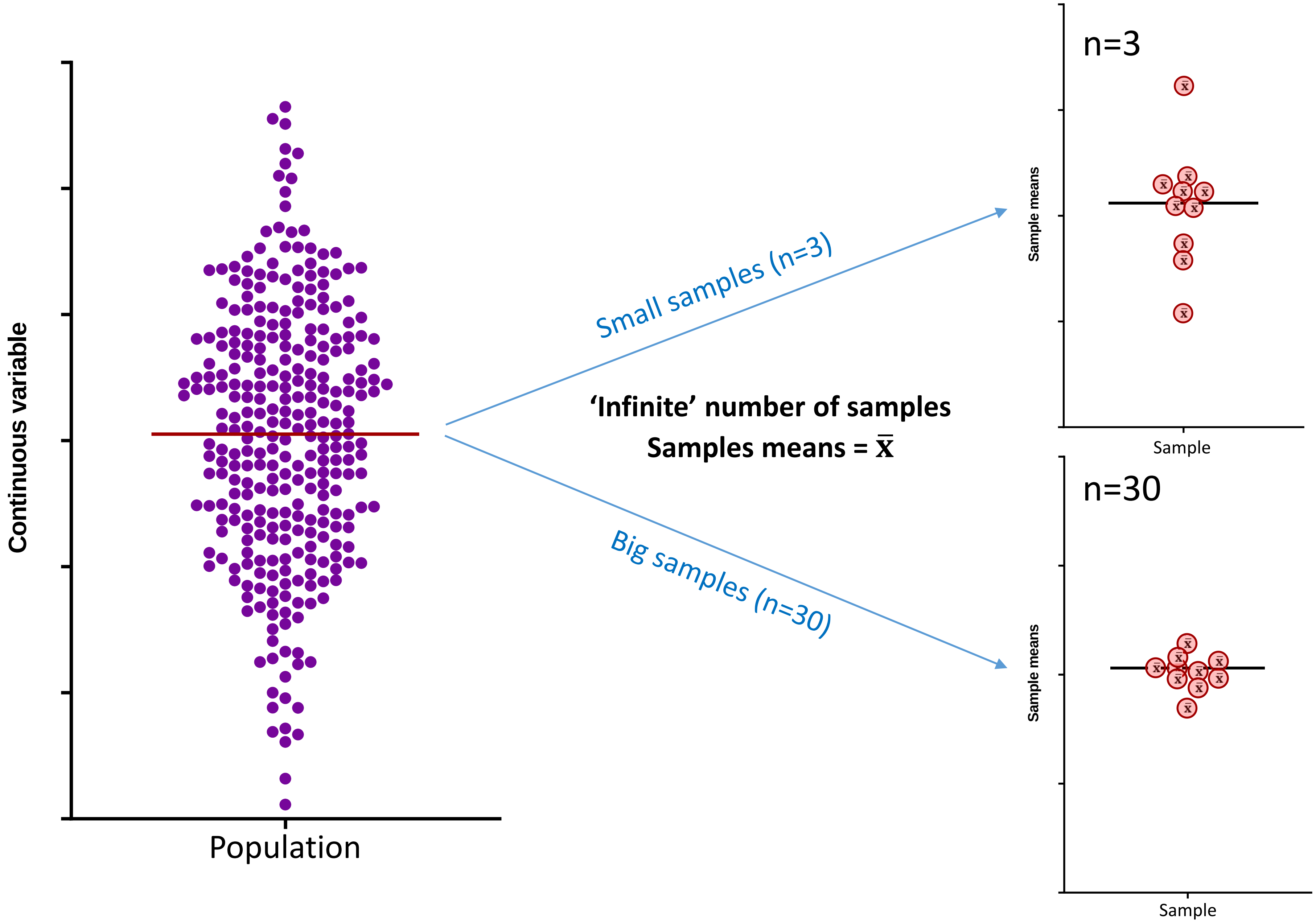
The sample size

- Most of the time, the output of a power calculation.

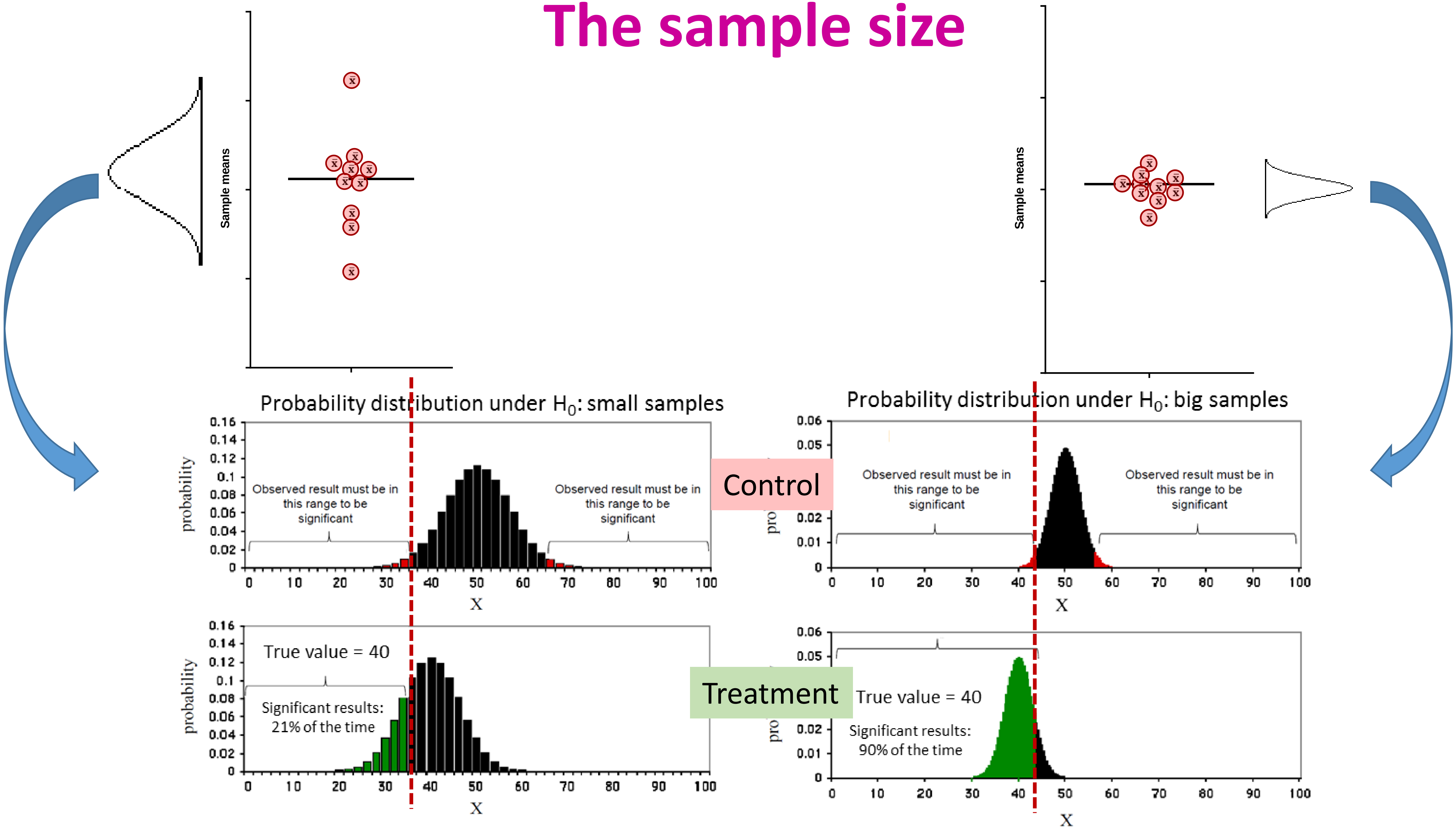
- **The bigger the sample, the bigger the power**
    - but how does it work actually?

- In reality it is difficult to reduce the variability in data, or the contrast between means,
    - most effective way of improving power:
        - **increase the sample size**.

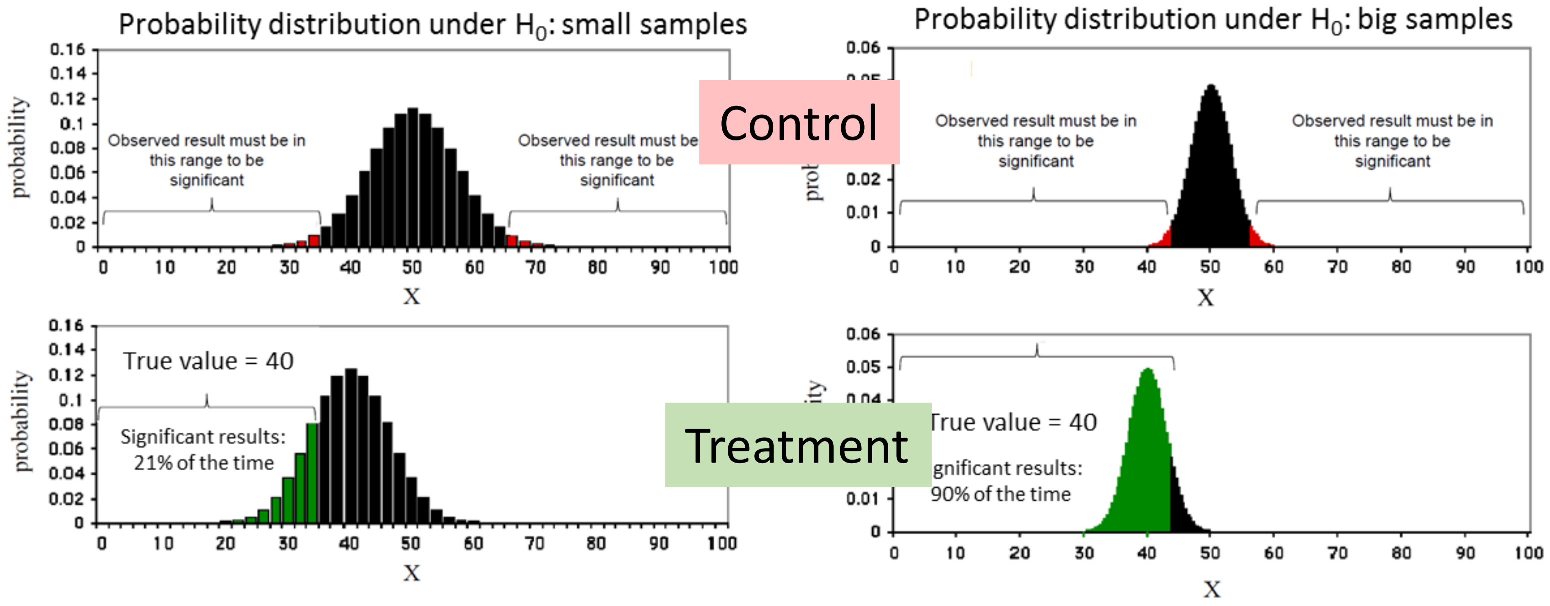
The sample size

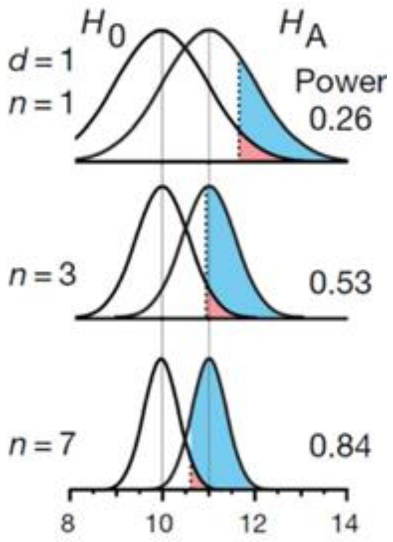


The sample size



The sample size



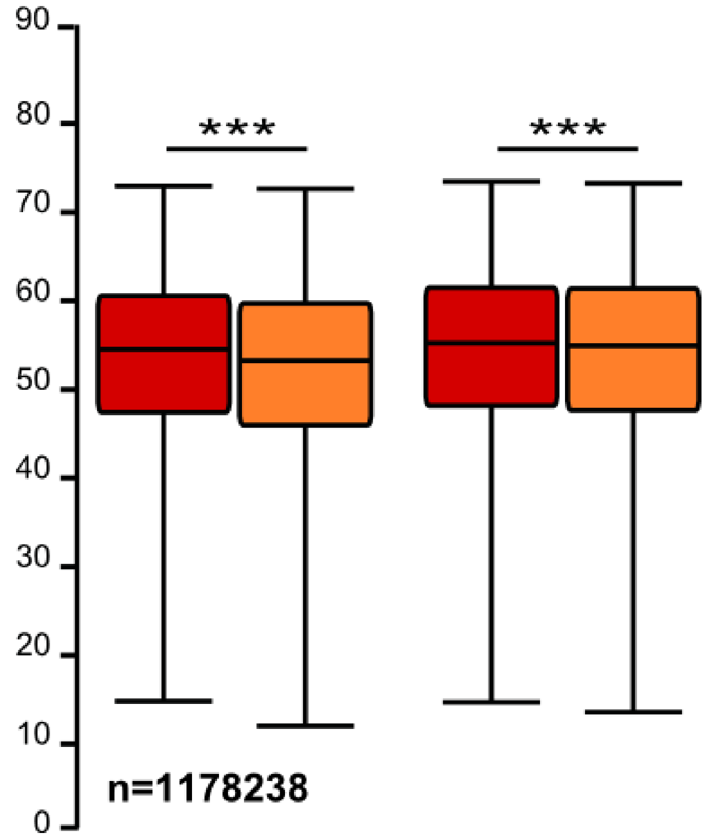


The sample size: the bigger the better?

- It takes huge samples to detect tiny differences but tiny samples to detect huge differences.

- What if the tiny difference is meaningless?
  - Beware of **overpower**
  - Nothing wrong with the stats: it is all about interpretation of the results of the test.

- Remember the important first step of power analysis
  - **What is the effect size of biological interest?**



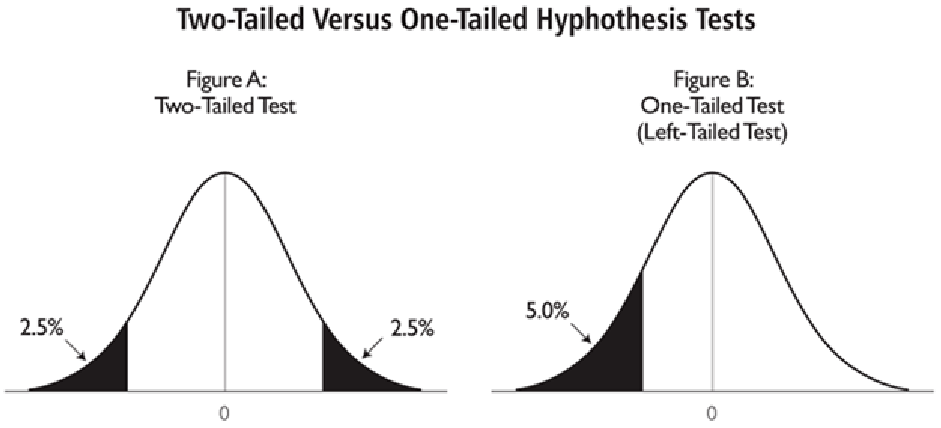
Power Analysis

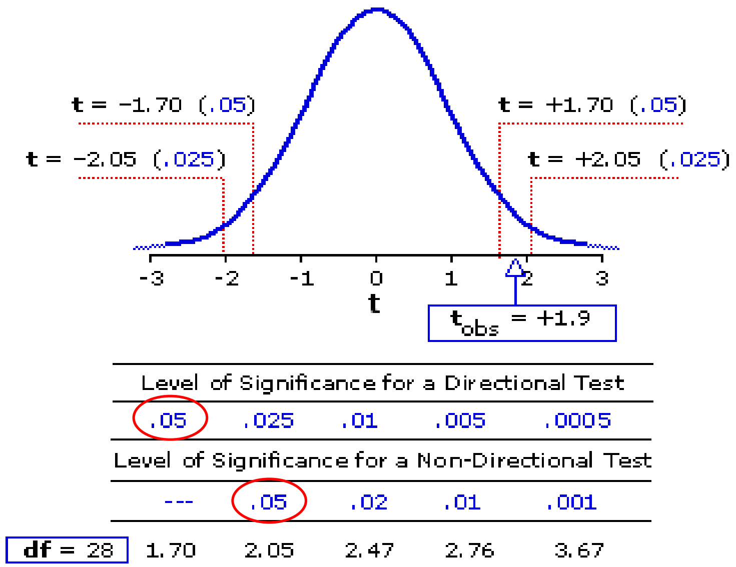
The power analysis depends on the relationship between 6 variables:

- the **effect size** of biological interest
- the **standard deviation**
- **the significance level (5%)**
- **the desired power of the experiment (80%)**
- the **sample size**
- the alternative hypothesis (ie one or two-sided test)

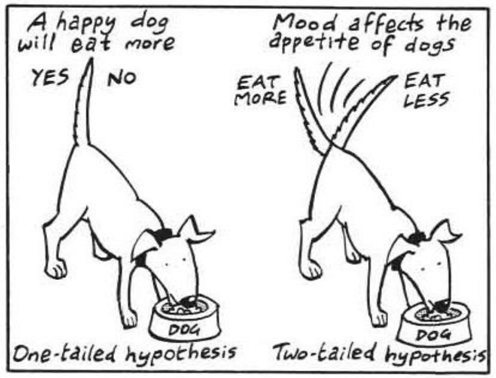
The alternative hypothesis: what is it?

- One-tailed or 2-tailed test? One-sided or 2-sided tests?





- Is the question:
  - Is the there a difference?
  - Is it bigger than or smaller than?



- Can rarely justify the use of a one-tailed test
- Two times easier to reach significance with a one-tailed than a two-tailed
  - Suspicious reviewer!

Hypothesis

↓

Experimental design
Choice of a Statistical test

↓

Power analysis

↓

Sample size

↓

Experiment(s)

↓

(Stat) analysis of the results

Power analysis

- Fix any **five of the variables**, a mathematical relationship is used to estimate the **sixth**

Difference of biological interest

\+ Variability in the data (standard deviation)

\+ Desired power of the experiment (80%)

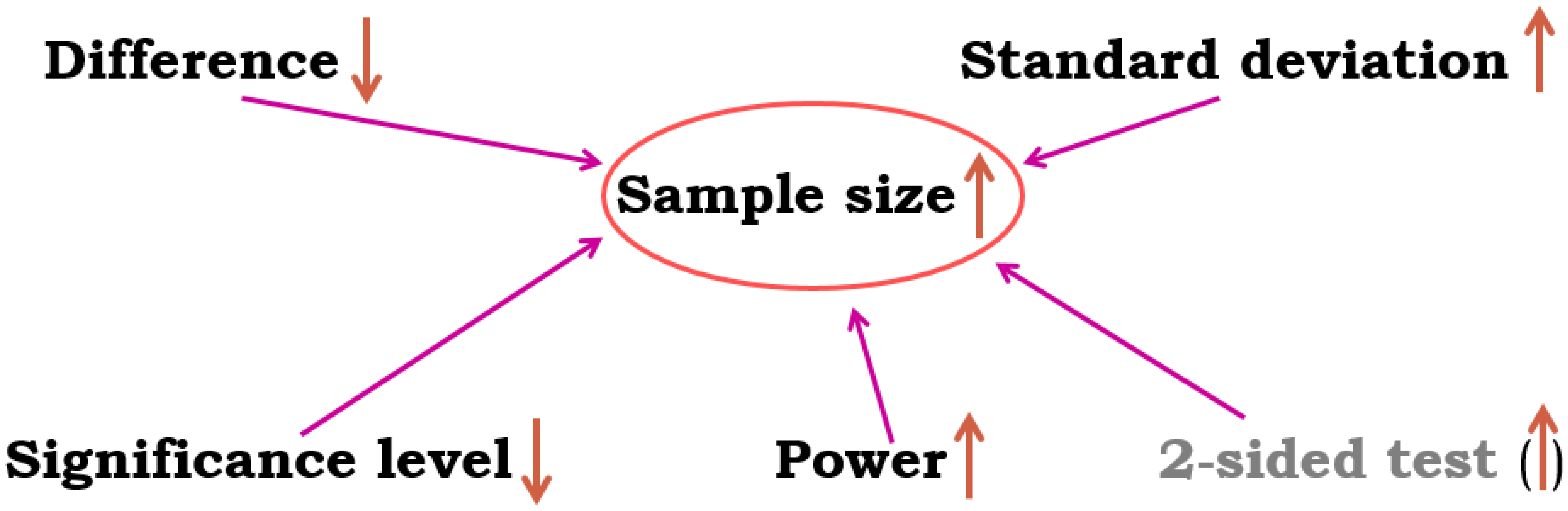
\+ Significance level (5%)

\+ Alternative hypothesis (ie one or two-sided test)

Appropriate sample size

- **Fix any five of the variables and a mathematical relationship can be used to estimate the sixth.**

e.g. What sample size do I need to have a 80% probability (**power**) to detect this particular effect (**difference** and **standard deviation**) at a 5% **significance level** using a **2-sided test**?



- **Good news**:

there are packages that can do the power analysis for you ... providing you have some prior knowledge of the key parameters!

difference + standard deviation = effect size

- **Free packages**:
  - R
  - **G*Power** and InVivoStat
  - Russ Lenth's power and sample-size page:
    - http://www.divms.uiowa.edu/~rlenth/Power/

- Cheap package: StatMate (~ $95)

- Not so cheap package: MedCalc (~ $495)

Power Analysis
Let’s do it

- **Examples of power calculations**:
  - Comparing 2 proportions: **Exercise 1**
  - Comparing 2 means: **Exercise 2**

Exercises 1 and 2

- Use the functions below to answer the exercises
  - Clue: exactly one of the parameters must be passed as NULL, and that parameter is determined from the others.
- Use **R Help** to find out how to use the functions
  - e.g. `?power.prop.test` in the console

Exercise 1

```
power.prop.test(n=NULL, p1=NULL, p2=NULL,
sig.level=NULL, power=NULL, alternative=c("two.sided", "one.sided"))
```

Exercise 2

```
power.t.test(n=NULL, delta=NULL, sd=1, sig.level=NULL, power=NULL,
type=c("two.sample", "one.sample", "paired"),
alternative=c("two.sided", "one.sided"))
```

Exercise 1:

- Scientists have come up with a solution that will reduce the number of lions being shot by farmers in Africa: painting eyes on cows' bottoms.
- Early trials suggest that lions are less likely to attack livestock when they think they're being watched
  - Fewer livestock attacks could help farmers and lions co-exist more peacefully.

- Pilot study over 6 weeks:
  - 3 out of 39 unpainted cows were killed by lions, none of the 23 painted cows from the same herd were killed.

- **Questions**:
  - Do you think the observed effect is meaningful to the extent that such a 'treatment' should be applied? Consider ethics, economics, conservation …
  - Run a power calculation to find out how many cows should be included in the study.
    - Clue 1: `power.prop.test()`
    - Clue 2: exactly one of the parameters must be passed as NULL, and that parameter is determined from the others.

http://www.sciencealert.com/scientists-are-painting-eyes-on-cows-butts-to-stop-lions-getting-shot

Exercise 1: *Answer*

- Scientists have come up with a solution that will reduce the number of lions being shot by farmers in Africa:
  - Painting eyes on the butts of cows
- Early trials suggest that lions are less likely to attack livestock when they think they're being watched
  - Less livestock attacks could help farmers and lions co-exist more peacefully.

- Pilot study over 6 weeks:
  - 3 out of 39 unpainted cows were killed by lions, none of the 23 painted cows from the same herd were killed.

```
power.prop.test(p1 = 3/39, p2 = 0, sig.level = 0.05, power = 0.8, alternative="two.sided")
```

```
     Two-sample comparison of proportions power calculation 

              n = 96.92364
             p1 = 0.07692308
             p2 = 0
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

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

Exercise 2:

- Pilot study: 10 arachnophobes were asked to perform 2 tasks:

Task 1: Group1 (n=5): to play with a big hairy tarantula spider with big fangs and an evil look in its eight eyes.
Task 2: Group 2 (n=5): to look at pictures of the same hairy tarantula.

- Anxiety scores were measured for each group (0 to 100).

Picture	Real Spider
25	45
35	40
45	55
40	55
50	65

- Use R to calculate the values for a power calculation
  - Data: **spider.data.csv**
  - Hint: you can use `group_by()` and `summarise()`
  - Or you can do it in Excel!
- Run a power calculation (assume balanced design and parametric test)
  - Clue 1: `power.t.test()`
  - Clue 2: choose the sd that makes more sense.

Exercise 2: *Answer*

```
spider.data %>%
  group_by(Group) %>%
    summarise(mean=mean(Scores), sd=sd(Scores))
```

Group <chr>	mean <dbl>	sd <dbl>
Picture	39	9.617692
Real	52	9.746794

2 rows

```
power.t.test(delta = 52 - 39, sd = 9.75, sig.level = 0.05, power = 0.8,
type = "two.sample", alternative = "two.sided")
```

```
     Two-sample t test power calculation

              n = 9.889068
          delta = 13
             sd = 9.75
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

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

- To reach significance with a t-test, providing the preliminary results are to be trusted,

and be confident in a difference between the 2 groups, we need about **10 arachnophobes** in each group.

Unequal sample sizes

- Scientists often deal with unequal sample sizes
  - No simple trade-off:
    - if one needs 2 groups of 30, going for 20 and 40 will be associated with decreased power.
  - **Unbalanced design = bigger total sample**
  - Solution:

    Step 1: power calculation for equal sample size

    Step 2: adjustment

$$N = \frac{2n(1+k)^2}{4k}$$

$$n_1 = \frac{N}{(1+k)}$$

$$n_2 = \frac{kN}{(1+k)}$$

- Cow example: balanced design: **n = 97**

but this time: unpainted group: 2 times bigger than painted one (k=2):

- Using the formula, we get a total:

$N=2*97*(1+2)^2/4*2 = 219$

Painted butts **(n_1)=73** Unpainted butts **(n_2)=146**

- Balanced design: **n = 2*97 = 194**
- Unbalanced design: **n= 70+140 = 219**

Non-parametric tests

- **Non-parametric tests**: do not assume data come from a Gaussian distribution.
  - Non-parametric tests are based on **ranking values** from low to high
  - Non-parametric tests almost always **less powerful**
- Proper power calculation for non-parametric tests:
  - Need to specify which **kind of distribution** we are dealing with
    - Not always easy
- Non-parametric tests never require more than 15% additional subjects providing that the distribution is not too unusual.
- **Very crude rule of thumb for non-parametric tests**:
  - Compute the sample size required for a parametric test and **add 15%.**

Sample Size: Power Analysis

- What happens if we ignore the power of a test?
  - Misinterpretation of the results

- p-values: never ever interpreted without context:
  - **Significant p-value (<0.05)**: exciting! Wait: what is the difference?
    - >= smallest meaningful difference: exciting
    - < smallest meaningful difference: not exciting
      - very big sample, too much power

  - **Not significant p-value (>0.05)**: no effect! Wait: how big was the sample?
    - Big enough = enough power: no effect means no effect
    - Not big enough = not enough power
      - Possible meaningful difference but we miss it