




## 第一节 假设检验

-  假设检验的基本思想和方法
-  假设检验的一般步骤
-  假设检验的两类错误



假设：我们想要去检验的前提或主张。

在假设检验中，我们把一个受关注的问题转化为针对一个或多个参数的假设。

Hypothesis - A premise or claim that we want to test.

In hypothesis testing, we turn a question of interest into hypotheses about the value of a parameter or parameters.

We create:

Sometimes referred to as the status quo hypothesis

Default/  
established

默认/现状

► A null hypothesis, denoted by  $H_0$

有时被称为现状假设

原假设，用 $H_0$ 表示。

► An alternative hypothesis, denoted by  $H_a$

备择假设，用 $H_a$ 表示。有时被称为研究假设

Sometimes referred to as the research hypothesis

Want to  
challenge  $H_0$

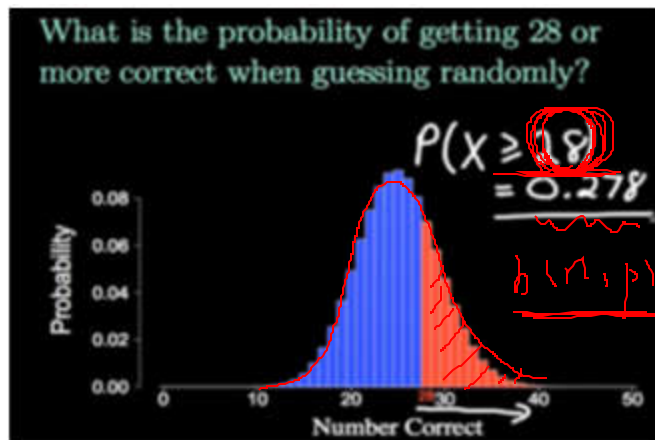
想挑战现状

We calculate an appropriate test statistic (based on sample data), and determine how much evidence there is against the null hypothesis.

If the evidence is strong enough (if it meets a certain significance level), we can reject the null hypothesis in favour of the alternative hypothesis.  $H_1$

计算一个合适的检验统计量（基于样本数据），并确定有多少证据反对原假设。如果证据足够强（如果它满足一定的显著性水平），我们可以拒绝原假设，从而支持备择假设。

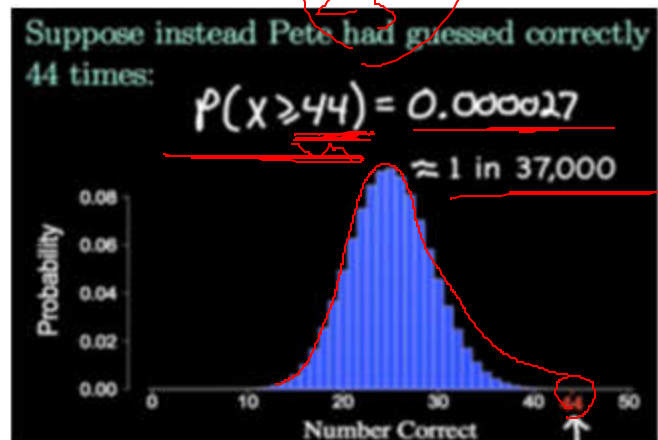
当随机猜测时，正确次数大于等于28的概率是多少？



$$H_0: p = 1/4 \quad \checkmark$$

$$H_1: p \neq 1/4$$

假设Peter答对了44次:



X  
✓



例如：男女大学生毕业后的平均工资不同吗？

Example:  
 Do men and women have different average salaries after graduating university?

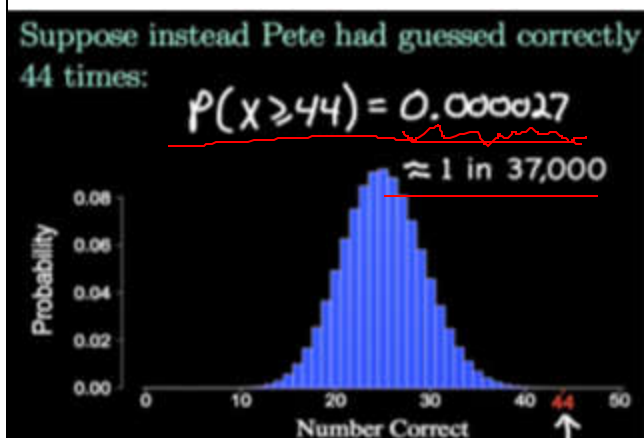
~~$H_0: \bar{x}_M = \bar{x}_F$~~        $H_0: \underline{\mu}_M = \underline{\mu}_F$

$H_a: \mu_M \neq \mu_F \leftarrow$

We make hypotheses about parameters,  
 \*never\* about statistics.

假设是关于参数的，而不是统计量的





拒绝/否定 $H_0$ 并不是说 $H_0$ 一定错，而只是说差异到了一定的显著程度，拒绝 $H_0$ 比较合理。

反之，接受 $H_0$ 并不是肯定 $H_0$ 一定对，而只是说差异还不够显著，还没有达到足以否定 $H_0$ 的程度。

所以假设检验又叫

“显著性检验”



## 二、假设检验的一般步骤

在上面的例子的叙述中，我们已经初步介绍了假设检验的基本思想和方法。

下面，我们再结合另一个例子，进一步说明假设检验的一般步骤。

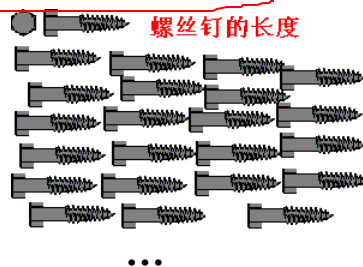


例2 某工厂生产的一种螺钉，标准要求长度是32.5毫米. 实际生产的产品，其长度 $X$ 假定服从正态分布  $N(\mu, \sigma^2)$ ,  $\sigma^2$  未知，现从该厂生产的一批产品中抽取6件，得尺寸数据如下：

32.56, 29.66, 31.64, 30.00, 31.87, 31.03

问这批产品是否合格？

分析：这批产品(螺钉长度)的全体组成问题的总体 $X$ . 现在要检验 $\mu$ 是否为32.5.



已知  $X \sim N(\mu, \sigma^2)$ ,  $\sigma^2$  未知.

第一步： 提出原假设和备择假设

$$H_0: \mu = 32.5 \Leftrightarrow H_1: \mu \neq 32.5$$

第二步： 取一检验统计量，在 $H_0$ 成立下  
求出它的分布

$$t = \frac{\bar{X} - 32.5}{S/\sqrt{6}} \sim t(5)$$

能衡量差异大小且分布已知



第三步:

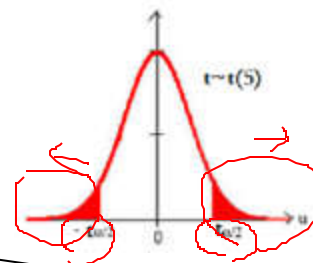
对给定的显著性水平  $\alpha = 0.01$ , 查表确定  
临界值  $t_{\alpha/2}(5) = t_{0.005}(5) = 4.0322$ , 使

$$P\{|t| > t_{\alpha/2}(5)\} = \alpha$$

即 “ $|t| > t_{\alpha/2}(5)$ ” 是一个小概率事件.

得否定域

$$W: |t| > 4.0322$$



小概率事件在一次  
试验中基本上不会  
发生.

A rejection region (a.k.a. critical region) in a Null Hypothesis Statistical Test is a part of the parameter space such that observing a result that falls under it will lead to the rejection of the null hypothesis.

在原假设统计检验中, 拒绝域 (又称临界区域) 是参数空间的一部分, 当观察到低于该区域的结果时, 将导致拒绝原假设。

得否定域  $W: |t| > 4.0322$

第四步:

将样本值代入算出统计量  $t$  的实测值,

$$|t| = 2.997 < 4.0322$$

没有落入  
拒绝域

故不能拒绝  $H_0$ .

上述利用  $t$  统计量得出得检验法称为  $t$  检验法。在实际中，正态总体的方差常为未知，所以我们常用  $t$  检验法来检验关于正态总体均值的检验问题。


与区间估计对比：

```
[m, l, u] = t_mean_confidence_interval(a, 0.99, 2)
print("Two-sided [(), ()]".format(l, u))

[t, p] = scipy.stats.ttest_1samp(a, 32.5)
print("1 sample t-test: t=(), p=()".format(t, p))
```

Two-sided [29.27885774126866, 32.97447559206467]  
1 sample t-test: t=-2.9967797470687896, p=0.030210851754886893

*Handwritten notes: A red 'p' is written next to the first print statement. A red arrow points from the 'p' in the second print statement to the 'p' in the output line. A red '0.03' is written next to the p-value in the output line.*



SPSS：分析 > 比较均值 > 单样本T检验





## 两个正态总体均值差的检验（ $t$ 检验）

我们还可以用 $t$ 检验法检验具有相同方差的两个正态总体均值差的假设。

例2 在平炉进行一项试验以确定改变操作方法的 建议是否会增加钢的得率，试验是在同一只平炉上进行的。每炼一炉钢时除操作方法外，其它条件都尽可能做到相同。先用标准方法炼一炉，然后用建议的新方法炼一炉，以后交替进行，各炼了10炉，其得率分别为

标准方法 78.1 72.4 76.2 74.3 77.4 78.4 76.0 75.5 76.7 77.3

新方法 79.1 81.0 77.3 79.1 80.0 79.1 79.1 77.3 80.2 82.1



$$\frac{X - Y - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

$$t_{0.05}(18) = 1.7341,$$

故拒绝域为

$$t = \frac{\bar{x} - \bar{y}}{s_w \sqrt{\frac{1}{10} + \frac{1}{10}}} \leq -t_{0.05}(18) = -1.7341,$$

现在由于样本观察值 $t = -4.295 < -1.7341$ , 所以拒绝  $H_0$ ,

即认为建议的新操作方法较原来的方法为优。



与区间估计对比:

```

a1 = [78.1, 72.4, 76.2, 74.3, 77.4, 78.4, 76.0, 75.5, 76.7, 77.3]
a2 = [79.1, 81.0, 77.3, 79.1, 80.0, 79.1, 79.1, 77.3, 80.2, 82.1]

[m, l, u] = mean_diff_confidence_interval v2(a1, a2) 95%
print("Two-sided [l, u]".format(l, u))

# equal_var : bool, optional
# If True (default), perform a standard independent 2 sample test that assumes equal population var.
t, p = scipy.stats.ttest_ind(a1, a2, equal_var=True)
print("ttest_ind: t = %g p = %g" % (t, p))

```

Two-sided [-4.765026326722848, -1.6349736732771865]  
ttest\_ind: t = -4.29574 p = 0.000435185

Handwritten notes: *paired*, *M1 ≠ M2*, *0*, *0.05*

## 假设检验的p值法 (The p-value method)

### p-value:

A p-value, or probability value, is a number describing how likely it is that your data would have occurred by random chance (i.e. that the null hypothesis is true).

p值, 也称概率值, 是一个数字, 描述数据随机发生的可能性 (即原假设为真)

The level of statistical significance is often expressed as a p-value between 0 and 1. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

统计显著性水平通常用0到1之间的p值表示。p值越小, 拒绝原假设的证据就越充分

当原假设成立时, 检验统计量至少比观察到的检验统计量更极端 (离谱) 的概率:

$p < 0.05$  \* 显著

$p < 0.01$  \*\* 极其显著

或 原假设可被拒绝的最小显著性水平。

## 假设检验的p值法 (The p-value method)

### p-value :

A p-value, or probability value, is a number describing how likely it is that your data would have occurred by random chance (i.e. that the null hypothesis is true).

The level of statistical significance is often expressed as a *p*-value between 0 and 1. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

当原假设成立时，检验统计量至少比观察到的检验统计量更极端（离谱）的概率：

$p < 0.05$  \* 显著  
 $p < 0.01$  \*\* 极其显著

或 原假设可被拒绝的最小显著性水平。



### p-hacking

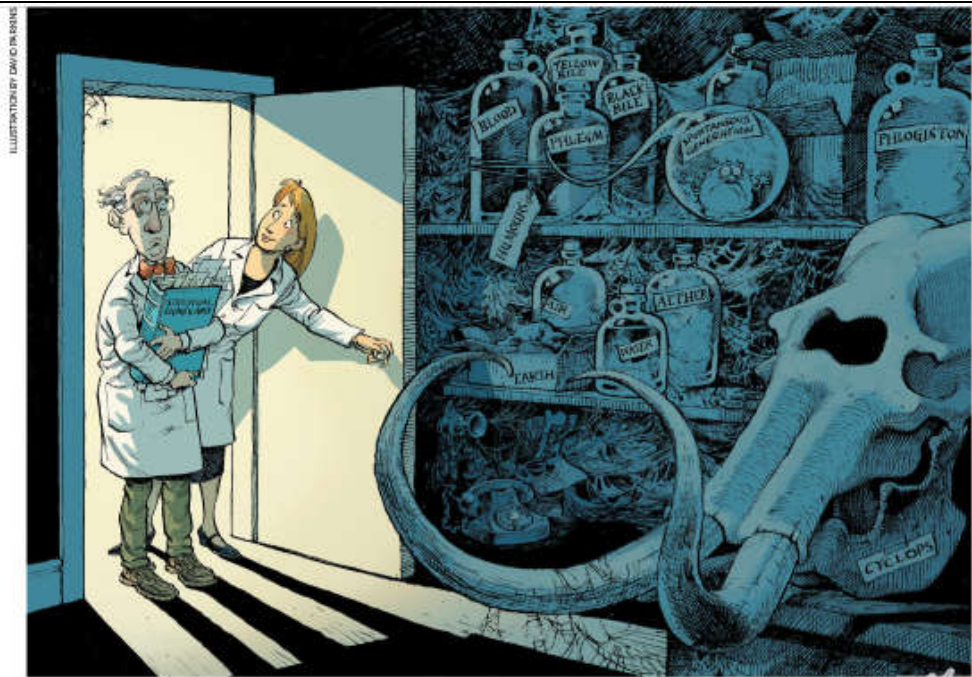
在追逐超低 p-value 的背景下，学者在面临这些选择做决定时会“非常微妙”，一切阻碍超低 p-value 诞生的数据都会被巧妙的避开。Dr. Harvey 将为了追求超低 p-value 而在因子研究中刻意选取的数据处理方法称为 p-hacking。

为了角逐在顶级期刊上发表文章，学者们过度追求因子在原假设下的低 p-value 值（即统计意义上“显著”）。

学者们在追逐 p-value 的道路上狂奔，却在科学的道路上渐行渐远。



## Discussion



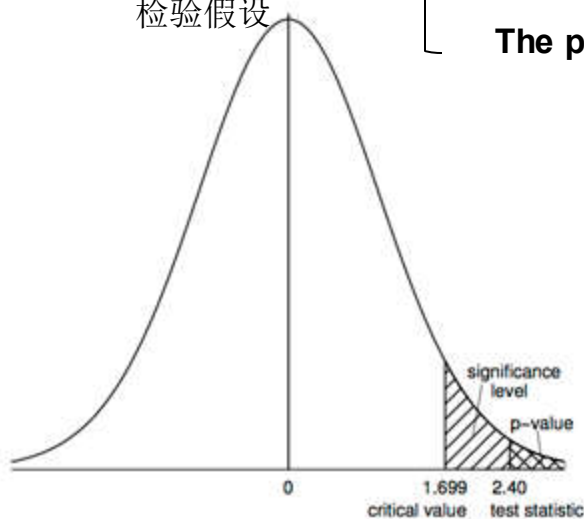
Retire statistical significance

### 小结：假设检验的2种方法

**Hypothesis Testing**  
检验假设

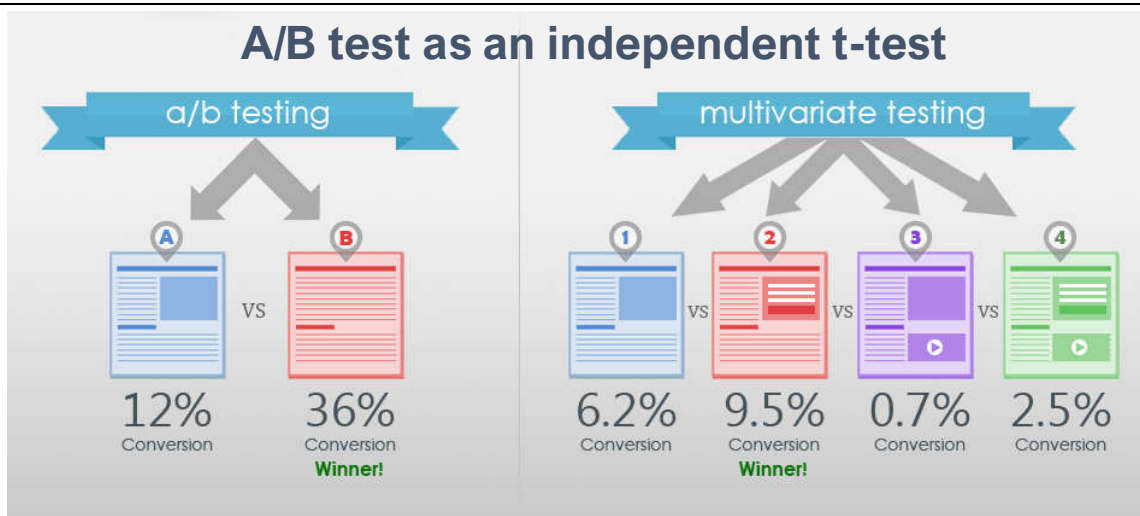
**The rejection region method** 拒绝域法

**The p-value method** P值法



SPSS: 分析 > 比较均值 > 独立样本T检验

SPSS: 分析 > 比较均值 > 配对样本T检验



A/B 测试是一种产品优化的方法，为同一个优化目标制定两个方案（比如两个页面），让一部分用户使用A 方案（称为控制组或对照组），同时另一部分用户使用 B 方案（称为变化组或试验组），统计并对比不同方案的转化率、点击量、留存率等指标，以判断不同方案的优劣并进行决策。

A/B测试的本质：

A/B测试中是用对照版本和试验版本这两个样本的数据来对两个总体是否存在差异进行检验，其本质是使用假设检验中的独立样本t检验。



除了以上对正态总体均值的假设检验，  
还可以对方差进行假设检验

利用以下检验统计量：

单个总体

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

多个总体

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$$



$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$$



方差齐次检验

ANOVA

在方差分析的F检验中，是以各个实验组内总体方差齐性为前提的，因此，按理应该在方差分析之前，要对各个实验组内的总体方差先进行齐性检验。如果各个实验组内总体方差为齐性，而且经过F检验所得多个样本所属总体平均数差异显著，这时才可以将多个样本所属总体平均数的差异归因于各种实验处理的不同所致；如果各个总体方差不齐，那么经过F检验所得多个样本所属总体平均数差异显著的结果，可能有一部分归因于各个实验组内总体方差不同所致。

简单地说就是在进行两组或多组数据进行比较时，先要使各组数据符合正态分布，另外就是要使各组数据的方差相等（齐性）。



### 三、假设检验的两类错误

假设检验会不会犯错误呢？

由于作出结论的依据是下述

小概率原理

不是一定不发生

小概率事件在一次试验中基本上不会发生。



如果 $H_0$ 成立，但统计量的实测值落入否定域，从而作出否定 $H_0$ 的结论，那就犯了“以真为假”的错误。

如果 $H_0$ 不成立，但统计量的实测值未落入否定域，从而没有作出否定 $H_0$ 的结论，即接受了错误的 $H_0$ ，那就犯了“以假为真”的错误。





## 假设检验的两类错误

	实际情况	
决策	$H_0$ 为真	$H_0$ 不真
接受 $H_0$	正确	第二类错误
拒绝 $H_0$	第一类错误	正确

Decision is:	The Null Hypothesis is	
	True	False
Accept $H_0$	(1- $\alpha$ ) Confidence Level	$\beta$
Reject $H_0$	$\alpha$	(1- $\beta$ ) Power of the test

犯两类错误的概率:

$$P\{\text{拒绝}H_0|H_0\text{为真}\} = \alpha,$$

$$P\{\text{接受}H_0|H_0\text{不真}\} = \beta.$$

显著性水平 $\alpha$ 为犯第一类错误的概率.



## 一类错误(type I error)和二类错误(type II error)

**Type I error**  
(false positive)



**Type II error**  
(false negative)





## 两类错误的概率的关系

两类错误是互相关联的，当样本容量固定时，一类错误概率的减少导致另一类错误概率的增加。

要同时降低两类错误的概率  $\alpha, \beta$  或者要在  $\alpha$  不变的条件下降低  $\beta$ ，需要增加样本容量。



		假设为真 Hypothesis true	假设为假 Hypothesis false
接受假设 Accept hypothesis		Type II error 取伪	
拒绝假设 Reject hypothesis		Type I error 弃真	



通常情况：控制1类错误，不关心2类错误

Many textbooks and instructors will say that the Type 1 (false positive) is worse than a Type 2 (false negative) error. The rationale boils down to the idea that if you stick to the status quo or default assumption, at least you're not making things worse.

许多教科书和教师会说，类型1（弃真）比类型2（取伪）错误更糟糕。其基本原理可以归结为这样一种（保守的）观点：如果你坚持现状或默认假设，至少你不会让事情变得更糟。

实际情况下，应评估两类错误的现实代价，以权衡1类和2类。

对于2类错误（ $\beta$ ），该如何度量和控制？ **Power Analysis !**

$1 - \beta$



通常情况：控制1类错误，不关心2类错误

Many textbooks and instructors will say that the Type 1 (false positive) is worse than a Type 2 (false negative) error. The rationale boils down to the idea that if you stick to the status quo or default assumption, at least you're not making things worse.

许多教科书和教师会说，类型1（弃真）比类型2（取伪）错误更糟糕。其基本原理可以归结为这样一种（保守的）观点：如果你坚持现状或默认假设，至少你不会让事情变得更糟。

实际情况下，应评估两类错误的现实代价，以权衡1类和2类。

对于2类错误（ $\beta$ ），该如何度量和控制？ **Power Analysis !**

$1 - \beta$



## 效应量 (effect size) (选学)

## Power & Effect Size

效能 / 势

Power

- The ability of a statistical test to detect a relationship or difference

效能：统计检验发现某种关系或差异的能力

- The probability of rejecting a null hypothesis when it is false – and therefore should be rejected.

效能：当一个原假设为假时拒绝它的概率——因此应该被拒绝

- Jacob Cohen is the father of power analysis

- Power is described as  $1 - \beta$

- Acceptable power is .80 or higher – some say > .70 is adequate and > .90 is excellent

- Power analysis is usually done a priori, but can also be done afterward - controversial

- Jacob Cohen是效能分析之父
- 把效能描述为 $1 - \beta$
- 可接受的效能为大于等于.80-有人认为大于.70是可接受的，并且大于.90是很优秀。
- 效能分析通常是事先进行的，但也可以事后进行



如何提高效能

## How to improve power (APRONS)

- Power is a function of

1) alpha level

2) sample size (most dependent on this)

3) effect size

4) the type of statistical test being conducted

5) the type of design used

- Relax alpha level (.10 or .15)

- Use a parametric statistic

- Increase the reliability of the measure

- Use one tailed tests

- Increase the sample size (N)

- Increase the sensitivity of the design and or analysis

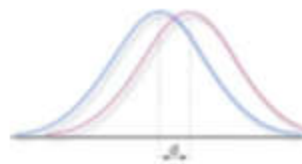
- 效能与以下因素有关：

- 1)  $\alpha$  水平
- 2) 样本容量（主要决定因素）
- 3) 效应量
- 4) 正在进行统计的统计检验的类型
- 5) 试验的设计

- 放宽  $\alpha$  的约束（取  $\alpha$  为.1或.15）
- 使用参数统计量
- 增加测量的可靠性
- 使用单边检验（one-sided/one-tailed）
- 增加样本容量（N）
- 提高设计或分析的灵敏度



## 效应量 Effect size



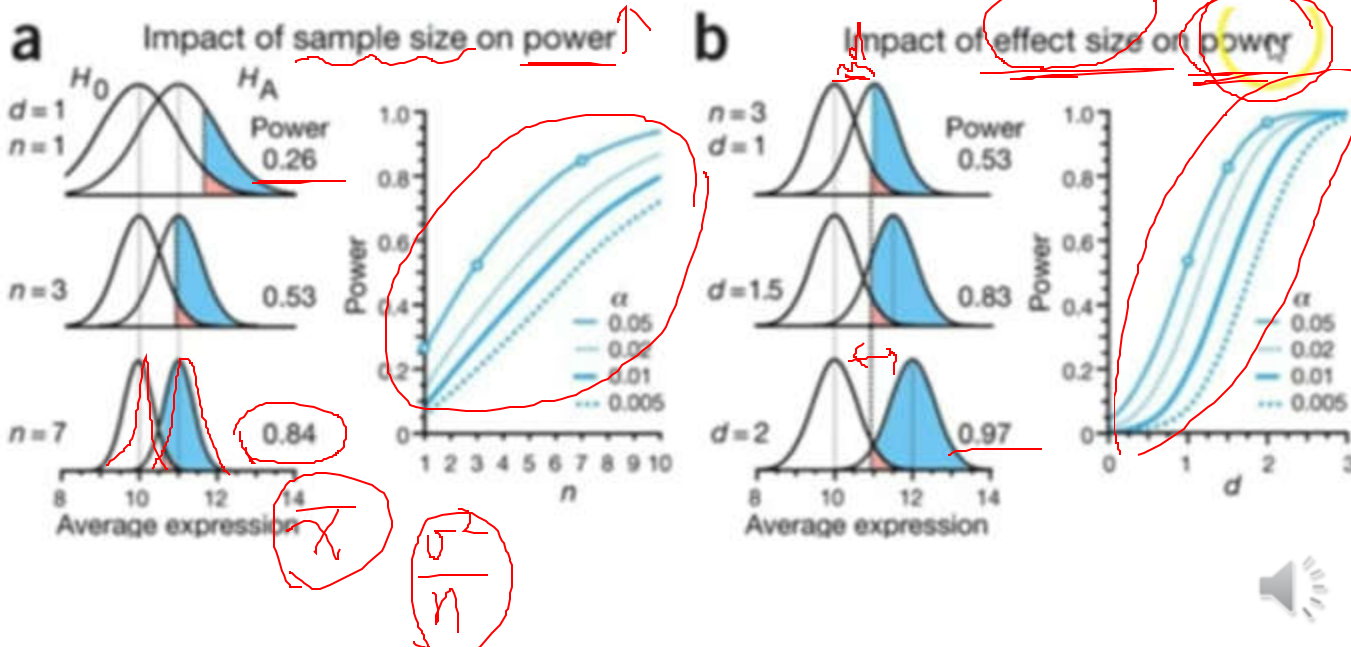
- Effect size is a quantitative measure of the *strength of a phenomenon*. 效应量是对现象强度的定量测量
- Effect size emphasizes the **size** of the difference or relationship 效应量强调区别或联系的强度/大小
- Examples:
  - the correlation between two variables (specifically  $r^2$ ) 相关分析中两个变量之间的相关性  $R^2$ 
    - $r=.1$  weak,  $r=.5$  moderate,  $r=.7$  strong,  $r=.9$  very strong
  - the regression coefficient in a regression ( $B_0, B_1, B_2$ ) 回归中的回归系数
    - Relative to model and field 与模型和领域相关
  - the mean differences in t tests (use Cohen's D) T检验中的均值差异
    - $d = .2$  is small;  $r = .5$  is medium;  $r = .8$  is large
  - The mean differences in ANOVA (use eta) 方差分析的平均差异 (使用  $\eta$ )
    - $.01$  is small,  $.06$  medium,  $.14$  large

Eta squared is a measure of effect size that is commonly used in ANOVA. Eta squared =  $SS_{\text{effect}} / SS_{\text{total}}$



样本容量对效能的影响

效应量对效能的影响



小结：

假设检验的原理  
拒绝域法 /  $p$ 值法  
两类错误  
效能分析（选学）

