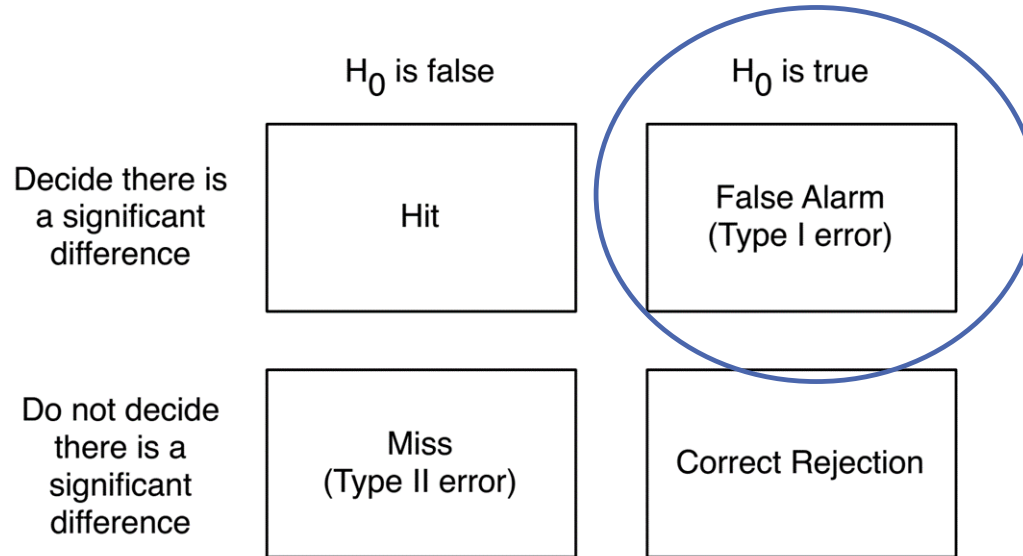


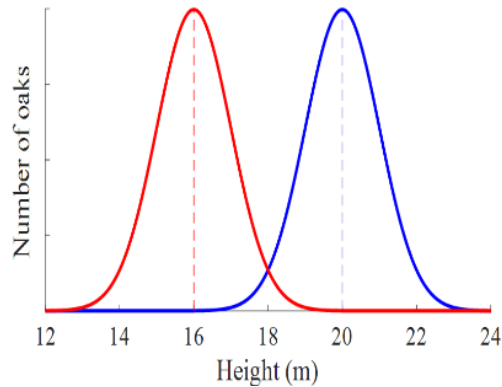
UNDERSTANDING STATISTICS & EXPERIMENTAL DESIGN

1. Basic Probability Theory
2. Signal Detection Theory (SDT)
3. SDT and Statistics I and II
4. Statistics in a nutshell
5. Multiple Testing
6. ANOVA
7. Experimental Design & Statistics
8. Correlations & PCA
9. Meta-Statistics: Basics
10. Meta-Statistics: Too good to be true
11. Meta-Statistics: How big a problem is publication bias?
12. Meta-Statistics: What do we do now?

Power & Power analysis (book ch.7)



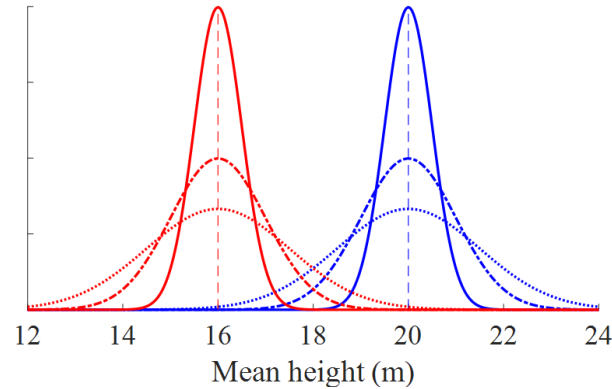
Population distributions



$$t = \frac{(\bar{x}_{North} - \bar{x}_{South})}{\frac{s}{\sqrt{n/2}}}$$

Sampling distributions

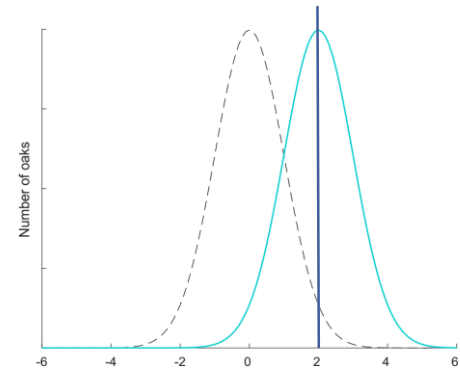
$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$



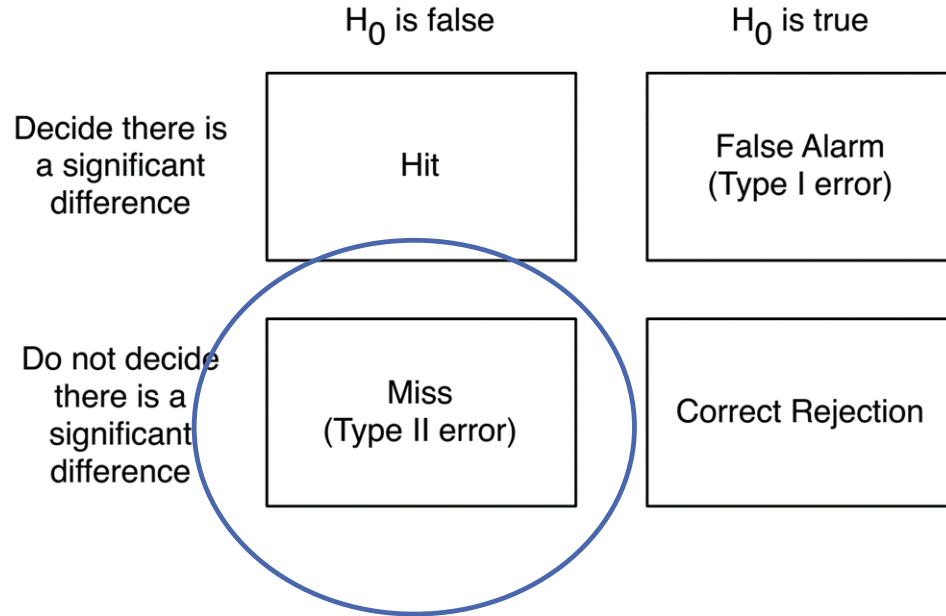
$$t = d * \sqrt{n/2}$$

Null + Difference distribution

$$s_{\bar{x}_{North} - \bar{x}_{South}} = s \sqrt{\frac{2}{n}}$$



$$p = (data|H_0)$$



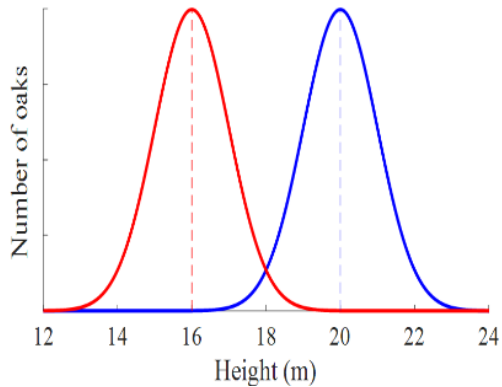
Implication IId: The Null and the Alternative hypothesis

Null hypothesis is true



$$t = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}} * \sqrt{n}$$

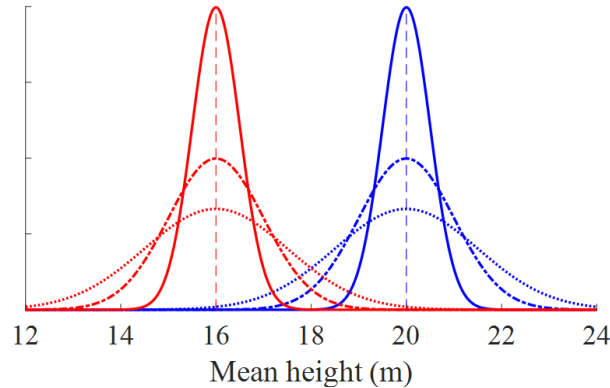
Population distributions



$$t = \frac{(\bar{x}_{North} - \bar{x}_{South})}{\frac{s}{\sqrt{n/2}}}$$

Sampling distributions

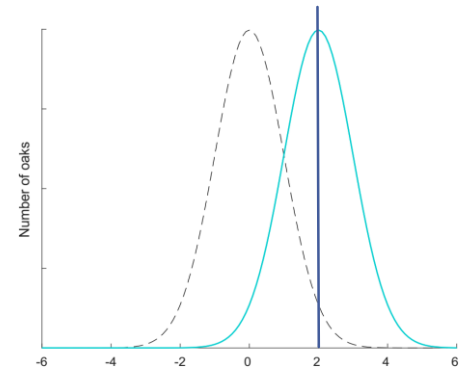
$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$



$$t = d * \sqrt{n/2}$$

Null + Difference distribution

$$s_{\bar{x}_{North} - \bar{x}_{South}} = s \sqrt{\frac{2}{n}}$$



$$p = (data|H_0)$$

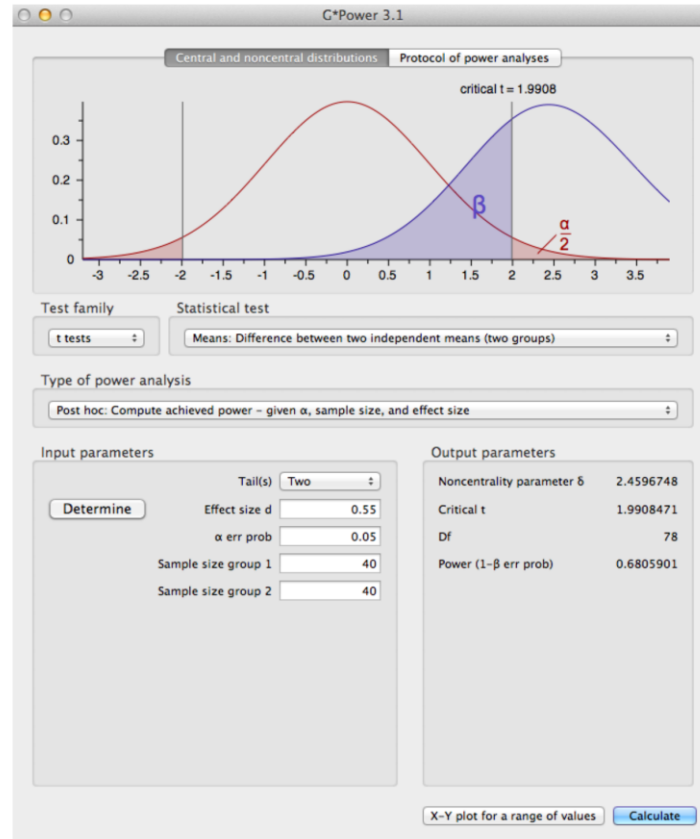
1. The sample size
2. The level of statistical significance required
3. The minimum size of effect that it is reasonable to expect
4. The test you are using

Acceptable risk of a Type II error is often set at 1 in 5, i.e., a probability of 0.2 (β).

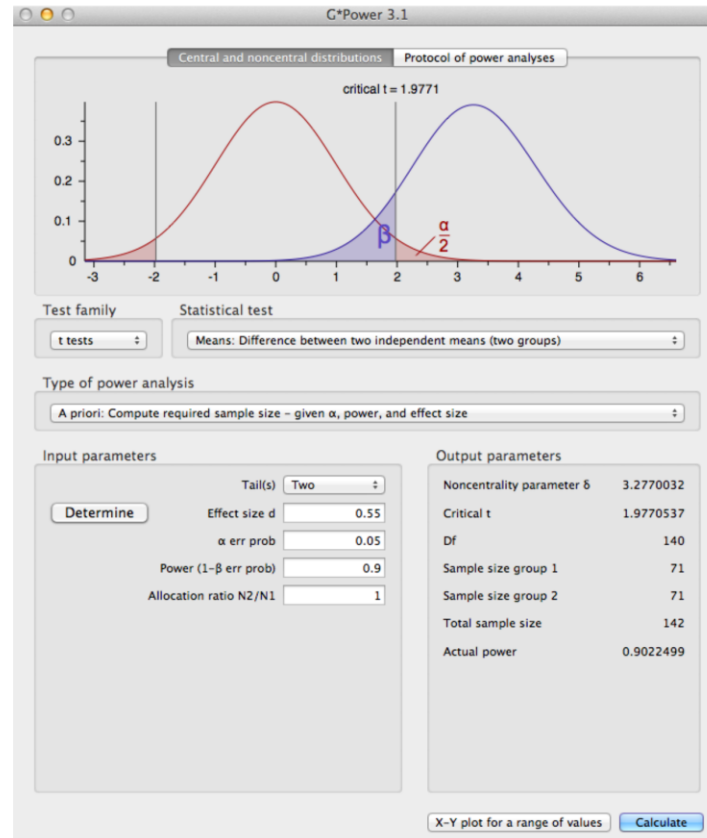
The conventionally uncontroversial value for “adequate” statistical power is therefore set at $1 - 0.2 = 0.8$.

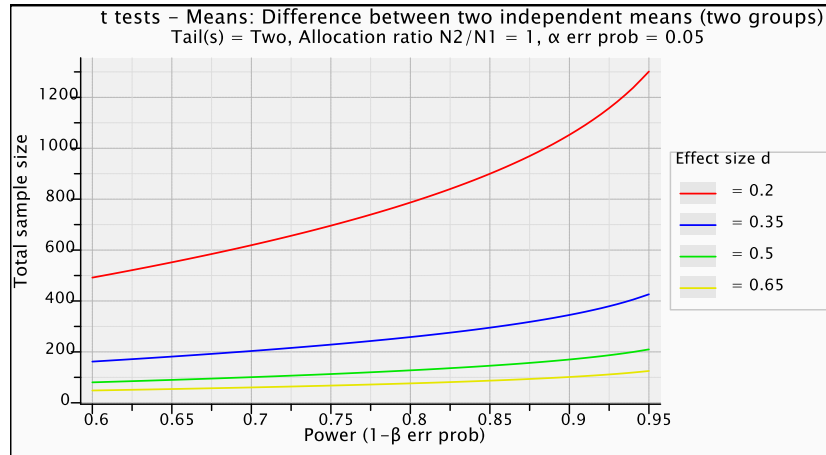
A *retrospective power analysis* is used in order to know whether the studies you are interpreting were well enough designed.

If you fail to reject the null hypothesis you might want to know what chance you had of finding a significant result – defending the failure

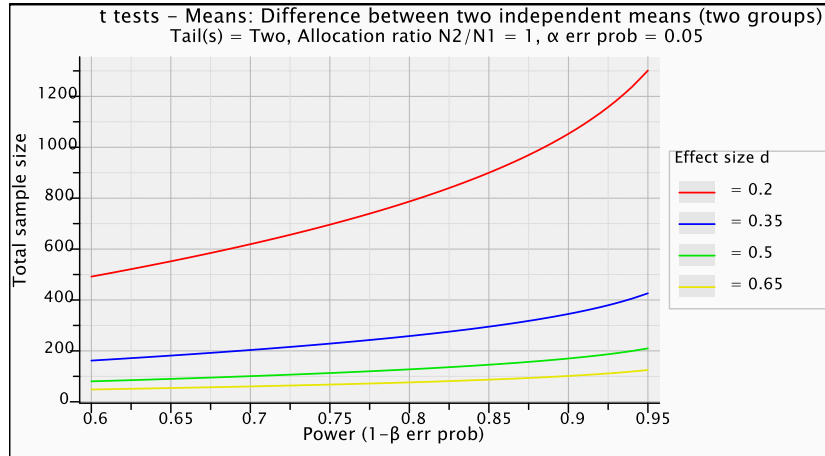


A *prospective power analysis* is used before collecting data, to consider *design sensitivity*





Why are there so many significant results with so low power, both low n and d' ?



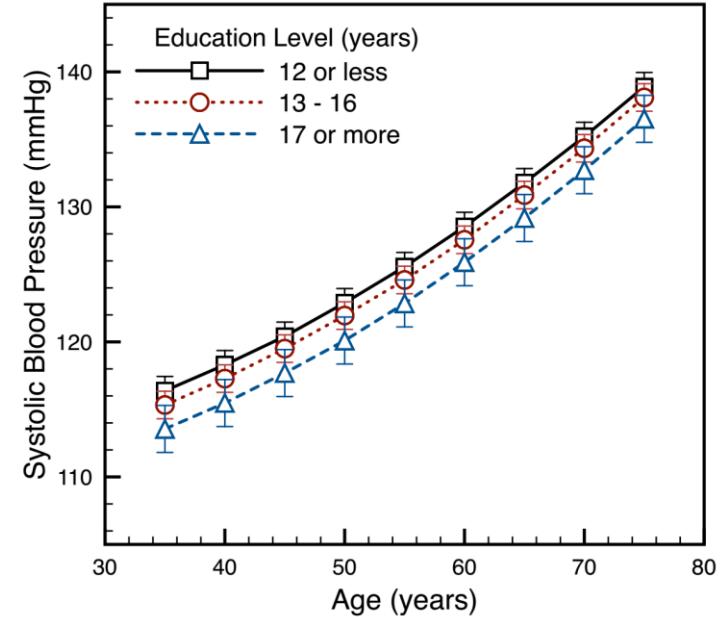
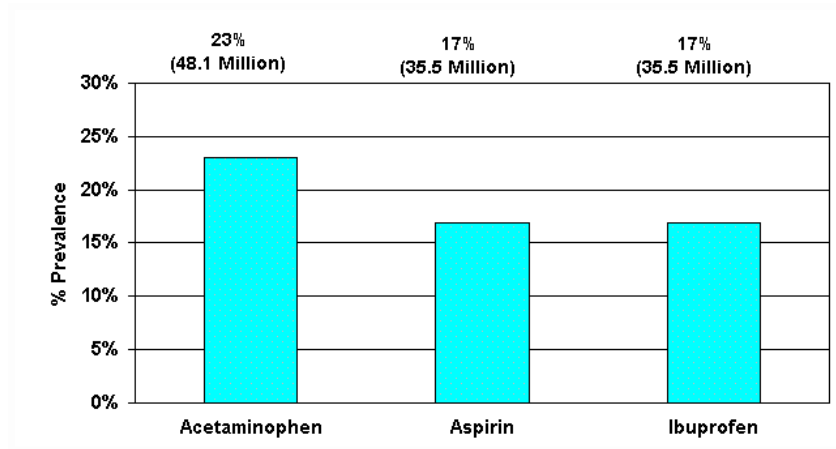
The classic approach: terminology & metrics

- A **metric** on a set X is a function (called the distance function or simply **distance**)
- $d : X \times X \rightarrow \mathbb{R}^+$ (where \mathbb{R}^+ is the set of non-negative real numbers). For all x, y, z in X , this function is required to satisfy the following conditions:
 - $d(x, y) \geq 0$ (**non-negativity**)
 - $d(x, y) = 0$ if and only if $x = y$ (**identity of indiscernibles**. Note that condition 1 and 2 together produce **positive definiteness**)
 - $d(x, y) = d(y, x)$ (**symmetry**)
 - $d(x, z) \leq d(x, y) + d(y, z)$ (**subadditivity / triangle inequality**).

Metrics: Scaling

- Nominal: no order (countries, genes, therapies)
- Ordinal: rank order (military ranks)
- Interval: addition makes sense (no 0), ratios make no sense (eg temperature)
- Ratio: addition & division make sense (with 0, eg body weight)

Metrics: Scaling



Design

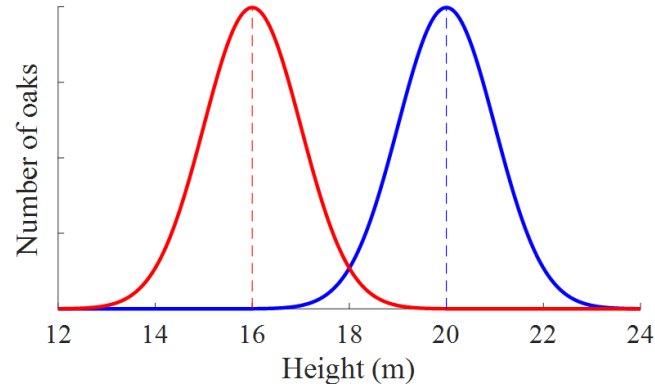
- Experimental design: split up observers randomly into two groups
- Cohorts: groups are defined by “natural” labels, e.g. patients vs. controls

The classic approach: Null hypothesis testing

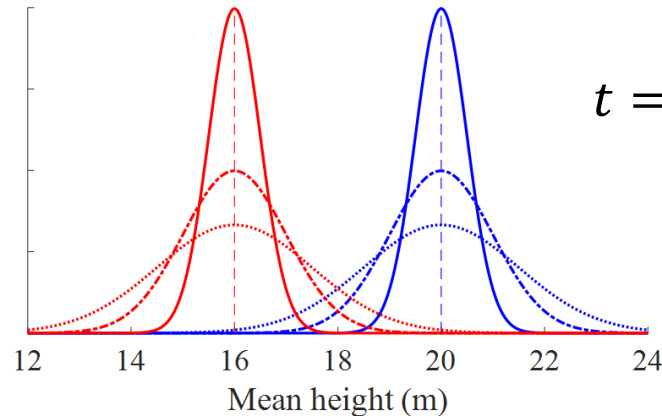
The Null Hypothesis Significance Testing (NHST) approach

- Formulate H_1 (alternative-hypothesis, the one matching your strategy)
- Choose a statistical model, that represents the alternative hypothesis
- Formulate H_0 (null-hypothesis)
- Assume that H_0 is true (normally H_0 is that there is no effect)
- Compute p value
- If $p < 0.05$, reject H_0 and accept H_1

Population distribution



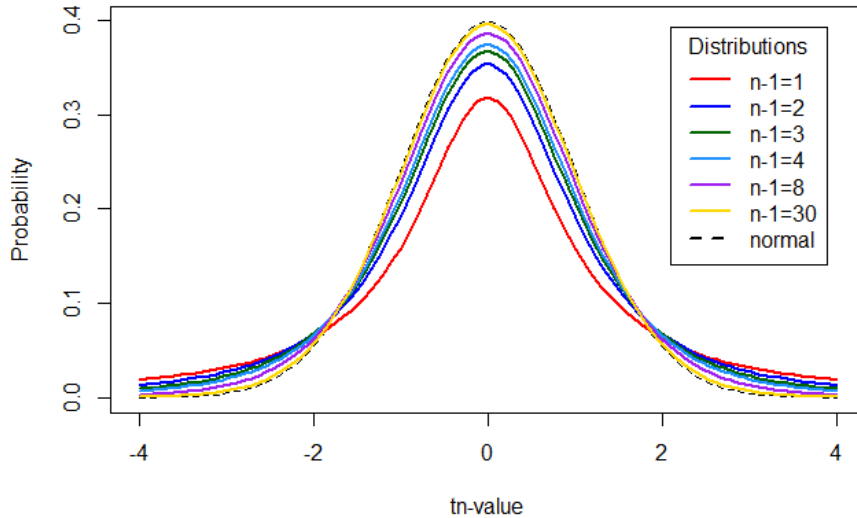
Family (n)
Sampling distributions



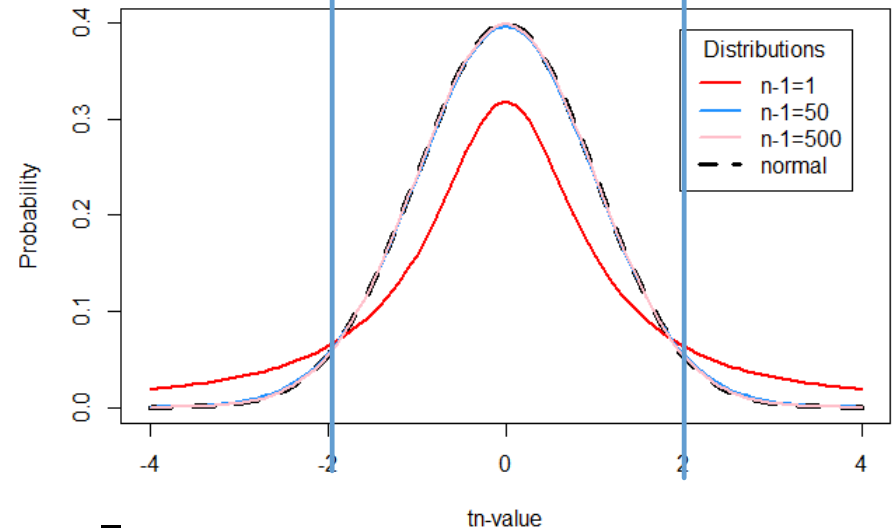
$$t = \frac{\bar{x}_{North} - \bar{x}_{South}}{s/\sqrt{n/2}}$$

The t distributions for the Null hypothesis

Student's t distributions with various degrees of freedom and the normal distribution

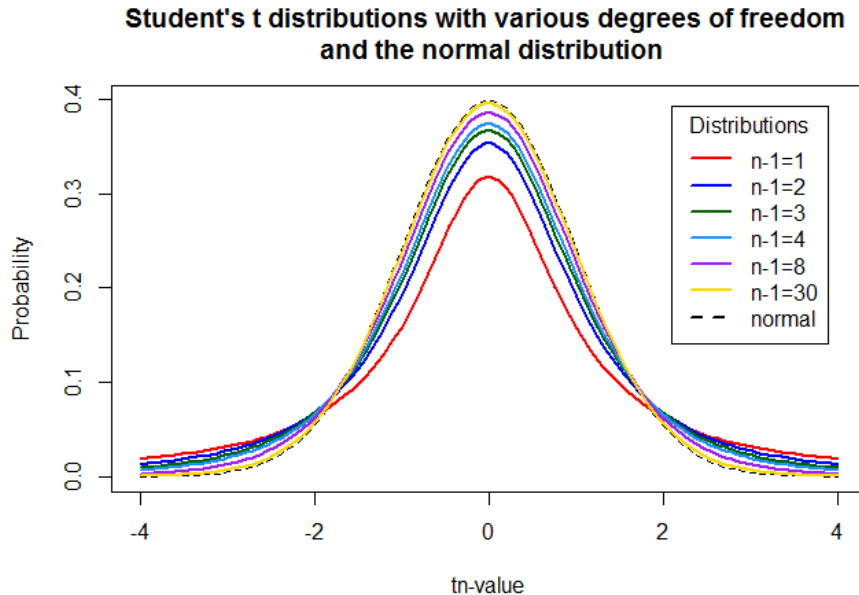


Student's t distributions with various degrees of freedom and the normal distribution



$$t = \frac{\bar{x}_{North} - \bar{x}_{South}}{s/\sqrt{n/2}}$$

The t distributions



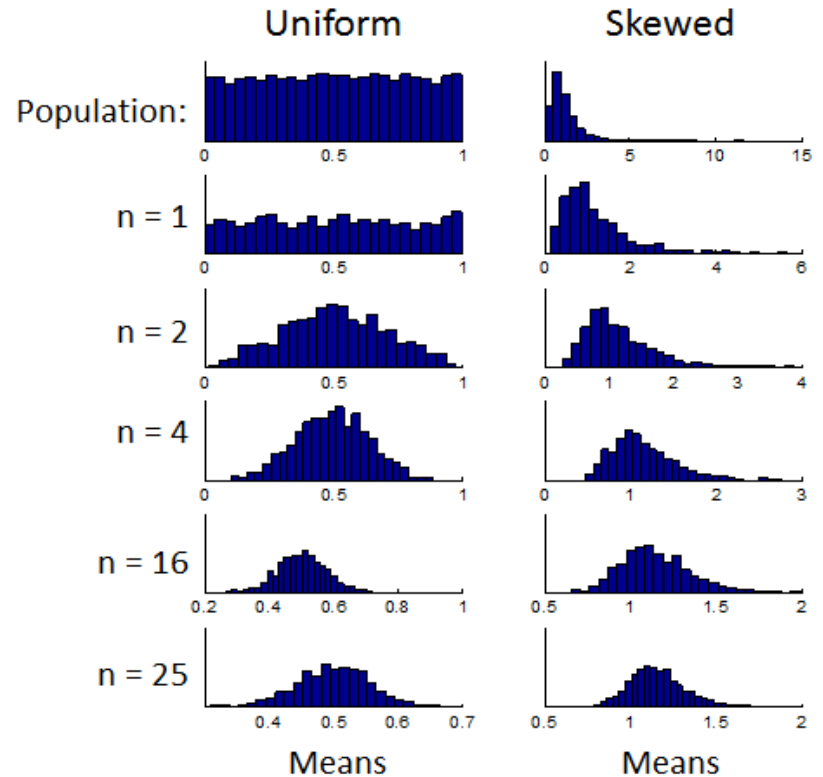
When the population variance is unknown (the usual case) and sample size is small ($n < 100$, the usual case), the t-distribution is used. The t distribution is a short, fat relative of the normal. The shape of t depends on n . As n becomes infinitely large, t becomes normal.

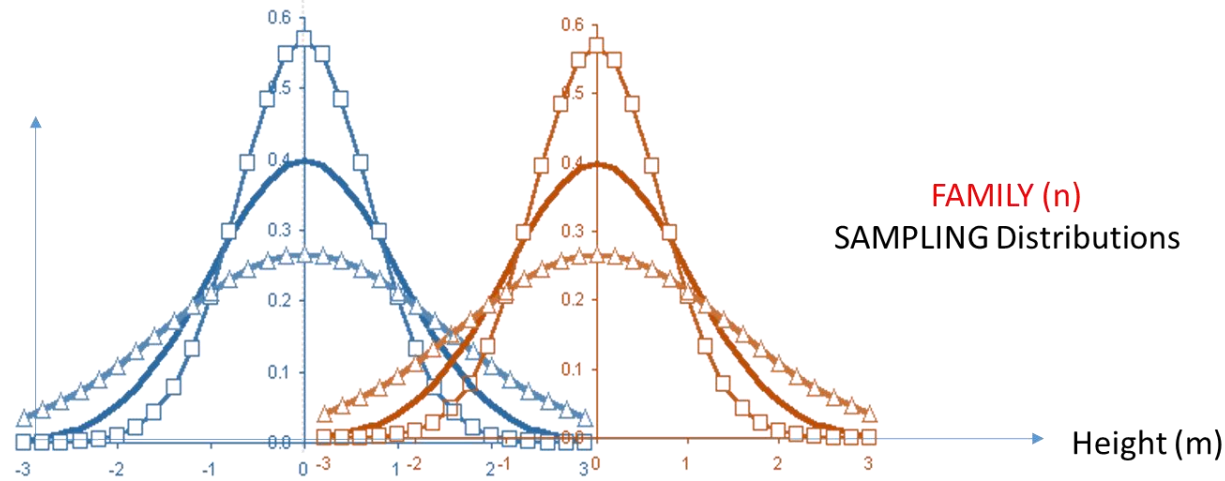
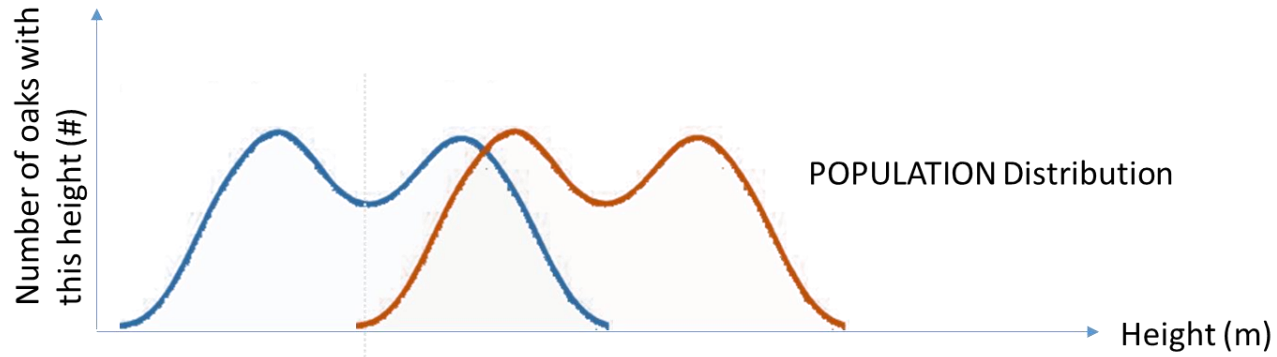
For $n \rightarrow \infty$:

- $\bar{x} \rightarrow \mu$, with $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
- $s_{\bar{x}} \rightarrow 0$, with $\frac{s}{\sqrt{n}}$ (std error of the mean: SEM)
- Sampling distribution: Gaussian: $N(\mu, SEM)$

Central Limit Theorem (CLT)

- The population distribution, i.e., **raw scores**, may have any arbitrary shape.
- However, the distribution of the **sample means** tends to be bell-shaped, regardless of how raw scores are distributed.
- This is important because certain statistical tests assume that data are normally distributed (i.e. Gaussian, or “bell-shaped”).





The classic approach: types of t-tests

- Independent samples *t*-test: 2 independent groups; eg trees in the above example, participants randomly assigned to a group
- Independent samples *t*-test (*Welsh test*): 2 independent groups, with different variances and sample sizes
- One sample *t*-test: 1 group is compared with a standardized distribution, eg IQ test
- Repeated measures *t*-test: 1 group, two conditions, eg same participants in a pre/post test, or related participants, such husband-wife, brother-sister
- One- and two-sided t-tests

- Two sample t-test

$$t = \frac{\bar{x}_{North} - \bar{x}_{South}}{s/\sqrt{n/2}}$$

- One sample t-test

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}$$

- Repeated measures t-test: a one sample t-test with $\mu_0 = 0$

Brother	Sister	Diff	$(D - \bar{D})^2$
5	7	2	1
7	8	1	0
3	3	0	1
$x_1 = 5$	$x_2 = 6$	$\bar{D} = 1$	

$$s_D = \sqrt{\frac{\sum (D - \bar{D})^2}{3 - 1}} = 1$$

$$t = \frac{\bar{x} - \mu}{s_D / \sqrt{3}} = \frac{1 - 0}{1 / \sqrt{3}}$$

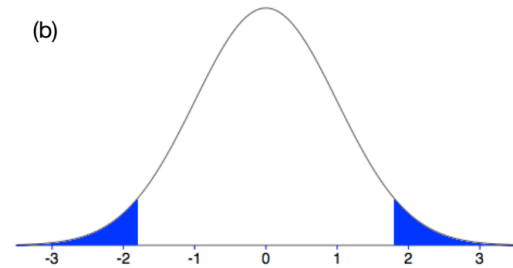
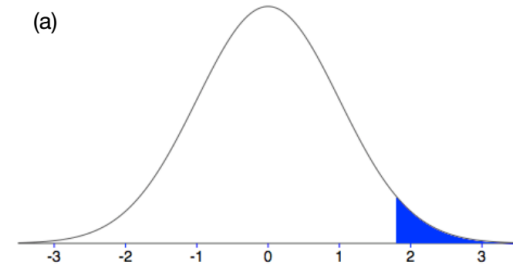
- The independent two sample t-test (*Welsh test*)

$$t = \frac{\bar{x}_{North} - \bar{x}_{South}}{\sqrt{\frac{s_{north}^2}{n_{north}} + \frac{s_{south}^2}{n_{south}}}} \quad \text{when } n_{north} \neq n_{south}, s_{north} \neq s_{south}$$

One and two sided t-tests

The two tailed t-test assumes that there is a difference between, for example, therapy A and B

The one tailed t-test assumes that therapy A is better than B (or therapy B is better than A) and is usually not justified and looks fishy; advantage: higher power

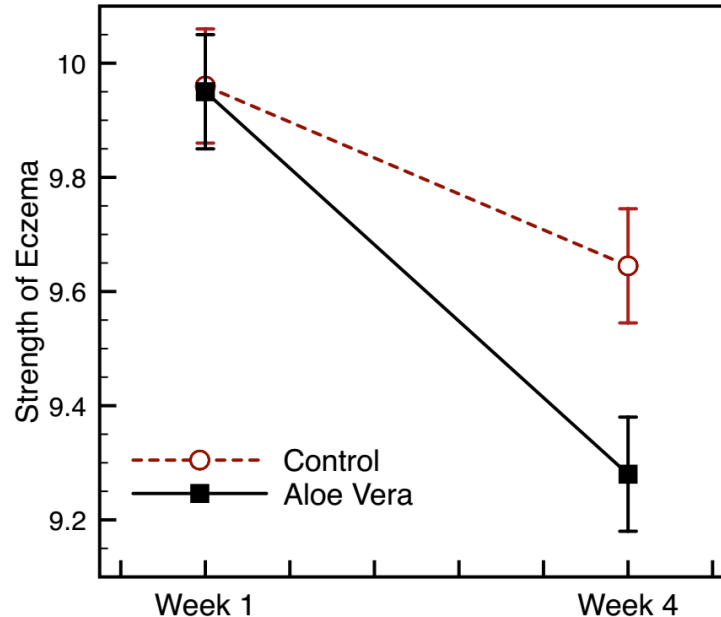


What test to choose?

Implication IIIb: A difference of significance is not a significant difference!

This is the same situation whenever a control group is used. It makes usually very little sense to state: “there was an effect in condition A but not B”.

Never ever compare a significant result with a Null result.....



Koch C, Dölle S, Metzger M, et al. Docosahexaenoic acid (DHA) supplementation in atopic eczema: a randomized, double-blind, controlled trial. Br J Dermatol 2008;158:786-792.

Assumptions

- Normality (for $n < 100$)
- Equal Variance & sample size
- Independent variable is ratio scaled
- Independent samples
- Identical distribution

	$n_1 = n_2 = 5$		$n_1 = 5, n_2 = 25$	
	$\sigma_2 = 1$	$\sigma_2 = 5$	$\sigma_2 = 1$	$\sigma_2 = 5$
$\sigma_1 = 1$	0.050	0.074	0.052	0.000
$\sigma_1 = 5$	0.073	0.051	0.383	0.047

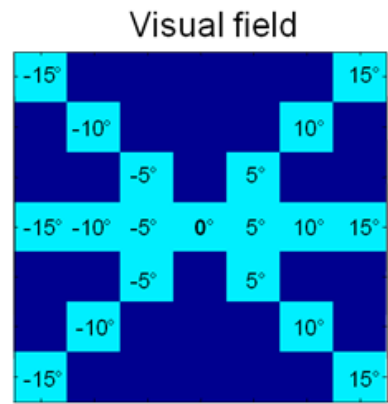
Table 1: Type I error rates for 10,000 simulated t -tests with different population standard deviations and sample sizes.

<http://shinyapps.org/apps/p-hacker/>

- Normality (for $n < 100$)
- Equal Variance & sample size
- Dependent variable is ratio scaled
- Independent samples
- Identical distribution

**Assumptions of all statistical tests: iid:
identically and independently distributed**

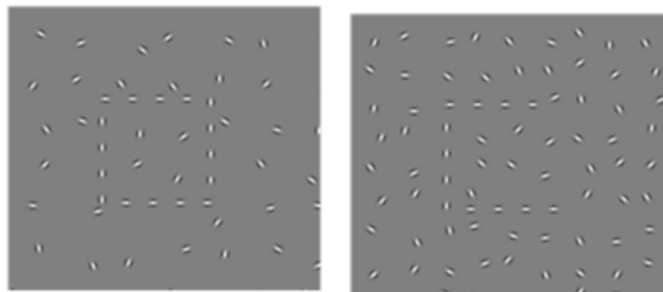
- Independent samples
- Identical Distribution



Left Field intact Right Field intact

UP (optic tract) FJ (occipital-parietal-temporal)
 HF (occipital)

Stimuli

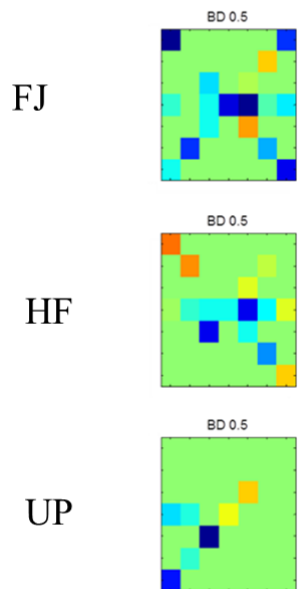


Square Fragment

Task: Is there a square?

One Sample T-Test

	t	df	p
FA rate	-2.499	29.00	0.018



Example

	Cardiovascular Disease	Cancer	Accident	Other
Reference Groups				
US Population Ages 55–64, (<i>n</i> = 338, 127)	27%	34%	5%	35%
Non-Flight Astronauts, (<i>n</i> = 35)	9%*	29%	53%*	9%*
Astronaut Groups				
All Flight Astronauts, (<i>n</i> = 42)	17%	31%	43%*	10%*
Low Earth Orbit Astronauts, (<i>n</i> = 35)	11%*	31%	49%*	9%*
Apollo Lunar Astronauts, (<i>n</i> = 7)	43% <u>±±</u>	29%	14% <u>^</u>	14%

From: [Apollo Lunar Astronauts Show Higher Cardiovascular Disease Mortality: Possible Deep Space Radiation Effects on the Vascular Endothelium](#)

Confidence Intervals

Calculating Confidence Intervals

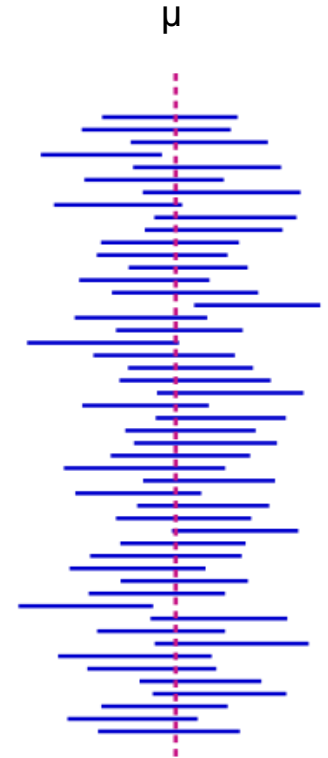
$$\mu = \bar{x} \pm t * \frac{s}{\sqrt{n}}$$

\bar{x} = sample mean

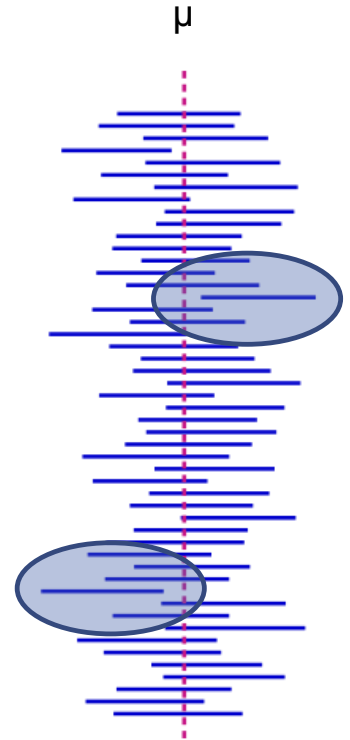
t determines how sure you want to be that μ lies in that interval, e.g. for a 95% CI, $t(.975, df=inf) = 1.96$

$\frac{s}{\sqrt{n}}$ = standard error of the sample mean

- Each blue line depicts the Confidence Interval of one individual sample (e.g., one experiment with n observations, or one patch of forest with n trees), centered around the respective sample mean \bar{x}
- Since the CI is centered on \bar{x} the CI will have equally long 'legs' to both sides.
- This length, like \bar{x} , is likely to vary from one sample to another due to chance.
- Yet, the sample means $\bar{x}_1 \dots \bar{x}_z$ will be randomly, but normally distributed around μ .



- Each blue line depicts the Confidence Interval of one individual sample (e.g., one experiment with n observations, or one patch of forest with n trees), centered around the respective sample mean \bar{x}
- Since the CI is centered on \bar{x} , the CI will have equally long 'legs' to both sides.
- This length, like \bar{x} , is likely to vary from one sample to another due to chance.
- Yet, the sample means $\bar{x}_1 \dots \bar{x}_z$ will be randomly, but normally distributed around μ .
- If Null hypothesis is true and if you continue taking samples (experiment) over and over again until the end of days, you will find that α (e.g. 5%) of these CIs do not contain μ



How is this useful? Example:

Some friend told you that Chinese men are 150cm tall on average. You don't believe him and travel to China and measure 100 men at random. This sample gives you an average height of 170cm and 10cm variance.

Claim: $\mu_0 = 150\text{cm}$

Observation: $\bar{x} = 170\text{cm}$ $s = 10\text{cm}$ $n = 100$

Choice: $\alpha = 0.05$ $\mu_1 = \bar{x} \pm t * \frac{s}{\sqrt{n}} = 170 \pm 1.96 * \frac{10}{\sqrt{100}} = 170 \pm 1.96$

95% CI for $\mu = [168.04; 171.96]$

150 is not included in the 95% CI. In fact, it doesn't even come close

Example:

You're a researcher and working on a highly debated issue: It has been suggested, that playing action video games can help dyslexics children read better.

You review the literature and indeed you find ten studies that investigated the phenomenon. You evaluate the studies and they are all methodologically sound and set up in the same way.

Example

Reported on the right are the pre-test / post-test differences and test statistics (t, p) per study.

What do you conclude?

- A. The evidence is equivocal, we need more research.
- B. All of the mean differences show a positive effect of the intervention, therefore, we have consistent evidence that the treatment works.
- C. Five of the studies show a significant result ($p < .05$), but the other 5 do not. Therefore, the studies are inconclusive: some suggest that the intervention is better than alpha, but others suggest there's no difference. The fact that half of the studies showed no significant effect means that the treatment is not (on balance) more successful in reducing symptoms than the control.

Study	Difference between		
	Means	t	p
Study 1	4.193	3.229	0.002*
Study 2	2.082	1.743	0.086
Study 3	1.546	1.336	0.187
Study 4	1.509	0.890	0.384
Study 5	3.991	2.894	0.006*
Study 6	4.141	3.551	0.001*
Study 7	4.323	3.745	0.000*
Study 8	2.035	1.479	0.155
Study 9	6.246	4.889	0.000*
Study 10	0.863	0.565	0.577

Example

Reported on the right are the pre-test / post-test differences, test statistics (t, p), **and confidence intervals** per study.

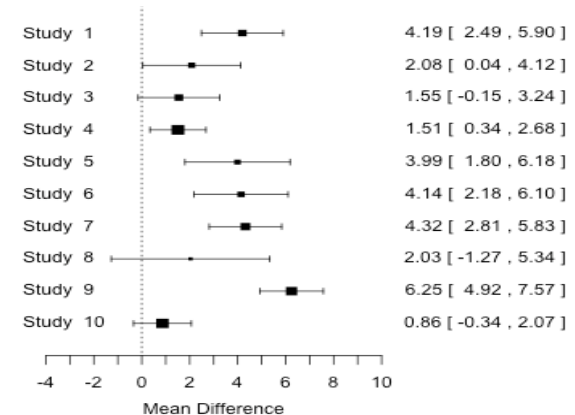
What do you conclude?

A. The evidence is equivocal, we need more research.

B. All of the mean differences show a positive effect of the intervention, therefore, we have consistent evidence that the treatment works.

C. Five of the studies show a significant result ($p < .05$), but the other 5 do not. Therefore, the studies are inconclusive: some suggest that the intervention is better than alpha, but others suggest there's no difference. The fact that half of the studies showed no significant effect means that the treatment is not (on balance) more successful in reducing symptoms than the control.

Study	Difference between Means	t	p
Study 1	4.193	3.229	0.002*
Study 2	2.082	1.743	0.086
Study 3	1.546	1.336	0.187
Study 4	1.509	0.890	0.384
Study 5	3.991	2.894	0.006*
Study 6	4.141	3.551	0.001*
Study 7	4.323	3.745	0.000*
Study 8	2.035	1.479	0.155
Study 9	6.246	4.889	0.000*
Study 10	0.863	0.565	0.577



Example

What added value do the p values have over the CI information?

Exactly *none*, that's why they're unnecessary!

What added value do the CIs have over plain p values?

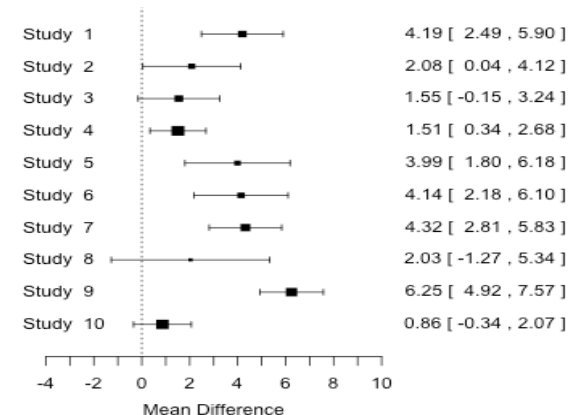
They contain additional information that is crucial: it changes your conclusion!

What does that tell us?

p values *alone* can be deceiving!

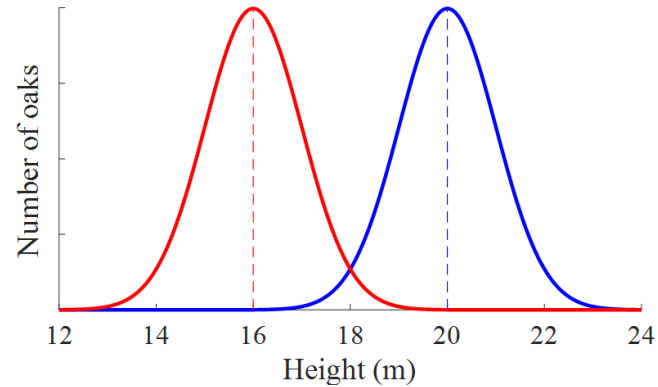
Dichotomous true / not-true decisions based on whether $p < .05$ are not very smart.

Study	Difference between Means	t	p
Study 1	4.193	3.229	0.002*
Study 2	2.082	1.743	0.086
Study 3	1.546	1.336	0.187
Study 4	1.509	0.890	0.384
Study 5	3.991	2.894	0.006*
Study 6	4.141	3.551	0.001*
Study 7	4.323	3.745	0.000*
Study 8	2.035	1.479	0.155
Study 9	6.246	4.889	0.000*
Study 10	0.863	0.565	0.577

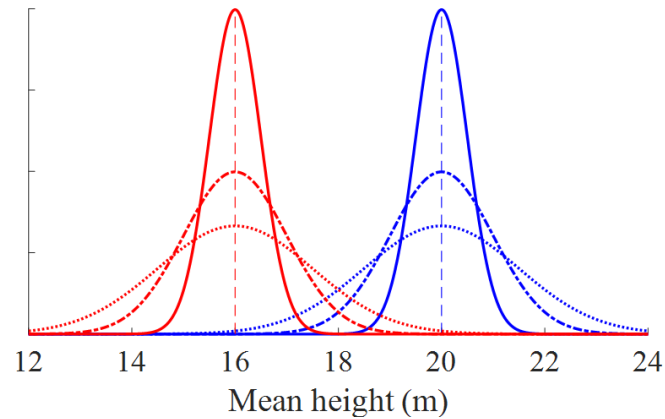


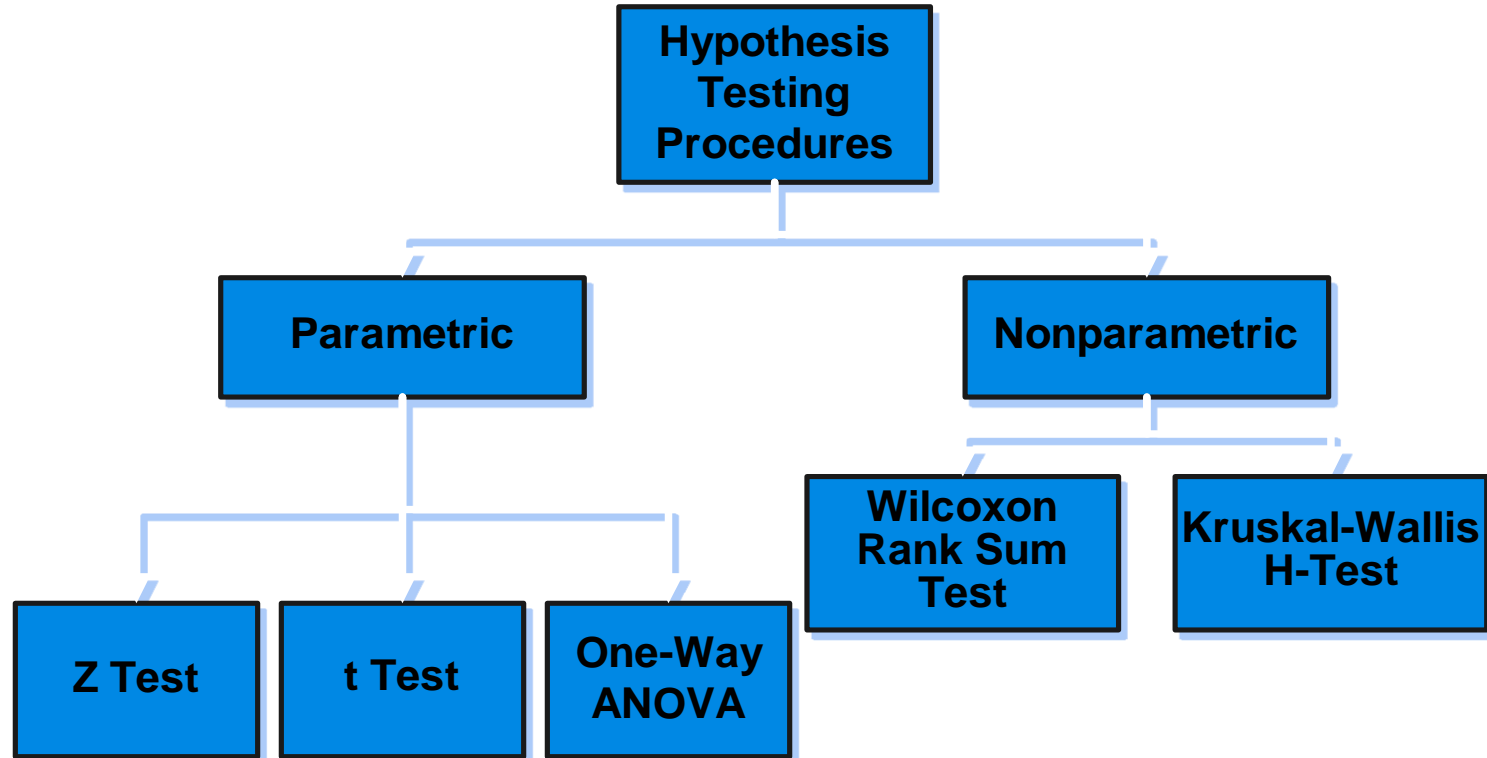
Non-Parametric Tests

Population distribution



Family (n)
Sampling distributions





1. Sign Test corresponds to one sample test
2. Wilcoxon Rank Sum Test corresponds to independent t-test
3. Man-Whitney U-test corresponds to repeated measures t-test

You're an analyst for Chef-Boy-R-Dee. You've asked 7 people to rate a new ravioli on a 5-point scale (1 = terrible,..., 5 = excellent) The ratings are:

2 5 3 4 1 4 5

At the .05 level, is there evidence that the median rating is **at least 3**?

- $H_0: \eta = 3$
- $H_1: \eta < 3$
- $\alpha = 0.05$
- Test statistic:

$S = 2$ (Ratings 1&2 are $< \eta = 3$: **2, 5, 3, 4, 1, 4, 5**)

Is observing 2 or less a small prob event?

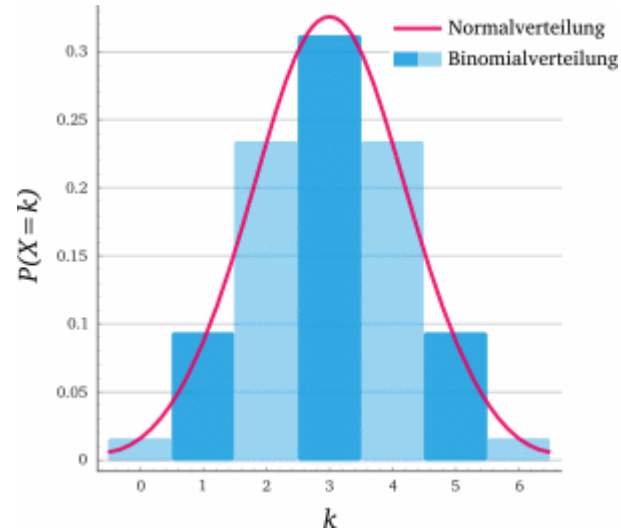
p-value: $p(x > 2) = 1 - p(x \leq 1) = 0.9297$

(Binomial table, $n = 7, p = 0.50$)

Decision: Do not reject at $\alpha = 0.05$

Conclusion: There is No evidence for Median < 3

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$



Example:

The table below shows the hours of relief provided by two analgesic drugs in 12 patients suffering from arthritis. Is there any evidence that one drug provides longer relief than the other?

Case	Drug A	Drug B	Case	Drug A	Drug B
1	2.0	3.5	7	14.9	16.7
2	3.6	5.7	8	6.6	6.0
3	2.6	2.9	9	2.3	3.8
4	2.6	2.4	10	2.0	4.0
5	7.3	9.9	11	6.8	9.1
6	3.4	3.3	12	8.5	20.9

Solution:

In this case our null hypothesis is that the median difference is zero. Our actual differences (Drug B - Drug A) are:

$$+1.5, +2.1, +0.3, -0.2, +2.6, -0.1, +1.8, -0.6, +1.5, +2.0, +2.3, +12.4$$

Our actual median difference is 1.65 hours. We have $r^+ = 9, r^- = 3, n = 12, r = \max(r^-, r^+) = 9$. Therefore our two-sided p-value (from binomial tables) is $p = 0.146$. We would conclude that there is no evidence for a difference between the two treatments. Note that the Wilcoxon Signed Rank Sum Test would also be appropriate in this case and is a more powerful test because it takes account of the magnitude of the differences as well as the sign.

Example

The observations in a sample of size n are x_1, x_2, \dots, x_n

The null hypothesis is that the population median is equal to some value M . Suppose that r_+ of the observations are greater than M and r_- are smaller than M . Values of x which are exactly equal to M are ignored; the sum $r_+ + r_-$ may therefore be less than n — we will denote it by n' .

Under the null hypothesis we would expect half the x 's to be above the median and half below. Therefore, under the null hypothesis both r_+ and r_- follow a binomial distribution with $p = 0.5$ and $n = n'$.

The test procedure is as follows:

1. Choose $r = \max(r_-, r_+)$.
2. Use tables of the binomial distribution to find the probability of observing a value of r or higher assuming $p = 0.5$ and $n = n'$. If the test is one-sided, this is your p -value.
3. If the test is a two-sided test, double the probability obtained in (2) to obtain the p -value.

1. Used With All Scales
2. Easier to Compute
3. Make Fewer Assumptions
4. Need Not Involve, Population Parameters
5. Results May Be as Exact as Parametric Procedures

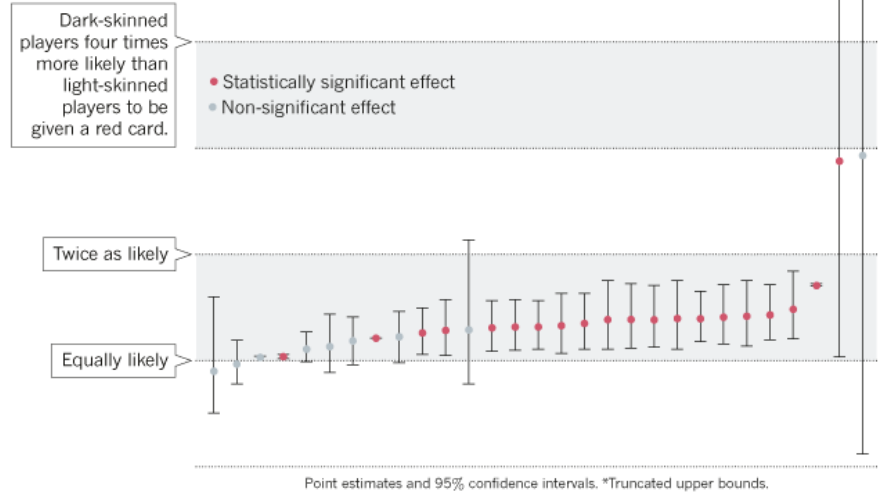
1. May Waste Information Parametric model more efficient if data permit
2. Difficult to Compute by hand for Large Samples
3. Tables Not Widely Available

What test to choose?

20/29 found effect, 1 even opposite effect

ONE DATA SET, MANY ANALYSTS

Twenty-nine research teams reached a wide variety of conclusions using different methods on the same data set to answer the same question (about football players' skin colour and red cards).

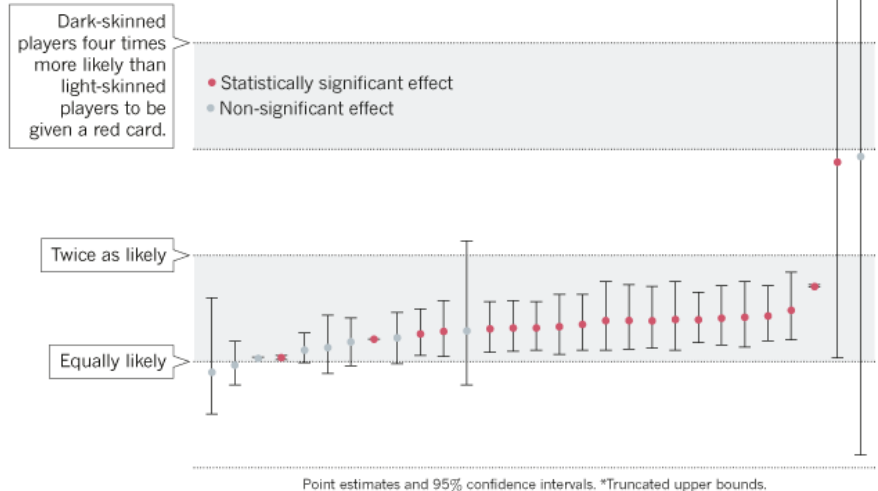


Solution: inverse labels

20/29 found effect, 1 even opposite effect

ONE DATA SET, MANY ANALYSTS

Twenty-nine research teams reached a wide variety of conclusions using different methods on the same data set to answer the same question (about football players' skin colour and red cards).



Bias by choosing statistical test and expectation

Obviously, different tests, different results
Choice of tests depends on expectation: Bias

ONE DATA SET, MANY ANALYSTS

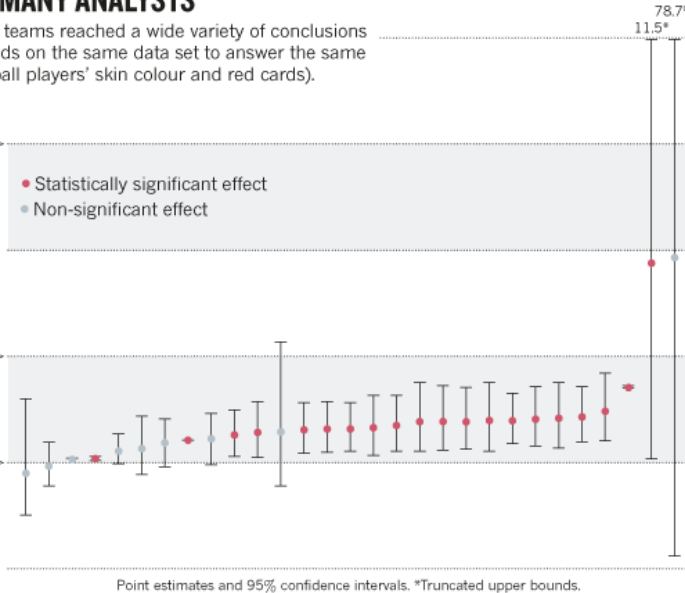
Twenty-nine research teams reached a wide variety of conclusions using different methods on the same data set to answer the same question (about football players' skin colour and red cards).

Dark-skinned players four times more likely than light-skinned players to be given a red card.

• Statistically significant effect
• Non-significant effect

Twice as likely

Equally likely



Good & Bad Statistical Practice

- Specific research question and output variable(s)
- Choose statistical test beforehand
- Choose alpha (0.05) and beta level (0.2)
- Choose effect size and compute sample size n
- Run the experiment without any changes
- Compute statistics
- No post-hoc hypothesis testing
- Assumptions: iid, stats model?
- Always: Effect size? Ceiling, Floor effects?

Bad Practice

- Measure as many variable as possible
- See what happens: interesting hot spots?
- Create Post-hoc hypothesis
- Remove “outliers”- depending on post-hoc hypothesis
- Try various statistical tests
.....until significant

Actually this is all fine, if you treat your experiment as an exploratory experiment- and repeat.....

Discuss the robustness of a t hypothesis test to various violations. Consider the impact of violating population normality, homogeneity of population variance, and fixed sample size. How do these violations affect control of Type I error? What can be done to reduce the impact of these violations?

Take Home Messages

1. For a t -test, make sure your data are iid distributed and the y -axis is a ratio-scale.
2. Data should be Gaussian distributed or n should be large.

END Class 4