Lecture 13: 12/23/2020

Power function and Two-Sample Problems

Review:

Power =

$1-\beta$ = Pr(reject $H_0 | H_a$ is true)

Decision we make	State of Nature	
	H_o is true	H_o is false
Accept H_o	ok	Type II error probability β
Reject H_o	Type I error probability α	ok

Power at $H_a: \mu = \mu_a$

- Step 1: Derive the rejection rule based on $H_0: \mu = \mu_0$

$$\Pr(\text{reject } H_0 \mid \mu = \mu_0) = \alpha$$

- Step 2: Calculate the Type II error and power at $H_a: \mu = \mu_a$

$$\Pr(\text{reject } H_0 \mid \mu = \mu_a) = 1-\beta\,, \quad \Pr(\text{Accept } H_0 \mid \mu = \mu_a) = \beta$$

Exercises: $X_i \sim^{iid} N(\mu, \sigma^2)$ for $i = 1, \ldots, 100$, $\sigma^2 = 5^2$ is known and α = 0.05.

Case 1: $H_0: \mu = \mu_0$ vs. $H_a: \mu = \mu_a > \mu_0$.

Power at $H_a: \mu = \mu_a$

$$\Pr(\bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} \mid \mu = \mu_a)$$ "rejection rule 不要化簡,才容易計算"

$$= \Pr(\frac{\bar{X} - \mu_a}{\sigma/\sqrt{n}} > \frac{\mu_0 - \mu_a}{\sigma/\sqrt{n}} + z_\alpha \mid \mu = \mu_a)$$

$$= \Pr(Z > \frac{\mu_0 - \mu_a}{\sigma/\sqrt{n}} + z_\alpha)$$

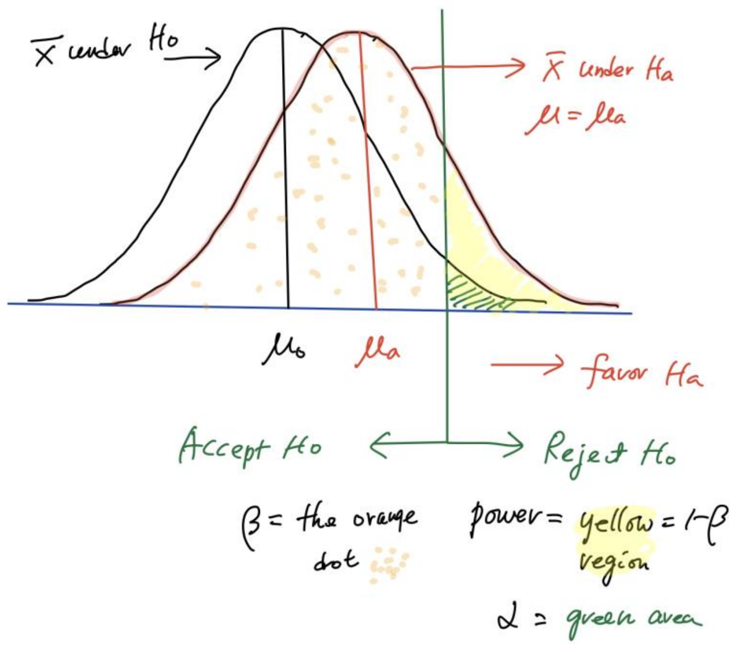
Case 2: $H_0: \mu = \mu_0$ vs. $H_a: \mu = \mu_a < \mu_0$.

Power at $H_a: \mu = \mu_a$

$$\Pr(\bar{X} < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}} \mid \mu = \mu_a)$$

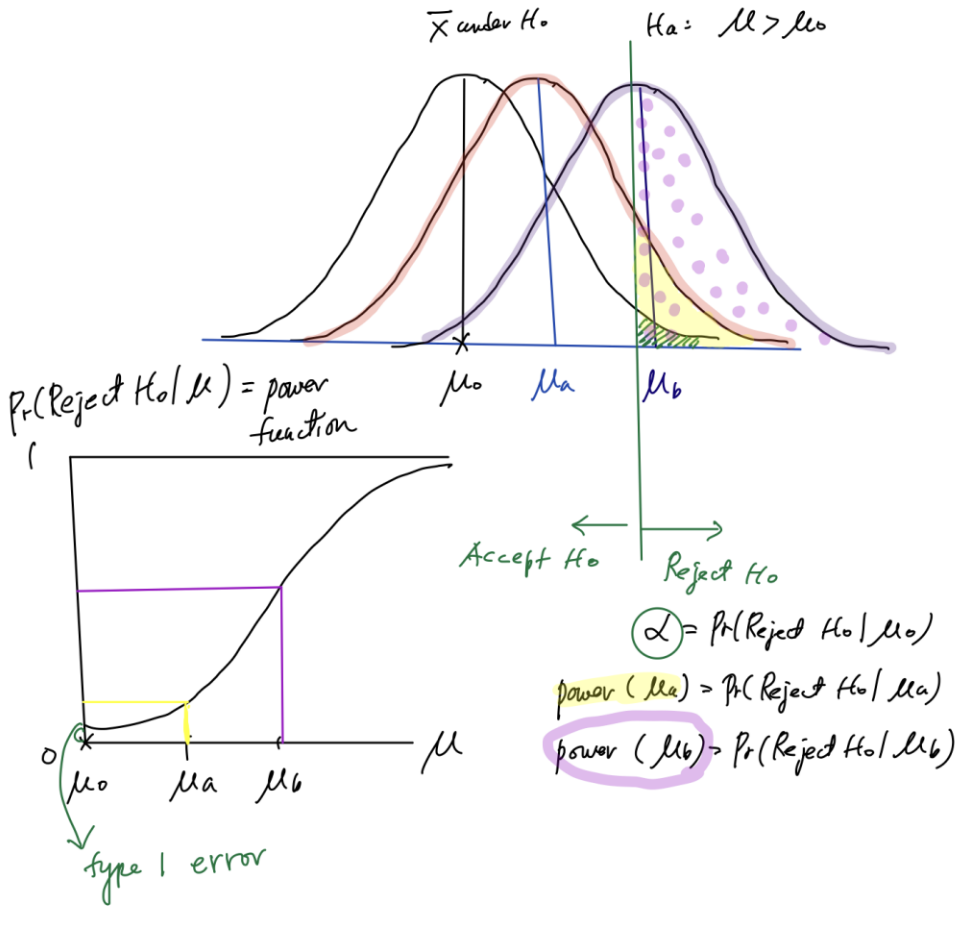
$$= \Pr(\frac{\bar{X} - \mu_a}{\sigma/\sqrt{n}} < \frac{\mu_0 - \mu_a}{\sigma/\sqrt{n}} - z_\alpha \mid \mu = \mu_a)$$

$$= \Pr(Z < \frac{\mu_0 - \mu_a}{\sigma/\sqrt{n}} - z_\alpha)$$

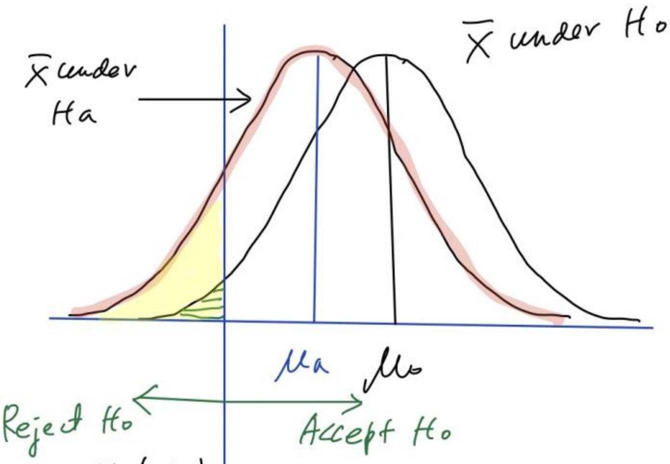


When the curve under H_a moves away from H_0, the power increases

The lower-left plot: the power function for 右尾檢定 (right-sided test)



Case (ii)



$\alpha = Pr(\text{Reject } H_0 | \mu_0)$

$1-\beta(\mu_a) = \text{power}(\mu_a)$

$= Pr(\text{Reject } H_0 | \mu_a)$

Case 3: $H_0 : \mu = \mu_0$ vs. $H_a : \mu = \mu_a \neq \mu_0$.

→ **雙尾的 power 要特別注意，有兩塊機率 (大 & 小)**

Step 1: 先由方法一算出 reject H_0 的公式

Rejection rule: if $\bar{X} < \mu_0 - z_{\alpha/2}\dfrac{\sigma}{\sqrt{100}}$ or $\bar{X} > \mu_0 + z_{\alpha/2}\dfrac{\sigma}{\sqrt{100}}$

Power at $H_a : \mu = \mu_a$

$$\Pr(\bar{X} > \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \mid \mu = \mu_a) + \Pr(\bar{X} < \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \mid \mu = \mu_a)$$

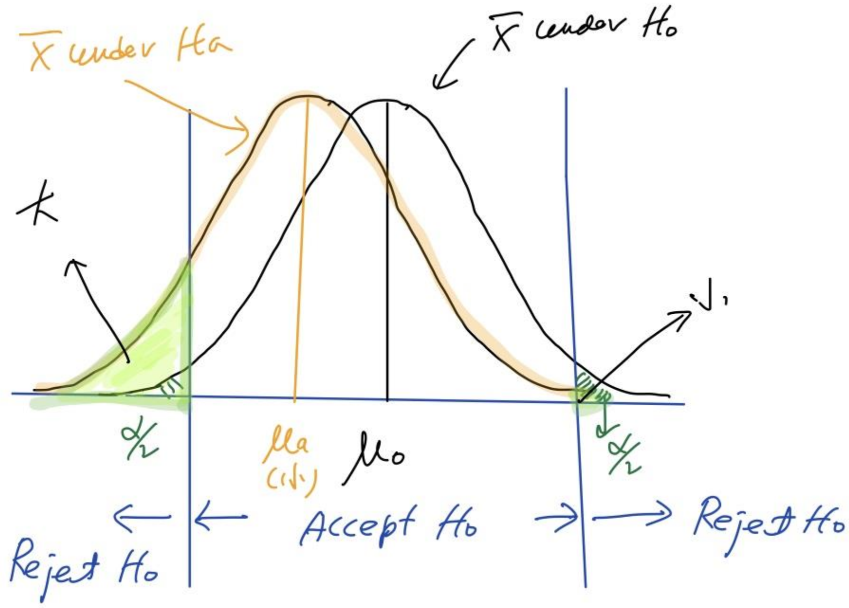
$$= \Pr(\frac{\bar{X}-\mu_a}{\sigma/\sqrt{n}} > \frac{\mu_0-\mu_a}{\sigma/\sqrt{n}} + z_{\alpha/2} \mid \mu = \mu_a) + \Pr(\frac{\bar{X}-\mu_a}{\sigma/\sqrt{n}} < \frac{\mu_0-\mu_a}{\sigma/\sqrt{n}} - z_{\alpha} \mid \mu = \mu_a)$$

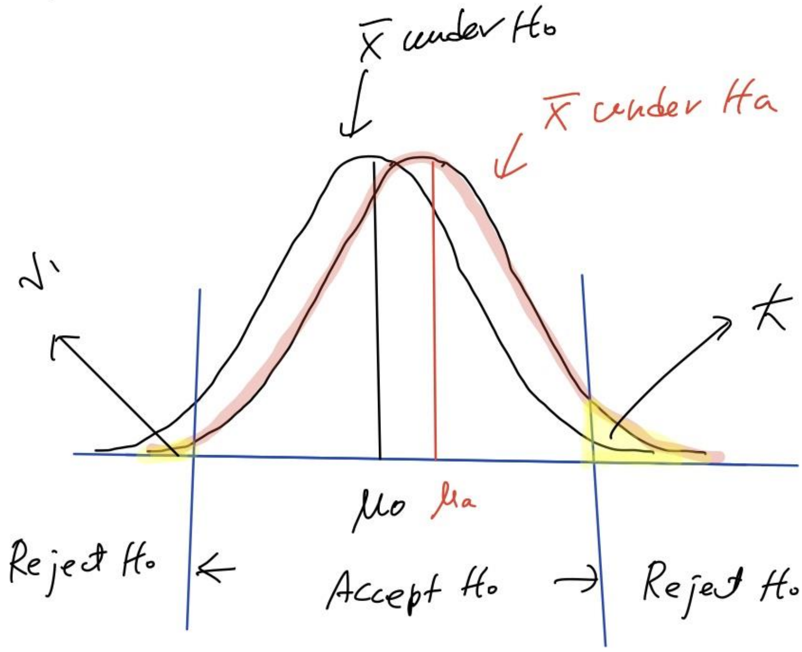
$$= \Pr(Z > \frac{\mu_0-\mu_a}{\sigma/\sqrt{n}} + z_{\alpha/2}) + \Pr(Z < \frac{\mu_0-\mu_a}{\sigma/\sqrt{n}} - z_{\alpha/2})$$

兩塊機率的大小

If $\mu_a > \mu_0$, $\Pr(Z > \dfrac{\mu_0-\mu_a}{\sigma/\sqrt{n}} + z_{\alpha/2})$ is larger;

If $\mu_a < \mu_0$, $\Pr(Z < \dfrac{\mu_0-\mu_a}{\sigma/\sqrt{n}} - z_{\alpha/2})$ is larger;





Example: $H_0: \mu = 3$ vs. $H_a: \mu = \mu_a \neq \mu_0 = 4$.

$n = 100, \sigma = 5, \alpha = 5\%$

Rule of accepting H_0: $3 - 1.96\frac{5}{10} < \bar{X} < 3 + 1.96\frac{5}{10}$ → used to compute β

Rule of rejecting H_0: $\bar{X} < 3 - 1.96\frac{5}{10}$ or → $\bar{X} > 3 + 1.96\frac{5}{10}$

At $\mu_a = 2$

$$\beta = \Pr(\text{Accept } H_0 \mid \mu_a = 2)$$

$$= \Pr(3 - 1.96\frac{5}{10} < \bar{X} < 3 + 1.96\frac{5}{10} \mid \mu_a = 2)$$

$$\text{Power} = \Pr(Z > \frac{(3-2)}{5/10} + 1.96) + \Pr(Z < \frac{(3-2)}{5/10} - 1.96) \quad +$$

$$= \Pr(Z > 3.96) + \Pr(Z < 0.04) \quad (\text{主要的值在第二項})$$

When $\mu_a = 4$.

$$\beta = \Pr(3 - 1.96\frac{5}{10} < \bar{X} < 3 + 1.96\frac{5}{10} \mid \mu_a = 4)$$

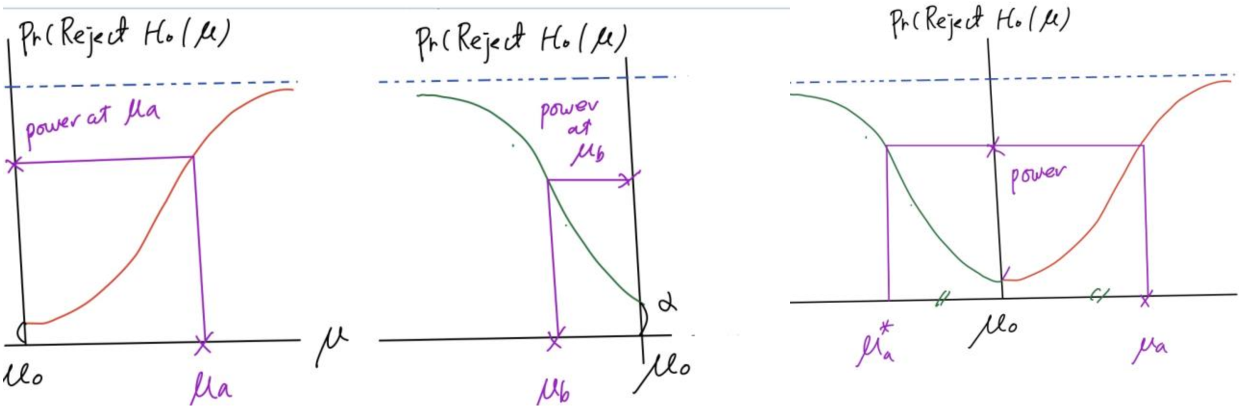
$$\text{Power} = \Pr(\bar{X} > 3 + 1.96\frac{5}{10} \mid \mu_a = 4) + \Pr(\bar{X} < 3 - 1.96\frac{5}{10} \mid \mu_a = 4) \quad +$$

$$= \Pr(Z > \frac{(3-4)}{5/10} + 1.96) + \Pr(Z < \frac{(3-4)}{5/10} - 1.96) \quad +$$

$$= \Pr(Z > -0.04) + \Pr(Z < -3.96) \quad (\text{主要值在第一項})$$

Power as a function of μ

Case 1: $H_a: \mu > \mu_0$ Case 2: $H_a: \mu < \mu_0$ Case 3: $H_a: \mu \neq \mu_0$



Remarks:

- $\alpha = \Pr(\text{reject } H_0 \mid \mu = \mu_0)$
- power = $\Pr(\text{reject } H_0 \mid \mu = \mu_a)$
- Both are the probabilities of rejecting H_0 but they are derived under different curves.
-

決定 power 的因子

- Distance between μ_0 and μ_a:
  - Larger value of $|\mu_a - \mu_0|$ → larger value of power
- The value of σ:
  - Smaller value of σ, larger value of power
- The sample size n:
  - Larger value of n, larger value of power

重要題型：給定 Power 的目標，決定最小樣本

當研究者心目中有個 μ_a (研究者認為真正的 μ 值)，希望檢定的 power 在此 μ_a 值時要達到某個標準 (如 power 至少九成). 可以求“最小樣本數”.

Case 1: $H_0: \mu = 3$ vs. $H_a: \mu > 3$.

$$\Pr(\bar{X} > 3 + 1.645\frac{5}{\sqrt{n}} \mid \mu = 4) \ge 0.8 \rightarrow \Pr(Z > \frac{3-4}{5/\sqrt{n}} + 1.645) \ge 0.8$$

$$\frac{3-4}{5/\sqrt{n}} + 1.645 \approx -0.84 \quad (\text{查表得到 } \Pr(Z > -0.84) \approx 0.8)$$

$$\rightarrow 5(1.645 + 0.84) \approx \sqrt{n} \rightarrow 12.425 \approx \sqrt{n}$$

樣本數至少要 155. (自行檢查老師有沒有算錯)

Case 2: $Power = \Pr(\bar{X} < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}} \mid \mu = \mu_a) \ge K \quad (0 < K < 1)$

$$\Pr(\frac{\bar{X} - \mu_a}{\sigma/\sqrt{n}} < \frac{\mu_0 - \mu_a}{\sigma/\sqrt{n}} - z_\alpha) = \Pr(Z < \frac{\mu_0 - \mu_a}{\sigma/\sqrt{n}} - z_\alpha) \ge K \rightarrow \text{畫圖}$$

Case 3: $Power = \Pr(\bar{X} < \mu_0 - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \mid \mu = \mu_a) + \Pr(\bar{X} > \mu_0 + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \mid \mu = \mu_a) \ge K$

$$Power = \Pr(Z < \frac{\mu_0 - \mu_a}{\sigma/\sqrt{n}} - z_{\alpha/2}) + \Pr(\bar{X} > \frac{\mu_0 - \mu_a}{\sigma/\sqrt{n}} + z_{\alpha/2}) \ge K$$

→ 不好算最小 n，不會在考試為難你們☺

Remarks:

- 現在探討的問題是最基本的題目，因此只教一個檢定方法. 目前課堂學到的方法有最好的 power.
- 複雜的問題，往往存在不只一種檢定方法，此時就要比較它們的“power”.
- 當一個檢定在“所有的情形” (所有的 μ_a 值) 比起其他的檢定方法都有最大的 power. 這種 test 叫做 “UMP” (uniformly most power test). 但遇到稍複雜一點的情形，就不存在這種方法.
- Power 低的檢定，若具有“其他優點” (如所需假設比較弱)，還是有市場競爭力. 基本上統計軟體讓你選用的檢定，都有其適用時機.

Example: loss of sweetness

Let $X = X_{before} - X_{after}$ $\sigma = 1$

$H_0 : \mu = 0$ (no loss of sweetness after storage for a period of time)

$H_a : \mu = \mu_a = 0.8$

Objective: find the smallest sample size → Pr(Reject $H_0 \mid \mu = \mu_a$) = target value

The bigger these differences, the greater the loss of sweetness.

Suppose we know that for any cola, the sweetness loss scores vary from taster to taster according to a Normal distribution with standard deviation $\sigma = 1$. The mean μ for all tasters measures loss of sweetness and is different for different colas.

Here are the sweetness losses for a new cola, as measured by 10 trained tasters:

2.0 0.4 0.7 2.0 −0.4 2.2 −1.3 1.2 1.1 2.3

Most are positive. That is, most tasters found a loss of sweetness. But the losses are small, and two tasters (the negative scores) thought the cola gained sweetness. The average sweetness loss is given by the sample mean $\bar{x} = 1.02$. Are these data good evidence that the cola lost sweetness in storage?

EXAMPLE 15.7 Sweetening colas: planning a study

Let's illustrate typical answers to these questions in the example of testing a new cola for loss of sweetness in storage (Example 14.5, page 363). Ten trained tasters rated the sweetness on a 10-point scale before and after storage, so that we have each taster's judgment of loss of sweetness. From experience, we know that sweetness loss scores vary from taster to taster according to a Normal distribution with standard deviation about

Rejection rule: reject if

$$\bar{X} > \mu_0 + 1.645 \frac{\sigma}{\sqrt{n}} \quad (\mu_0 = 0,\ \sigma = 1)$$

power at $\mu = \mu_a = 0.8$

$$Pr\left(\bar{X} > 0 + 1.645 \frac{1}{\sqrt{n}} \mid \mu_a = 0.8\right)$$

$$= Pr\left(\frac{\bar{X} - \mu_a}{\sigma/\sqrt{n}} > \frac{0 - 0.8}{1/\sqrt{n}} + 1.645\right)$$

$$= Pr\left(Z > -0.8\sqrt{n} + 1.645\right) = 0.9 \text{ (target}$$

Based on normal table

$\Pr(Z > -1.285) \approx 0.9$ → $-0.8\sqrt{n} + 1.645 \approx -1.285$ → $\sqrt{n} \approx 3.66$ → $n \approx 13.42$

Take $n = 14$

EXAMPLE 15.9 Finding power: use software

We asked Minitab to find the number of observations needed for the one-sided z test to have power 0.9 against several specific alternatives at the 5% significance level when the population standard deviation is $\sigma = 1$. Here is the table that results:

Difference	Sample Size	Target Power	Actual Power
0.1	857	0.9	0.900184
0.2	215	0.9	0.901079
0.3	96	0.9	0.902259
0.4	54	0.9	0.902259
0.5	35	0.9	0.905440
0.6	24	0.9	0.902259
0.7	18	0.9	0.907414
~~0.8~~	14	0.9	0.911247
0.9	11	0.9	0.909895
1.0	9	0.9	0.912315

思考一：目前的檢定規則為何如此訂定？

- 目前通行的準則: fix α and minimize β
  - α = Pr(reject $H_0 | H_0$ is true) → 型一錯誤機率
  - β = Pr(accept $H_0 | H_0$ is false) → 型二錯誤機率
  - 理想: both α and β are minimized → impossible!!! 魚與熊掌不可兼得
- 因為研究者通常是想要 reject H_0，所以可以預期他認為H_0 不太可能為真，這是為何 fix α,而不是要 fix β 的理由 (因為 H_0 不太可能，萬一它是對的，那就認了. 把接受這種錯誤的機率固定在α)
- 為何設 $\alpha = 0.05$，而不是$\alpha = 0.3$?
  - 令 $\alpha = 0.05$ (而不是$\alpha = 0.3$) 會使拒絕H_0 這件事<u>更不容易達到</u>.
    - ◆ $\alpha = 0.05$ 的拒絕標準很嚴格
    - ◆ $\alpha = 0.3$ 的拒絕標準很寬鬆
  - 因為 H_0 通常是<u>現有或是過去通行的理論</u>，所以要否定它，必需<u>證據很</u>

強烈才行. (所以才會把α 的值設得很小，不容易拒絕 H_0)

- 常見的現象：雖然新方法看來是有效，但是證據尚不具有統計顯著性.

  (There is no significance evidence to reject H_0).

Grifols Offers Hope Of A Treatment For Alzheimer
The Corner Economic - 2018年10月29日
... high level of mistakes and Morgan Stanley analysts argue that the lack of statistical significance in some of the results invites caution until we ...

AHA: CABG Still Better for Multivessel Disease in Diabetes
MedPage Today - 2018年11月12日
... was observed in the survival curves comprised of only patients with extended follow-up, and the lack of statistical significance in this cohort is ...

rmatologyTin

Alternative medicines may treat psoriasis, but proceed cautiously
Dermatology Times (press release) (blog) - 2018年9月25日
... on the Mediterranean diet, gluten-free diet, micronutrient supplementation and orange-peel extract due to a lack of statistical significance.

思考二：統計顯著性與樣本數

- n增加到很大時，$\bar{X} \sim N(\mu_0, \dfrac{\sigma^2}{n})$ density 會變得很尖(集中). 此時，拒絕 H_0 的機率會隨n增大而增加. 所以即使 $|\mu_a - \mu_0|$ 很小，卻可能因為樣本很大，使得差異具有 "統計顯著性".
- 換言之當樣本很小時，即使真實的差異是明顯的 ($|\mu_a - \mu_0|$ 很大)，仍無法用統計方法予以拒絕.
- "科學的顯著性" (反映在 $|\bar{x}_{obs} - \mu_0|$ 的值) 與 "統計的顯著性" (反映在 (反映在 p-value) 並不完全等價.

Example: $H_0: \mu = \mu_0$ vs. $H_a: \mu > \mu_0$

$$\Pr(\bar{X} > \bar{x}_{obs} \mid \mu = \mu_0) = \Pr(\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} > \frac{\bar{x}_{obs} - \mu_0}{\sigma / \sqrt{n}})$$

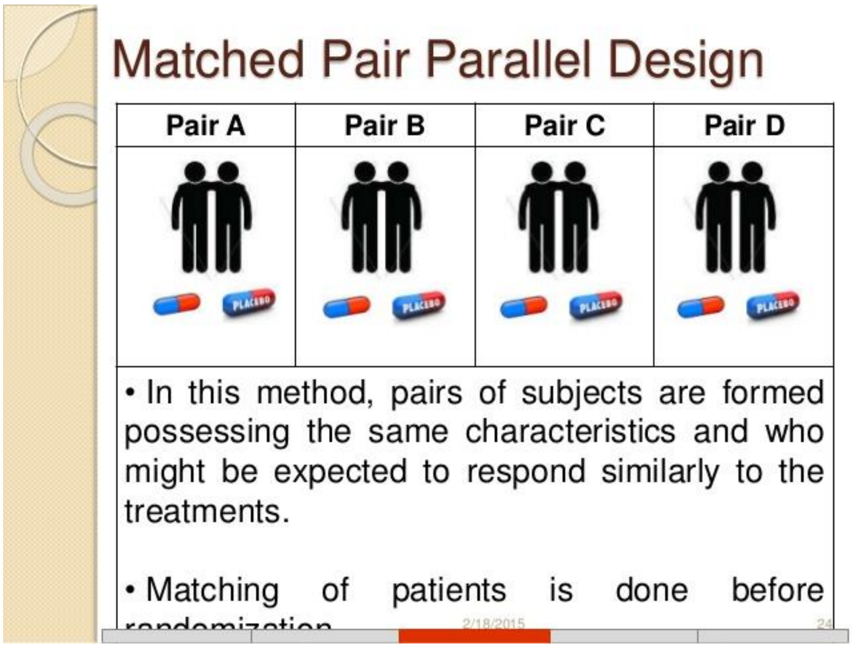
$$= \Pr(Z > \frac{\bar{x}_{obs} - \mu_0}{\sigma / \sqrt{n}}) = \Pr(Z > \frac{\sqrt{n}(\bar{x}_{obs} - \mu_0)}{\sigma})$$

Small $\bar{x}_{obs} - \mu_0$, but very large $\sqrt{n}$ → small p value → reject $H_0: \mu = \mu_0$

Large $\bar{x}_{obs} - \mu_0$, but very small $\sqrt{n}$ → large p value → fail to reject H_0

Topic: “matched pairs” procedures – 用單樣本的方法做雙樣本的比較

目的：去掉個體的異質性



Example: Coke versus Pepsi

Pepsi wanted to demonstrate that Coke drinkers prefer Pepsi when they taste both colas blind.

The subjects, all of whom said they were Coke drinkers tasted both colas from glasses without brand markings.

Since the response may depend on which cola is tasted first, the order of tasting should be chosen at random for each subject.

When more than half the Coke drinkers chose Pepsi, Coke claimed that the experiment was biased.

Pepsi glasses were marked M and Coke glasses were marked Q. Coke claimed that the results could just mean that people like the letter M better than the letter Q.

Matched pairs design is OK, but any distinction other than Coke vs Pepsi should have been avoided.

10 個品嚐員嚐可樂兩次 (樣本一：新鮮；樣本二：放了一個月)

因為是“同一個人”紀錄兩次分數，當紀錄差值時，可視為單樣本.

Ex2:

香味和學習. 受測者做兩次實驗, 一次戴有香精的口罩, 一次戴沒有味道的, 考試. 兩次實驗的次序經過 randomization (目的: 去除學習效應).

Original data: $(X_i, Y_i)(i=1,...,n)$ → take two measurements from the same subject

Data based on difference: $D_i = Y_i - X_i \quad (i=1,...,n)$

Assumption on $D_i \sim N(\mu_D, \sigma_D^2)$

Hypothesis testing

$H_0 : \mu_D = 0$ vs. H_a : three possible cases → same as before

Distribution theory: $\dfrac{\bar{D}-0}{S_D/\sqrt{n}} \sim^{H_0} t_{n-1}$

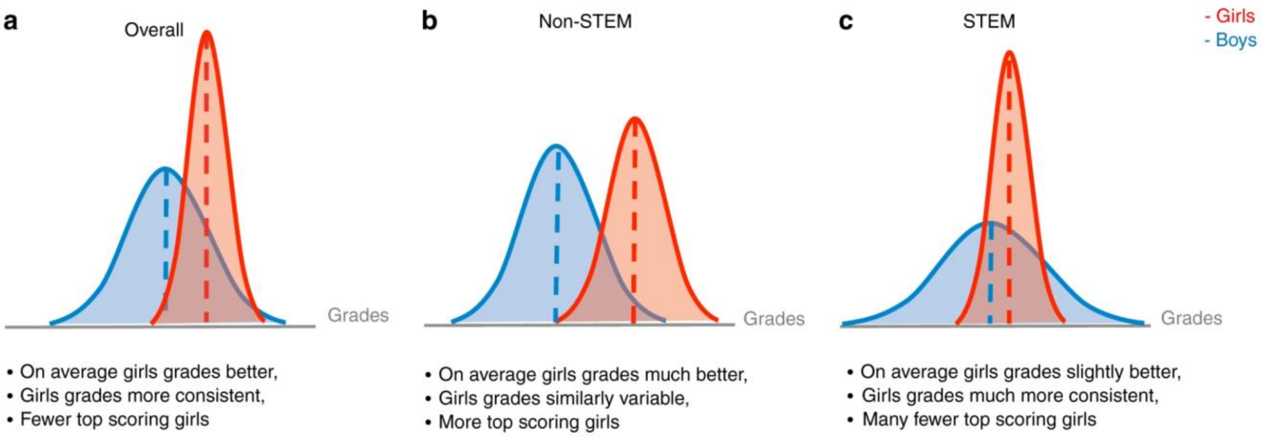
where $\sum_{i=1}^{n} D_i / n = \bar{D}$ and $\sum_{i=1}^{n} (D_i - \bar{D})^2 / (n-1) = S_D^2$

Topic: Two-sample Comparison

Assume: sample 1 and sample 2 are independent $X_{1i} \perp X_{2j} \quad \forall i, j$

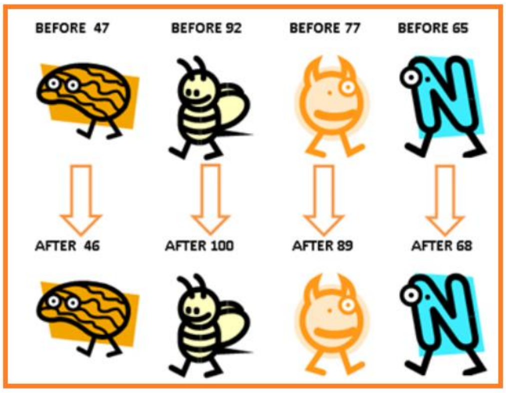
Example: compare gender difference

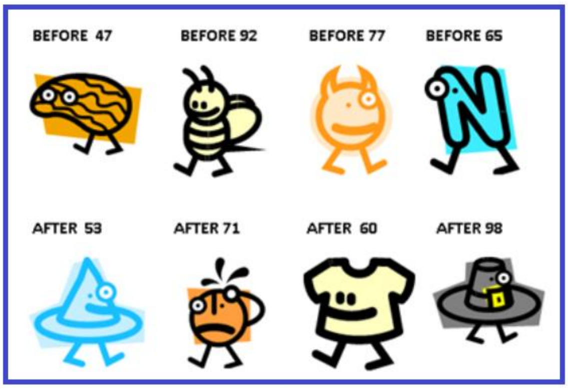
From: Gender differences in individual variation in academic grades fail to fit expected patterns for STEM



Compare matched-pair design and two-sample study

左: matched pair 右: two sample (兩個樣本無關係)





Hypothesis Testing for two-sample comparison

Case 1: $H_0 : \mu_1 - \mu_2 = 0$ versus $H_a : \mu_1 - \mu_2 > 0$

Case 2: $H_0 : \mu_1 - \mu_2 = 0$ versus $H_a : \mu_1 - \mu_2 < 0$

Case 3: $H_0 : \mu_1 - \mu_2 = 0$ versus $H_a : \mu_1 - \mu_2 \neq 0$

Construction of the test statistic:

1. Use $\bar{X}_1 - \bar{X}_2$ as the point estimator of $\mu_1 - \mu_2$
2. The distribution theory of $\bar{X}_1 - \bar{X}_2$ when the two populations are normal

$$E(\bar{X}_1 - \bar{X}_2) = E(\frac{X_{11} + ... + X_{1n_1}}{n_1}) - E(\frac{X_{21} + ... + X_{2n_2}}{n_2}) = \mu_1 - \mu_2$$

$$Var(\bar{X}_1 - \bar{X}_2) = Var(\frac{X_{11} + ... + X_{1n_1}}{n_1}) - 2Cov(\bar{X}_1, \bar{X}_2) + Var(\frac{X_{21} + ... + X_{2n_2}}{n_2})$$

$$= Var(\frac{X_{11} + ... + X_{1n_1}}{n_1}) + Var(\frac{X_{21} + ... + X_{2n_2}}{n_2}) \text{ (因為兩樣本獨立)}$$

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, Var(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1})$$

Under $H_0 : \mu_1 = \mu_2$

$$\bar{X}_1 - \bar{X}_2 \sim N(0, Var(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1}) \rightarrow \frac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}} \sim N(0,1)$$

Two-sample Z test – when σ_1^2 and σ_2^2 are known

Case 1: $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 - \mu_2 > 0$

method 1: reject H_0 if $obs\{\bar{X}_1 - \bar{X}_2\} > 0 + z_\alpha \cdot SE(\bar{X}_1 - \bar{X}_2)$,

$\Leftrightarrow$ reject if $obs\{\bar{X}_1 - \bar{X}_2\} > 0 + z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

method 1: p-value = $\Pr(Z > \frac{obs\{\bar{X}_1 - \bar{X}_2\} - 0}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}})$

reject if p-value < α

Case 2: $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 - \mu_2 < 0$

method 1: reject H_0 if $obs\{\bar{X}_1 - \bar{X}_2\} < 0 - z_\alpha \cdot SE(\bar{X}_1 - \bar{X}_2)$,

$\Leftrightarrow$ reject if $obs\{\bar{X}_1 - \bar{X}_2\} < 0 - z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

method 2: p-value = $\Pr(Z < \dfrac{obs\{\bar{X}_1 - \bar{X}_2\}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}})$ → reject if p-value < α

Case 3: $H_0: \mu_1 = \mu_2$ versus $H_a: \mu_1 - \mu_2 \neq 0$

method 1: reject H_0 if $\left|\dfrac{obs\{\bar{X}_1 - \bar{X}_2\}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}\right| > z_{\alpha/2}$

method 2: p-value $= 2\Pr(Z > \left|\dfrac{obs\{\bar{X}_1 - \bar{X}_2\}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}\right|)$

reject if p-value < α

Case 3 (method 3): $(1-\alpha)\cdot 100\%$ 信賴區間 for $\mu_1 - \mu_2$:

$$obs\{\bar{X}_1 - \bar{X}_2\} \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Rule: 0 falls outside $obs\{\bar{X}_1 - \bar{X}_2\} \pm z_{\alpha/2} \cdot \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$, then reject H_0

Example: Case 1

Example

The alkalinity, in milligrams per litre, of water in the upper reaches of rivers in a particular region is known to be normally distributed with a standard deviation of 10 mg/l. Alkalinity readings in the lower reaches of rivers in the same region are also known to be normally distributed, but with a standard deviation of 25 mg/l.

Ten alkalinity readings are made in the upper reaches of a river in the region and fifteen in the lower reaches of the same river with the following results.

Upper reaches	91	75	91	88	94	63	86	77	71	69
Lower reaches	86	95	135	121	68	64	113	108	79	62
	143	108	121	85	97					

, the claim that the true mean alkalinity of water in the lower reaches of this river is greater than that in the upper reaches.

Sample 1: Lower $\sigma_1 = 25$, $\bar{x}_{1,obs} = 99.0$, $n_1 = 15$

Sample 2: Upper $\sigma_2 = 10$, $\bar{x}_{2,obs} = 80.5$, $n_2 = 10$

$H_0 : \mu_1 = \mu_2$ vs. $H_0 : \mu_2 < \mu_1$

Reject if $obs\{\bar{X}_1 - \bar{X}_2\} > 0 + z_\alpha \cdot SE(\bar{X}_1 - \bar{X}_2) = 1.645 * 7.18 = 11.81$

$$SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{51.67} = 7.18$$

Since $\bar{x}_{1,obs} - \bar{x}_{2,obs} = 18.5 > 11.81$ → reject $H_0 : \mu_1 = \mu_2$

p-value $= \Pr(Z > \frac{99.0 - 80.5}{\sqrt{25^2 / 15 + 10^2 / 10}}) = \Pr(Z > 2.57) = 0.0051$

→ reject $H_0 : \mu_1 = \mu_2$

Example: case 3

The amount of a certain trace element in blood is known to vary with a standard deviation of 14.1 ppm (parts per million) for male blood donors and 9.5 ppm for female donors. $\sigma_1 = 14.1$ and $\sigma_2 = 9.5$

Random samples of 75 male and 50 female donors yield concentration means of 28 and 33 ppm, respectively. $n_1 = 75$ and $n_2 = 50$

$\bar{x}_{1,obs} = 28$, $\bar{x}_{2,obs} = 33$

Please test the population means of concentrations of the element are the same for men and women?

p-value $= 2\Pr(Z > |\frac{28 - 33}{\sqrt{14.1^2 / 75 + 9.5^2 / 50}}|) = 2\Pr(Z > 1.1221) = 2*(1-0.8686)=0.26$

Fail to reject $H_0 : \mu_1 = \mu_2$

Two-sample T test

when σ_1^2 and σ_2^2 are both unknown

Fact: When the two populations are normal (即使是常態母體，也不是真的 t 分配)

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \sim^{approximately} T_{v^*}, \text{ where } v^* = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(S_2^2/n_2\right)^2}{n_2 - 1}}$$

Option 1: 使用最精確的近似 計算 v^* (非整數，所以需要軟體)

Option 2: 保守原則 (當沒有軟體的時候)

Conservative principle: choose $df = \min(n_1 - 1, n_2 - 1)$ → **訂更嚴格的標準**

Since $df < v^*$, make it less likely to reject $H_0: \mu_1 = \mu_2$ using Option 2.

Case 1: $H_0: \mu_1 = \mu_2$ versus $H_a: \mu_1 - \mu_2 > 0$

method 1: reject H_0 if $obs\{\bar{X}_1 - \bar{X}_2\} > 0 + t_{df,\alpha} \cdot \hat{S}E(\bar{X}_1 - \bar{X}_2)$,

$$\Leftrightarrow \text{ reject if } obs\{\bar{X}_1 - \bar{X}_2\} > 0 + t_{df,\alpha}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

method 1: p-value = $\Pr(T_{df} > \frac{obs\{\bar{X}_1 - \bar{X}_2\} - 0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}})$

reject if p-value < α

Case 2: $H_0: \mu_1 = \mu_2$ versus $H_a: \mu_1 - \mu_2 < 0$

method 1: reject H_0 if $obs\{\bar{X}_1 - \bar{X}_2\} < 0 - t_{df,\alpha,} \cdot \hat{S}E(\bar{X}_1 - \bar{X}_2)$,

$$\Leftrightarrow \text{ reject if } obs\{\bar{X}_1 - \bar{X}_2\} < 0 - t_{df,\alpha}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

method 2: p-value = $\Pr(T_{df} < \frac{obs\{\bar{X}_1 - \bar{X}_2\}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}})$

reject if p-value < α

Case 3: $H_0: \mu_1 = \mu_2$ versus $H_a: \mu_1 - \mu_2 \neq 0$

method 1: reject H_0 if $\left| \dfrac{obs\{\bar{X}_1 - \bar{X}_2\}}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \right| > t_{df,\alpha/2}$

method 2: p-value $= 2\Pr(T_{df} > \left| \dfrac{obs\{\bar{X}_1 - \bar{X}_2\}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \right|)$

reject if p-value < α

Case 3 (method 3): $(1-\alpha)\cdot 100\%$ 信賴區間 for $\mu_1 - \mu_2$:

$$obs\{\bar{X}_1 - \bar{X}_2\} \pm t_{df,\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Rule: 0 falls outside $obs\{\bar{X}_1 - \bar{X}_2\} \pm t_{df,\alpha/2} \cdot \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$, then reject H_0

Option 3: 假設 $\sigma_1 = \sigma_2 = \sigma$ → 有些書的作者不建議

$$Var(\bar{X}_1 - \bar{X}_2) = Var\left(\frac{X_{11} + \ldots + X_{1n_1}}{n_1}\right) + Var\left(\frac{X_{21} + \ldots + X_{2n_2}}{n_2}\right)$$

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \quad = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim N(0,1)$$

For estimating σ, use $S_{pooled} = \sqrt{\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2/(n_1-1) + \sum_{j=1}^{n_2}(X_{2j} - \bar{X}_2)^2/(n_2-1)}$

Conclusion: If two samples are from two independent normal distributions with the same variance:

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_{pooled}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim^{exactly} t_{n_1+n_2-2}$$

nature

View all Nature Research journals | Search | Login

Explore our content | Journal information | Subscribe | Sign up for alerts | RSS feed

nature > editorials > article

EDITORIAL · 20 MARCH 2019

It's time to talk about ditching statistical significance

Looking beyond a much used and abused measure would make science harder, but better.

Some statisticians are calling for *P* values to be abandoned as an arbitrary threshold of significance. Credit: Erik Dreyer/Getty

Fans of The Hitchhiker's Guide to the Galaxy know that the answer to life, the Universe and everything is 42. **The joke, of course, is that truth cannot be revealed by a single number.**

And yet this is the job often assigned to P values: a measure of how surprising a result is, given assumptions about an experiment, including that no effect exists. *Whether a P value falls above or below an arbitrary threshold demarcating 'statistical significance' (such as 0.05) decides whether hypotheses are accepted, papers are published and products are brought to market.* But **using P values as the sole arbiter of what to accept as truth can also mean that some analyses are biased, some false positives are overhyped and some genuine effects are overlooked. 不應該只看 p-value!**

Change is in the air. In a Comment in this week's issue, **three statisticians call for scientists to abandon statistical significance.** The authors do not call for P values themselves to be ditched as a statistical tool — rather, **they want an end to their use as an arbitrary threshold of significance. More than 800 researchers have added their names as signatories.** A series of related articles is being published by the American Statistical Association this week (R. L. Wasserstein et al. Am. Stat. https://doi.org/10.1080/00031305.2019.1583913; 2019). "The tool has become the tyrant (暴君)," laments one article.

Statistical significance is so deeply integrated into scientific practice and evaluation that extricating it would be painful. Critics will counter that arbitrary gatekeepers are better than unclear ones, and that the more useful argument is over which results should count for (or against) evidence of effect. There are reasonable viewpoints on all sides;
Nature is not seeking to change how it considers statistical analysis in evaluation of papers at this time, but we encourage readers to share their views.

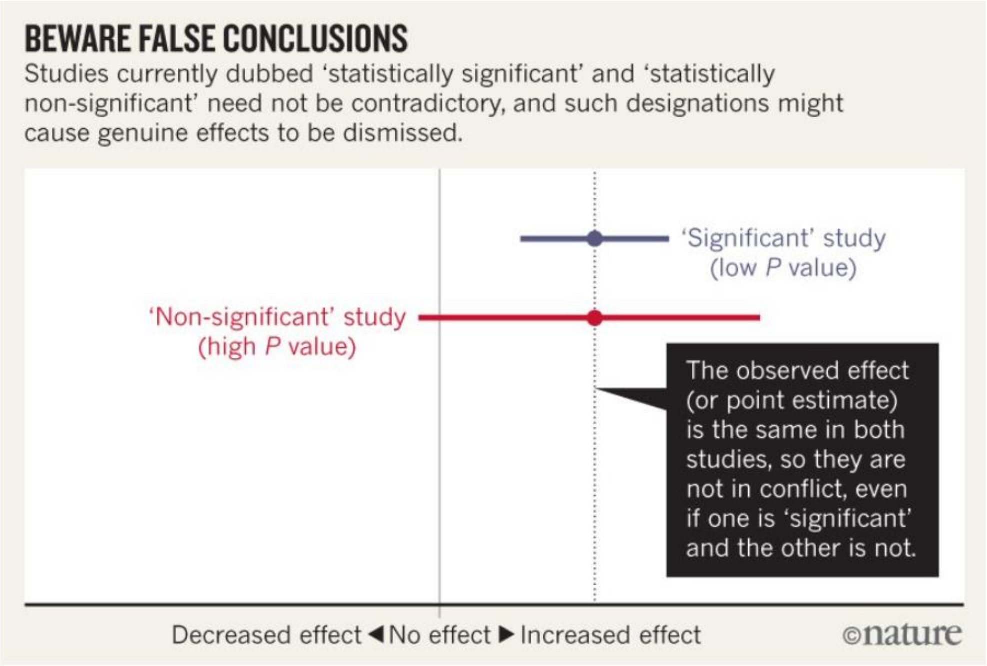
If researchers do discard statistical significance, what should they do instead? They can start by educating themselves about statistical misconceptions. Most important will be the courage to consider uncertainty from multiple angles in every study. Logic, background knowledge and experimental design should be considered alongside P values and similar metrics to reach a conclusion and decide on its certainty.

When working out which methods to use, **researchers should also focus as much as possible on actual problems.** People who will duel to the death over abstract theories on the best way to use statistics often agree on results when they are presented with concrete scenarios.

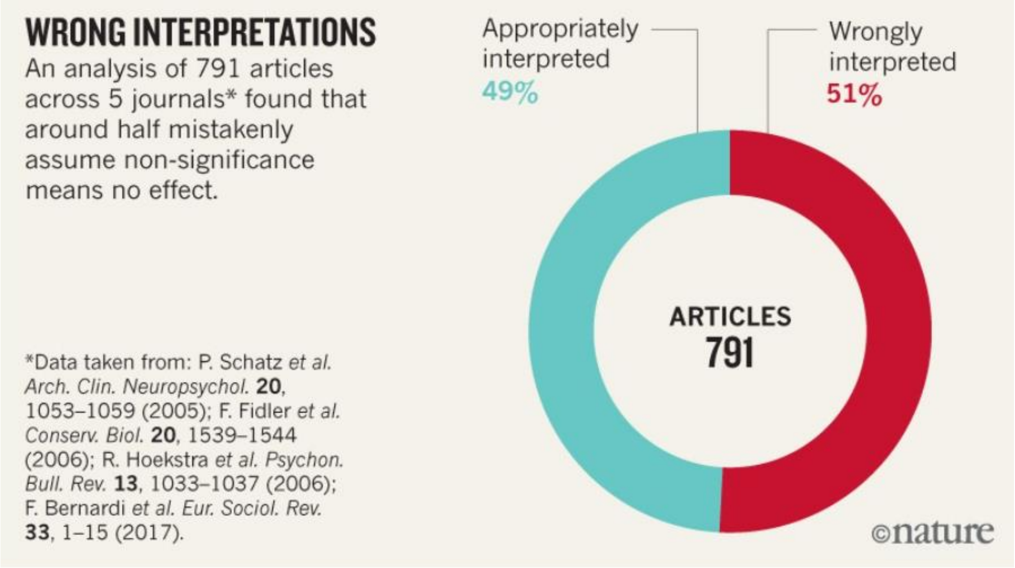
Researchers should seek to analyse data in multiple ways to see whether different analyses converge on the same answer. Projects that have crowdsourced analyses of a data set to diverse teams suggest that this approach can work to validate findings and offer new insights.

In short, be sceptical, pick a good question, and try to answer it in many ways. **It takes many numbers to get close to the truth.**

https://www.nature.com/articles/d41586-019-00857-9



Source: V. Amrhein *et al.*



Source: V. Amrhein *et al.*