上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

上海交通大学生物医学工程学院
SJTU School of Biomedical Engineering

生物医学统计概论(BI148-2)

Fundamentals of Biomedical Statistics

授课：林关宁

2022 春季

课程内容安排

上课日期	章节	教学内容	教学要点	作业	随堂测	学时
2.16	1	数据可视化，描述性统计	1. 课程介绍 & 数据类型	作业1 (8%)	测试1 (8%)	2
2.23			2. 描述性统计Descriptive Statistics & 数据常用可视化			2
3.2			3. 大数定理 & 中心极限定理			2
3.9			4. 常用概率分布			2
3.16	2	推断性统计，均值差异检验	5. 统计推断基础-1：置信区间 Confidence Interval *	作业2 (10%)	测试2 (10%)	2
3.23			6. 统计推断基础-2：假设检验，I及II类错误，统计量，p-值			2
3.30			7. 数值数据的均值比较-1： 单样本及双样本t-检验，效应量，功效			2
4.6			8. 数值数据的均值比较-2：One-Way ANOVA，正态性检验			2
4.13			9. 数值数据的均值比较-3：Two-Way ANOVA			2
4.20	3	比例差异检验	10. 样本和置信区间预估 *	作业3 (6%)	测试3 (6%)	2
4.27			11. 类别数据的比例比较-1： 单样本比例推断			2
5.7			12. 类别数据的比例比较-2： 联立表的卡方检验			2
5.11	4	协方差，相关分析，回归分析	13. 相关分析（Pearson r, Spearman rho, Kendal’ s tau) *	作业4 (6%)	测试4 (6%)	2
5.18			14. 简单回归分析			2
5.25			15. 多元回归Multiple Regression			2
6.1	5	Course Summary	16. 课程总结 *			2
			Total	30%	30%	32

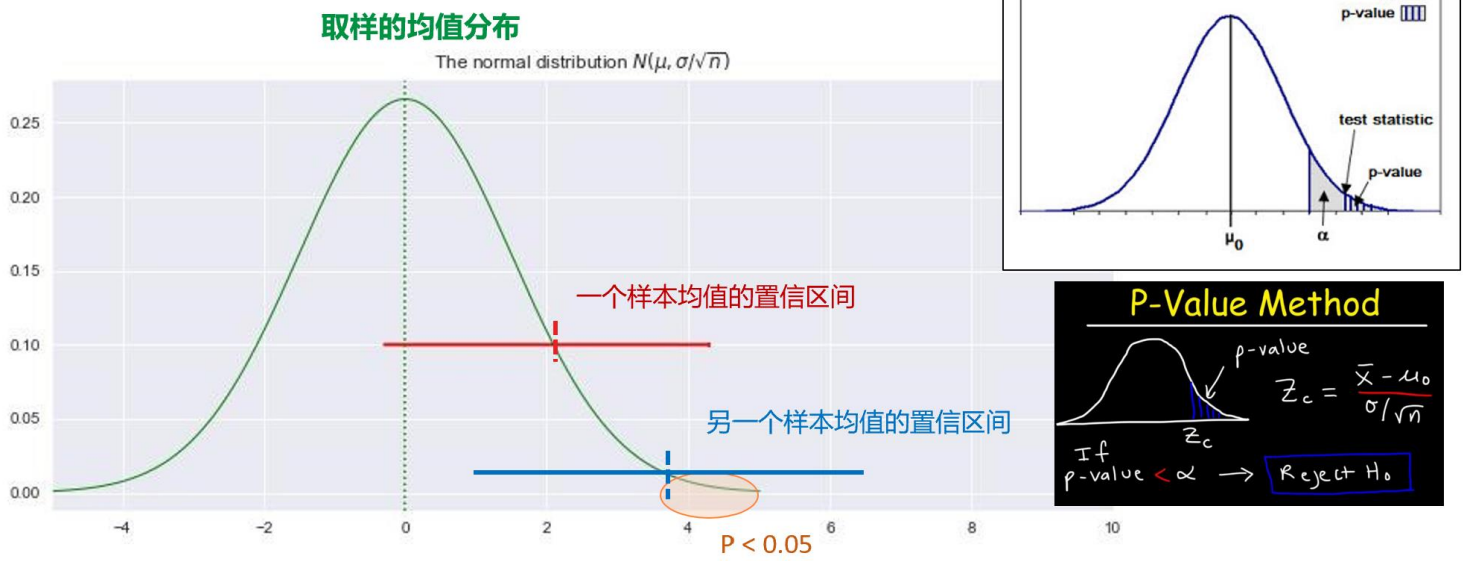
*** 随堂测试**

上节课回顾

- **假设检验和置信区间之间的区别?**

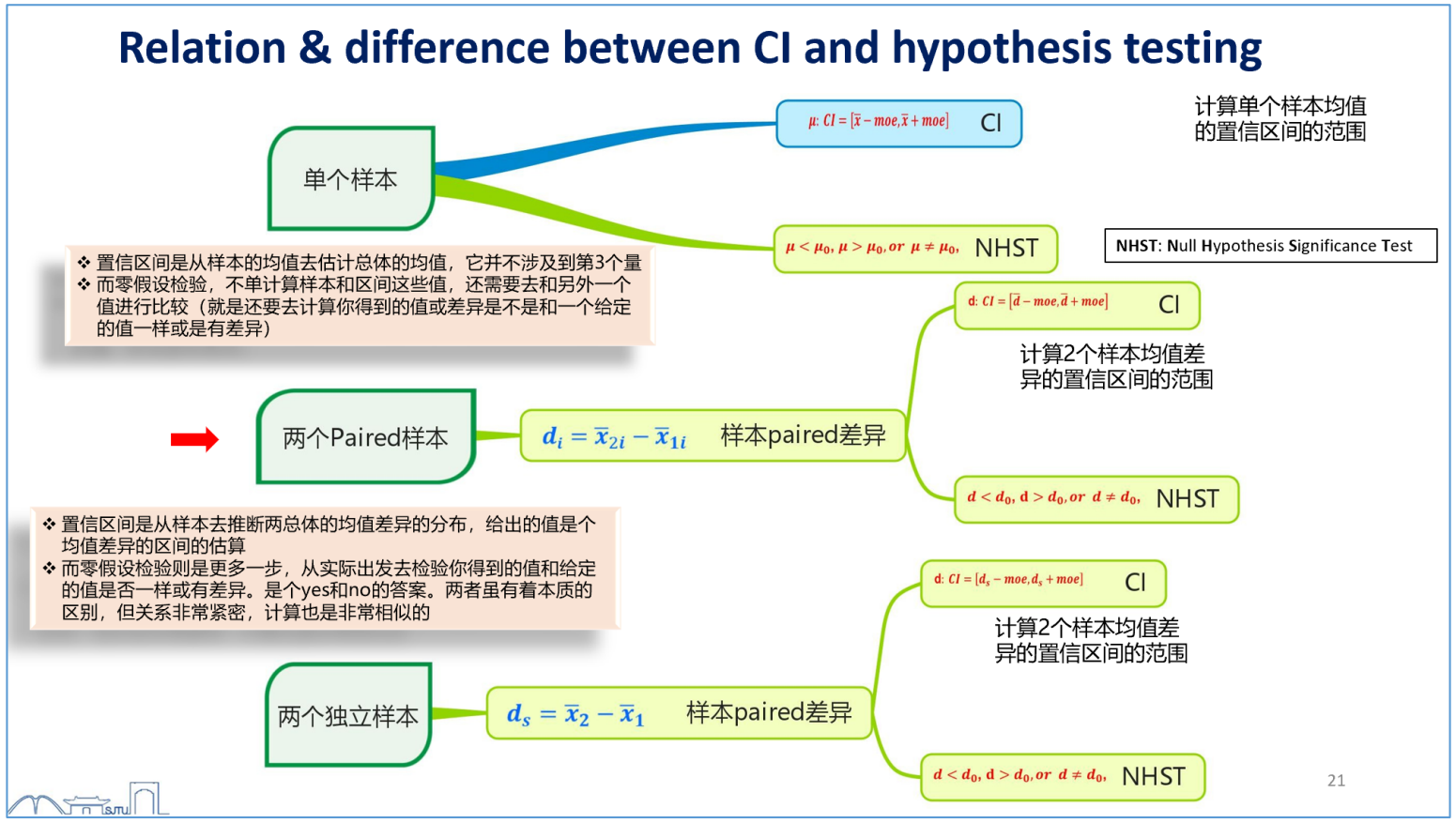
- 假设检验与相应置信区间之间的关系由显著性水平 α 定义
- 这两种方法**基于相同的推理逻辑，只是视角不同**。
- 假设检验方法询问 $\bar{x}$ 是否距离 μ_0 足够远，可以认为是寻找极端的概率，
- 而置信区间方法询问 μ_0 是否接近足以让 $\bar{x}$ 变得可信。在这两种情况下，“足够远”和“足够近”都由 α 定义

Relation between CI and hypothesis testing



上节课回顾

- 假设检验和置信区间之间的区别?
- 单样本均值检验，双样本均值检验，各自在检验什么?



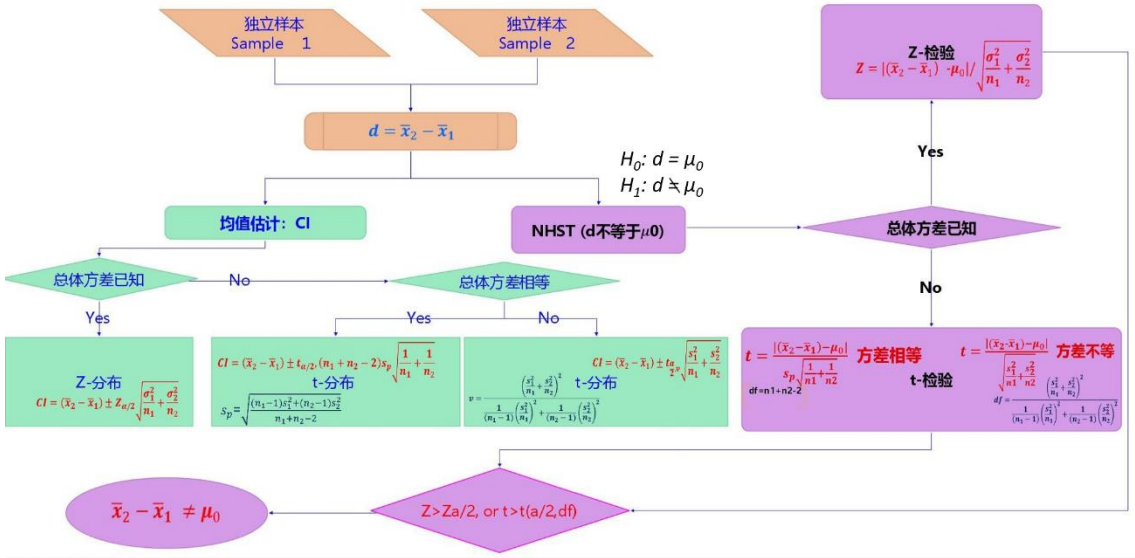
上节课回顾

- 假设检验和置信区间之间的区别?
- 单样本均值检验，双样本均值检验，各自在检验什么?
- Z-test, T-test 的在单，双样本检验上的使用条件是什么?

T 检验的统计假设检验框架

- 是否有一个样本要与指定值进行比较？做一个样本 t 检验。例如，假设大家都知道橡子的平均质量是10克，你想测试一下它们来自酸雨森林的橡子质量是否有显著的不同。
- 是否有两个相互比较的独立样品吗？做一个独立的样本 t 检验。例如，假设你从一个上风森林和一个煤电厂的下风森林中采集橡子样本，你想测试两个样本中橡子的平均质量是否相同。
- 是否有两个相依样本是从同一个人或物体中提取的？做配对样本 t 检验。例如，假设你测量了当地一个森林中50棵树的橡子平均质量，你想看看在发电厂从煤转化为天然气前后的橡子的质量是否有差异。

Two independent samples hypothesis test (two-tailed)



上节课回顾

- 假设检验和置信区间之间的区别?
- 单样本均值检验，双样本均值检验，各自在检验什么?
- Z-test, T-test 的在单，双样本检验上的使用条件是什么?
- 什么是效应量（effect size）和功效（power）？

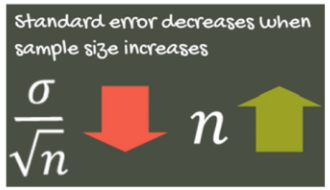
Practical significance – Effect size 效应量

之前我们讨论了如何使用假设检验法去识别统计上有显著的差异。如果 p 值小于 α （通常为0.05），则结果具有统计学意义。
意思就是：当假设的总体参数和观察到的样本统计之间的差异大到足以得出不可能是**通常形式发生**的结论时，结果会被认为**具有统计学意义（statistically significance）**。

实际意义（Practical significance）是指 **效应量（Effect size）**，也就是所谓的效果大小。**实际意义不受样本量的直接影响。**

Note:

- 统计显著性是直接受样本量的影响。因为样本量与标准误差（样本分布的标准差）之间是存在一种相反的关系。**因此在样本量很大的情况下，非常小的差异在统计学上也会出现显著性。**
- **因此，当样本量够大而得到的统计学结果显著时，还必须检查效应量，有必要报告效应量大小。**



39

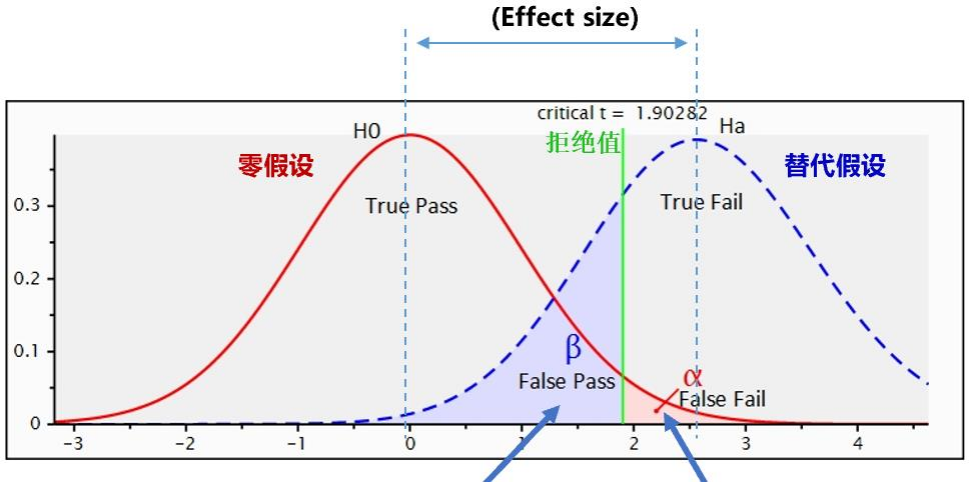
- 做完一个假设检验之后，如果结果具有统计显著性，那么还**需要继续计算其 effect size 效应量**；如果结果不具有统计显著性，并且还需要继续进行决策的话，那么**需要计算功效**
- **功效（power）：**当备择假设为真时，拒绝零假设的概率，记作1-β

上节课回顾

- 假设检验和置信区间之间的区别?
- 单样本均值检验，双样本均值检验，各自在检验什么?
- Z-test, T-test 的在单，双样本检验上的使用条件是什么?
- 什么是效应量（effect size）和功效（power）？
- 请解释显著水平（α），效应量（effect size），样本量（sample size）和功效（power）之间的关系?

假设检验的功效受以下三个因素影响：

- **显著性水平 (α)**： 其他条件保持不变，显著性水平越低，功效就越小。
- **效应量**（两总体之间的差异）：其他条件保持不变，总体参数的真实值和估计值之间的差异越大，功效就越大。也可以说，效应量越大，功效就越大。
- **样本量 (*n*)**：其他条件保持不变，样本量越大，功效就越大。



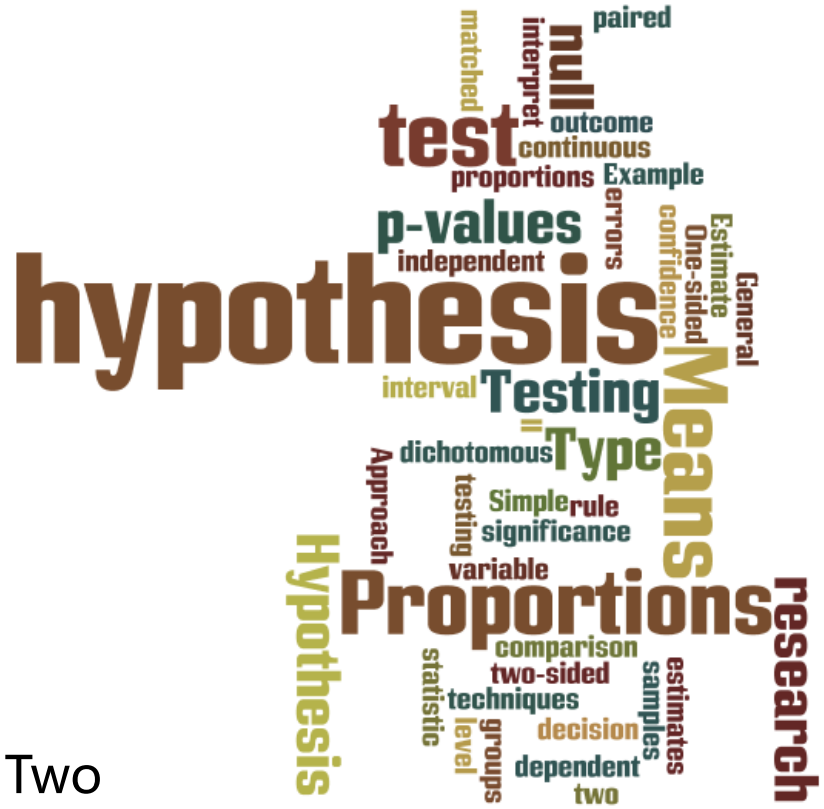
课程内容安排

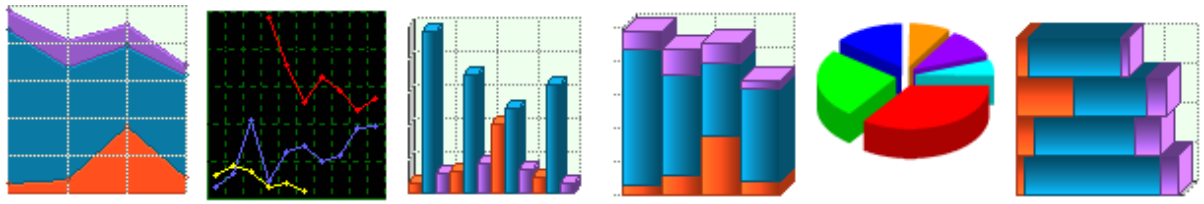
上课日期	章节	教学内容	教学要点	作业	随堂测	学时
2.16	1	数据可视化，描述性统计	1. 课程介绍 & 数据类型	作业1 (8%)	测试1 (8%)	2
2.23			2. 描述性统计Descriptive Statistics & 数据常用可视化			2
3.2			3. 大数定理 & 中心极限定理			2
3.9			4. 常用概率分布			2
3.16	2	推断性统计，均值差异检验	5. 统计推断基础-1：置信区间 Confidence Interval *	作业2 (10%)	测试2 (10%)	2
3.23			6. 统计推断基础-2：假设检验，I及II类错误，统计量，p-值			2
3.30			7. 数值数据的均值比较-1： 单样本及双样本t-检验，效应量，功效			2
4.6			8. 数值数据的均值比较-2：One-Way ANOVA，正态性检验			2
4.13			9. 数值数据的均值比较-3：Two-Way ANOVA			2
4.20	3	比例差异检验	10. 样本和置信区间预估 *	作业3 (6%)	测试3 (6%)	2
4.27			11. 类别数据的比例比较-1： 单样本比例推断			2
5.7			12. 类别数据的比例比较-2： 联立表的卡方检验			2
5.11	4	协方差，相关分析，回归分析	13. 相关分析（Pearson r, Spearman rho, Kendal’ s tau) *	作业4 (6%)	测试4 (6%)	2
5.18			14. 简单回归分析			2
5.25			15. 多元回归Multiple Regression			2
6.1	5	Course Summary	16. 课程总结 *			2
			Total	30%	30%	32

*** 随堂测试**

单元2内容（week 5-9）

- **总体、样本**
  - 参数、统计量
- **置信区间**
  - 区间估计、总体方差
- **均值差异**
  - One-sample 均值估计
  - Two-sample 均值差异
- **零假设检验**
  - H_0 vs H_a，Type I & II errors，P-value，One-sample vs Two sample test
- **样本量、效应量、功效（Sample size, Effect size, Power）**
- **方差分析**
  - ANOVA test

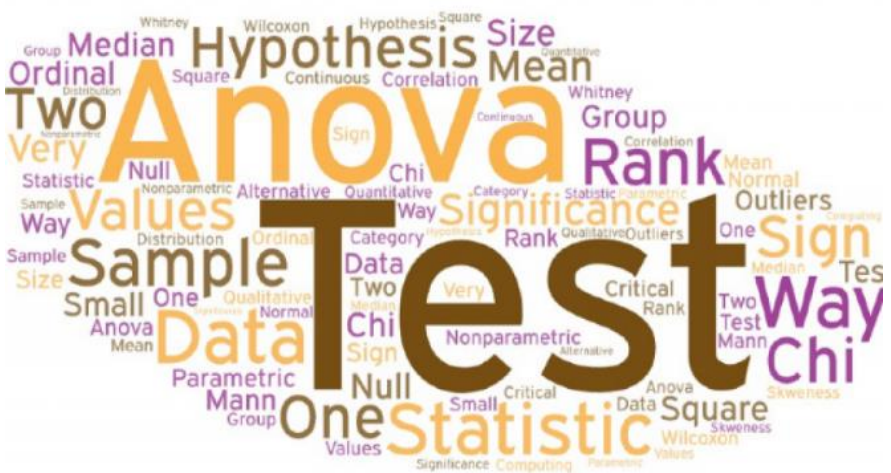




Week 8-9

方差分析

back



判断数据正态分布

在统计学中，正态检验主要用于检验一个数据集是否服从正态分布。常用的t检验、方差分析（ANOVA）等参数检验都有一个共同的前提条件：**样本数据必须服从正态分布**，即样本数据必须来源于一个正态分布的总体，若样本数据不服从正态分布，就不能用以上参数检验对数据进行分析，而应该使用非参数检验（如卡方检验、置换检验等）。

因此在对数据进行统计分析之前，第一步就需要对数据进行正态性检验，以检验该数据来自正态分布总体的概率有多大，再选择对应的参数或非参数检验方法进行分析

判断数据正态分布 – 方法 1

- **正态性检验：偏度和峰度**

1. 偏度（Skewness）：描述数据分布不对称的方向及其程度

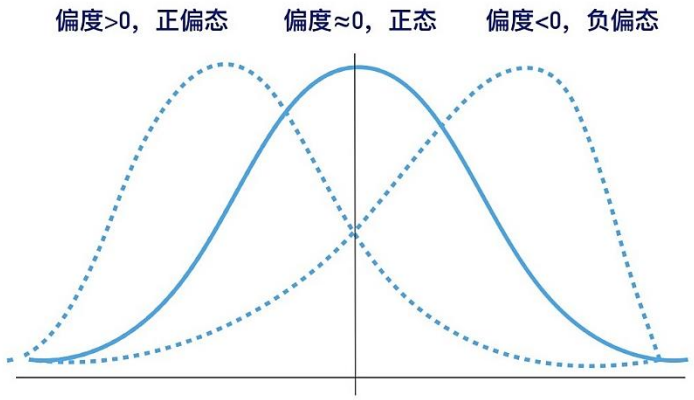


图1. 数据分布的偏态特征

当偏度≈0时，可认为分布是对称的，服从正态分布；
当偏度>0时，分布为右偏，即拖尾在右边，峰尖在左边，也称为正偏态；
当偏度<0时，分布为左偏，即拖尾在左边，峰尖在右边，也称为负偏态；
注意：数据分布的左偏或右偏，指的是数值拖尾的方向，而不是峰的位置，容易引起误解。

2、峰度（Kurtosis）：描述数据分布形态的陡缓程度

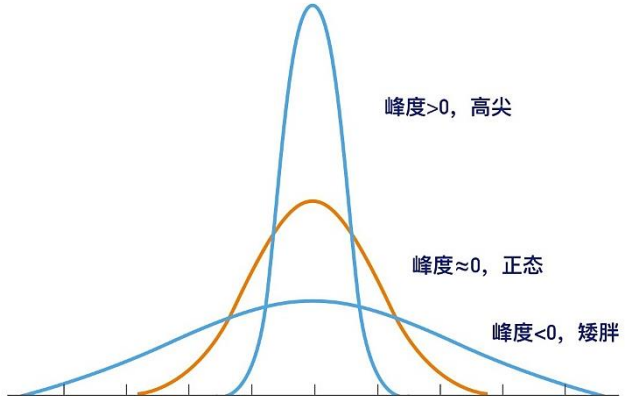


图2. 数据分布的峰态特征

当峰度≈0时，可认为分布的峰态合适，服从正态分布（不胖不瘦）；
当峰度>0时，分布的峰态陡峭（高尖）；
当峰度<0时，分布的峰态平缓（矮胖）；

利用偏度和峰度进行正态性检验时，可以同时计算其相应的Z评分（Z-score），即：偏度Z-score=偏度值/标准误，峰度Z-score=峰度值/标准误。在α=0.05的检验水平下，若Z-score在±1.96之间，则可认为资料服从正态分布

单选题 1分

设置

结果解读

Statistics

BMI

N	Valid	180
	Missing	0
Skewness		.194
Std. Error of Skewness		.181
Kurtosis		.373
Std. Error of Kurtosis		.360

是否正态分布?

A 是

B 否

提交

结果解读

是否正态分布？

Statistics

BMI

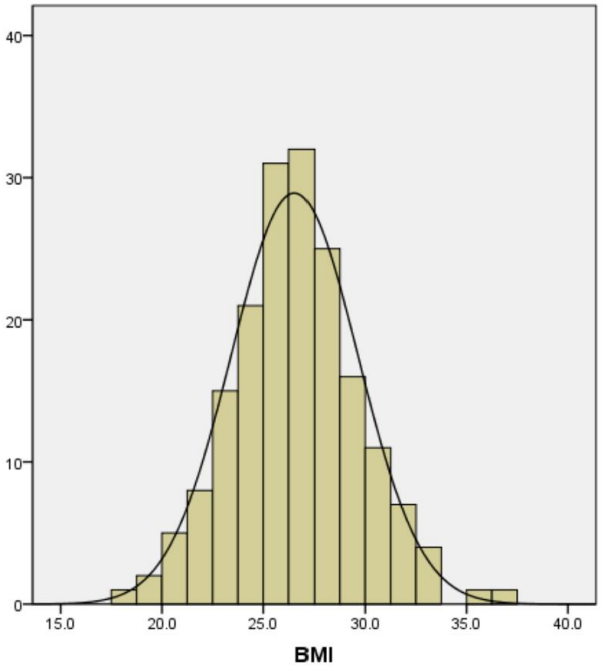
N	Valid	180
	Missing	0
Skewness		.194
Std. Error of Skewness		.181
Kurtosis		.373
Std. Error of Kurtosis		.360

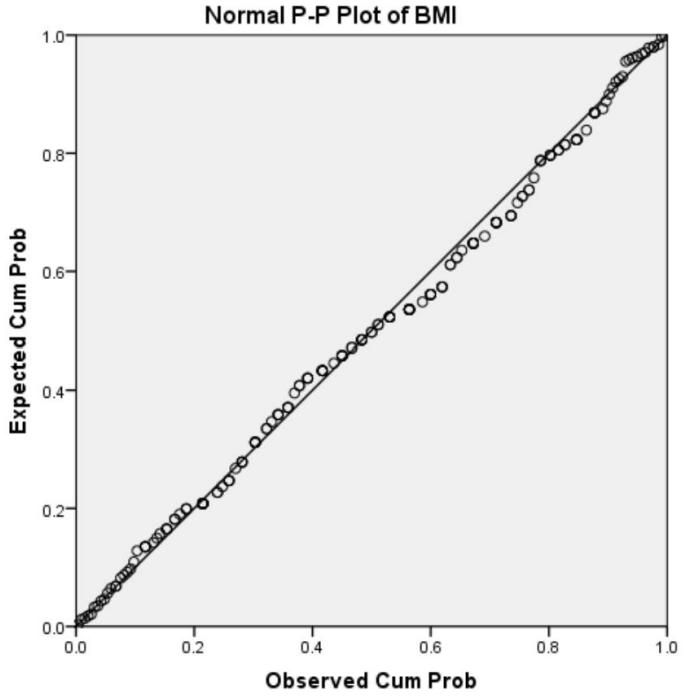
对变量BMI进行了基本的统计描述，同时给出了其分布的偏度值0.194（标准误0.181），Z-score = 0.194/0.181 = 1.072，峰度值0.373（标准误0.360），Z-score = 0.373/0.360 = 1.036。偏度值和峰度值均≈0，Z-score均在±1.96之间，**可认为数据服从正态分布**

判断数据正态分布 – 方法 2

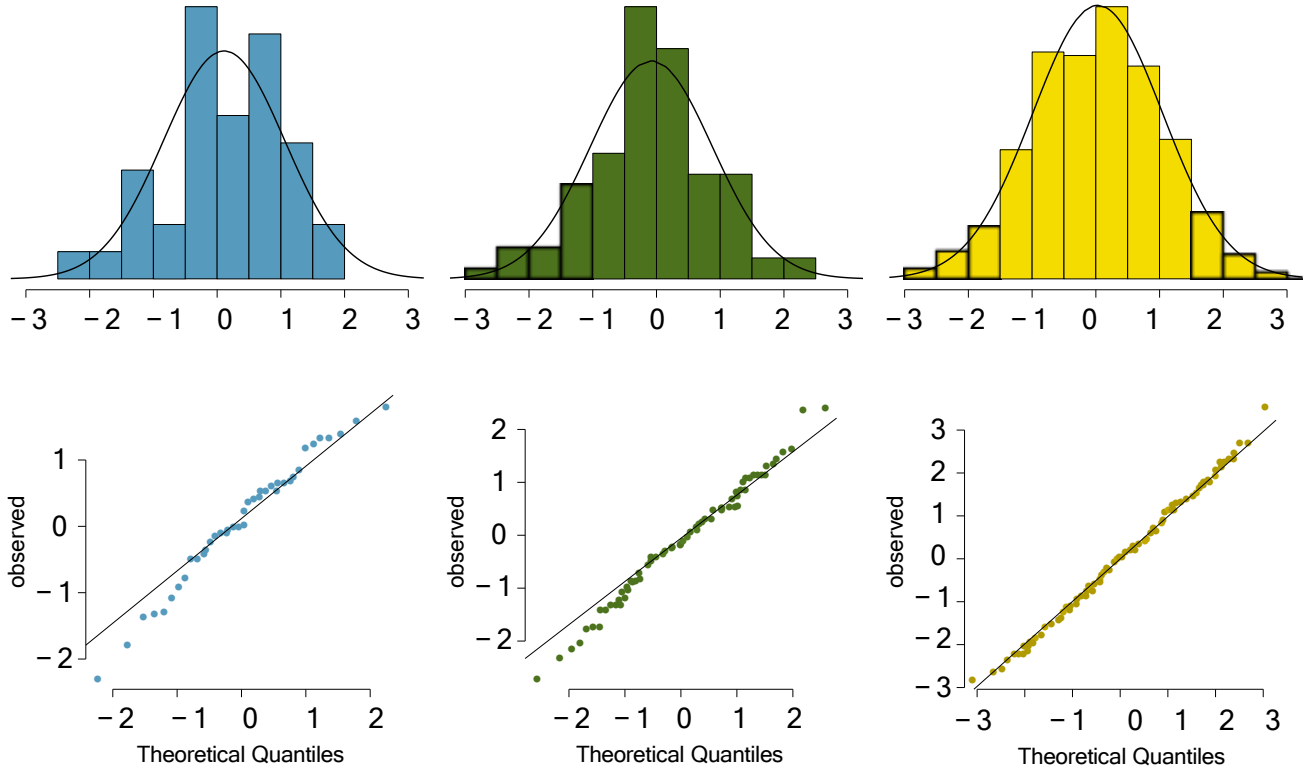
- **正态性检验：图形判断**

直方图+QQ plot





Normal probability plots (Q-Q plots)



If points fall on or near the line, data closely follow a normal distribution.

- Difficult to evaluate in small datasets
- Plots show three simulated normal datasets: from L to R, $n = 40$, $n = 100$, $n = 400$

Exponential Distribution 指数分布

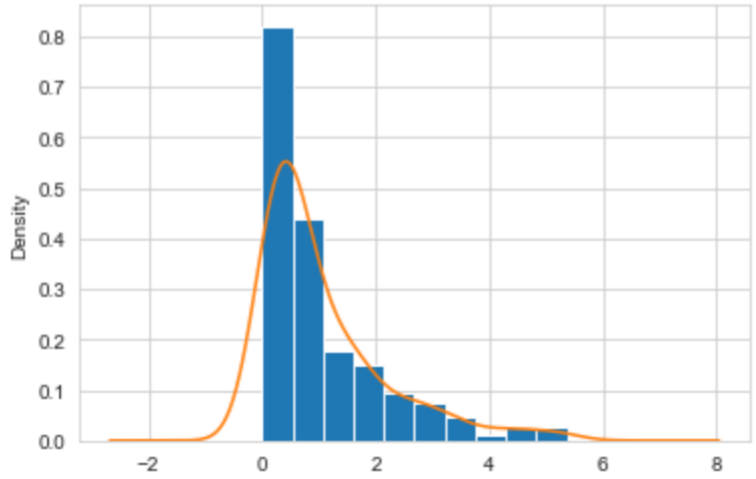
如果我们用理论正态分布绘制一个指数分布的变量，图表如下所示。

```
#Creating exponentional datapoints

np_exp = pd.Series(np.random.exponential(scale=1.0, size=200))
```

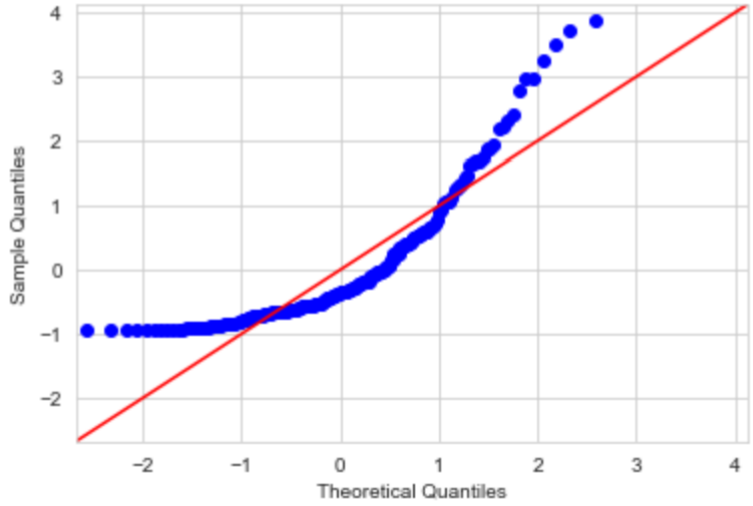
```
#Plotting Exponential datapoints

fig, ax = plt.subplots()
np_exp.plot.hist(ax=ax,density=True)
np_exp.plot.kde(ax=ax)
plt.show()
```




```
#QQ plot for exponential distribution with normal distribution

sm.qqplot(np_exp,fit=True,line='45',dist=stats.norm)
plt.show()
```



判断数据正态分布 – 方法 3

正态性检验属于非参数检验，零假设为“样本来自的总体与正态分布**无显著性差异**，即符合正态分布”，也就是说P>0.05才能说明数据符合正态分布

通常正态分布的检验方法有两种，一种是Shapiro-Wilk检验，适用于小样本资料（样本量≤5000），另一种是Kolmogorov–Smirnov检验，适用于大样本资料（>5000）

Shapiro-Wilk Test

```
for i in columns:
    print ([i])
    a,b= stats.shapiro(df[[i]])
    print ("Statistics", a, "p-value", b)
    if b < alpha:
        print("The null hypothesis can be rejected")
    else:
        print("The null hypothesis cannot be rejected")
```

```
['Length']
Statistics 0.9696492552757263 p-value 7.377629143455696e-29
The null hypothesis can be rejected
['Diameter']
Statistics 0.970473051071167 p-value 1.6414551656441721e-28
The null hypothesis can be rejected
['Height']
Statistics 0.8896051645278931 p-value 0.0
The null hypothesis can be rejected
['Whole weight']
Statistics 0.9722852110862732 p-value 1.0143092052495122e-27
The null hypothesis can be rejected
['Shucked weight']
Statistics 0.962066650390625 p-value 9.362563887244321e-32
The null hypothesis can be rejected
['Viscera weight']
Statistics 0.9681379199028015 p-value 1.774917175594725e-29
The null hypothesis can be rejected
['Shell weight']
Statistics 0.9704266786575317 p-value 1.568463383716868e-28
The null hypothesis can be rejected
['Rings']
Statistics 0.9311649203300476 p-value 3.2763058745146386e-40
The null hypothesis can be rejected
```

Kolmogorov–Smirnov test

```
for i in columns:
    print ([i])
    a,b= stats.kstest(df[[i]], 'norm')
    print ("Statistics", a, "p-value", b)
    if b < alpha:
        print("The null hypothesis can be rejected")
    else:
        print("The null hypothesis cannot be rejected")
```

```
['Length']
Statistics 0.7924638449586122 p-value 0.0
The null hypothesis can be rejected
['Diameter']
Statistics 0.7421538891941353 p-value 0.0
The null hypothesis can be rejected
['Height']
Statistics 0.8707618877599821 p-value 0.0
The null hypothesis can be rejected
['Whole weight']
Statistics 0.9976396556000835 p-value 0.0
The null hypothesis can be rejected
['Shucked weight']
Statistics 0.9316245531838383 p-value 0.0
The null hypothesis can be rejected
['Viscera weight']
Statistics 0.7763727075624006 p-value 0.0
The null hypothesis can be rejected
['Shell weight']
Statistics 0.8425515750696722 p-value 0.0
The null hypothesis can be rejected
['Rings']
Statistics 1.0 p-value 0.0
The null hypothesis can be rejected
```

注意事项

- 事实上，Shapiro-Wilk检验及Kolmogorov-Smirnov检验从实用性的角度，远不如图形工具进行直观判断好用。在使用这两种检验方法的时候要注意，当样本量较少的时候，检验结果不够敏感，即使数据分布有一定的偏离也不一定能检验出来；而当样本量较大的时候，检验结果又会太过敏感，只要数据稍微有一点偏离，P值就会<0.05，检验结果倾向于拒绝原假设，认为数据不服从正态分布。**所以，如果样本量足够多，即使检验结果P<0.05，数据来自的总体也可能是服从正态分布的。**

- 因此，在实际的应用中，往往会出现这样的情况，明明直方图显示分布很对称，但正态性检验的结果P值却<0.05，拒绝原假设认为不服从正态分布。此时建议大家不要太刻意追求正态性检验的P值，一定要**参考直方图、Q-Q图等图形工具来帮助判断。**很多统计学方法，如T检验、方差分析等，与其说要求数据严格服从正态分布，不如说“数据分布不要过于偏态”更为合适。

- 有专家根据经验提出，标准差超过均值的1/2时提示数据不服从正态分布，或者四分位间距与标准差的比值在1.35左右时提示服从正态分布，这些可以作为正态性检验的一个粗略判断依据（仅供参考）

在街上看到两个人，一个看起来稍微胖点，一个看起来稍微瘦点，问谁比较重？

假设检验：
H0: 两人没差别
H1: 看起来胖的比较重

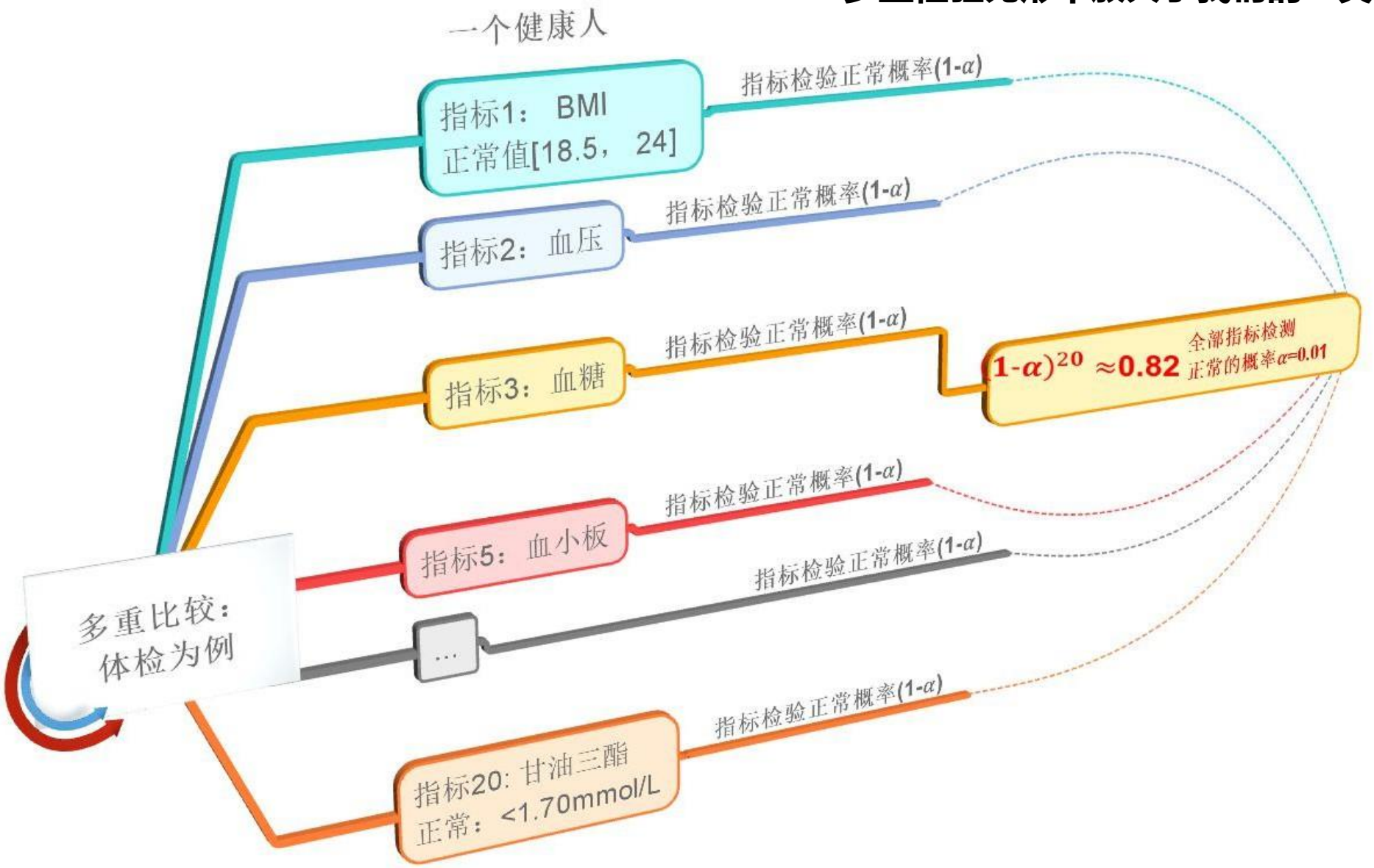
如果我们获得 P value < 1%，就是说我们拒绝了零假设，并做出了一个判断（看的胖的人比看的瘦重），但这个判断犯错的概率是1%（这里就是假阳性率，False positive rate）。虽然可能犯错（比如看的胖的其实是穿的比较厚），因为是这个属于小概率事件，我们就忍了吧，于是接受了这个判断。

这只是一次判断，

那假如我们要做很多次判断，对路上的每个人都想判断一次呢？怎么算p值？

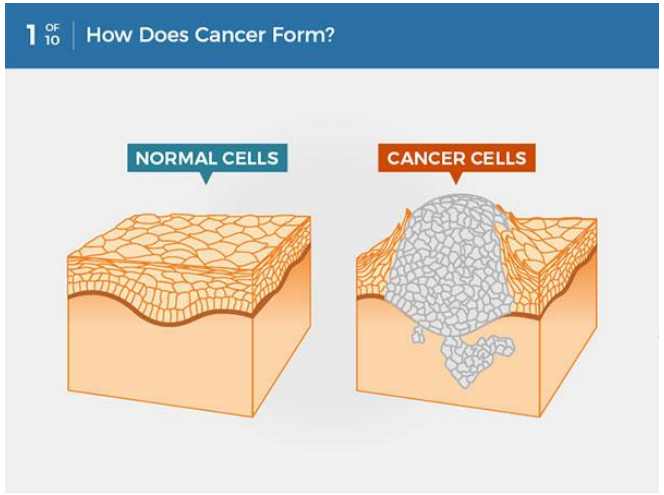
多重比较的问题（issues with multiple testing）

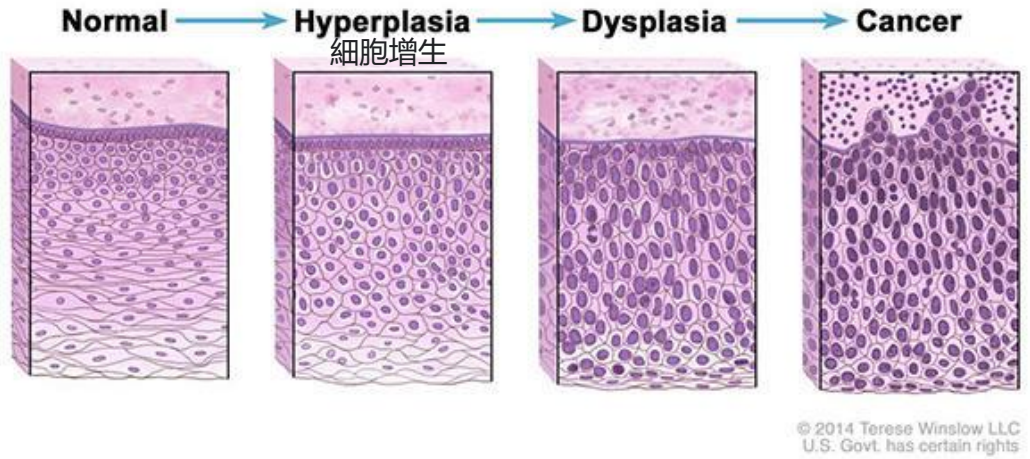
多重检验无形中放大了我们的一类错误的概率

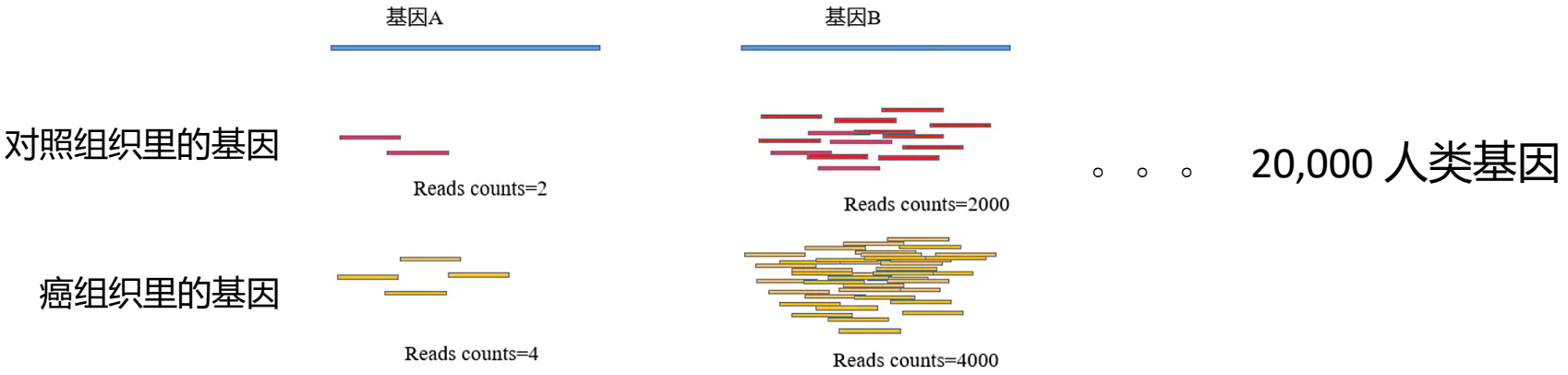


多重比较的问题（issues with multiple testing）

在很多科学实验中，在某些情况下，我们要做多次判断。
例如，要判断癌组织样本和正常组织的对应的20,000个基因的表达量是否在组间存在差异：基因A是否有差异？基因B是否有差异？基因C是否有差异？.....，如此下去，我们要进行20,000次的比较：







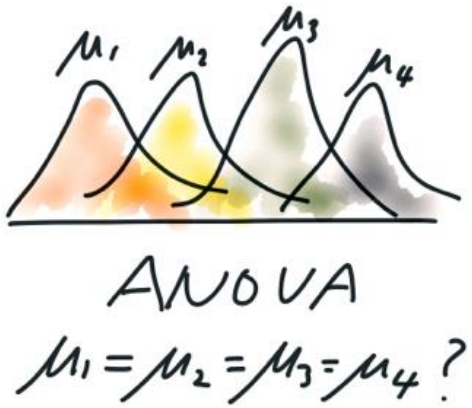
多重比较的问题（issues with multiple testing）

在很多科学实验中，在某些情况下，我们要做多次判断。
例如，要判断癌组织样本和正常组织的对应的20,000个基因的表达量是否在组间存在差异：基因A是否有差异？基因B是否有差异？基因C是否有差异？.....，如此下去，我们要进行20,000次比较。

- 如果我们以p value 1% (假阳性的概率是1%)来作为阈值，并假设每次判断都是彼此独立的，**那么即使这20,000个基因实际上都没有差异，我们也可能会得出有200个差异基因的结论（阳性结果的错误率为100%，也就是之后要提到的FDR (False Discovery Rate)值为100%）**。
- 也就是说，**一个小效率事件就在多次反复尝试后，变成了一个多次出现的事件**（“常在河边走，怎能不湿鞋”）。如果这20,000个基因中有200个基因真实存在差异的，在 p vlaue为1%的阈值标准下，我们可能会得出399个基因有差异的结论（阳性结果的错误率，即FDR值约为50%）。
- 可以看到，在进行多次检验后（也就是所说的多重检验，multiple test），基于单次比较的检验标准将变得过于宽松，使得阳性结果中的错误率（FDR值）已经大到令人不可忍受的地步。

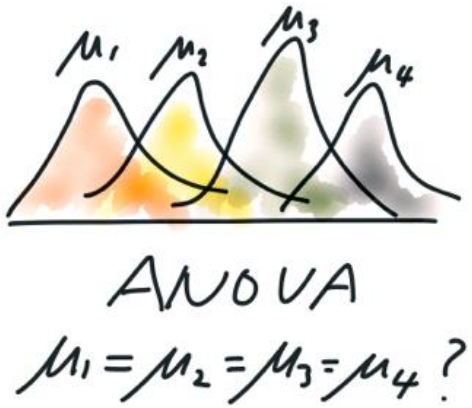
方差分析（Analysis of Variance, ANOVA）

- **研究分类型自变量对数值型应变量的影响**
  - 1个或多个分类型自变量
  - 1个数值型因变量
- **通过检验多个总体均值是否相等来判断是否有显著影响**
  - 通过分析数据的误差 -> 判断各总体**均值**是否相等



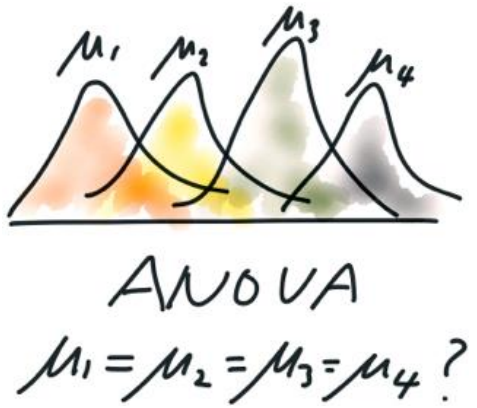
方差分析（Analysis of Variance, ANOVA）

- **研究分类型自变量对数值型应变量的影响**
  - 1个或多个分类型自变量
  - 1个数值型因变量
- **通过检验多个总体均值是否相等来判断是否有显著影响**
  - 通过分析数据的误差 -> 判断各总体**均值**是否相等
- ✓ **方差分析 Vs 假设检验**
  - 假设检验：一次只能研究2个样本
    - 需要比较的次数随因素的数量增多而增加
    - Type I error 发生的可能性增大
  - 方差分析：**同时分析多个样本**
    - 提高检验效率
    - 将所有信息结合一起，增加了分析的可靠性



方差分析（Analysis of Variance, ANOVA）

- **研究分类型自变量对数值型应变量的影响**
  - 1个或多个分类型自变量
  - 1个数值型因变量
- **通过检验多个总体均值是否相等来判断是否有显著影响**
  - 通过分析数据的误差 -> 判断各总体**均值**是否相等
- **方差分析 Vs 假设检验**
  - 假设检验：一次只能研究2个样本
    - 需要比较的次数随因素的数量增多而增加
    - Type I error 发生的可能性增大
  - 方差分析：**同时分析多个样本**
    - 提高检验效率
    - 将所有信息结合一起，增加了分析的可靠性
- ✓ **有单因素方差分析和双因素方差分析**
  - 单因素方差分析（One-way ANOVA）
    - 涉及**1个**分类型自变量对数值因变量影响
  - 双因素方差分析（Two-way ANOVA）
    - 涉及**2个**自变量对数值因变量影响
    - 它分为只考虑主效应（main effect）的双因素方差分析和考虑交互效应（interaction）的双因素方差分析



Tests of Between-Subjects Effects

Dependent Variable: Score on Beck's Depression Inventory

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	3687.775[a]					
Intercept	117629.949	1	117629.949	3199.098	.000	.972
gender	332.202	1	332.202	9.035	.003	.089
medicine	2519.647	3	839.882	22.842	.000	.427
gender * medicine	567.784	3	189.261	5.147	.002	.144
Error	3382.815	92	36.770			
Total	128105.000	99				
Corrected Total	7070.590	98				

MAIN EFFECTS: GENDER AND MEDICINE

INTERACTION EFFECT: GENDER * MEDICINE

a. R Squared = .522 (Adjusted R Squared = .485)

方差分析的一些概念

- **因素或因子（factor)**
  - 所要检验的对象。比如：要分析行业对投诉次数是否有影响， 行业是要检验的因素
- **水平或处理（treatment)**
  - 因素的不同表现或取值：零售业、 旅游业、 航空公司、 家电制造业就是因素的处理
- **观察值**
  - 在每个因素处理下得到的样本数据。 每个行业被投诉的次数就是观察值
- **实验**
  - 涉及一个因素的多水平， 可称为单因子多处理的试验
- **总体**
  - 因子的每一个处理看作是一个总体
- **样本数据**
  - 观察值可以看作是从着多个总体中抽取的样本数据

简单说就是：分类变量是因子或因素，而分的类别就可以称为水平或处理，观察值则是数值型变量。试验就是就是分类的过程，总体其实就是水平，样本数据就是观测值。

方差分析的一些术语

- **处理误差**（treatment error）
  - 因素的不同处理造成观测数据的误差
- **总平方和**（sum of squares for total，SST）
  - 反映全部观测数据误差大小的平方和
- **处理平方和**（treatment sum of squares，SSA）
  - 反映处理误差大小的平方和
- **误差平方和**（sum of squares of error，SSE）
  - 反映随机误差大小的平方和，也称组内平方和（within-group sum of squares）
- **均方**（mean square，MS）
  - 也称方差（variance），数据误差大小的平方和除以相应的自由度的结果
- **主效应**（main effect）
  - 因素对因变量的单独影响
- **交互效应**（interaction）
  - 一个因素和另一个因素联合产生的对因变量的附加效应
- **可重复双因素分析**（two-factor with replication）
  - 考虑交互性
- **多重比较**（multiple test）
  - 用于分析哪些处理之间有显著差异，即具体水平的分析，就是找哪两个均值不相等，多重比较方法多
- **事后分析/因果分析**（post-hoc analysis）

Source of Variation	df	Sum of Squares (SS)	Mean Sum of Squares (MSS)	F-test	p-value
Treatment	k-1	SSTr	MSTr=SSTr/(k-1)	F=MSTr/MSE	
Error	N-k	SSE	MSE=SSE/(N-k)		
Total	N-1	SSTo			

.

Source of Variation	Degrees of Freedom		Sums of Squares	Mean Squares	F	*P*
Between subjects	5		601.5	...	...	...
Treatment	...	1	541.5	541.5	36.1*	0.0039
Error subjects within treatment	...	4	60.0	15.0	...	...
Within subjects	18	...	1707.0	...	...	...
Time	...	3	1348.5	449.5	27.1†	0.0001
Treatment×time	...	3	159.2	53.1	3.2†	0.0626
Error	...	12	199.3	16.6	...	...
Total	23	...	2308.5	...	...	...

*The denominator is the mean square error due to subjects within treatment.
†The denominator is the mean square error.

方差分析背后的理念

Is the variability in the sample means large enough that it seems unlikely to be from chance alone? 样本均值的变异性是否足够大，以至于不太可能仅仅是偶然出现的现象?

Compare two quantities:

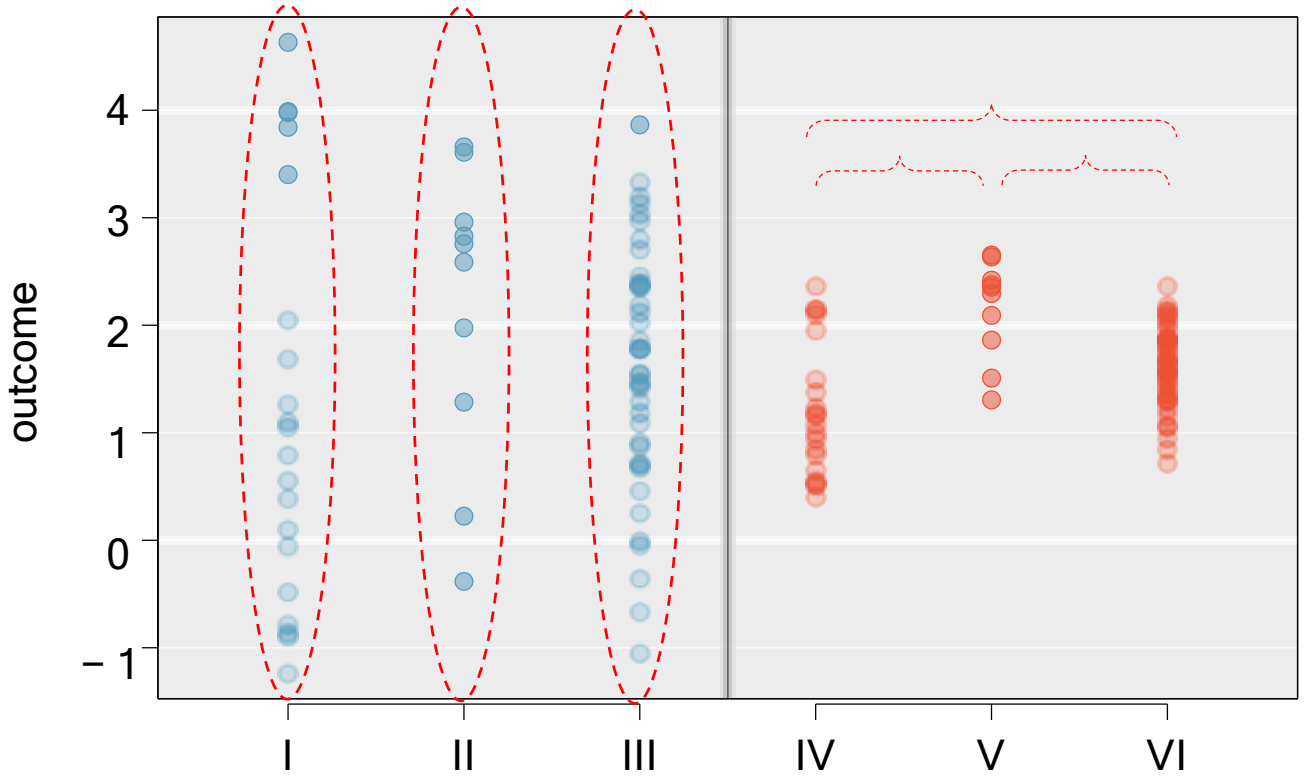
各组平均值之间的差异有多大，即各组平均值与总体平均值的差异有多大？

- Variability between groups 组间可变性 (*MSG*): how different are the group means from each other, i.e., how much does each group mean vary from the overall mean?
- Variability within groups 组内可变性 (*MSE*): how variable are the data within each group?

MSG denotes mean square between groups, while *MSE* denotes mean square error. MSG表示组间的均方误差，而MSE表示均方误差

(参考书 5.5)

方差分析背后的理念



很难看出平均数的差异，每个组内的变异性都很高

- **I, II, and III**: difficult to discern differences in means, variability within each group is high 组内变异高
- **IV, V, and VI**: appears to be differences in means, these differences are large relative to variance within each group 组间看的出在均值上是有差异的，组间的差异相对比组内的差异更大

似乎是平均数的差异，这些差异相对于各组内的方差来说是很大的

方差分析的理念

Under the null hypothesis, there is no real difference between the groups; thus, any observed variation in group means is due to chance. 在零假设成立下，组间是没有真正的差异；因此，观察到的总体平均数的任何变化都是偶然的

- Think of all observations as belonging to a single group. 所有的观察值都看作是属于同一个总体
- Variability between group means should equal variability within groups 组间有差异意味着组内是不存在差异的

The *F-statistic* is the test statistic for ANOVA. F统计量是方差分析的检验统计量

$$F = \frac{\text{variance between groups}}{\text{variance within groups}} = \frac{MSG}{MSE}$$

组间差异

组内差异

- When the population means are equal, the *F*-statistic is approximately 1. 当总体均值相等时，F统计量约为1
- When the population means differ, *F* will be larger than 1. Larger values of *F* represent stronger evidence against the null. 当总体均值不同时，F将大于1。F的值越大，则表示拒绝零假设的证据越有力。
- The *F* statistic follows an *F* distribution, with two degrees of freedom,

  df_1 and df_2; $df_1 = n_{groups} - 1$, $df_2 = n_{obs} - n_{groups}$. F统计量遵循F分布，有两个自由度$df_1$和$df_2$
- The *p*-value for the *F*-statistic is the probability *F* is larger than the *F*-statistic.

  p值是得到的F值>给定的F统计量的概率

使用方差分析 ANOVA 的前提

It is important to check whether the assumptions for conducting ANOVA are reasonably satisfied: 进行方差分析前需要考虑，数据假设是否合理满足前提要求

1. Observations independent within and across groups 组内和组间的观察值都是独立的观察值
   - Think about study design/context
2. Data within each group are nearly normal 每组内的观察值基本符合正态分布
   - Look at the data graphically, such as with a histogram 以图形方式查看数据，例如使用直方图
   - Normal Q-Q plots can help… 也可以直接使用 Q-Q图
3. Variability across groups is about equal 每组的组内差异类似
   - Look at the data graphically 直接画图观察
   - Numerical rule of thumb: ratio of largest variance to smallest variance < 3 is considered “about equal”

   经验：最大方差与最小方差之比<3被认为“大致相等”

t-test & ANOVA （Analysis of Variance）

ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the t-test beyond two means.

- **What are they**
  - *t*-test: determines whether ***two*** populations are statistically different from each other
  - ANOVA determines whether ***three or more*** populations are statistically different from each other
- **When do they used**
  - (approximately) normal distributions or when the sample number is high (e.g., > 30 per group)
- **How do they work**
- **What type of statistical value do I get**
  - The *t*-test and ANOVA produce a test statistic value (“t” or “F”, respectively)
  - converted into a “p-value”
  - A p-value is the probability that the null hypothesis – that both (or all) populations are the same – is true.
    - In other words, a lower p-value reflects a value that is more significantly different across populations.

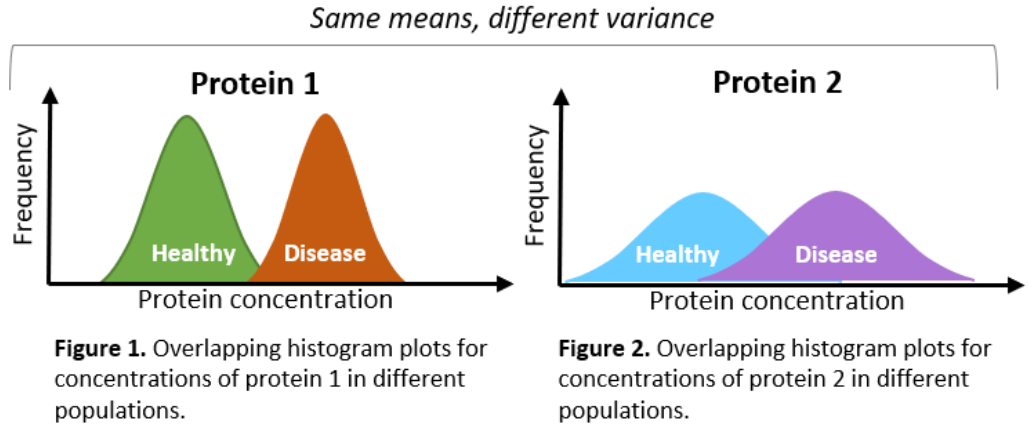


Figure 1. Overlapping histogram plots for concentrations of protein 1 in different populations.

Figure 2. Overlapping histogram plots for concentrations of protein 2 in different populations.

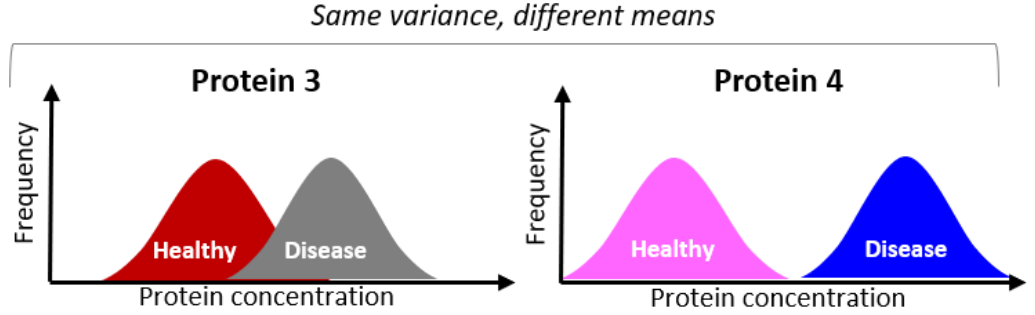


Figure 3. Overlapping histogram plots for concentrations of protein 1 in different populations.

Figure 4. Overlapping histogram plots for concentrations of protein 2 in different populations.

小结

- **如何判断数据的正态性**
  - 偏度和峰度
  - 图形法（Q-Q）
  - 非参数检验法
- **多重检验的问题**
- **ANOVA**
  - 单因素方差分析（one-way ANOVA）
  - 组内差异，组间差异

谢谢，下周见！