Lecture 7

Power and sample size in *t*-tests and proportion tests

t检验和比例检验中的检验力和样本

Detecting true differences

检测真正的差异

假设两个组在变量均值上确实不同，或者两个群体之间的某些比例确实不同

- Suppose that two groups *truly* differ in a variable mean; or that some proportion truly differs between two populations
  - elephants and mice truly differ in mean weight
  - proportion of obese people truly differs between India and Argentina

为了证明两组之间存在显著差异，样本量应该多大？

- Question: to demonstrate that two groups significantly differ, how large should my sample size be if…
  - the species differ by 5,000 kg? or 1kg?
  - the proportion of obese people differs by least 50%? Or 5%
- Common sense suggests that: 组间的真实差异越小
  - the smaller the true difference between groups…
  - the larger the sample size required to detect it 检测他所需要的样本量越大

计算适当的样本量可以避免收集的数据小于和超出所需

- Calculating appropriate sample size avoids two problems:
  - collecting less data than needed to test a hypothesis
  - collecting more data than needed!

Types of error: Type I

错误地拒绝了一个"真"的零假设(无差异)

- Type I error: when you incorrectly reject a ‘true’ null hypothesis (of no difference)
  - test says that groups are different (P<0.05), but in fact they are similar
- It is a false positive 你认为发现了显著差异，但实际并不存在
  - You think you found a significant difference, but it does not really exist
- Probability of type I error is the significance level α
  - = probability of obtaining samples with significantly different means (P < 0.05) purely by chance

第一类错误的概率是显著性水平=纯粹偶然获得具有显著差异的样本的概率

Type II error 接受了一个错误的零假设

- Type II error is the opposite: it is when you accept a 'wrong' null hypothesis
  - test says that groups don't differ (test returns P > 0.05), but if fact they are different!

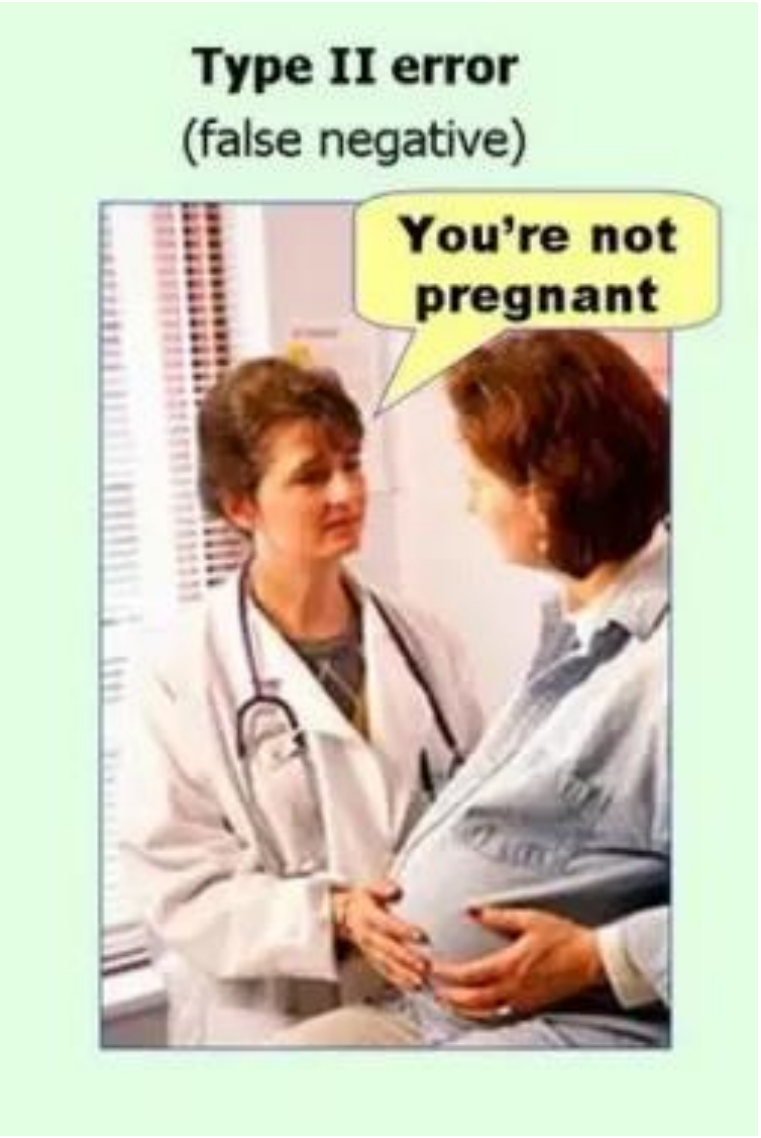
- Type II error is a false negative: the difference exists, but you are not able to see it

当测试不够powerful，无法识别真正差异时，就会发生二类错误

- Type II error occurs when your test is not *powerful* enough to identify true difference
  - test is 'myopic', or does not have enough resolution to detect that level of difference (or *effect size*) between the groups

效应大小：群体对变量的真实反映

- Effect size: the true effect of group on a variable
  - =the effect of species on weight
  - =the effect of country on being obese



Statistical power = test resolution

统计功效=测试分辨率

识别真正差异的能力，在差异检验中获得p<0.05的能力
拒绝错误零假设的能力，避免第二类错误的能力

- = power to identify a true difference
- = power to obtain P < 0.05 in a test of differences
- = power to reject a wrong null hypothesis
- = power to avoid Type II error

决定统计功效的因素

Factors determining statistical power

例如在t检验中，平均值之间的较大差异由较大的t值决定

- In a t-test for example, differences between means are determined by large t values

- Look at the formula: what makes t small and t-test 'short-sighted' or powerless? 自己的值和测试值之间差异很小
  - *Small effect size* (the difference between X and mu, or δ)
    - more difficult to detect true difference of 1 than 10
  - *Large standard deviations* sd很大
    - larger overlap between groups 组之间的重叠很大
  - *Small sample size* 样本量小
    - makes it more difficult to obtain a P <0.05

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

- Calculations of statistical power of a test must take all those factors into account

Statistical power

1-二类错误的概率

- Statistical power β = 1 - (probability of type II error)

- Example: If power of test is β=0.9=90%, chance of type II error is 0.1=10%

- Parameters determining power:
  - δ = delta or the real difference between sample means　delta或者样本均值之间的实际差值
  - σ = standard deviation　标准差
  - n = sample size　样本量
  - α = significance level (=0.05 by default; modified with “sig.level=0.01”)　显著性水平，默认0.05 可以用后面那个参数修改

统计能力至少80%，理想情况下90%

- You should design tests with power of at least 80%; ideally, 90%
  - Power of 80%: you have a chance of 4 in 5 of detecting a true difference between groups

- Notice that significance level (α) and statistical power (β) are different things
  - You want a power of 90% of obtaining P < α = 0.005

非中心t分布

Noncentral *t*-distribution

可以通过估计非中心参数V来调整t分布以计算第二类错误的概率

- It is possible to adapt the *t*-distribution to calculate probability of type II error by estimating the noncentral parameter *v* (noo)

V类似于t统计量，并且依赖于特定的检验

- *v* is similar to *t*-statistic and depends on the specific test (one-sample *t*-test, two-sample *t*-test, paired *t*-test, two-proportions test)
  - but generally speaking, it is the test statistic to estimate the probability of an effect size under a given sample size, confidence level (P value) and standard variation

但一般来说，他是在给定样本容量，置信水平(p值)和标准差的情况下，估计效应大小的概率的检验统计量

单样本t检验的功效

Power of one-sample *t*-test

确定组均值和检验值之间的真实差异或效应大小的概率是多少

- Question is: what is the probability of identifying a true difference or effect size δ between a group mean and a test value?
  - = what is the power of the test?

- In one-sample t-tests, noncentral parameter ν is

$$\nu = \frac{\delta}{\frac{\sigma}{\sqrt{n-1}}}$$

- Noncentral parameter ν is the difference δ between mean and test value divided by sem (standard error of mean

Power of one-sample *t*-test

- We use function *power.t.test*

> *power.t.test(delta, sd, n, power)*　　需要计算delta或者power，填剩下三个就行

Example:

- We have height measurements for 20 Agta women from the Philippines
- What is the power of a one-sample test to demonstrate that the height of Agta women truly differs by 5 cm (at least) from the mean height of a neighbouring population ?
  - (assume sd of height is 7)

Parameters:

- sample size: n=20
- effect size: δ=5 cm
- standard deviation: σ=7
- To run test, we enter the 3 parameters (in any order) and the 4th (=power β) is calculated

Power of one-sample *t*-test

- For one-sample t-test, add ***type="one.sample"*** 加上类型

```
>power.t.test(n=20, delta=5, sd=7, type="one.sample")

One-sample t test power calculation
  n = 20
  delta = 5
  sd = 7
  sig.level = 0.05
  power = 0.8575538
  alternative = two.sided
```

检验力是0.86，这是足够好的

- Conclusion: for a true difference of 5cm, a sample of 20 Agta women would provide a t-test with power β=0.86 (which is good enough)

(for a one-tailed test, add ***alt="one.sided"***) 单侧测试加参数

计算样本量

power test不能检验数据是否足够大或呈现正态分布
(假定正态分布)

Calculating sample size

比如上面这个例子20，就可以将样本量增加到30
然后测正态性，如果满足就可以用t-test

- But with a sample of only n=20, you should not run a t-test!
  - sample is too small to prove that distribution is normal
  - (and we are assuming that effect size is going to be large = 5 cm)

所以如果不需要事先证明数据是正态分布的，可以用power test计算出t test所需要的样本量

- The power test shows the required sample for a t-test *if you did not need to prove in advance that the variable is normally distributed!*
  - although power test says n=20 is ok, it may not pass Shapiro test

如果用power test计算出来的样本量太小，需要增加样本量，或者用wIicoxon test

- So: if the required sample size for the desired power is too small
  - design the test with a larger sample to prove with histograms, Shapiro test etc. that the distribution is normal 增加样本量然后检验正态分布
  - or if a larger sample is not possible, run a Wilcoxon test with the calculated small sample size; if power of the equivalent t-test is high, it is likely that the Wilcoxon test will able to accept or reject the null hypotheses

可以对wIicoxon检验测试其power，但不是本节课的内容，下载mkpower包

  - Note: it is possible to estimate sample size of Wilcoxon tests, but the approach is empirical and based on simulations
    - too advanced for this module
    - but if you are curious, install and try package *MKpower*

Calculating sample size

- If we want a power of 90% for a difference of 4 cm , which sample size do we need?

```
> power.t.test(power=0.9, sd=7, delta = 4, type="one.sample")
  One-sample t test power calculation
           n = 34.15781
       delta = 4
          sd = 7
   sig.level = 0.05
       power = 0.9
 alternative = two.sided
```

- So what's the sample size needed? 34? 35? 需要样本量为35
  - If the estimate is 34.157 (minimum), then you need n= 35
  - Calculating power shows that the answer to the question 'what is a large sample' depends on which effect size we want to detect "什么是大样本"取决于我们想要的检测效应大小

Break

Two-sample *t*-tests

- In the case of a two-sample test, noncentral parameter is

$$\nu = \frac{\delta}{\frac{\sigma}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}$$

- n_1 and n_2 are sizes of the two samples
  - test assumes n1= n2 = the smallest of the two　两个中小的那个
    - If n1 = 30 and n2 = 40, it assumes n1= n2 = 30
  - variance assumed similar in the two groups, i.e. we enter only one standard deviation　方差假定两组相似，只输入一个sd
  - to be conservative, if you have two standard deviations, enter the *largest*!　保守起见，如果有两个sd，输入最大的

Sample size, two-sample t-test

- Let's say we want to test for a true difference of 7 *cm* between two groups
- For a power of 80%, **a true difference (=effect size) of 5**, what is the required sample size?
  - (two-sample *t*-test is default; no need to specify type) 无需指定类型

```
> power.t.test(power=0.8,delta=5, sd=7)

     Two-sample t test power calculation

              n = 31.75716
          delta = 5
             sd = 7
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

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

是每个组

- You need a minimum sample of 32 per group, 64 in total)

```
> power.t.test(power=0.8,delta=2, sd=7)
```

Exercises

- Based on the previous example, what is the sample size required if the effect size was only 2 cm?

- If the available sample you have for your test is n= 40 in group 1 and 60 in group 2, what is the power of your test (still assuming sd = 7)?

```
假设有40个(取小的那个)
> power.t.test(n=40,delta=2, sd=7)
```

Paired t-tests

加类型paired

- For a paired t-test, add ***type="paired"***

```
> power.t.test(power=0.8, delta=2, sd=7,type="paired")

     Paired t test power calculation

              n = 98.08684
          delta = 2
             sd = 7
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

- Sample size　样本量是two-sample t-test的一半，和单样本t-test样本量相同
  - ~ half the size of a two-sample t-test
  - (note: paired test requires same size as a one-sample t-test; try replacing type 'paired' with 'one.sample'

两个比例检验的功效

Power of two-proportion tests

- To calculate power and sample sizes in two-proportion tests, we use function *power.prop.test*
  - based on a binomial approximation to a normal distribution

- Limitations: 仅适用于独立比例，不是单样本比例检验的功效
  - only works for independent proportions
    - *not a power test for one-sample proportion tests*
  - it cannot be used when sample size is smaller than 5 (a limitation of model distribution)

当样本量小于5时不能使用

只需要两个参数

- Only two parameters needed: 将其替换为比例以比例二，不需要sd
  - δ is replaced by p_1=proportion 1 and p_2=proportion 2.
  - standard deviation not required

Power of two-proportion tests

- Example: Which sample size do we need to detect a difference of 15% (let's say between 25% and 10%) in preference for hybrid cars between two countries?

```
> power.prop.test(power=0.9, p1=0.1, p2=0.25)

     Two-sample comparison of proportions power calculation

              n = 132.7557
             p1 = 0.1
             p2 = 0.25
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

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

每组需要的样本量

- One-tailed option is available

单侧是可以选择的

Notes

如何估计一类变量的标准差？如果已经有数据集，根据数据集估计，如果没有在文献中搜索

- What to do about estimating standard deviation of a type of variable?
  - estimate it from your own dataset (when you already have one)
  - if you don’t have a sample yet, search in literature

通常无法实现适当的样本量，可以进行估计，或者使用可用样本量来调查测试的功效

- Appropriate sample sizes are often not achievable
  - you may estimate that a sample size of 100 Neanderthal fossil skulls would be required for a comparison of brain size with humans; or data from 100 countries etc., but data may not be available!
  - in this case, you can still take the available sample size and investigate the power of the test

Notes

关于ANOVA检验力的检验函数

- There is a test for power of ANOVA tests

> power.anova.test(groups= , between.var = , within.var = , n=)

  - Additional difficulty is how to estimate both between-group and within group variance 组间方差和组内方差难以估计(如果没有自己的数据集的话)
  - If you already have a sample, run ANOVA and use sample sizes and estimated mean squared between and within group variances to estimate power of your test

小样本运行非参数检验的方法

- Remember: power calculation may suggest very small sample sizes
  - if required sample size is small, run non-parametric tests (Wilcoxon test instead of t-test, Kruskal-Wallis instead of ANOVA)
  - …unless there is evidence (from other sources) that the type of variable in your study is always or very likely to be normally distributed 除非可以证明他们是正态分布的
  - so…safer to run non-parametric tests with small sample sizes

所以小样本量用非参数检验更加安全

最小相关差异和统计显著性

Minimal relevant difference vs. statistical significance

- What if you have *no idea* about the expected difference δ between two groups? 选择您认为值得做为科学发现报告的最小值
  - select the minimum value that you may consider worth being reported as a scientific finding!
  - = ‘minimal relevant difference’ or ‘smallest meaningful difference’

- For example, you will not get a medical award for identifying a significant difference ($P < 0.05$) of *1 second* in life expectancy between people who eat parsnips everyday and people who don’t!

- If you want to test whether eating some vegetable or anything else affects lifespan, design the test around a relevant ‘effect size’
  - you may decide that a relevant effect on life expectancy should be at least one year?

总之比如干什么可以延长一周寿命，这是大家不关心的，如果是一年，可以研究

- In summary: a difference may be statistically significant, and yet scientifically irrelevant!

差异在统计学上可能是显著的，但在科学上是不相关的

Other tests

pwr包有一系列关于估计power,sample sizes的函数(好像更加准确，两个组分开的)

- Package *pwr* has a series of functions to estimate power, sample sizes etc. of
  - t-tests
  - Proportion (chi-square) tests, one and two independent proportions
  - ANOVAs
  - Correlations

Summary

测试应该至少有80%的检验力，如果可以，尝试90%

- Your test should have a power of at least 80%; if possible, try 90%

在设计实验、收集或者分析数据时，首先进行检验力和样本量测试

- When designing experiments, collecting or analysing data, run power and sample size tests first

测试应该瞄准更值得报道的组间差异

- When you design a test, aim at a difference between groups that is worth reporting

结果并不完全由统计显著性来定义

- A ‘result’ is not defined exclusively by statistical significance; relevance of finding (or effect size) is as important, and this can be determined by statistical power

- If you power test says you need a very small sample for a high power: either use a larger sample anyway (to be able to prove that variable is normal) or run a non-parametric alternative…it is safer

如果检验力测试说需要非常小的样本，那么需要增加样本，或者用非参数替代方法

```
> power.t.test(power=0.9, sd=400, delta = 3300*0.05)
> power.t.test(power=0.9, sd=400, delta = 3300*0.1)

> power.prop.test(power=0.8, p1=0.4724409, p2=0.5277778)
```

Exercises

- Suppose that the average baby girl in the UK is born weighing 3300 g.
  - Which sample size do you need to show (with a probability of 90%) that birth in boys is at least 5% different from birth size in girls? (assume sd=400 g in both boys and girls)
  - And for a difference of at least 10% between the two groups?

- From Lecture 5: proportion tests failed to identified a difference in proportion of born boys between rural gypsies and rural Hungarians
  - Calculate the sample size required for a two-independent proportions test to have an 80% chance of detecting the observed difference in proportions of boys in gypsies and non-gypsies

Table 2. *Sex ratios at birth for each population*

	number of sons per 100 daughters			
	rural populations		urban populations	
	Gypsy	Hungarian	Gypsy	Hungarian
A. all children				
sample size	254	216	239	224
males/100 females	89.3	111.8	89.7	113.3
B. first-born children only				
sample size	87	85	77	102
males/100 females	81.3	157.6	94.3	131.8

重叠越多需要越多样本量去区分两个的差异，重叠越少需要的样本量越少