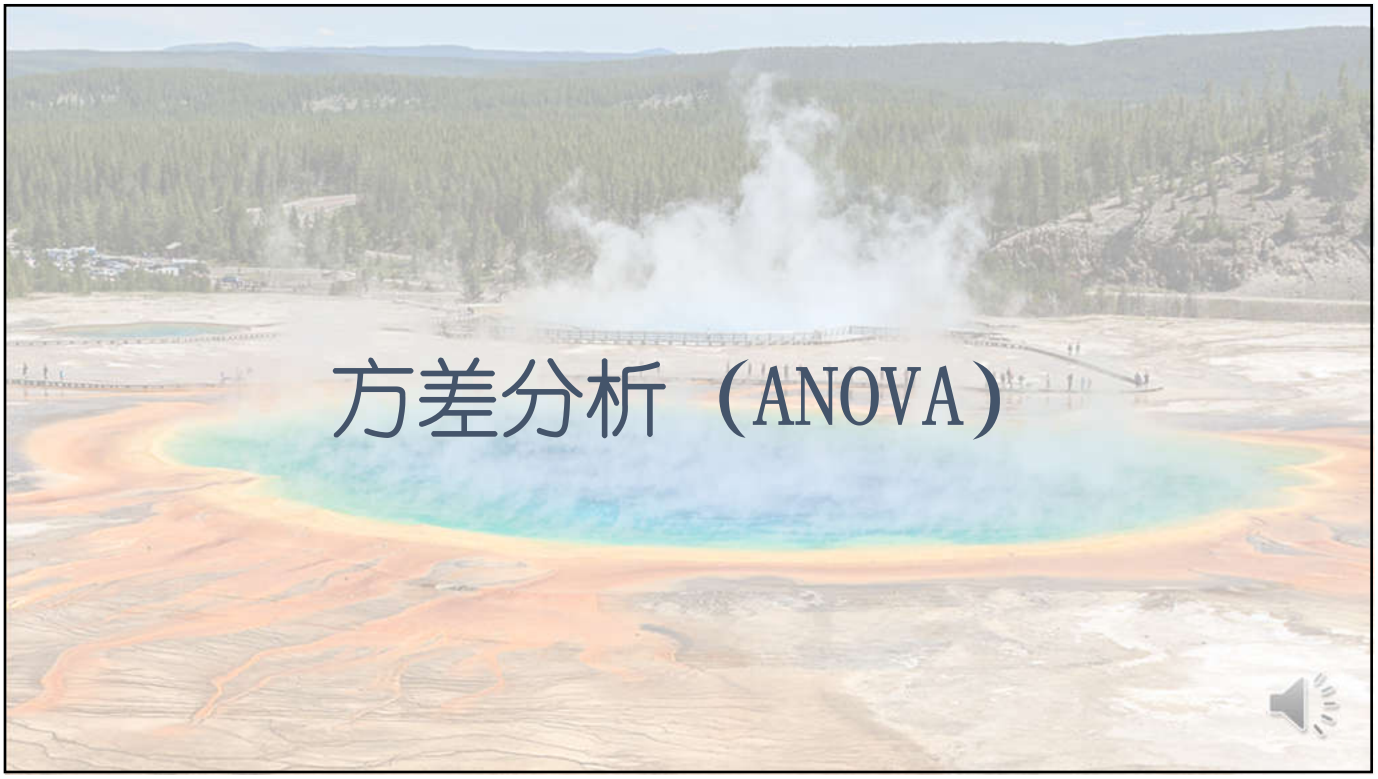
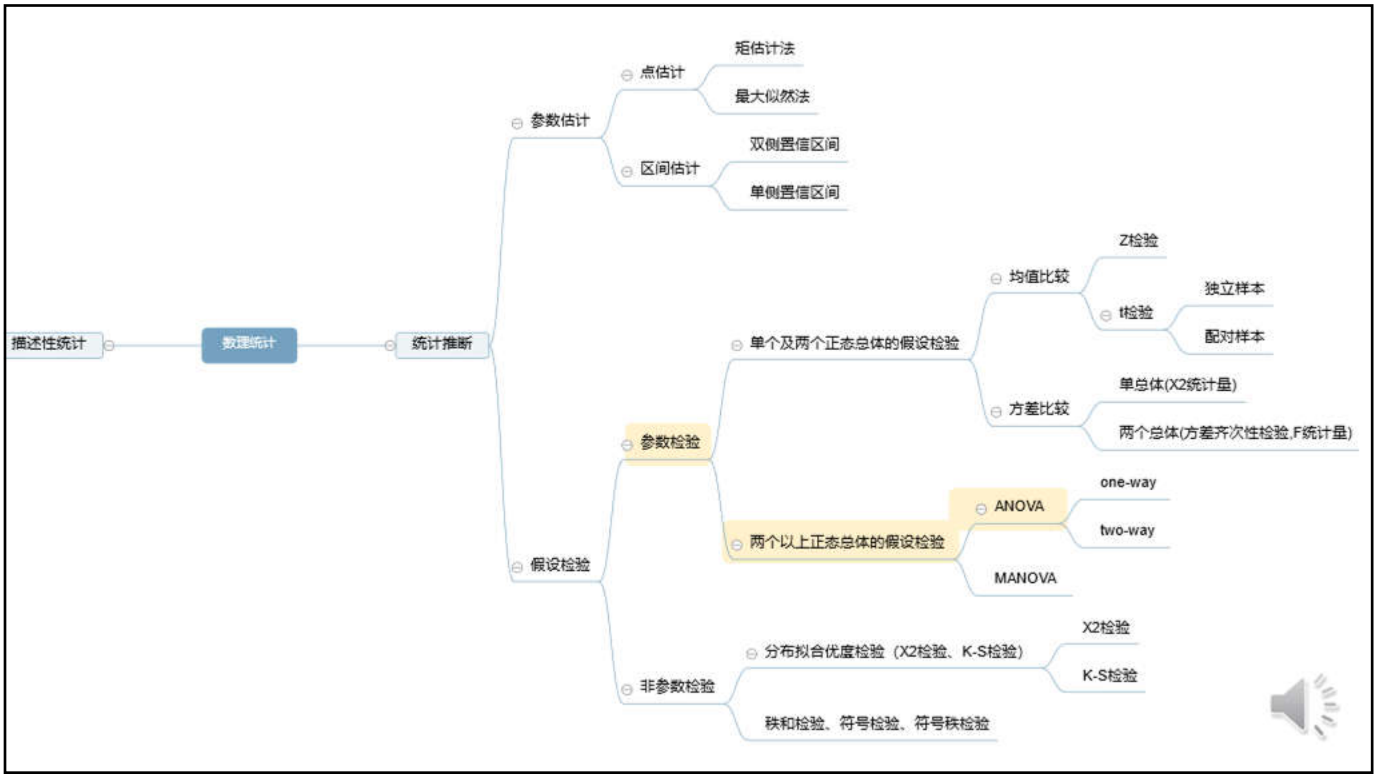


方差分析（ANOVA）

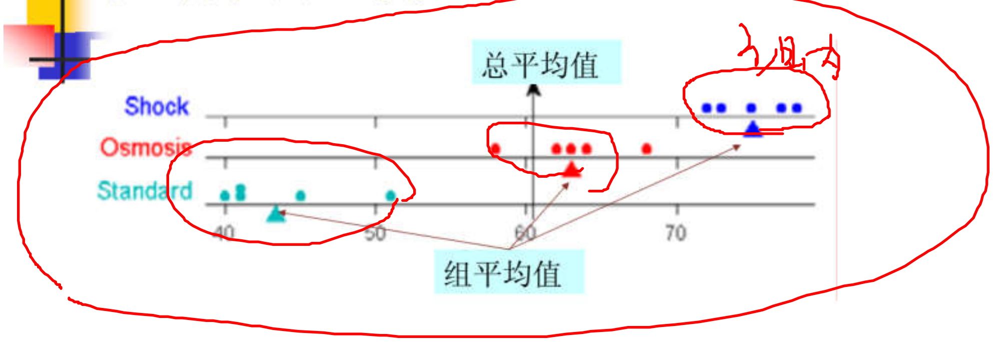


- 数理统计
  - 描述性统计
  - 统计推断
    - 参数估计
      - 点估计
        - 矩估计法
        - 最大似然法
      - 区间估计
        - 双侧置信区间
        - 单侧置信区间
    - 假设检验
      - 参数检验
        - 单个及两个正态总体的假设检验
          - 均值比较
            - Z检验
            - t检验
              - 独立样本
              - 配对样本
          - 方差比较
            - 单总体(X2统计量)
            - 两个总体(方差齐次性检验,F统计量)
        - 两个以上正态总体的假设检验
          - ANOVA
            - one-way
            - two-way
          - MANOVA
      - 非参数检验
        - 分布拟合优度检验（X2检验、K-S检验）
          - X2检验
          - K-S检验
        - 秩和检验、符号检验、符号秩检验

学习方法是否影响考试成绩？

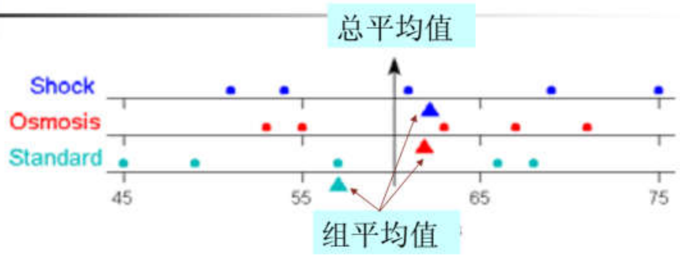
- 考虑三种学习方法
  - 标准方法(Standard)
  - 潜移默化(Osmosis)
  - 突击法(Shock)
- 15 个学生参加试验，随机分配为三组
- 八星期后得到他们的考试成绩

学习方法比较



总平均值与组平均值差别大于组内差别

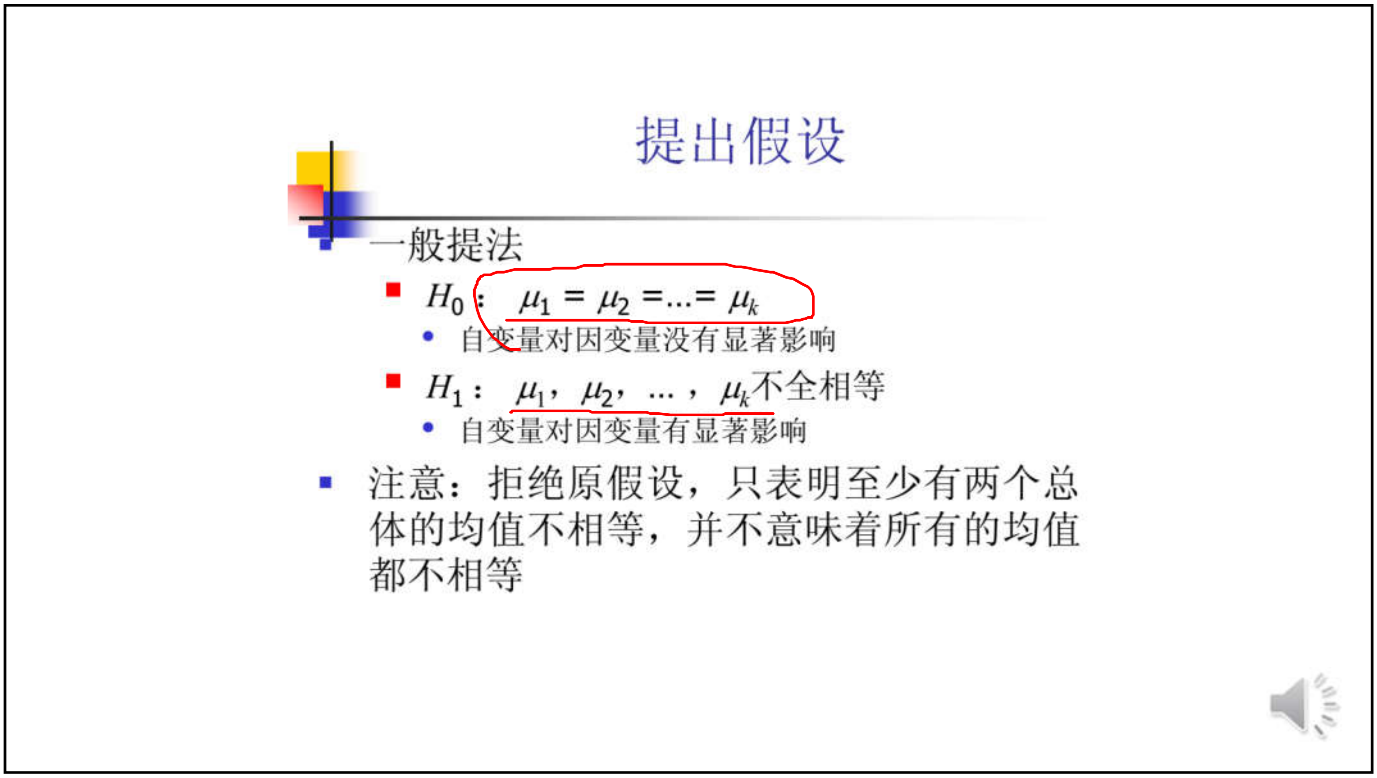
学习方法比较



总平均值与组平均值差别小于组内差别

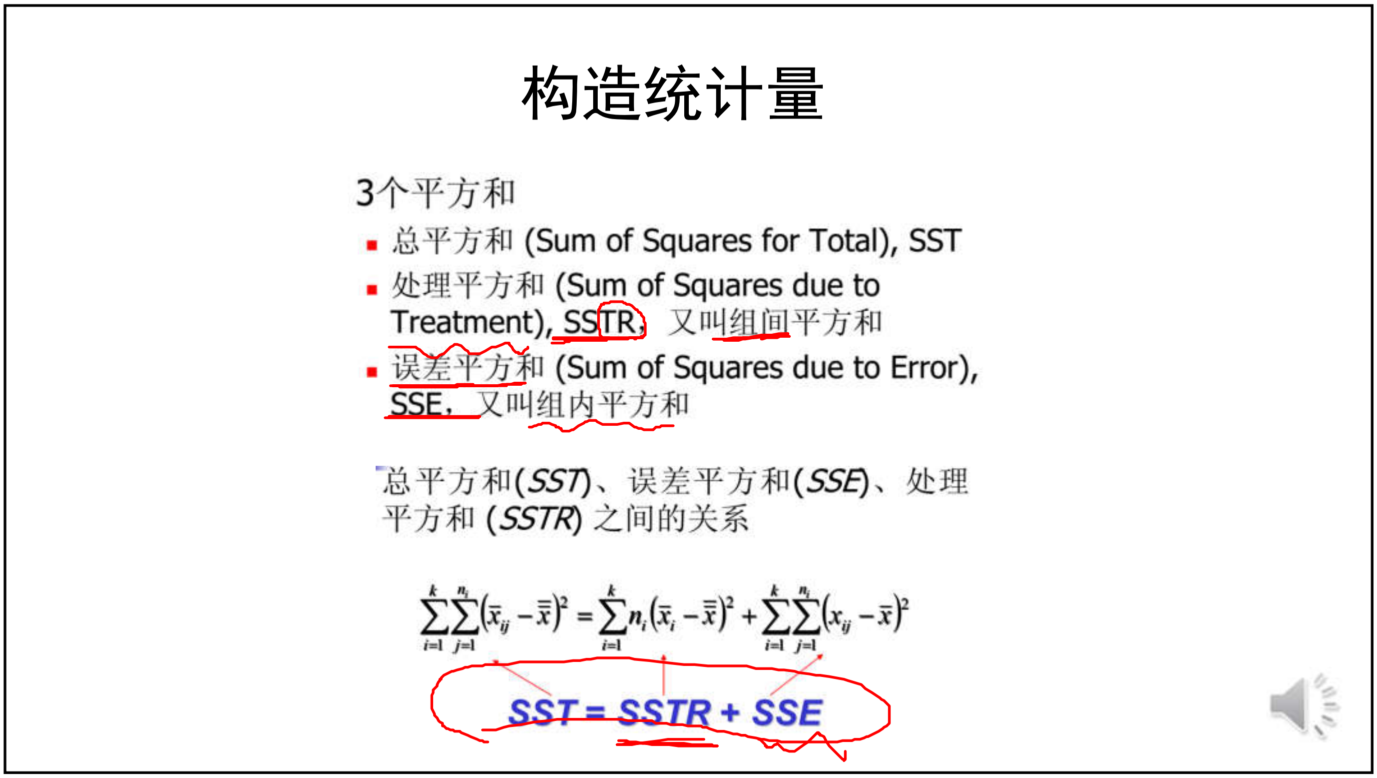
方差分析的基本思想和原理

- 比较两类误差，以检验均值是否相等
- 比较的基础是方差比
- 如果系统（处理）误差明显地不同于随机误差，则均值就是不相等的；反之，均值就是相等的
- 误差是由各部分的误差占总误差的比例来测度的



提出假设

- 一般提法
  - H_0： $\mu_1 = \mu_2 = \ldots = \mu_k$
    - 自变量对因变量没有显著影响
  - H_1： μ_1，μ_2，…，μ_k不全相等
    - 自变量对因变量有显著影响
- 注意：拒绝原假设，只表明至少有两个总体的均值不相等，并不意味着所有的均值都不相等



构造统计量

3个平方和

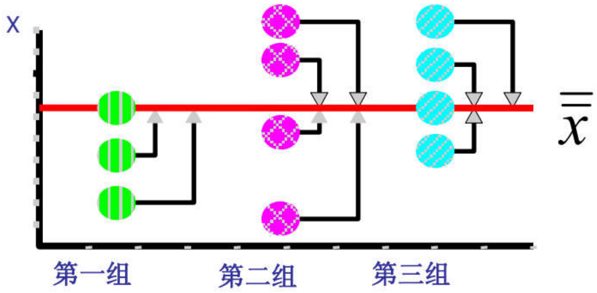
- 总平方和 (Sum of Squares for Total), SST
- 处理平方和 (Sum of Squares due to Treatment), SSTR，又叫组间平方和
- 误差平方和 (Sum of Squares due to Error), SSE，又叫组内平方和

总平方和(*SST*)、误差平方和(*SSE*)、处理平方和 (*SSTR*) 之间的关系

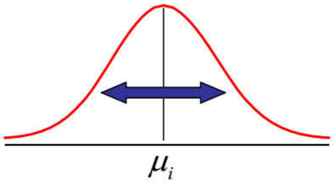
$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{x}_{ij}-\bar{\bar{x}}\right)^2=\sum_{i=1}^{k}n_i\left(\bar{x}_i-\bar{\bar{x}}\right)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}\right)^2$$

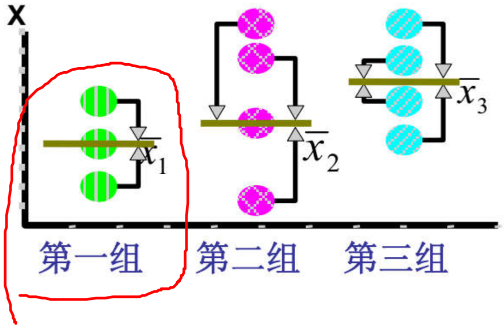
SST* = *SSTR* + *SSE

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{\bar{x}}\right)^2$$

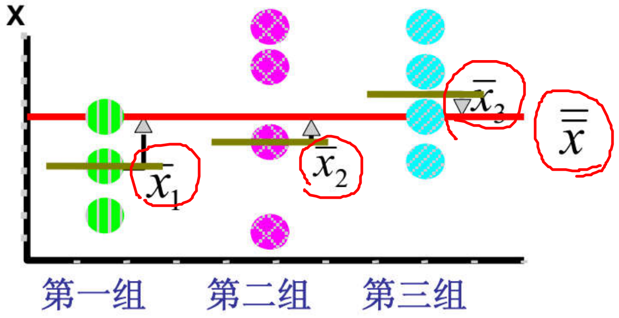


$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_i\right)^2$$





$$SSTR = \sum_{i=1}^{k} n_i \left(\bar{x}_i - \bar{\bar{x}}\right)^2$$



构造检验的统计量 (计算均方 MS)

- 处理均方：SSTR的均方，记为MSTR，计算公式为

$$MSTR = \frac{SSTR}{k-1}$$ 前例计算结果：$MSTR = \frac{1456.608696}{4-1} = 485.536232$

- 误差均方：SSE的均方，记为MSE，计算公式为

$$MSE = \frac{SSE}{n-k}$$ 前例计算结果：$MSE = \frac{2708}{23-4} = 142.526316$

构造检验的统计量
(计算检验统计量 *F*)

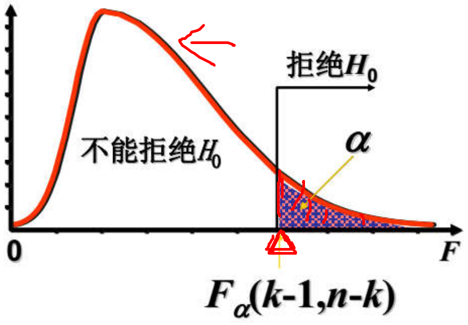
- 将*MSTR*和*MSE*进行对比，即得到所需要的检验统计量*F*
- 当H_0为真时，二者的比值服从分子自由度为k-1、分母自由度为 n-k 的 *F* 分布，即

$$F=\frac{MSTR}{MSE}\sim F(k-1,n-k)$$

前例计算结果：$F=\frac{485.536232}{142.526316}=3.406643$

构造检验的统计量
(*F*分布与拒绝域)

如果均值相等，
F=*MSTR*/*MSE*→1



F 分布

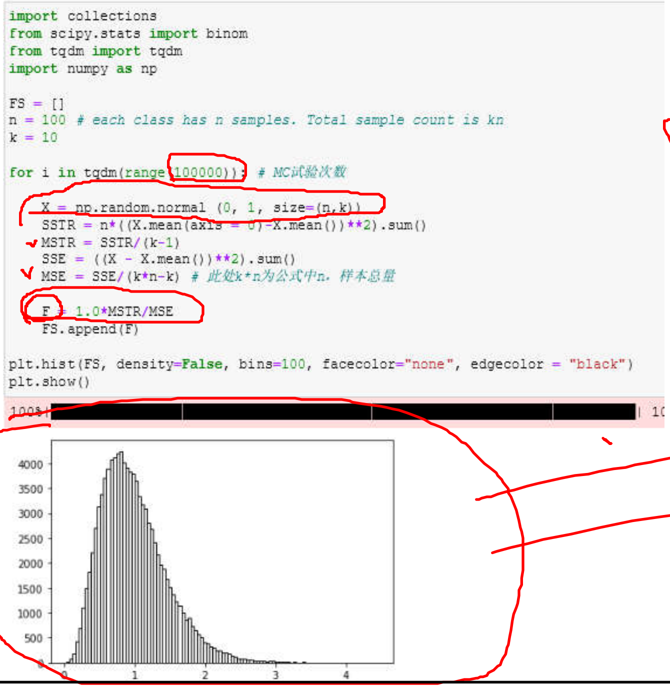
统计决策

将统计量的值F与给定的显著性水平α的临界值F_α进行比较，作出对原假设H_0的决策

- 根据给定的显著性水平α，在F分布表中查找与第一自由度$df_1 = k\text{-}1$、第二自由度$df_2 = n\text{-}k$ 相应的临界值 F_α
- 若$F > F_\alpha$，则拒绝原假设H_0，表明均值之间的差异是显著的，所检验的因子对观察值有显著影响
- 若$F < F_\alpha$，则不能拒绝原假设H_0，无证据支持表明所检验的因子对观察值有显著影响

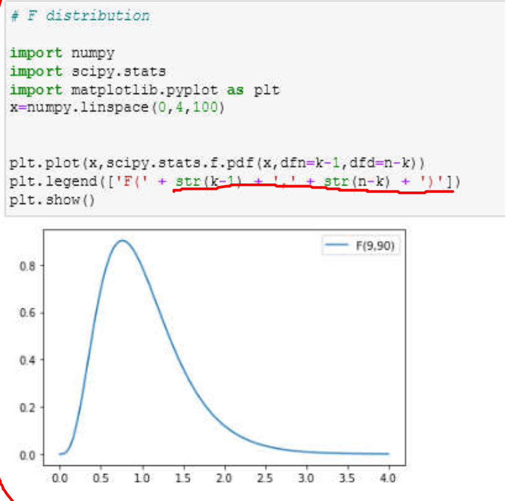
验证 ANOVA的F分布假设

F = MSTR/MSE ~ F(k-1, n-k)



```
import collections
from scipy.stats import binom
from tqdm import tqdm
import numpy as np

FS = []
n = 100 # each class has n samples. Total sample count is kn
k = 10

for i in tqdm(range(100000)): # MC试验次数

    X = np.random.normal (0, 1, size=(n,k))
    SSTR = n*((X.mean(axis = 0)-X.mean())**2).sum()
    MSTR = SSTR/(k-1)
    SSE = ((X - X.mean())**2).sum()
    MSE = SSE/(k*n-k) # 此处k*n为公式中n，样本总量

    F = 1.0*MSTR/MSE
    FS.append(F)

plt.hist(FS, density=False, bins=100, facecolor="none", edgecolor = "black")
plt.show()
```

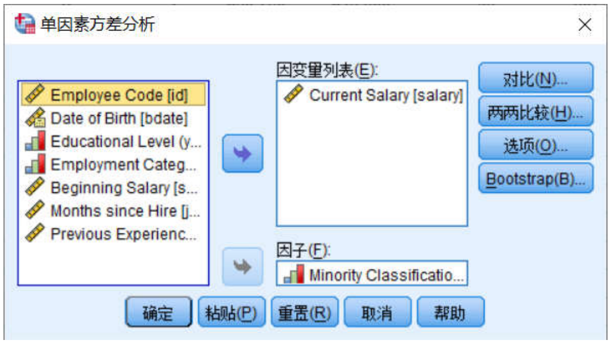
```
100%|████████████████████████████████████████| 10
```



```
# F distribution

import numpy
import scipy.stats
import matplotlib.pyplot as plt
x=numpy.linspace(0,4,100)


plt.plot(x,scipy.stats.f.pdf(x,dfn=k-1,dfd=n-k))
plt.legend(['F(' + str(k-1) + ',' + str(n-k) + ')'])
plt.show()
```

验证完毕，抽样分布的结果符合F分布

SPSS：分析 > 比较均值 > 单因素ANOVA



因变量列表(E): Current Salary [salary]

因子(F): Minority Classificatio...

Oneway

[数据集3] C:\Users\eleve\Desktop\ST_UG_2021\data\ch3\Employee data.sav

ANOVA

Current Salary

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4.337E9	1	4.337E9	15.326	.000
Within Groups	1.336E11	472	2.830E8		
Total	1.379E11	473			

MSTR
MSE
P

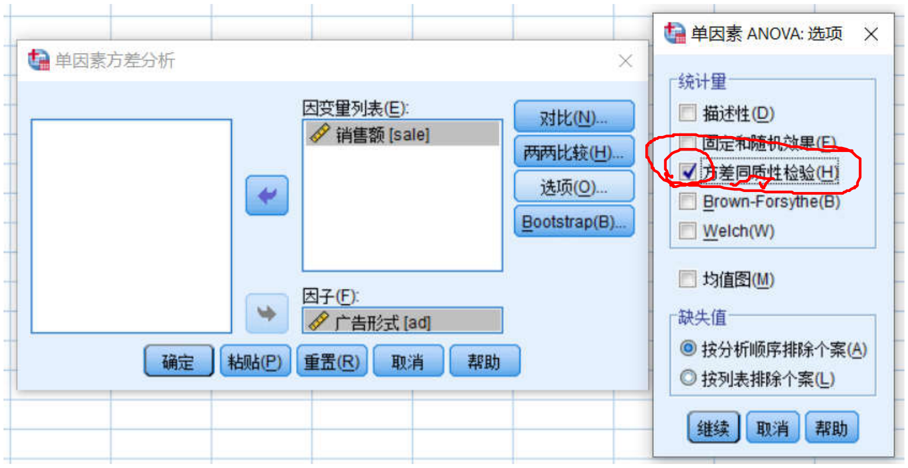
ANOVA注意事项 － 方差齐次检验

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）。

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

若组间方差不齐则不适用方差分析。但可通过对数变换、平方根变换、倒数变换、平方根反正弦变换等方法变换后再进行方差齐性检验,若还不行只能进行非参数检验.



单因素方差分析

因变量列表(E):
销售额 [sale]

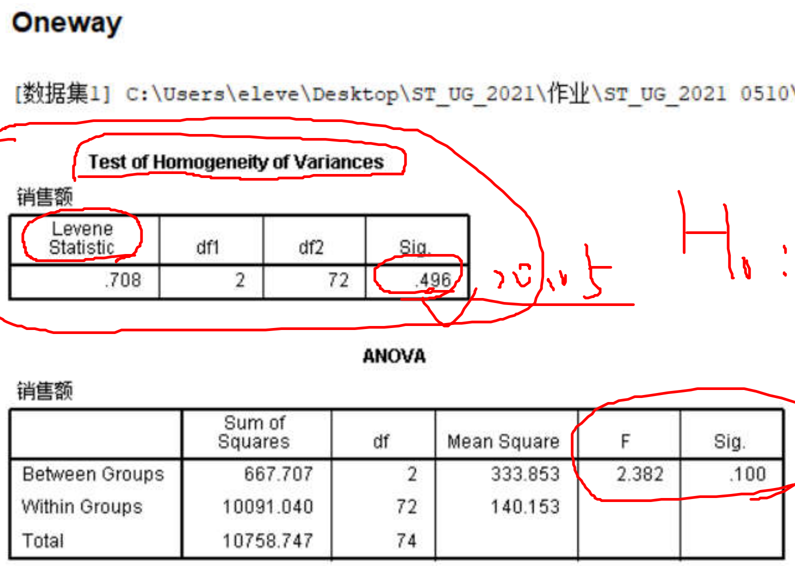
因子(F):
广告形式 [ad]

单因素 ANOVA: 选项

统计量
- □ 描述性(D)
- □ 固定和随机效果(E)
- ☑ 方差同质性检验(H)
- □ Brown-Forsythe(B)
- □ Welch(W)

□ 均值图(M)

缺失值
- ● 按分析顺序排除个案(A)
- ○ 按列表排除个案(L)



Oneway

[数据集1] C:\Users\eleve\Desktop\ST_UG_2021\作业\ST_UG_2021 0510\

Test of Homogeneity of Variances

销售额

Levene Statistic	df1	df2	Sig.
.708	2	72	.496

>0.05

$H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$

ANOVA

销售额

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	667.707	2	333.853	2.382	.100
Within Groups	10091.040	72	140.153		
Total	10758.747	74			

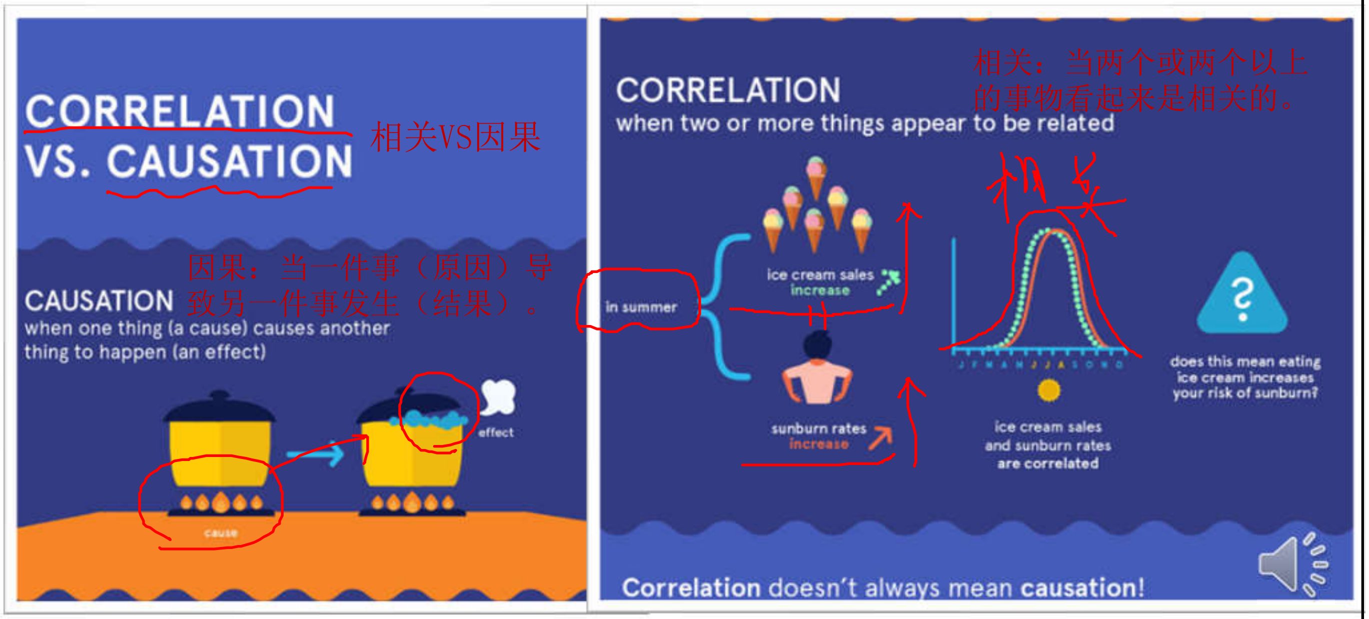
Levene检验是统计分析的一个组成部分。在进行其他统计分析（例如t检验和方差分析）之前，它可用于检验方差的齐次性/同质性。

Levene's test is an integral part of **statistical** analysis. It can be used to test the homogeneity of variance before we proceed with other **statistical** analyses such as the Student's t-test and Analysis of Variance (ANOVA).

Eta squared 是一种效应量指标 (effect size)，常用于方差分析中。代表的是通过方差分析能够得到解释的因变量变异程度 (sum of squares effect)在所有变异程度 (sum of squares total)中所占的比例。一般希望eta square值越大越好，>0.01 为较小效应量，>0.02为中等效应量，>0.083可视为大效应量

相关分析

因果 vs 相关


CORRELATION
VS. CAUSATION
CAUSATION
when one thing (a cause) causes another thing to happen (an effect)
cause
effect
CORRELATION
when two or more things appear to be related
in summer
ice cream sales increase
sunburn rates increase
J F M A M J J A S O N D
ice cream sales and sunburn rates are correlated
does this mean eating ice cream increases your risk of sunburn?
Correlation doesn't always mean causation!
相 关

相关VS因果

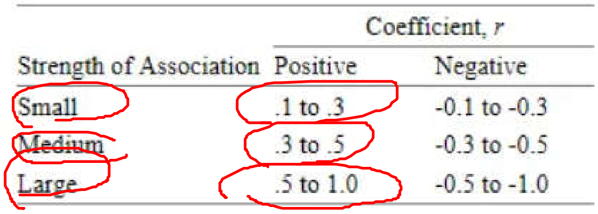
因果：当一件事（原因）导致另一件事发生（结果）。

相关：当两个或两个以上的事物看起来是相关的。

Pearson's r Correlation 皮尔逊相关系数

[-1, 1]

$|r|$



	Coefficient, *r*	
Strength of Association	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to -1.0

Pearson's r measure the linear correlation between two variables.

皮尔逊相关系数用来测度两个变量之间的线性相关性

Pearson's r2 (r square)

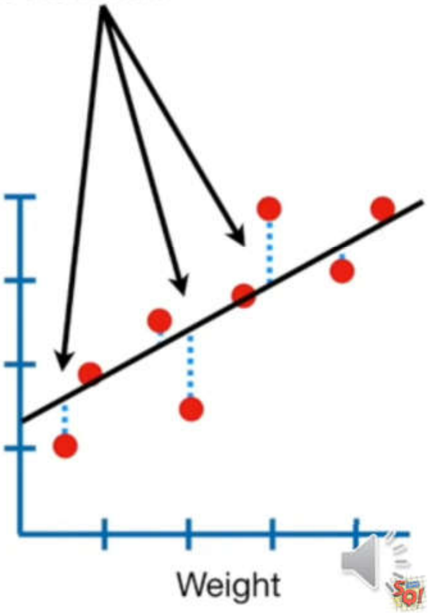
It is the 'percentage' of the independent variable x explains the dependent variable y. Value range [0,1]

Some researchers recommend r^2 over r

它是自变量x能够解释因变量y信息的“百分比”。取值范围[0,1]，一些研究者建议用 R2(r^2) 代替 r。

In brief, we square the residuals and then add them up. I call this **SS(fit)**, for "sum of squares for the Residuals around the best fitting line"…

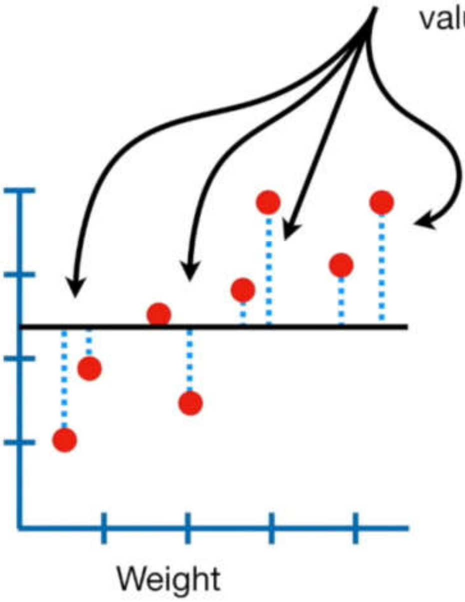
我们把残差平方和称为SS（拟合），即“最佳拟合线附近残差的平方和”。



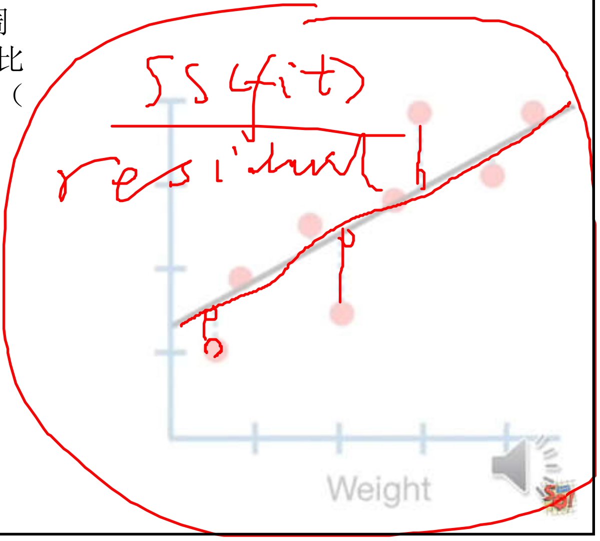
We also use R2 in linear regression as a measure of good fit

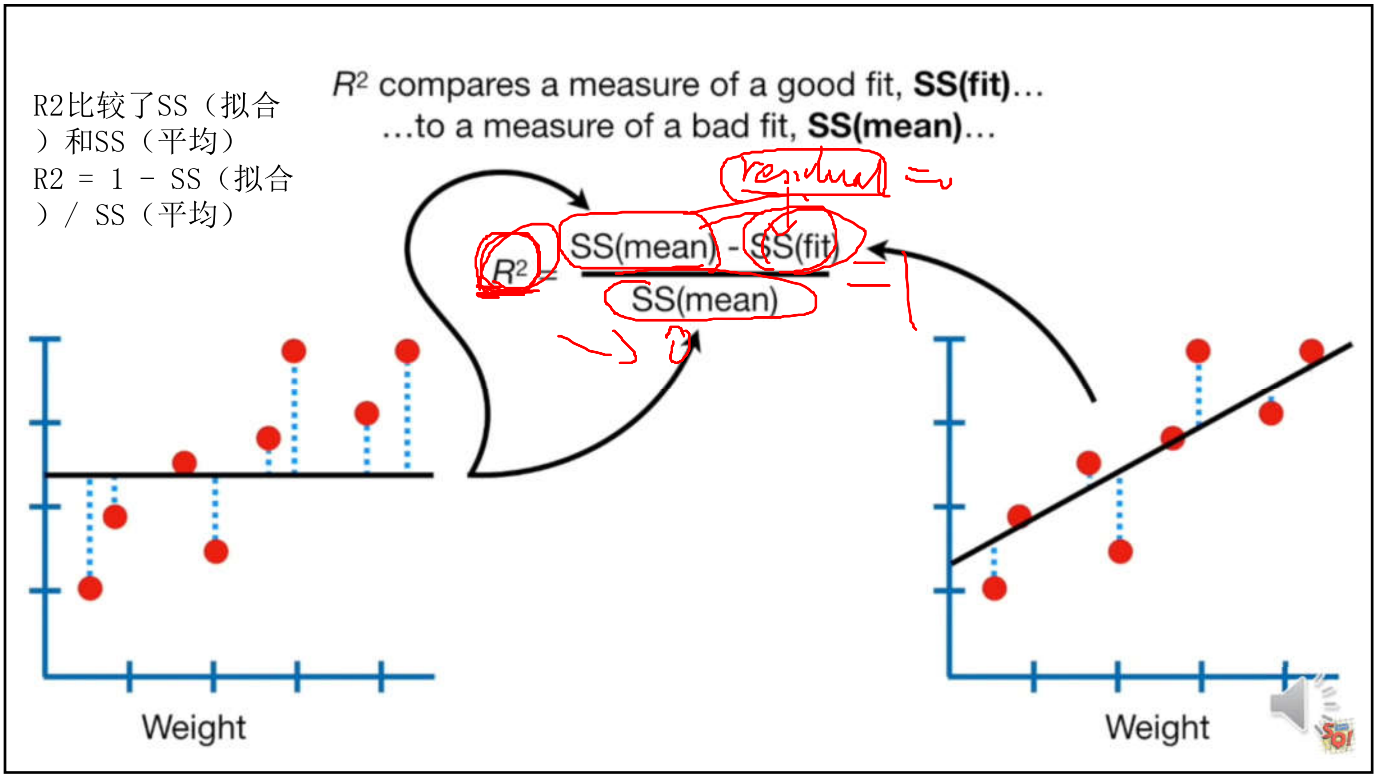
我们在线性回归中使用R2作为拟合优度的度量

…and we compare that to the sum of squared residuals around the worst fitting line, the mean of the y-axis values. This is called **SS(mean)**.



我们将其与最差拟合线（y轴值的平均值）周围的残差平方和进行比较。这就是所谓的SS（平均），即Y的方差





R2比较了SS（拟合）和SS（平均）
R2 = 1 - SS（拟合）/ SS（平均）

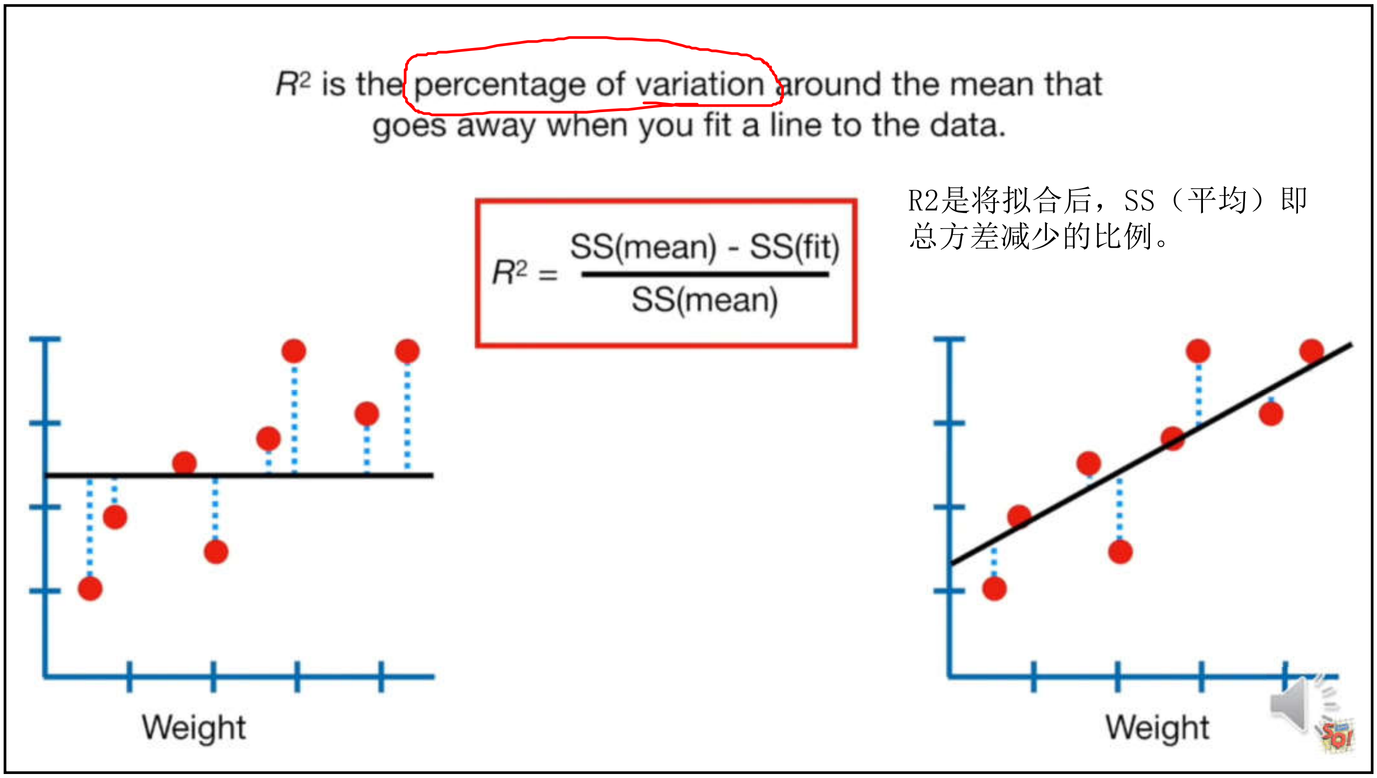
R^2 compares a measure of a good fit, **SS(fit)**…
…to a measure of a bad fit, **SS(mean)**…

$$R^2 = \frac{\text{SS(mean)} - \overset{\text{residual}}{\text{SS(fit)}}}{\text{SS(mean)}}$$

residual =0

=1

→ 0



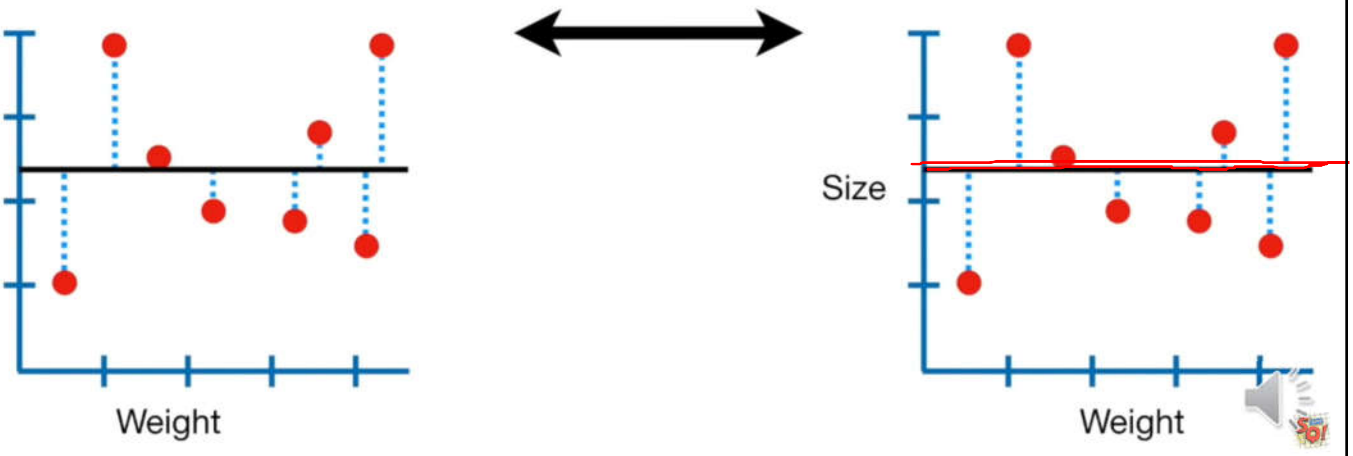
R^2 is the percentage of variation around the mean that goes away when you fit a line to the data.

$$\boxed{R^2 = \frac{\text{SS(mean)} - \text{SS(fit)}}{\text{SS(mean)}}}$$

R2是将拟合后，SS（平均）即总方差减少的比例。

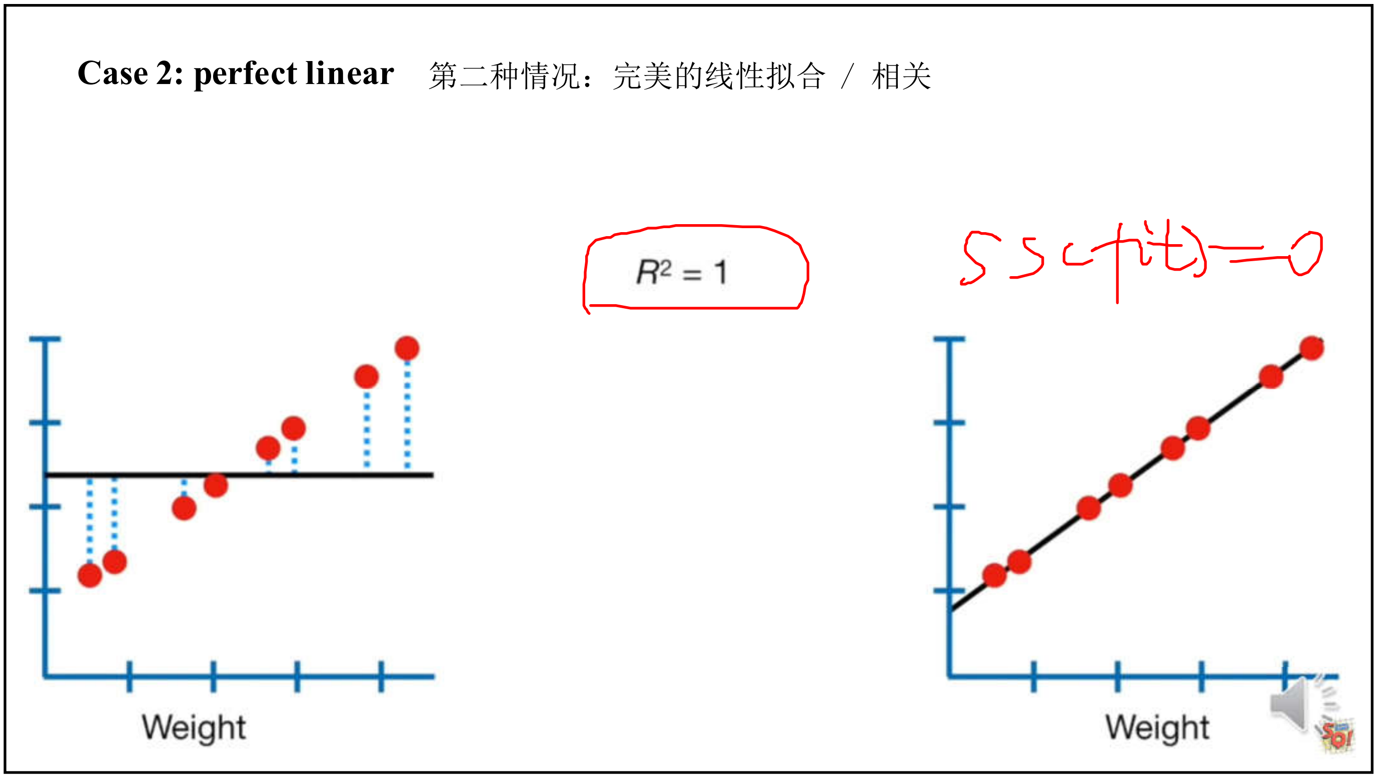
Case 1: non-linear 第一种情况：非线性 / 不相关

…in this case, SS(fit) = SS(mean)…



$R^2 = 0$

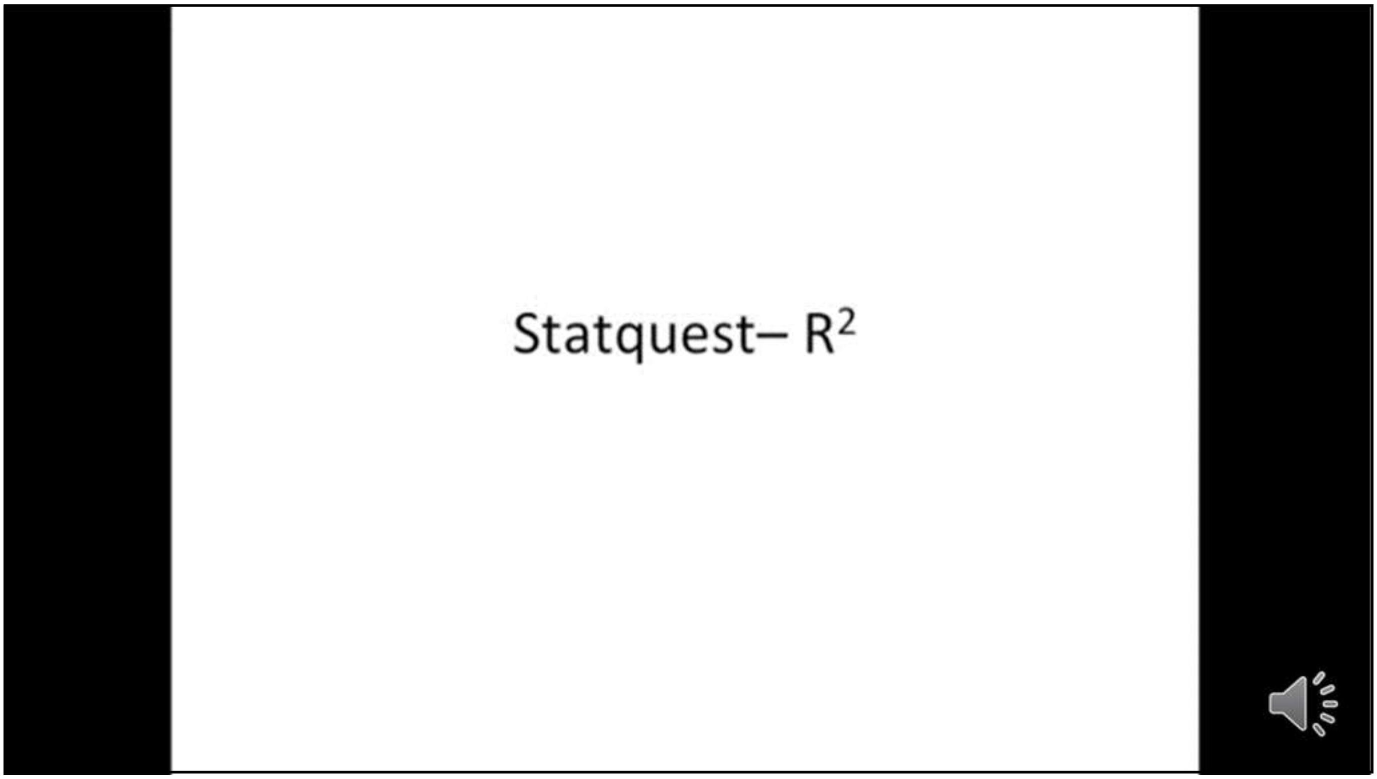
Weight
Weight



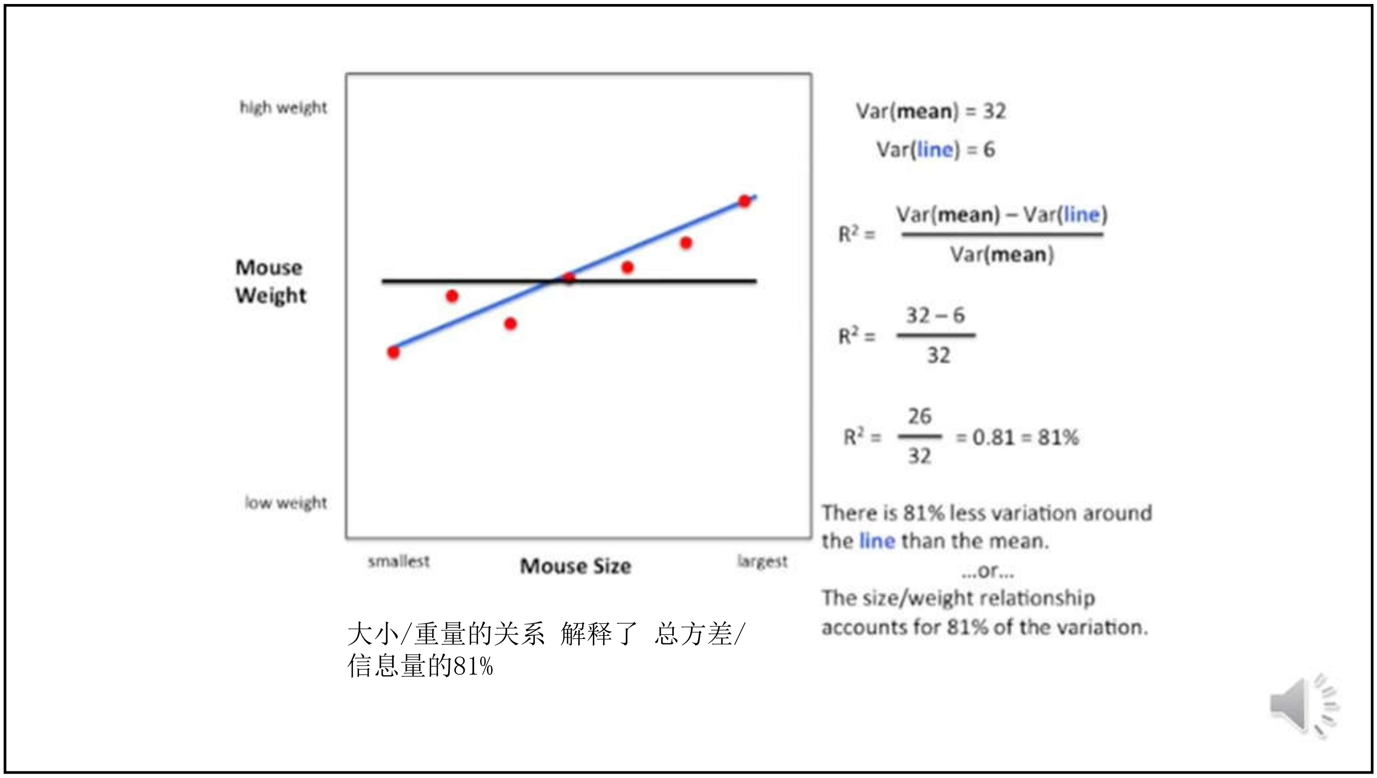
Case 2: perfect linear 第二种情况：完美的线性拟合 / 相关

$\boxed{R^2 = 1}$

$SS(fit) = 0$



Statquest– R^2



Var(**mean**) = 32

Var(line) = 6

$$R^2 = \frac{\text{Var(mean)} - \text{Var(line)}}{\text{Var(mean)}}$$

$$R^2 = \frac{32 - 6}{32}$$

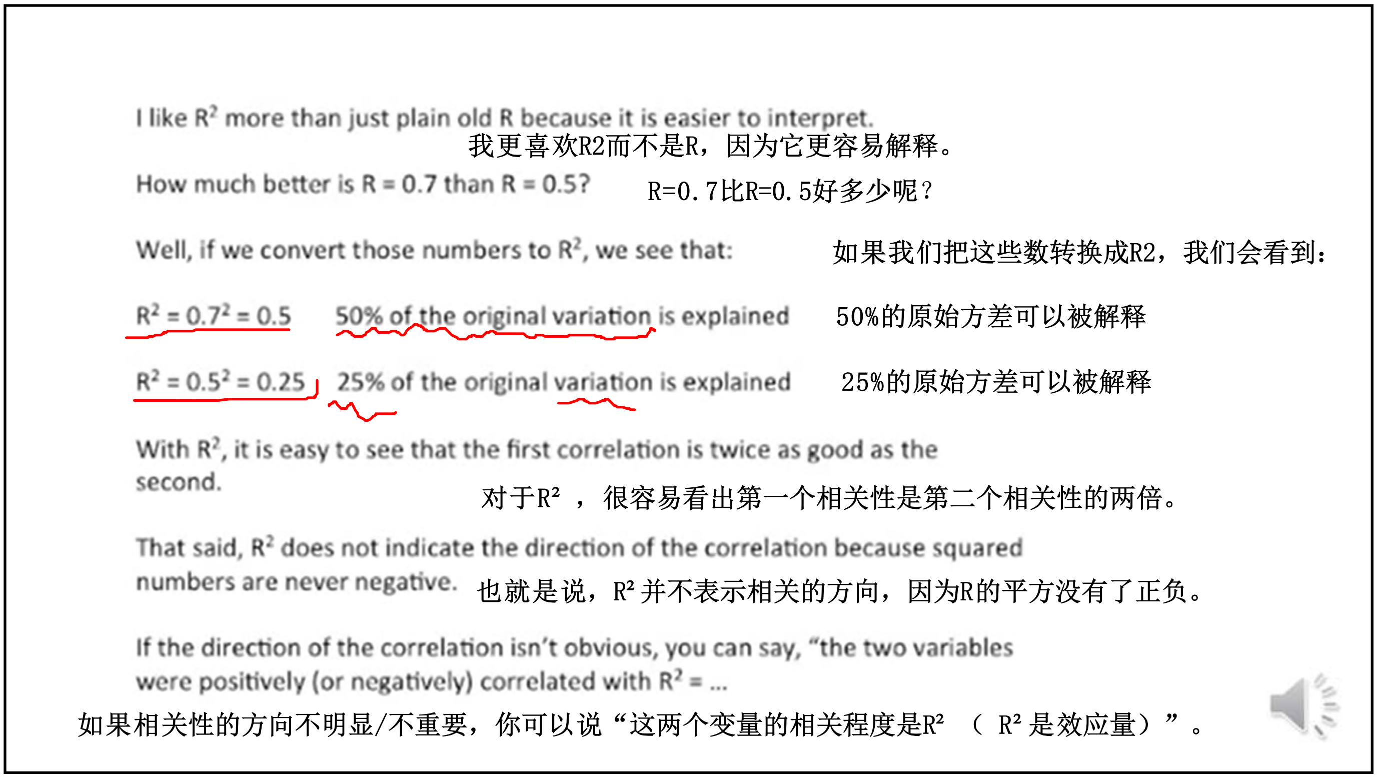
$$R^2 = \frac{26}{32} = 0.81 = 81\%$$

There is 81% less variation around the line than the mean.

...or...

The size/weight relationship accounts for 81% of the variation.

大小/重量的关系 解释了 总方差/信息量的81%



I like R^2 more than just plain old R because it is easier to interpret.

我更喜欢R2而不是R，因为它更容易解释。

How much better is R = 0.7 than R = 0.5?　R=0.7比R=0.5好多少呢？

Well, if we convert those numbers to R^2, we see that:　如果我们把这些数转换成R2，我们会看到：

$R^2 = 0.7^2 = 0.5$　50% of the original variation is explained　50%的原始方差可以被解释

$R^2 = 0.5^2 = 0.25$　25% of the original variation is explained　25%的原始方差可以被解释

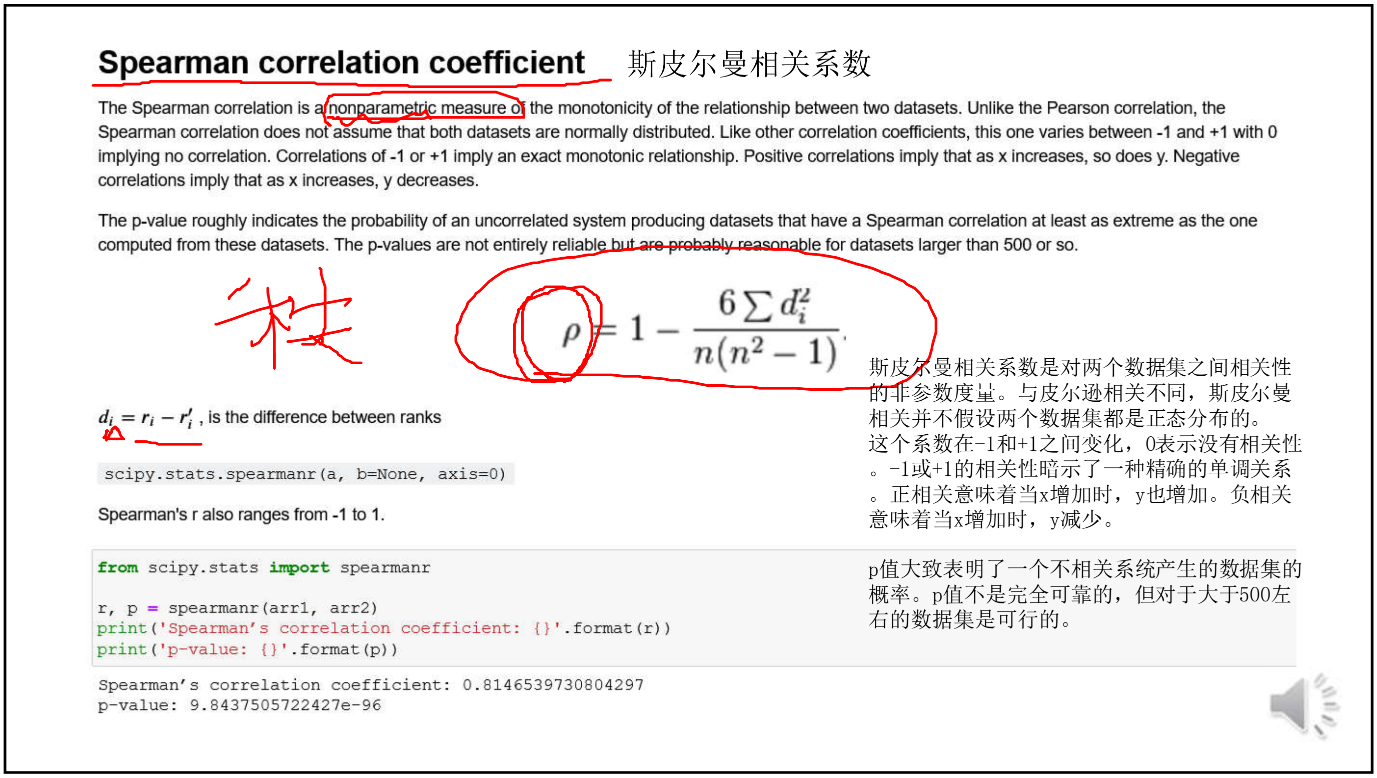
With R^2, it is easy to see that the first correlation is twice as good as the second.

对于R²，很容易看出第一个相关性是第二个相关性的两倍。

That said, R^2 does not indicate the direction of the correlation because squared numbers are never negative.　也就是说，R²并不表示相关的方向，因为R的平方没有了正负。

If the direction of the correlation isn't obvious, you can say, "the two variables were positively (or negatively) correlated with R^2 = ...

如果相关性的方向不明显/不重要，你可以说“这两个变量的相关程度是R²（R²是效应量）”。



Spearman correlation coefficient 斯皮尔曼相关系数

The Spearman correlation is a nonparametric measure of the monotonicity of the relationship between two datasets. Unlike the Pearson correlation, the Spearman correlation does not assume that both datasets are normally distributed. Like other correlation coefficients, this one varies between -1 and +1 with 0 implying no correlation. Correlations of -1 or +1 imply an exact monotonic relationship. Positive correlations imply that as x increases, so does y. Negative correlations imply that as x increases, y decreases.

The p-value roughly indicates the probability of an uncorrelated system producing datasets that have a Spearman correlation at least as extreme as the one computed from these datasets. The p-values are not entirely reliable but are probably reasonable for datasets larger than 500 or so.

秩

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)},$$

$d_i = r_i - r'_i$, is the difference between ranks

```
scipy.stats.spearmanr(a, b=None, axis=0)
```

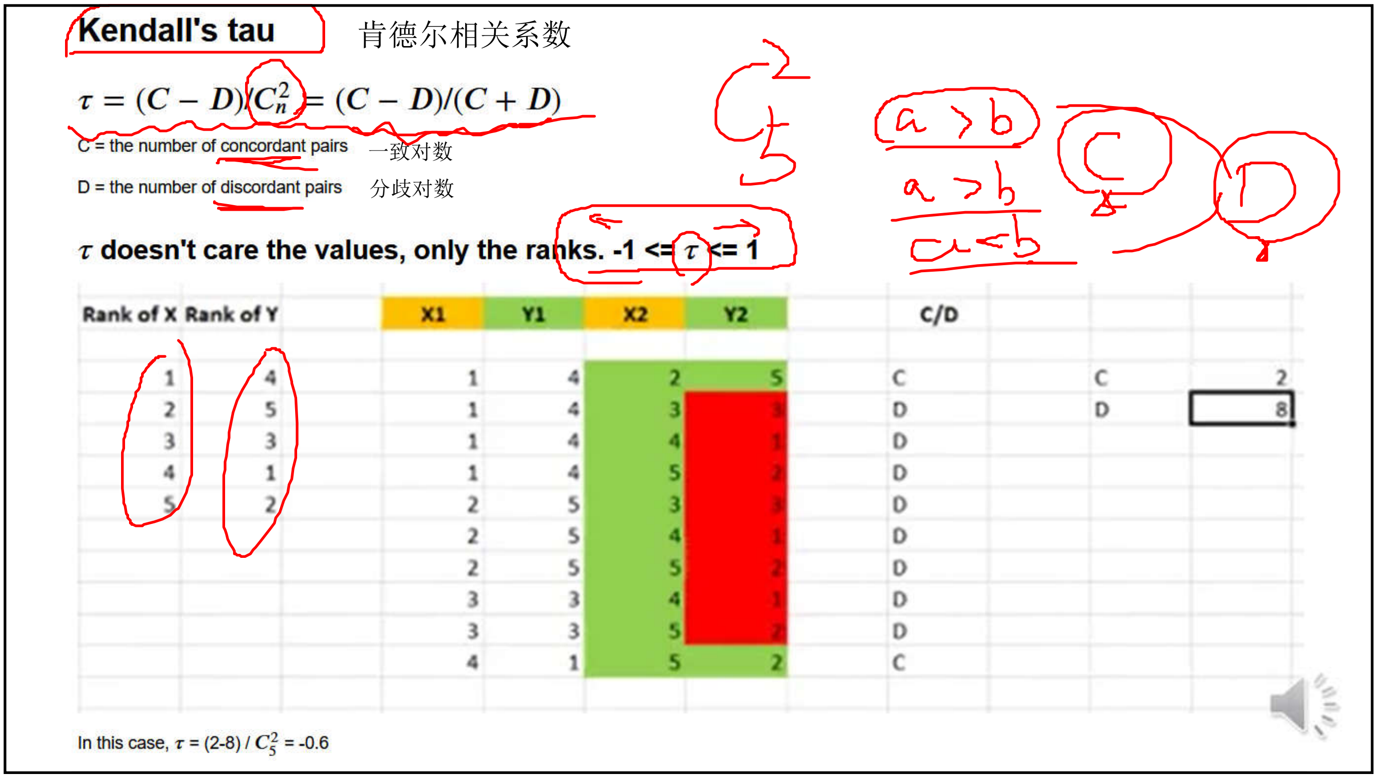
Spearman's r also ranges from -1 to 1.

```
from scipy.stats import spearmanr

r, p = spearmanr(arr1, arr2)
print('Spearman’s correlation coefficient: {}'.format(r))
print('p-value: {}'.format(p))
```

```
Spearman’s correlation coefficient: 0.8146539730804297
p-value: 9.8437505722427e-96
```

斯皮尔曼相关系数是对两个数据集之间相关性的非参数度量。与皮尔逊相关不同，斯皮尔曼相关并不假设两个数据集都是正态分布的。这个系数在-1和+1之间变化，0表示没有相关性。-1或+1的相关性暗示了一种精确的单调关系。正相关意味着当x增加时，y也增加。负相关意味着当x增加时，y减少。

p值大致表明了一个不相关系统产生的数据集的概率。p值不是完全可靠的，但对于大于500左右的数据集是可行的。



Kendall's tau 肯德尔相关系数

$$\tau = (C - D)/C_n^2 = (C - D)/(C + D)$$

C_5^2

C = the number of concordant pairs 一致对数

D = the number of discordant pairs 分歧对数

a > b

a > b

a < b

C

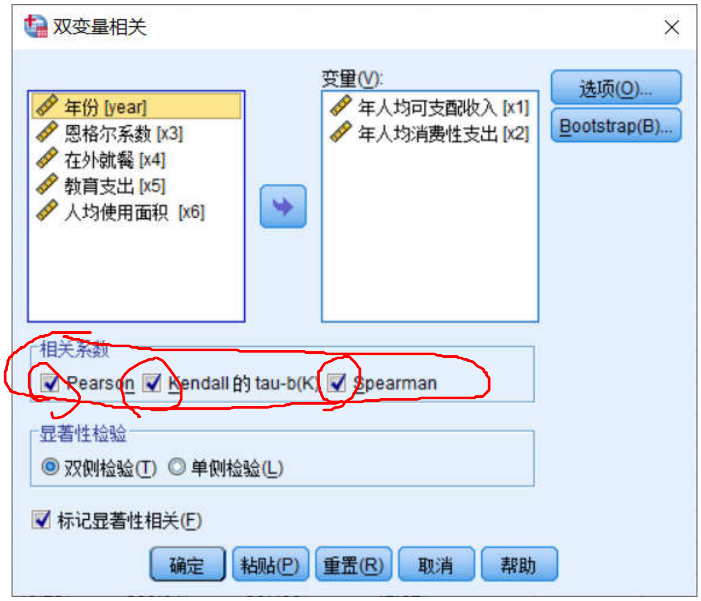
D

τ doesn't care the values, only the ranks. -1 <= τ <= 1

Rank of X	Rank of Y		X1	Y1	X2	Y2		C/D			
1	4		1	4	2	5		C		C	2
2	5		1	4	3	3		D		D	8
3	3		1	4	4	1		D			
4	1		1	4	5	2		D			
5	2		2	5	3	3		D			
			2	5	4	1		D			
			2	5	5	2		D			
			3	3	4	1		D			
			3	3	5	2		D			
			4	1	5	2		C			

In this case, τ = (2-8) / C_5^2 = -0.6

SPSS：分析 > 相关 > 双变量



双变量相关

- 年份 [year]
- 恩格尔系数 [x3]
- 在外就餐 [x4]
- 教育支出 [x5]
- 人均使用面积 [x6]

变量(V):

- 年人均可支配收入 [x1]
- 年人均消费性支出 [x2]

相关系数

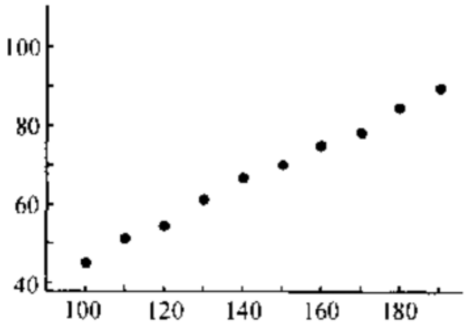
☑ Pearson ☑ Kendall 的 tau-b(K) ☑ Spearman

显著性检验

● 双侧检验(T) ○ 单侧检验(L)

☑ 标记显著性相关(F)

回归分析 & GLzM



$H_0:\quad b = 0,$

$H_1:\quad b \neq 0.$

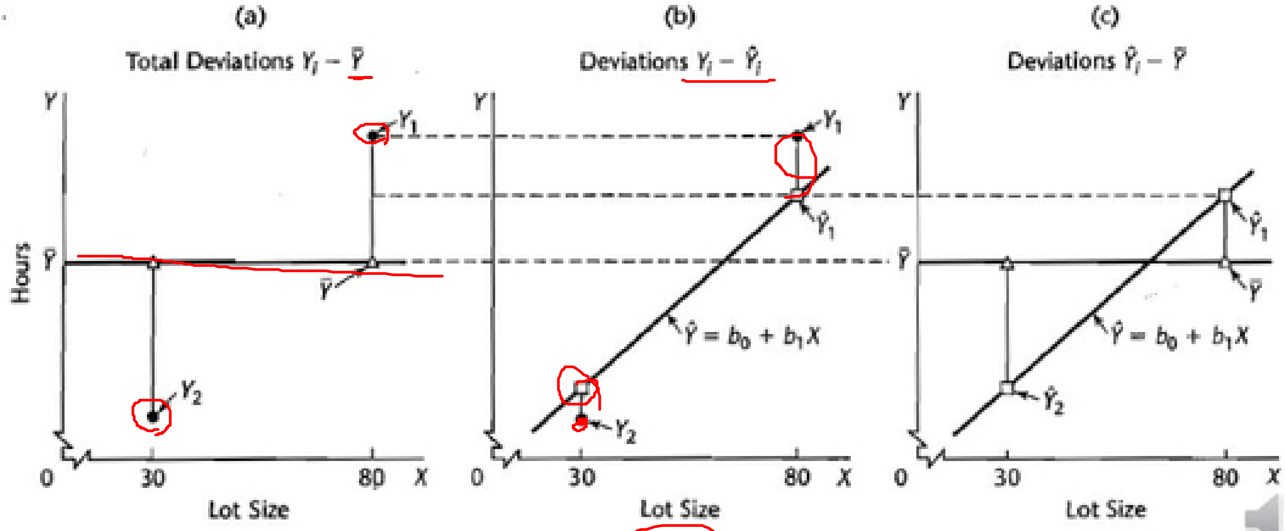
$$Y = a + bx + \varepsilon,\quad \varepsilon \sim N(0,\sigma^2). \qquad (3.2)$$

其中未知参数 a,b 及 σ^2 都不依赖于 x.(3.2) 称为**一元线性回归模型**,其中 b 称为回归系数.

(3.2) 式表明,因变量 Y 由两部分组成,一部分是 x 的线性函数 $a + bx$,另一部分 $\varepsilon \sim N(0,\sigma^2)$ 是随机误差,是人们不可控制的.

ANOVA & 线性回归

FIGURE 2.7 **Illustration of Partitioning of Total Deviations $Y_i - \bar{Y}$—Toluca Company Example (not drawn to scale; only observations Y_1 and Y_2 are shown).**

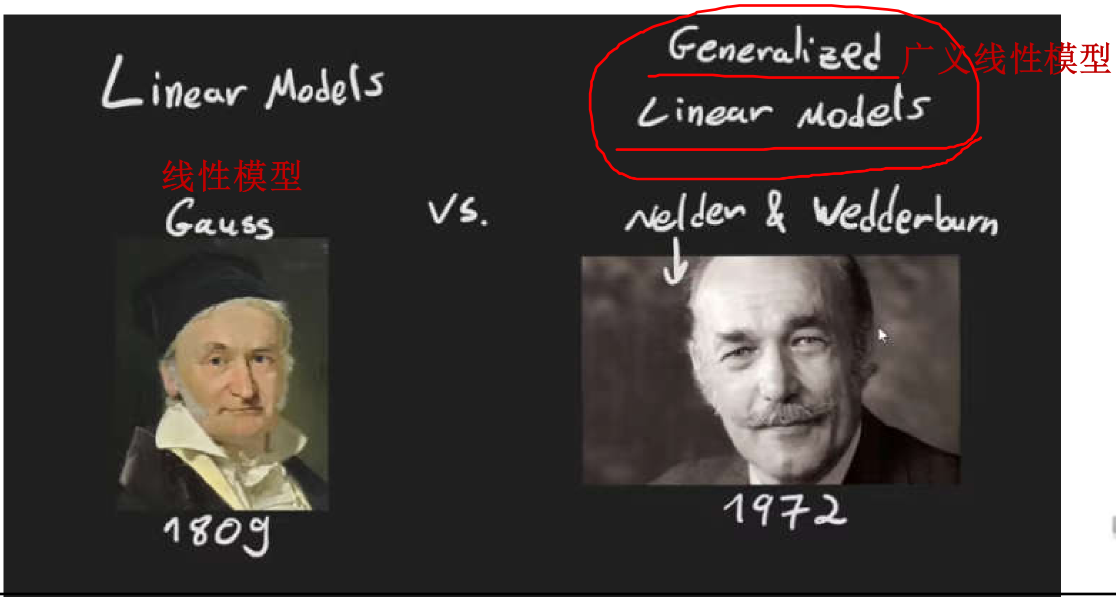

(a) Total Deviations $Y_i - \bar{Y}$

(b) Deviations $Y_i - \hat{Y}_i$

(c) Deviations $\hat{Y}_i - \bar{Y}$

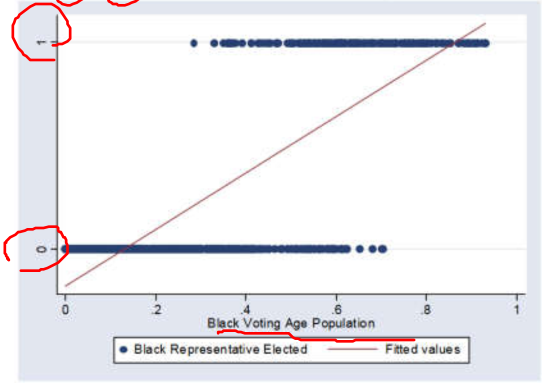
SST = SSE + SSR (SSTR)

广义线性模型
Generalized linear model



二元 Y

Dichotomous Dependent Vars.



The line doesn't fit the data very well.

And if we take values of Y between 0 and 1 to be probabilities, this doesn't make sense

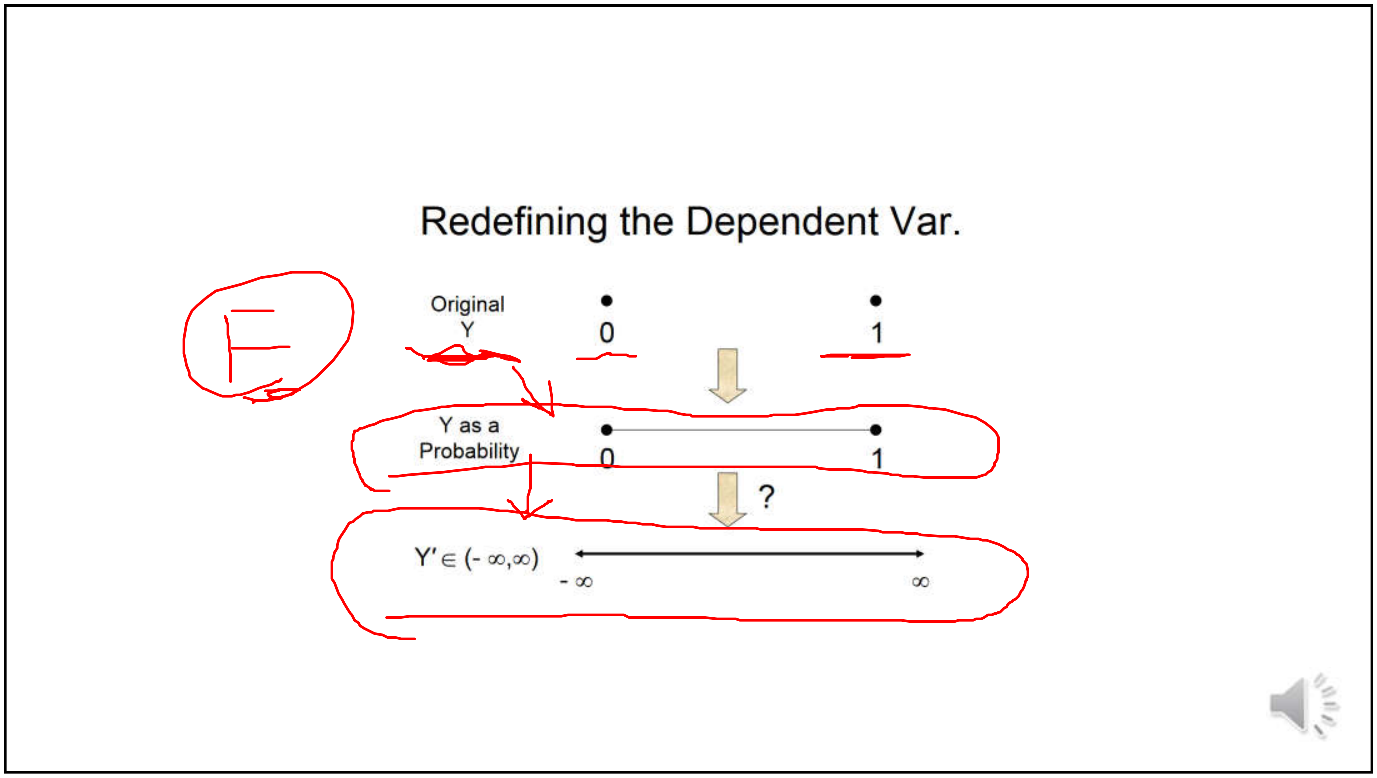
这条线与数据的拟合度不是特别好
如果我们把0到1之间的y值作为概率，是不合理的

重新定义因变量
Redefining the Dependent Var.

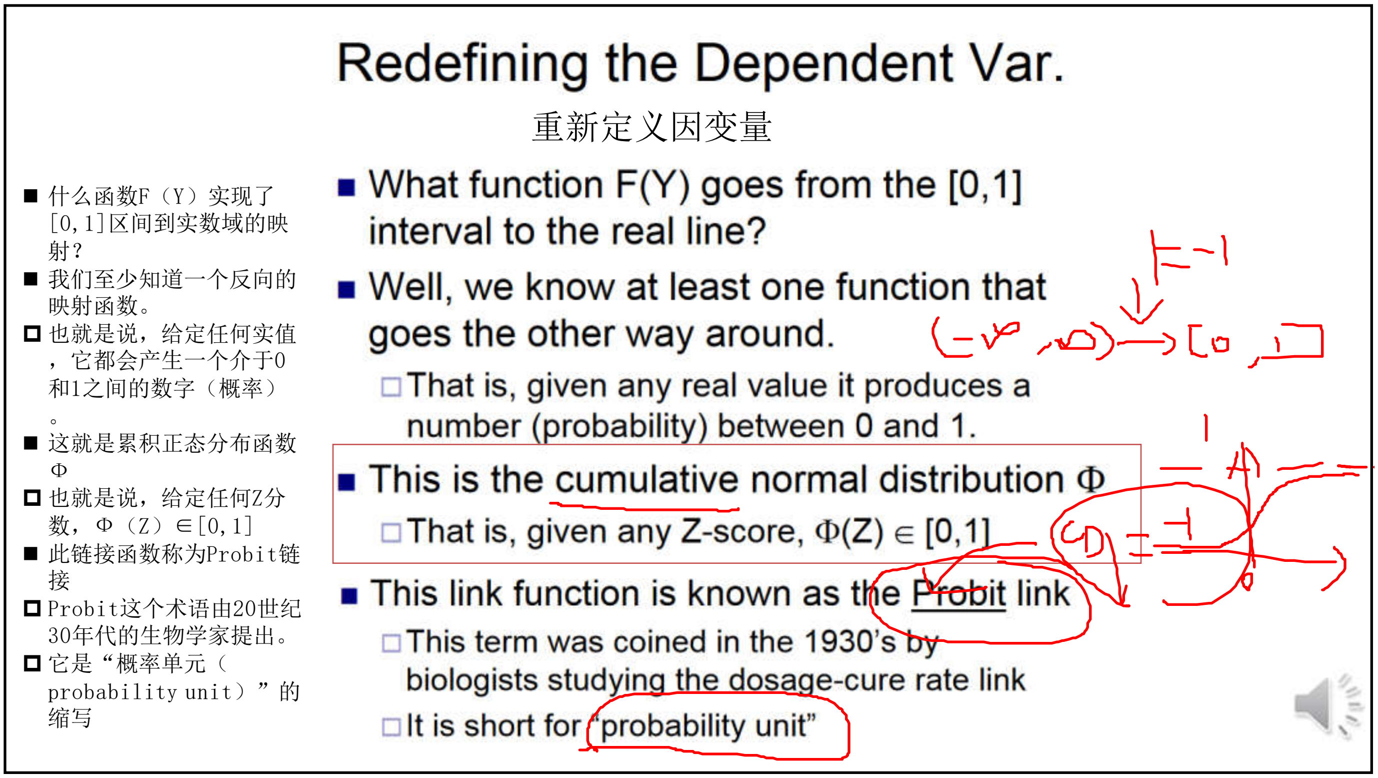
- How to solve this problem?
- We need to transform the dichotomous Y into a continuous variable Y′ ∈ (-∞,∞)
- So we need a link function F(Y) that takes a dichotomous Y and gives us a continuous, real-valued Y′
- Then we can run

$$F(Y) = Y' = \mathbf{X}\beta + \varepsilon$$

- 如何解决这个问题？
- 我们需要把二元取值的Y换成连续变量Y’
- 所以我们需要一个链接函数F（Y），将Y映射到一个连续的实值变量Y’
- 然后就可以继续使用线性方程了



Redefining the Dependent Var.



Redefining the Dependent Var.

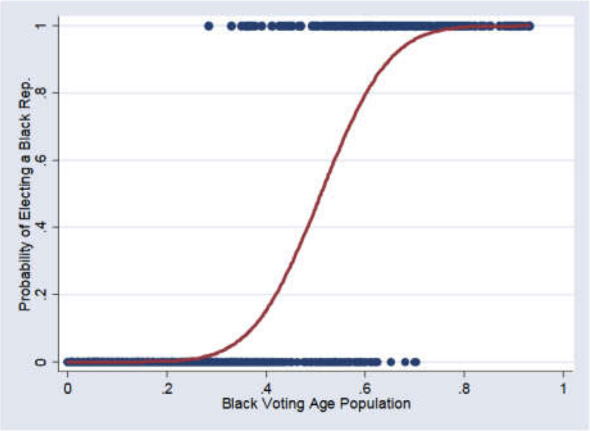
重新定义因变量

- 什么函数F（Y）实现了[0,1]区间到实数域的映射？
- 我们至少知道一个反向的映射函数。
  - 也就是说，给定任何实值，它都会产生一个介于0和1之间的数字（概率）。
- 这就是累积正态分布函数Φ
  - 也就是说，给定任何Z分数，Φ（Z）∈[0,1]
- 此链接函数称为Probit链接
  - Probit这个术语由20世纪30年代的生物学家提出。
  - 它是“概率单元（probability unit）”的缩写

- What function F(Y) goes from the [0,1] interval to the real line?
- Well, we know at least one function that goes the other way around. F^{-1} $(-\infty,\infty) \to [0,1]$
  - That is, given any real value it produces a number (probability) between 0 and 1.
- This is the cumulative normal distribution Φ
  - That is, given any Z-score, Φ(Z) ∈ [0,1]
- This link function is known as the Probit link
  - This term was coined in the 1930's by biologists studying the dosage-cure rate link
  - It is short for "probability unit"

$\Phi =$

Probit Estimation Probit估计



- This fits the data much better than the linear estimation
- Always lies between 0 and 1

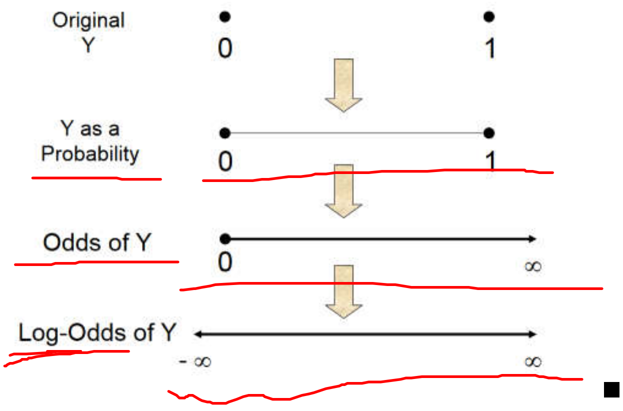
- 这比线性估计更符合数据的分布
- 始终介于0和1之间

Redefining the Dependent Var.

重新定义因变量

- Let’s return to the problem of transforming Y from {0,1} to the real line □ 让我们回到将Y从{0, 1}转换为实值的问题
- We’ll look at an alternative approach based on the odds 几率 □ 我们将根据几率寻找替代方法
- If some event occurs with probability p, then the odds of it happening are O(p) = p/(1-p) □ 如果某个事件发生的概率为p，那么它发生的几率是0（p）=p/（1-p）
  - □ p = 0 → O(p) = 0
  - □ p = ¼ → O(p) = 1/3 (“Odds are 1-to-3 against”)
  - □ p = ½ → O(p) = 1 (“Even odds”)
  - □ p = ¾ → O(p) = 3 (“Odds are 3-to-1 in favor”)
  - □ p = 1 → O(p) = ∞

Redefining the Dependent Var.



Logit Function logit函数

- This is called the logit function
  - □ logit(Y) = log[O(Y)] = log[y/(1-y)]
- Why would we want to do this?
  - □ At first, this was computationally easier than working with normal distributions
  - □ Now, it still has some nice properties that we'll investigate next time with multinomial dep. vars.
- The density function associated with it is very close to a standard normal distribution

- 我们为什么要这样做？
  - □ 首先，比使用正态分布更容易计算
  - □ 其次，它有一些很好的性质，我们下次将用多项式因变量来研究。
- 与之相关的概率密度函数非常接近标准正态分布

Logit Function

- This translates back to the original Y as:

$$\log\left(\frac{Y}{1-Y}\right) = \mathbf{X}\beta$$

$$\frac{Y}{1-Y} = e^{\mathbf{X}\beta}$$

$$Y = (1-Y)e^{\mathbf{X}\beta}$$

$$Y = e^{\mathbf{X}\beta} - e^{\mathbf{X}\beta}Y$$

$$Y + e^{\mathbf{X}\beta}Y = e^{\mathbf{X}\beta}$$

$$\left(1+e^{\mathbf{X}\beta}\right)Y = e^{\mathbf{X}\beta}$$

$$Y = \frac{e^{\mathbf{X}\beta}}{1+e^{\mathbf{X}\beta}}$$

Latent Variables 潜变量

- For the rest of the lecture we'll talk in terms of probits, but everything holds for logits too
- One way to state what's going on is to assume that there is a latent variable Y* such that

$$Y^* = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim N\left(0, \sigma^2\right) \longleftarrow \text{Normal = Probit}$$

- In a linear regression we would observe Y* directly
- In probits, we observe only

$$y_i = \begin{cases} 0 \text{ if } y_i^* \le 0 \\ 1 \text{ if } y_i^* > 0 \end{cases}$$

These could be any constant. Later we'll set them to ½.

这些可以是任意常数，稍后我们将设置为1/2

- 接下来将讨论probits（这些内容同样也适用于logits）
- 假设存在一个潜在变量Y*，
- 这样在线性回归中，我们将直接观察潜变量Y*。

通过*link function* 将latent variable $z = \theta^T x$*(即Y*=Xβ)，映射到 logit(Y)；*

反过来，即Y = sigmoid(z)

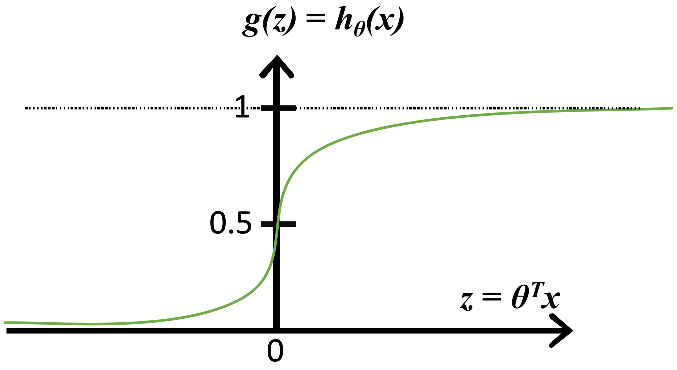
逻辑回归模型（Logistic Regression Model）

定义：$\boldsymbol{z = \theta^T x}$

z 的取值范围为（-∞,+∞）

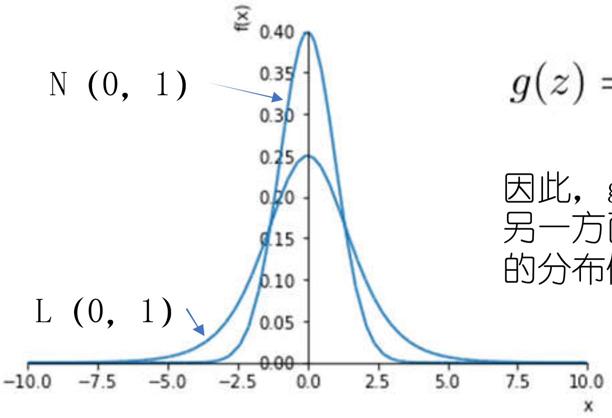
$$g(z) = \frac{1}{1+e^{-z}}$$

$g(z)$ 的取值范围为（0,1）



即：$h_\theta(x) = g(\ \theta^T x\ ) = \frac{1}{1+e^{-\theta^T x}}$

$h_\theta(x)$ 的输出值为什么具有概率的含义？



logistic distribution 与 normal distribution 的PDF

$$g(z) = \frac{1}{1+e^{-z}}$$

为logistic distribution的CDF

因此，g(z)具有概率的意义，g(z)反映了P(Z <= z)。另一方面，也暗含了Logistic Regression对隐变量z的分布假设，z ~ Logistic(0,1)

logistic distribution 与 normal distribution 的PDF形态相似，具有相同的标准差。但前者的Kurtosis（峰度）大于后者，即具有更长的“尾巴”，离群点也更多。

Summary 小结

In statistics, a **generalized linear model** (**GLM**) is a flexible generalization of ordinary linear regression. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a *link function* and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

在统计学中，广义线性模型（GLM or GzLM）是普通线性回归的灵活推广。GLM通过允许线性模型通过链接函数与因变量相关。

There are three components in generalized linear models.
广义线性模型的3个模块：

1. **Linear predictor** 线性预测模型
2. **Link function** 链接函数
3. **Probability distribution** 概率分布

Link function Linear predictor

$$\ln \lambda_i = b_0 + b_1 x_i$$

$$y_i \sim \text{Poisson}(\lambda_i)$$

Probability distribution

Linear regression revisited 重新审视线性回归

Linear regression is used to predict the value of continuous variable *y* by the linear combination of explanatory variables *X*.

线性回归通过解释变量X的线性组合预测连续变量y的值。

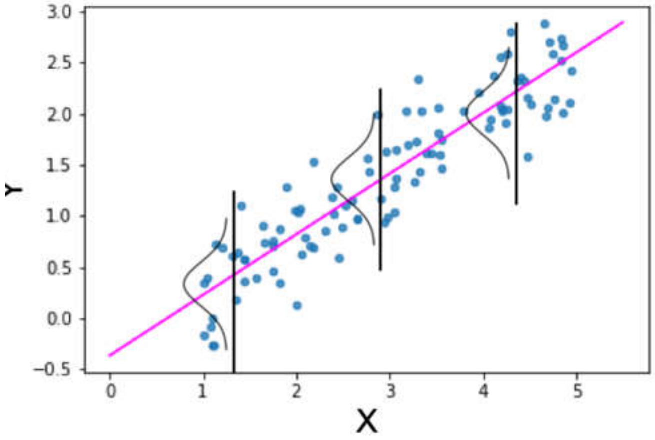
In the univariate case, linear regression can be expressed as follows; 在单变量情况下，线性回归可以表示为：

$$\mu_i = b_0 + b_1 x_i$$

$$y_i \sim \mathcal{N}(\mu_i, \varepsilon)$$

Link identity

Linear regression



Here, *i* indicates the index of each sample. Notice this model assumes normal distribution for the noise term. The model can be illustrated as follows;

i表示每个样本的索引。
注意这个模型假设噪声项为正态分布。

Poisson regression 泊松回归

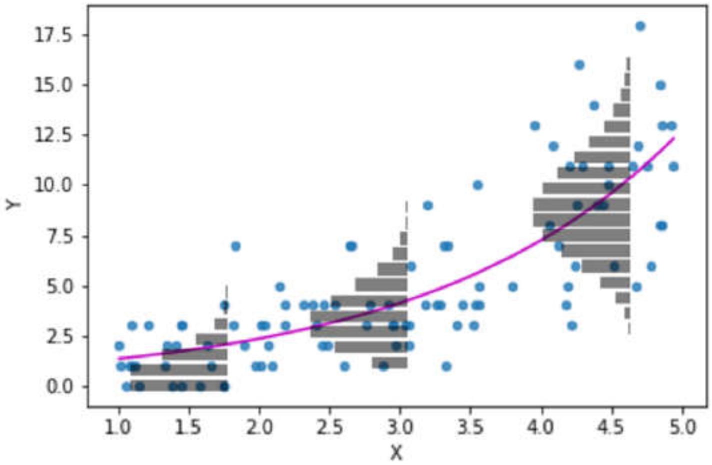
链接函数 线性预测模型

Link function Linear predictor

$$\ln \lambda_i = b_0 + b_1 x_i$$

$$y_i \sim \text{Poisson}(\lambda_\text{i})$$

Probability distribution



If you use **logit function** as the link function and **binomial / Bernoulli distribution** as the probability distribution, the model is called **logistic regression**.

如果使用logit函数作为链接函数，并使用二项式分布/伯努利分布作为概率分布，那么该模型为逻辑回归。

$$z_i = b_0 + b_1 x_i$$

$$q_i = \frac{1}{1 + \exp(-z_i)}$$

$$y_i \sim \text{Bern}(q_i)$$

logistic regression 逻辑回归

This is the list of probability distributions and their canonical link functions. 概率分布及其规范(cannonical)链接函数：

- Normal distribution: identity function
- Poisson distribution: log function
- Binomial distribution: logit function

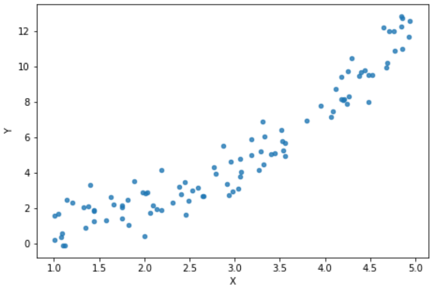
正态分布：恒等函数
泊松分布：对数函数
二项分布：logit函数

However, you don't necessarily use the canonical link function. Rather, the advantage of statistical modeling is that you can make any kind of model that fits well with your data.

但是，你不必使用规范关联函数。相反，统计建模的优点是，你可以创建任何类型的模型，以更好地符合您的数据。

Non-canonical link function example:

非规范链接函数举例：



This looks similar to the data I prepared for Poisson regression. However, if you see the data carefully, it seems the variance of *y* is constant with regard to *X*. Besides, *y* is continuous, not discrete.

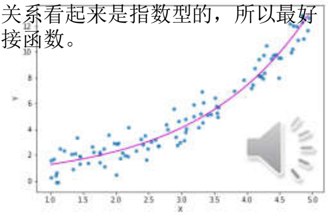
这与为泊松回归准备的数据类似。然而，如果仔细观察数据，y 相对于X的方差似乎是常数，而且y是连续的，而不是离散的。

Therefore, it's appropriate to use normal distribution here. As the relationship between *X* and *y* looks exponential, you had better choose the log link function.

因此，在这里使用正态分布是合适的。由于X和y之间的关系看起来是指数型的，所以最好选择log链接函数。

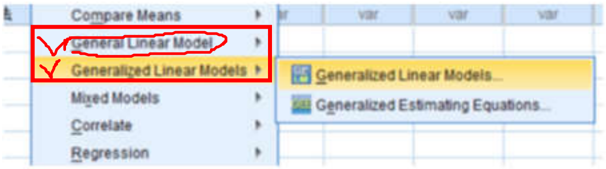
$$\ln \mu_i = b_0 + b_1 x_i$$

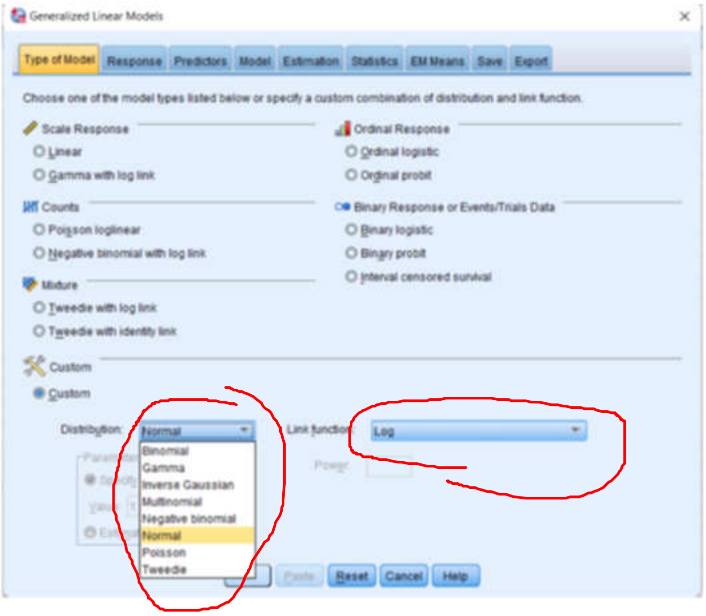
$$y_i \sim \mathcal{N}(\mu_i, \varepsilon)$$



GLM with non-canonical link function

GLM vs GLzM in SPSS


Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
var
var
var
Generalized Linear Models...
Generalized Estimating Equations...



Choose one of the model types listed below or specify a custom combination of distribution and link function.

● Custom

Distribution: Normal

Link function: Log

A generalized linear model specifying an identity link function and a normal family distribution is exactly equivalent to a (general) linear model.

GLM （一般线性模型）是GLzM（广义线性模型）的一个特例。

当广义线性模型GLzM使用恒等链接函数和一个正态分布时，就是一般线性模型GLM。

（选学内容）

ANCOVA | Two-way ANOVA | MANOVA?

都和回归分析密切相关

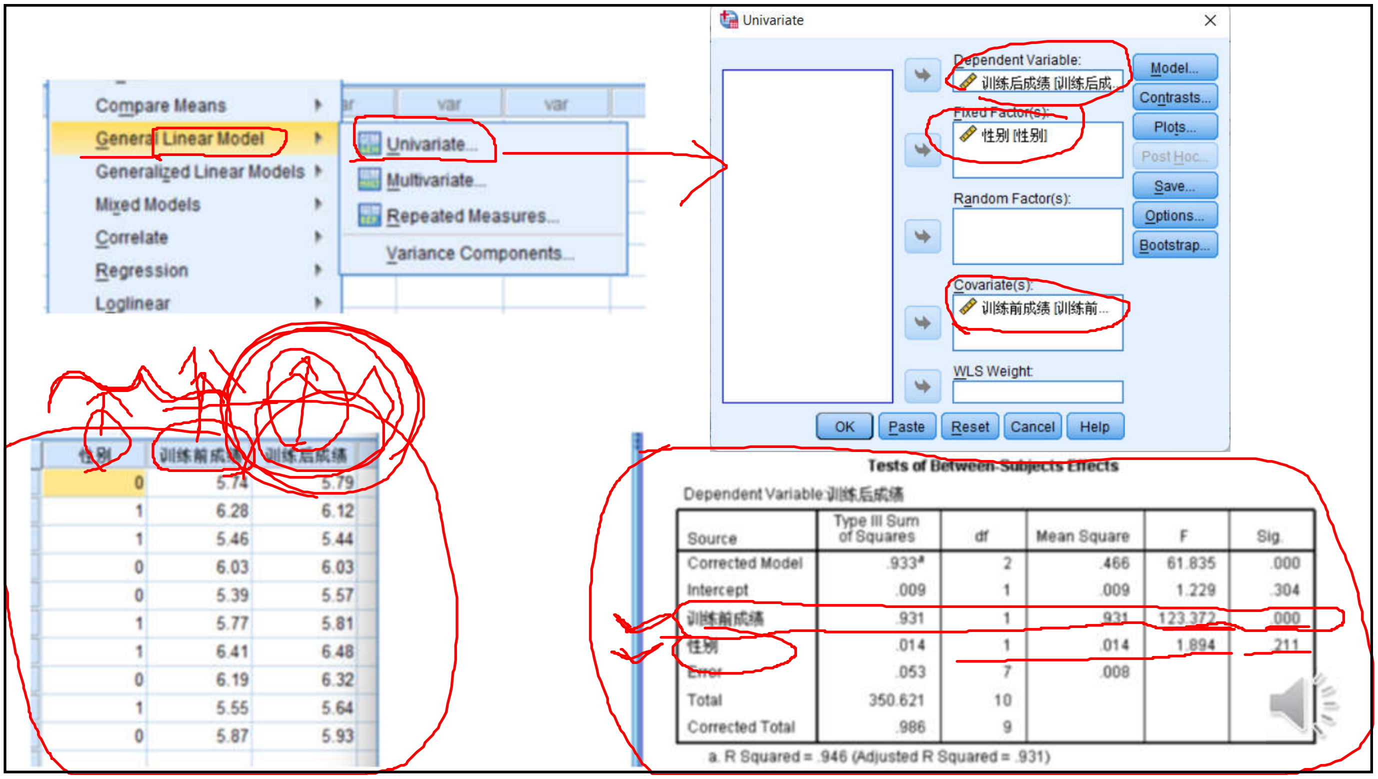
ANCOVA

When in a set of independent variable consist of both factor (categorical independent variable) and covariate (metric independent variable), the technique used is known as ANCOVA. The difference in dependent variables because of the covariate is taken off by an adjustment of the dependent variable's mean value within each treatment condition.

当一组自变量同时包含因子（类别自变量）和协变量（度量自变量）时，所用的技术称为协方差分析。核心是如何测量和分解因协变量引起的因变量差异。

适用条件：1个类别（categorical）因素和1个数值性（metric）因素

Basis for Comparison 对照表	ANOVA	ANCOVA (Analysis of Covariance) an extended form of ANOVA
Meaning	ANOVA is a process of examining the difference among the means of multiple groups of data for homogeneity. 方差分析是检查多组数据的平均值之间的差异	ANCOVA is a technique that remove the impact of one or more metric-scaled undesirable variable from dependent variable before undertaking research. 协方差分析（ANCOVA）是一种在进行研究之前从因变量中去除一个或多个度量、的影响的技术。
Uses	Both linear and non-linear model are used. 使用线性和非线性模型。	Only linear model is used. 仅使用线性模型
Includes	Categorical variable.类别变量	Categorical and interval variable. 类别变量和间隔变量。
Covariate 协变量	Ignored忽略	Considered考虑
BG variation	Attributes Between Group (BG) variation, to treatment. 组间/处理的差异。	Divides Between Group (BG) variation, into treatment and covariate. 将组间（BG）变异分为处理和协变量。
WG variation	Attributes Within Group (WG) variation, to individual differences.组内属性（WG），个体差异。	Divides Within Group (WG) variation, into individual differences and covariate.将组内（WG）变量分为个体差异和协变量。



Dependent Variable: 训练后成绩 [训练后成...

Fixed Factor(s): 性别 [性别]

Covariate(s): 训练前成绩 [训练前...

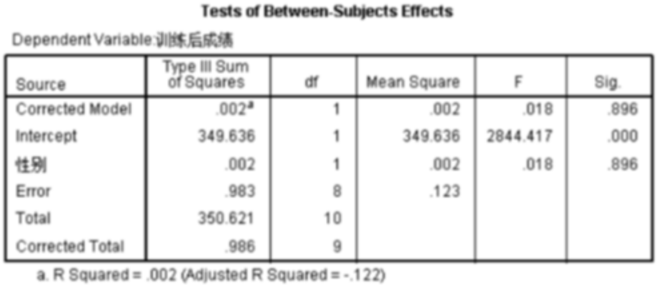
性别	训练前成绩	训练后成绩
0	5.74	5.79
1	6.28	6.12
1	5.46	5.44
0	6.03	6.03
0	5.39	5.57
1	5.77	5.81
1	6.41	6.48
0	6.19	6.32
1	5.55	5.64
0	5.87	5.93

Tests of Between-Subjects Effects

Dependent Variable:训练后成绩

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.933[a]	2	.466	61.835	.000
Intercept	.009	1	.009	1.229	.304
训练前成绩	.931	1	.931	123.372	.000
性别	.014	1	.014	1.894	.211
Error	.053	7	.008		
Total	350.621	10			
Corrected Total	.986	9			

a. R Squared = .946 (Adjusted R Squared = .931)



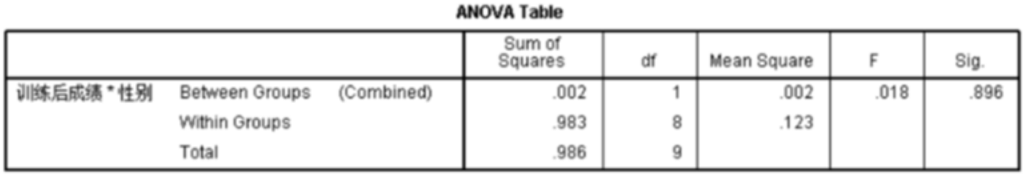
Tests of Between-Subjects Effects

Dependent Variable:训练后成绩

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.002[a]	1	.002	.018	.896
Intercept	349.636	1	349.636	2844.417	.000
性别	.002	1	.002	.018	.896
Error	.983	8	.123		
Total	350.621	10			
Corrected Total	.986	9			

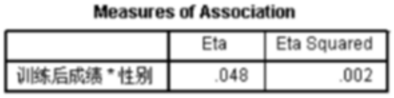
a. R Squared = .002 (Adjusted R Squared = -.122)

without covariate



ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
训练后成绩 * 性别	Between Groups	(Combined)	.002	1	.002	.018	.896
	Within Groups		.983	8	.123		
	Total		.986	9			

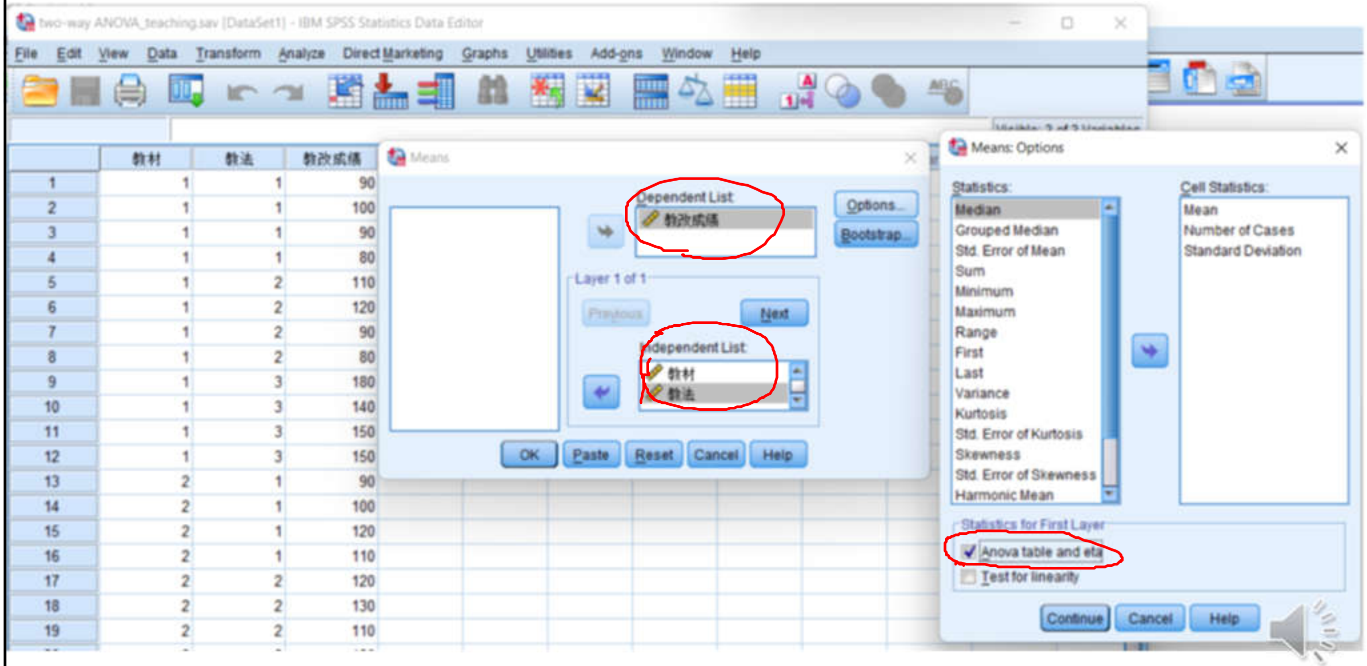


Measures of Association

	Eta	Eta Squared
训练后成绩 * 性别	.048	.002

Two-way ANOVA?：

two-way ANOVA_teaching.sav



	教材	教法	教改成绩
1	1	1	90
2	1	1	100
3	1	1	90
4	1	1	80
5	1	2	110
6	1	2	120
7	1	2	90
8	1	2	80
9	1	3	180
10	1	3	140
11	1	3	150
12	1	3	150
13	2	1	90
14	2	1	100
15	2	1	120
16	2	1	110
17	2	2	120
18	2	2	130
19	2	2	110

Dependent List: 教改成绩

Layer 1 of 1

Independent List: 教材, 教法

Cell Statistics: Mean, Number of Cases, Standard Deviation

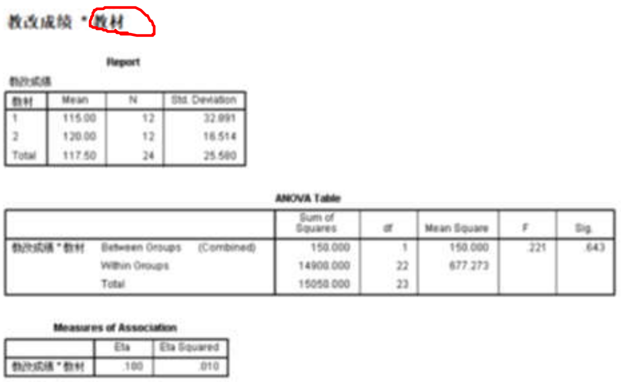
Statistics for First Layer

☑ Anova table and eta

□ Test for linearity

Not two-way ANOVA, but 2 one-way ANOVA tests. Not what we want

不是双因素方差分析，而是2个单因素方差分析分析。不是我们想要的



教改成绩 * 教材

Report

教改成绩

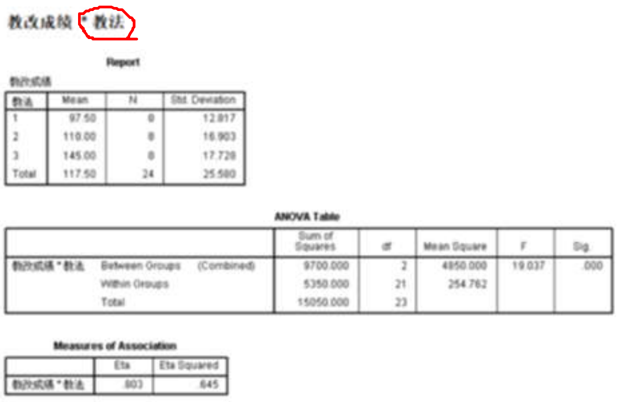
教材	Mean	N	Std. Deviation
1	115.00	12	32.891
2	120.00	12	16.514
Total	117.50	24	25.580

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
教改成绩 * 教材	Between Groups	(Combined)	150.000	1	150.000	.221	.643
	Within Groups		14900.000	22	677.273		
	Total		15050.000	23			

Measures of Association

	Eta	Eta Squared
教改成绩 * 教材	.100	.010



教改成绩 * 教法

Report

教改成绩

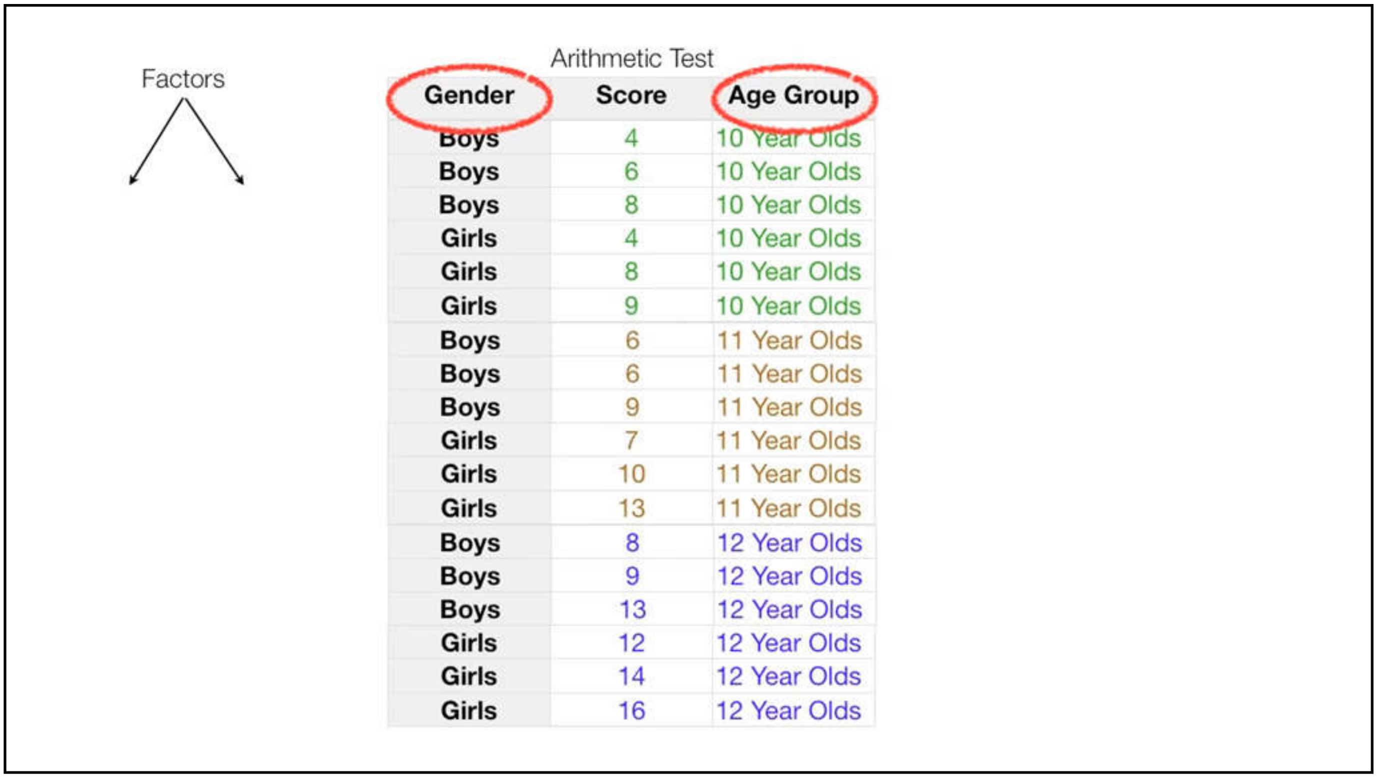
教法	Mean	N	Std. Deviation
1	97.50	8	12.817
2	110.00	8	16.903
3	145.00	8	17.728
Total	117.50	24	25.580

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
教改成绩 * 教法	Between Groups	(Combined)	9700.000	2	4850.000	19.037	.000
	Within Groups		5350.000	21	254.762		
	Total		15050.000	23			

Measures of Association

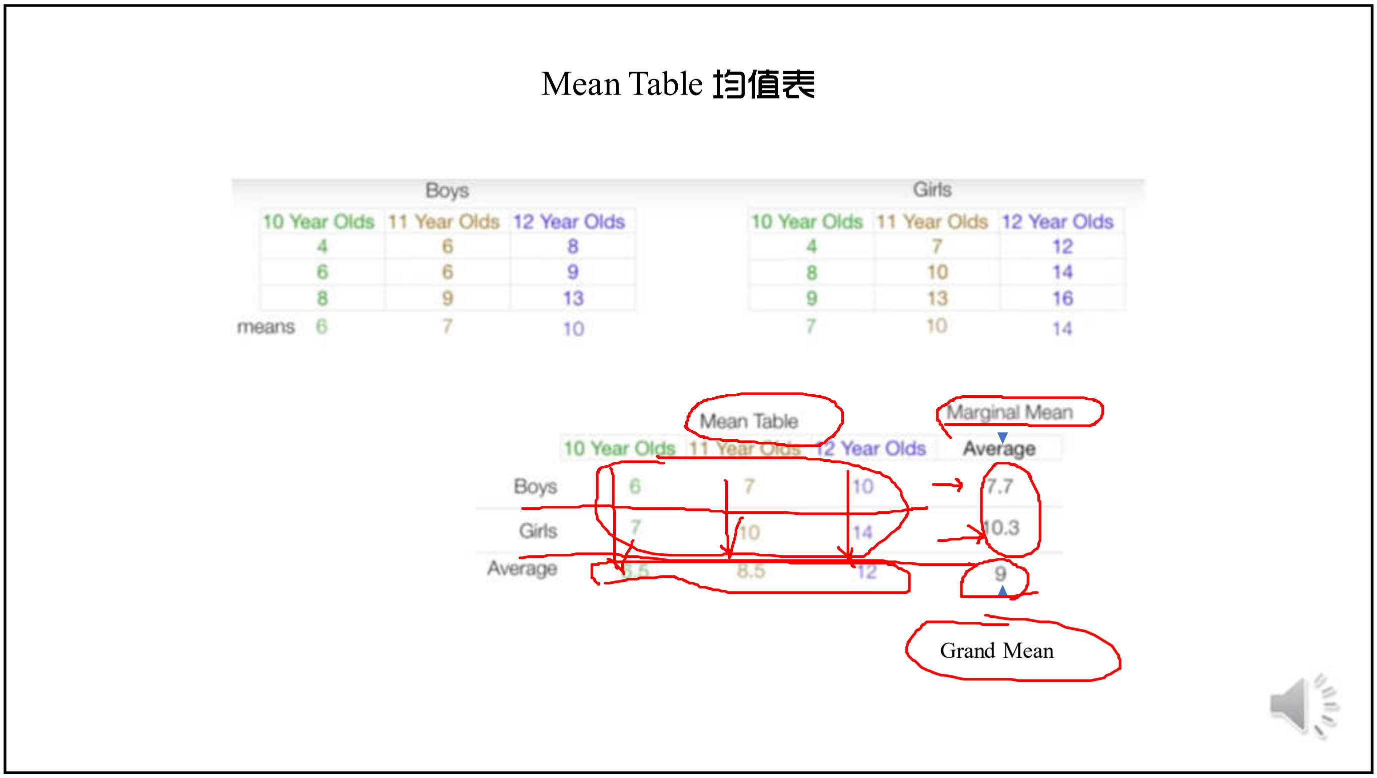
	Eta	Eta Squared
教改成绩 * 教法	.803	.645



Factors

Arithmetic Test

Gender	Score	Age Group
Boys	4	10 Year Olds
Boys	6	10 Year Olds
Boys	8	10 Year Olds
Girls	4	10 Year Olds
Girls	8	10 Year Olds
Girls	9	10 Year Olds
Boys	6	11 Year Olds
Boys	6	11 Year Olds
Boys	9	11 Year Olds
Girls	7	11 Year Olds
Girls	10	11 Year Olds
Girls	13	11 Year Olds
Boys	8	12 Year Olds
Boys	9	12 Year Olds
Boys	13	12 Year Olds
Girls	12	12 Year Olds
Girls	14	12 Year Olds
Girls	16	12 Year Olds



Mean Table 均值表

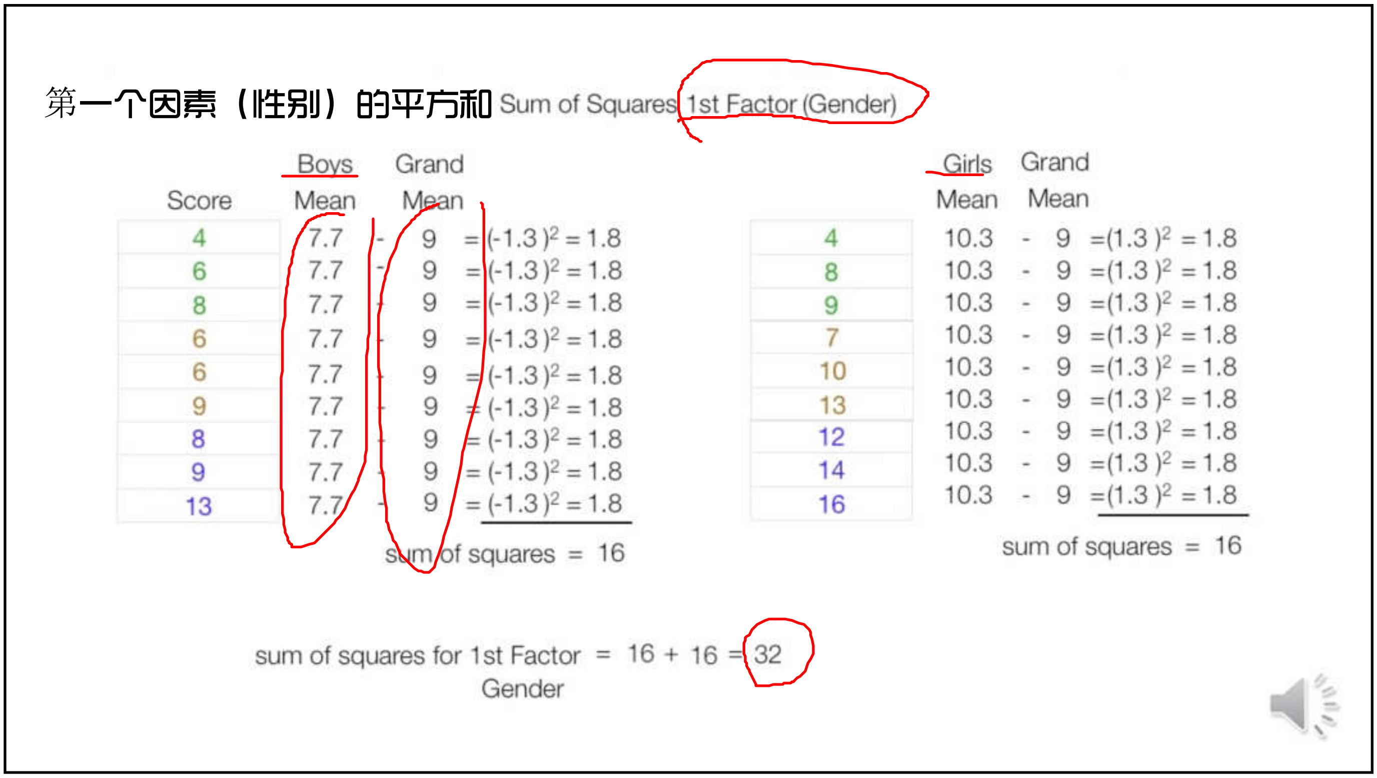
	Boys			Girls		
	10 Year Olds	11 Year Olds	12 Year Olds	10 Year Olds	11 Year Olds	12 Year Olds
	4	6	8	4	7	12
	6	6	9	8	10	14
	8	9	13	9	13	16
means	6	7	10	7	10	14

Mean Table

Marginal Mean

	10 Year Olds	11 Year Olds	12 Year Olds	Average
Boys	6	7	10	7.7
Girls	7	10	14	10.3
Average	6.5	8.5	12	9

Grand Mean



第一个因素（性别）的平方和 Sum of Squares 1st Factor (Gender)

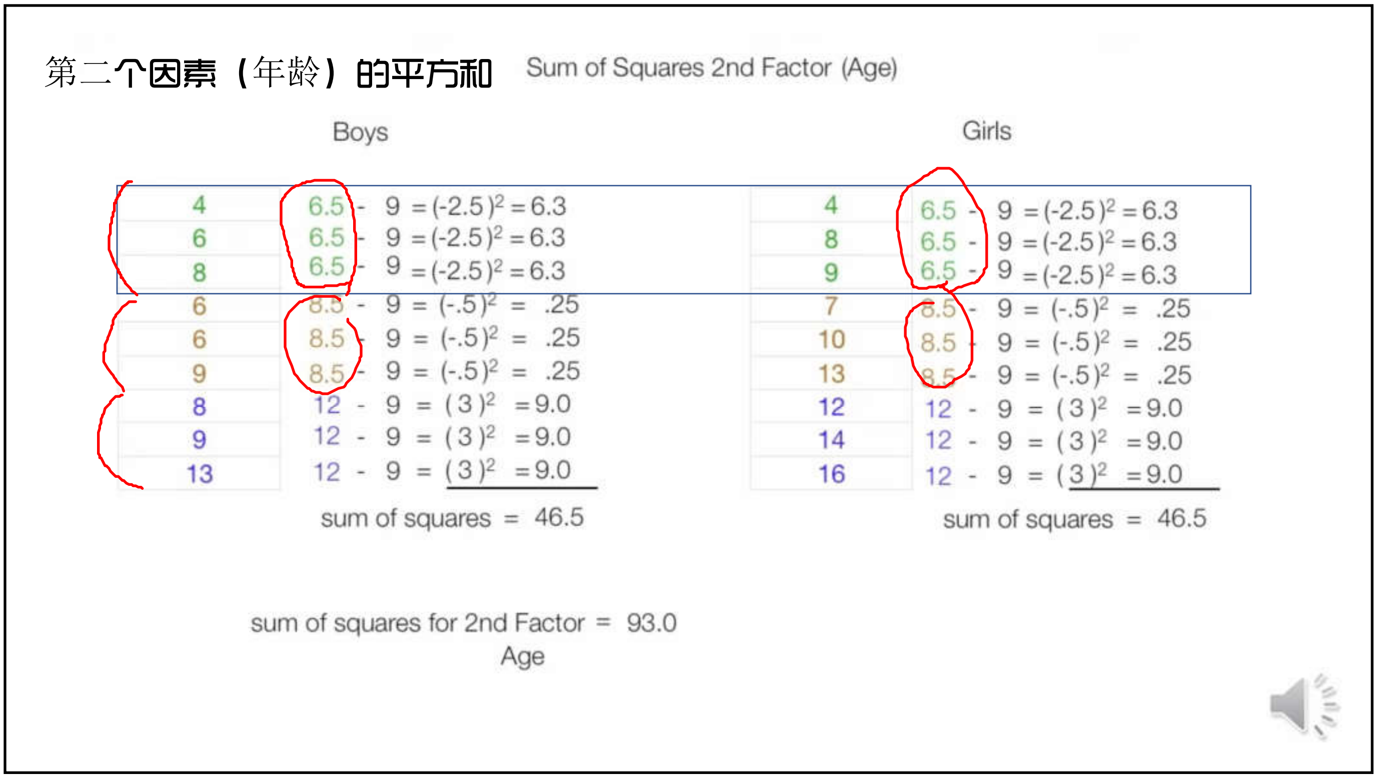
Score	Boys Mean		Grand Mean	
4	7.7	-	9	$= (-1.3)^2 = 1.8$
6	7.7	-	9	$= (-1.3)^2 = 1.8$
8	7.7	-	9	$= (-1.3)^2 = 1.8$
6	7.7	-	9	$= (-1.3)^2 = 1.8$
6	7.7	-	9	$= (-1.3)^2 = 1.8$
9	7.7	-	9	$= (-1.3)^2 = 1.8$
8	7.7	-	9	$= (-1.3)^2 = 1.8$
9	7.7	-	9	$= (-1.3)^2 = 1.8$
13	7.7	-	9	$= (-1.3)^2 = 1.8$

sum of squares = 16

	Girls Mean		Grand Mean	
4	10.3	-	9	$= (1.3)^2 = 1.8$
8	10.3	-	9	$= (1.3)^2 = 1.8$
9	10.3	-	9	$= (1.3)^2 = 1.8$
7	10.3	-	9	$= (1.3)^2 = 1.8$
10	10.3	-	9	$= (1.3)^2 = 1.8$
13	10.3	-	9	$= (1.3)^2 = 1.8$
12	10.3	-	9	$= (1.3)^2 = 1.8$
14	10.3	-	9	$= (1.3)^2 = 1.8$
16	10.3	-	9	$= (1.3)^2 = 1.8$

sum of squares = 16

sum of squares for 1st Factor = 16 + 16 = 32
Gender



第二个因素（年龄）的平方和 Sum of Squares 2nd Factor (Age)

Boys

4	6.5	-	9	$= (-2.5)^2 = 6.3$
6	6.5	-	9	$= (-2.5)^2 = 6.3$
8	6.5	-	9	$= (-2.5)^2 = 6.3$
6	8.5	-	9	$= (-.5)^2 = .25$
6	8.5	-	9	$= (-.5)^2 = .25$
9	8.5	-	9	$= (-.5)^2 = .25$
8	12	-	9	$= (3)^2 = 9.0$
9	12	-	9	$= (3)^2 = 9.0$
13	12	-	9	$= (3)^2 = 9.0$

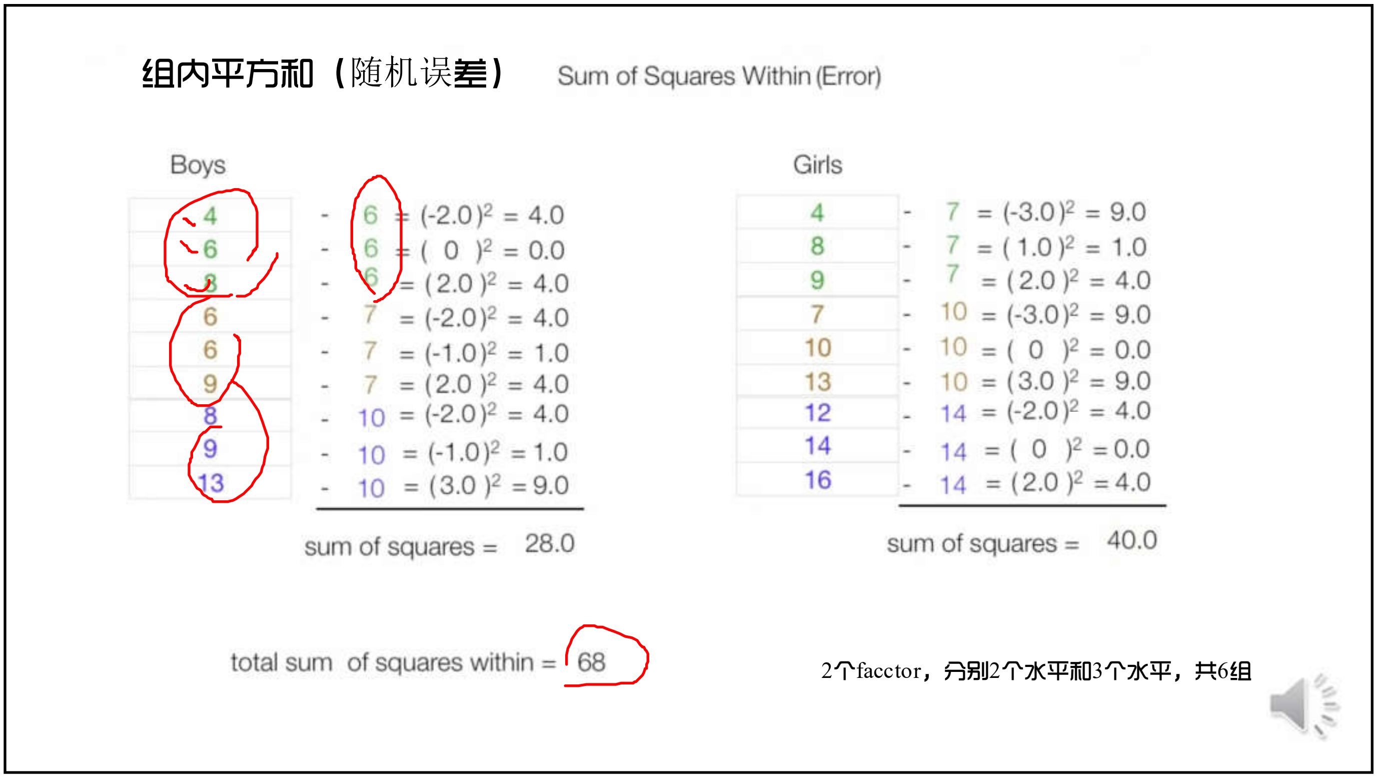
sum of squares = 46.5

Girls

4	6.5	-	9	$= (-2.5)^2 = 6.3$
8	6.5	-	9	$= (-2.5)^2 = 6.3$
9	6.5	-	9	$= (-2.5)^2 = 6.3$
7	8.5	-	9	$= (-.5)^2 = .25$
10	8.5	-	9	$= (-.5)^2 = .25$
13	8.5	-	9	$= (-.5)^2 = .25$
12	12	-	9	$= (3)^2 = 9.0$
14	12	-	9	$= (3)^2 = 9.0$
16	12	-	9	$= (3)^2 = 9.0$

sum of squares = 46.5

sum of squares for 2nd Factor = 93.0
Age



组内平方和（随机误差） Sum of Squares Within (Error)

Boys

4	- 6	= (-2.0)^2 = 4.0
6	- 6	= (0)^2 = 0.0
8	- 6	= (2.0)^2 = 4.0
6	- 7	= (-2.0)^2 = 4.0
6	- 7	= (-1.0)^2 = 1.0
9	- 7	= (2.0)^2 = 4.0
8	- 10	= (-2.0)^2 = 4.0
9	- 10	= (-1.0)^2 = 1.0
13	- 10	= (3.0)^2 = 9.0

sum of squares = 28.0

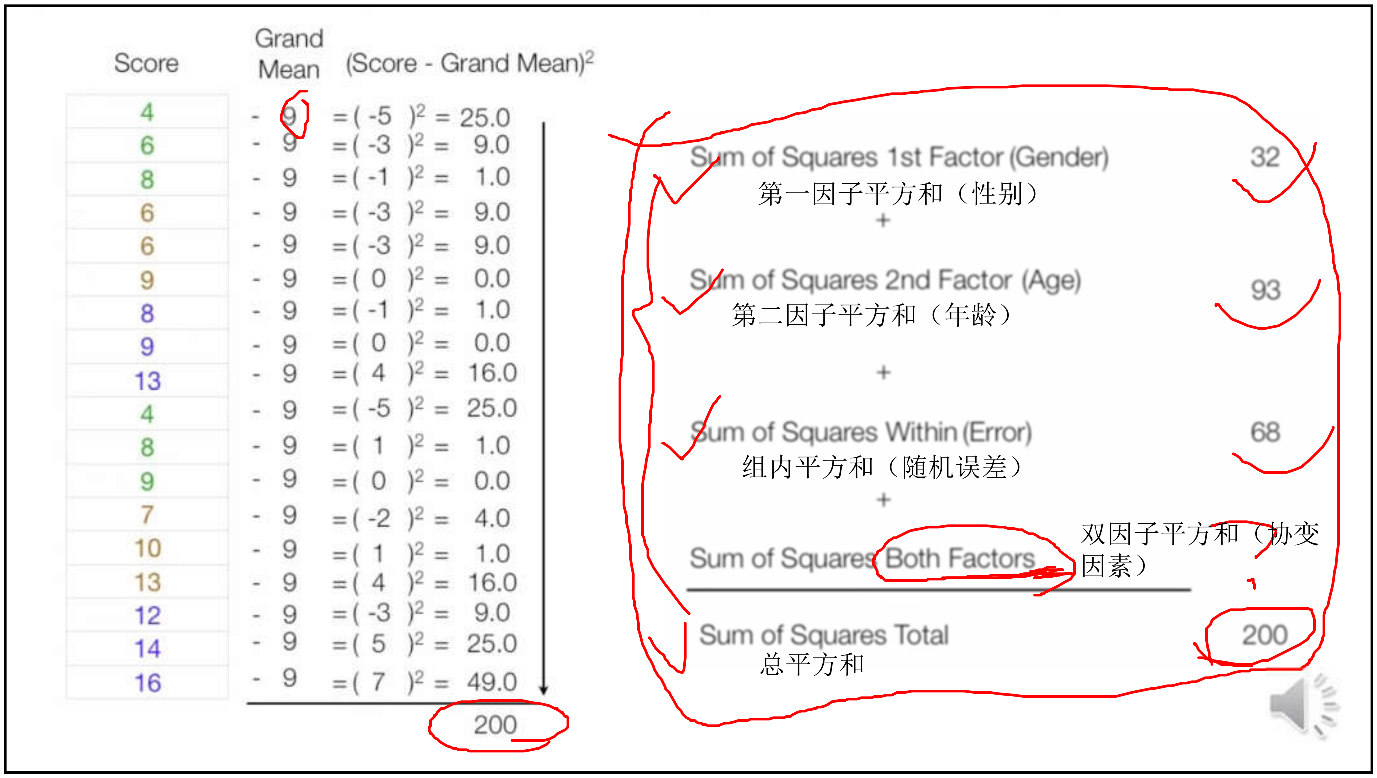
Girls

4	- 7	= (-3.0)^2 = 9.0
8	- 7	= (1.0)^2 = 1.0
9	- 7	= (2.0)^2 = 4.0
7	- 10	= (-3.0)^2 = 9.0
10	- 10	= (0)^2 = 0.0
13	- 10	= (3.0)^2 = 9.0
12	- 14	= (-2.0)^2 = 4.0
14	- 14	= (0)^2 = 0.0
16	- 14	= (2.0)^2 = 4.0

sum of squares = 40.0

total sum of squares within = 68

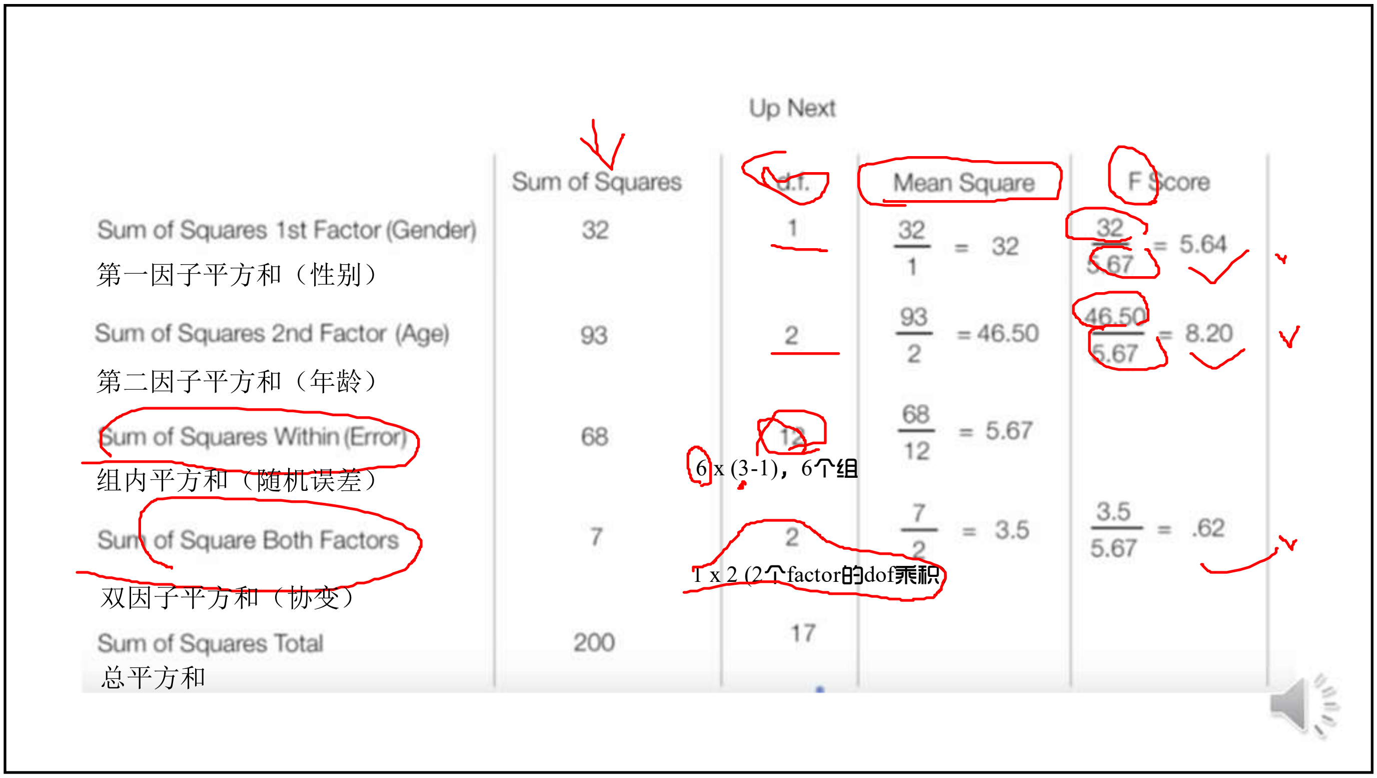
2个facctor，分别2个水平和3个水平，共6组



Score	Grand Mean	(Score - Grand Mean)^2
4	- 9	=(-5)^2 = 25.0
6	- 9	=(-3)^2 = 9.0
8	- 9	=(-1)^2 = 1.0
6	- 9	=(-3)^2 = 9.0
6	- 9	=(-3)^2 = 9.0
9	- 9	=(0)^2 = 0.0
8	- 9	=(-1)^2 = 1.0
9	- 9	=(0)^2 = 0.0
13	- 9	=(4)^2 = 16.0
4	- 9	=(-5)^2 = 25.0
8	- 9	=(1)^2 = 1.0
9	- 9	=(0)^2 = 0.0
7	- 9	=(-2)^2 = 4.0
10	- 9	=(1)^2 = 1.0
13	- 9	=(4)^2 = 16.0
12	- 9	=(-3)^2 = 9.0
14	- 9	=(5)^2 = 25.0
16	- 9	=(7)^2 = 49.0
		200

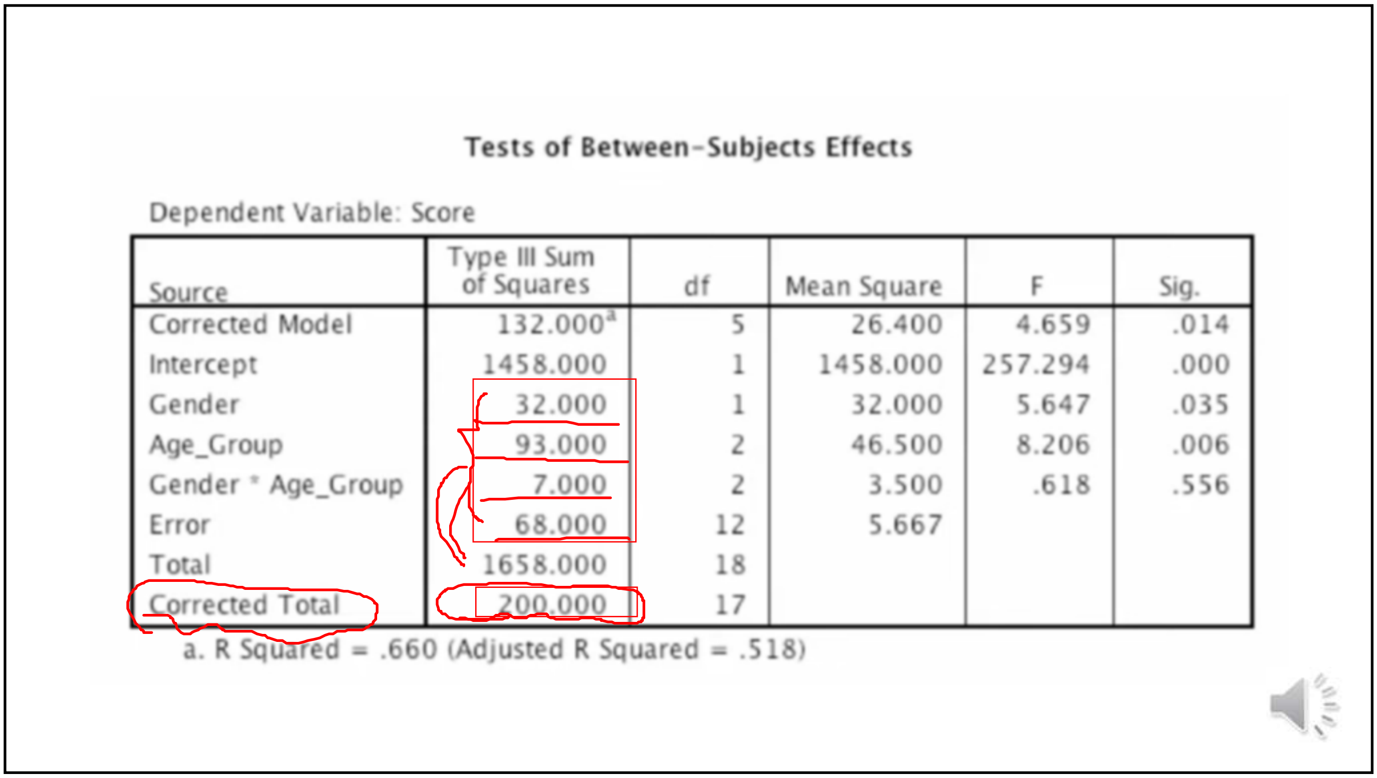
Sum of Squares 1st Factor (Gender) 32
第一因子平方和（性别）
+
Sum of Squares 2nd Factor (Age) 93
第二因子平方和（年龄）
+
Sum of Squares Within (Error) 68
组内平方和（随机误差）
+
Sum of Squares Both Factors
双因子平方和（协变因素）

Sum of Squares Total 200
总平方和



Up Next

	Sum of Squares	d.f.	Mean Square	F Score
Sum of Squares 1st Factor (Gender) 第一因子平方和（性别）	32	1	$\frac{32}{1} = 32$	$\frac{32}{5.67} = 5.64$
Sum of Squares 2nd Factor (Age) 第二因子平方和（年龄）	93	2	$\frac{93}{2} = 46.50$	$\frac{46.50}{5.67} = 8.20$
Sum of Squares Within (Error) 组内平方和（随机误差）	68	12 6 x (3-1)，6个组	$\frac{68}{12} = 5.67$	
Sum of Square Both Factors 双因子平方和（协变）	7	2 1 x 2 (2个factor的dof乘积)	$\frac{7}{2} = 3.5$	$\frac{3.5}{5.67} = .62$
Sum of Squares Total 总平方和	200	17		



Tests of Between-Subjects Effects

Dependent Variable: Score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	132.000[a]	5	26.400	4.659	.014
Intercept	1458.000	1	1458.000	257.294	.000
Gender	32.000	1	32.000	5.647	.035
Age_Group	93.000	2	46.500	8.206	.006
Gender * Age_Group	7.000	2	3.500	.618	.556
Error	68.000	12	5.667		
Total	1658.000	18			
Corrected Total	200.000	17			

a. R Squared = .660 (Adjusted R Squared = .518)

Establish Hypothesis

建立假设

REJECT

H_0: Gender will have no significant effect on students score.

H_1: Gender does have a significant effect on students score. $P<0.05$

Ho：性别对学生成绩没有显著影响。

H1：性别对学生成绩有显著影响。

REJECT

H_0: Age will have no significant effect on students score.

H_1: Age has a significant effect on students score. $P<0.05$

Ho：年龄对学生成绩没有显著影响。

H1：年龄对学生成绩有显著影响。

FAIL TO REJECT

H_0: Gender and Age interaction will have no significant effect on students score. $P>0.05$

Ho：性别和年龄的相互作用（协变）对学生成绩没有显著影响。

Multivariate analysis of variance (**MANOVA**)

多因变量方差分析

The purpose of a two-way ANOVA is to determine how two factors impact a response variable, and to determine whether or not there is an interaction between the two factors on the response variable. If you have three independent variables rather than two, you need a three-way ANOVA.双因素方差分析的目的是确定两个因素如何影响响应变量，并确定这两个因素对响应变量是否存在交互作用。如果你有三个自变量而不是两个，你需要一个三因素方差分析。

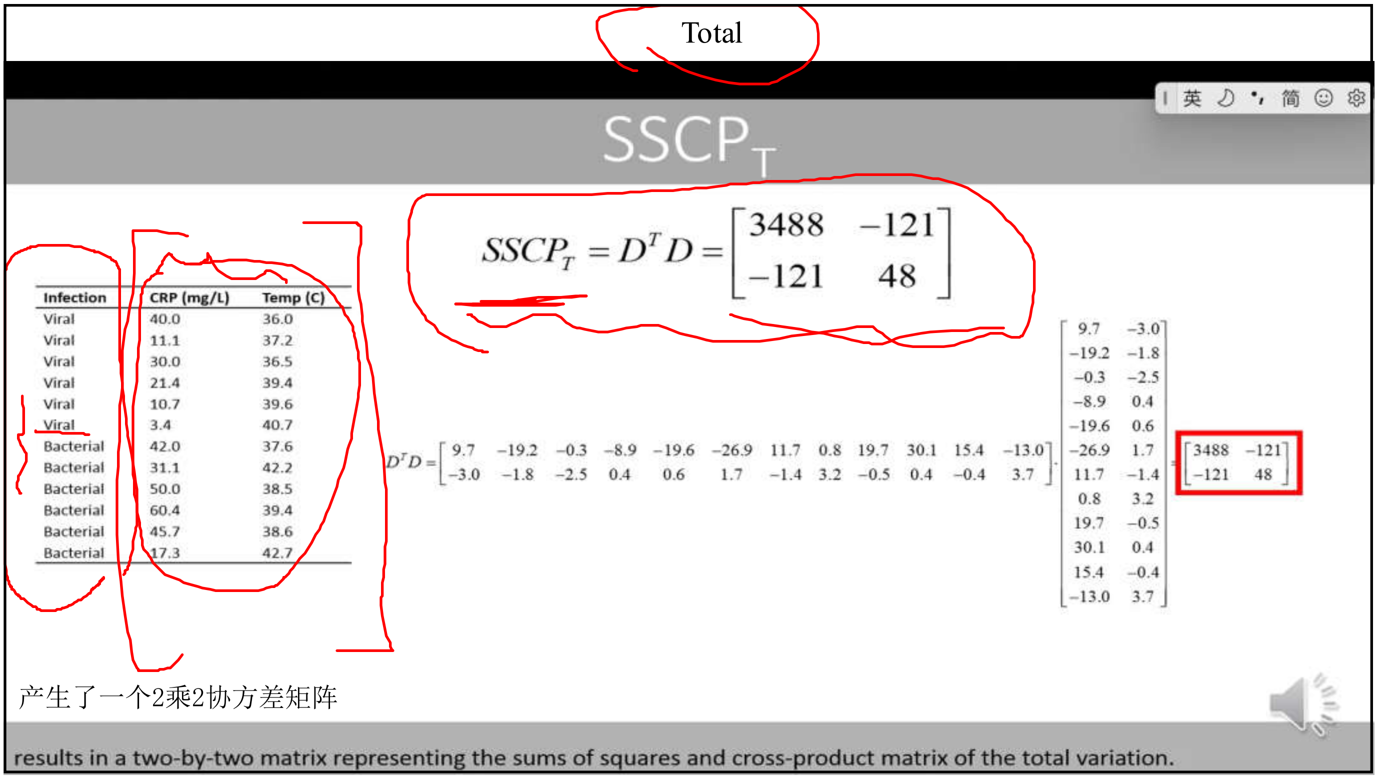
Two-way ANOVA concerns about two variables, while MANOVA concerns the differences in multiple variables simultaneously.

双因素方差分析关注两个变量，而多因变量方差分析同时关注多个变量的差异。

(2)

One/two/three-way ANOVA – 单/双/三因素方差分析 (factor)

MANOVA – 多因变量方差分析 (dependent variable)



Total

SSCP_T

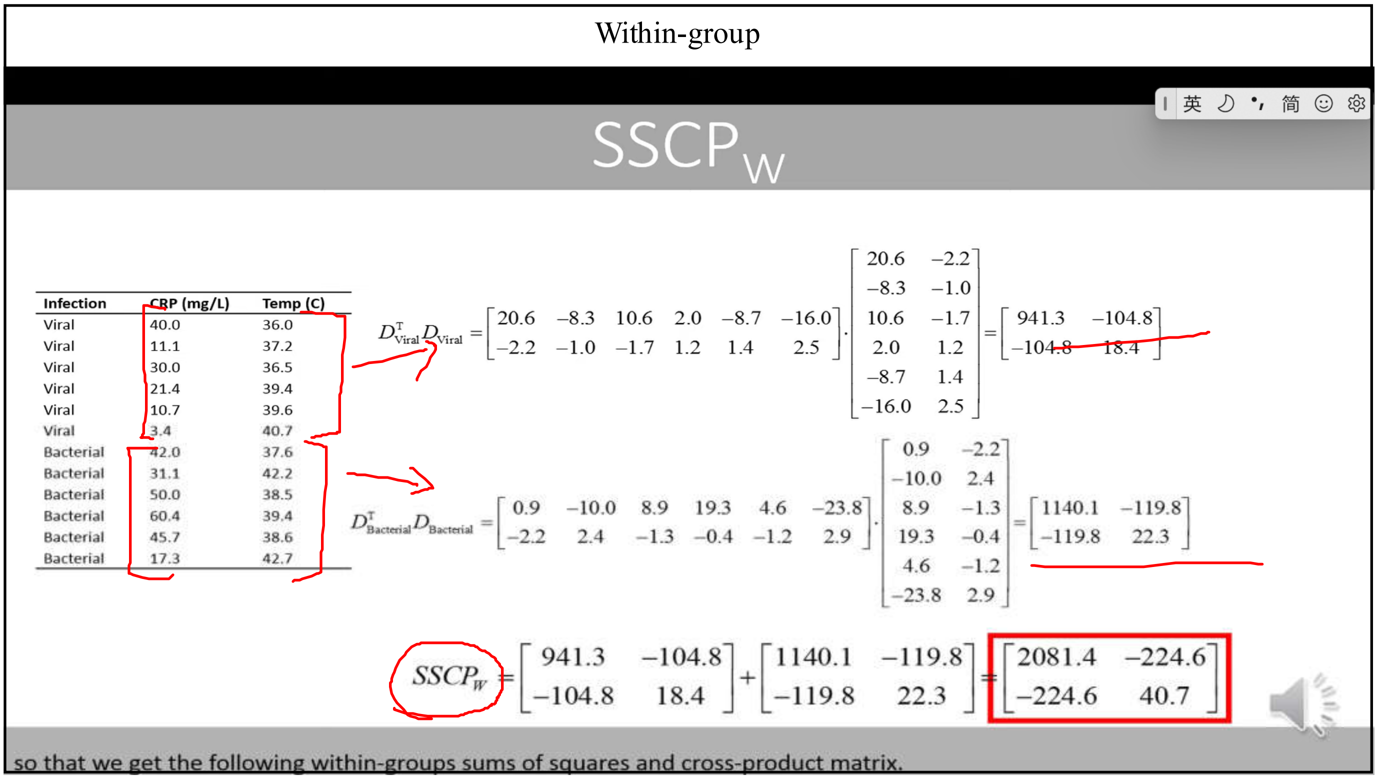
$$SSCP_T = D^T D = \begin{bmatrix} 3488 & -121 \\ -121 & 48 \end{bmatrix}$$

Infection	CRP (mg/L)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

$$D^T D = \begin{bmatrix} 9.7 & -19.2 & -0.3 & -8.9 & -19.6 & -26.9 & 11.7 & 0.8 & 19.7 & 30.1 & 15.4 & -13.0 \\ -3.0 & -1.8 & -2.5 & 0.4 & 0.6 & 1.7 & -1.4 & 3.2 & -0.5 & 0.4 & -0.4 & 3.7 \end{bmatrix} \cdot \begin{bmatrix} 9.7 & -3.0 \\ -19.2 & -1.8 \\ -0.3 & -2.5 \\ -8.9 & 0.4 \\ -19.6 & 0.6 \\ -26.9 & 1.7 \\ 11.7 & -1.4 \\ 0.8 & 3.2 \\ 19.7 & -0.5 \\ 30.1 & 0.4 \\ 15.4 & -0.4 \\ -13.0 & 3.7 \end{bmatrix} = \boxed{\begin{bmatrix} 3488 & -121 \\ -121 & 48 \end{bmatrix}}$$

产生了一个2乘2协方差矩阵

results in a two-by-two matrix representing the sums of squares and cross-product matrix of the total variation.



Within-group

SSCP_W

Infection	CRP (mg/L)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

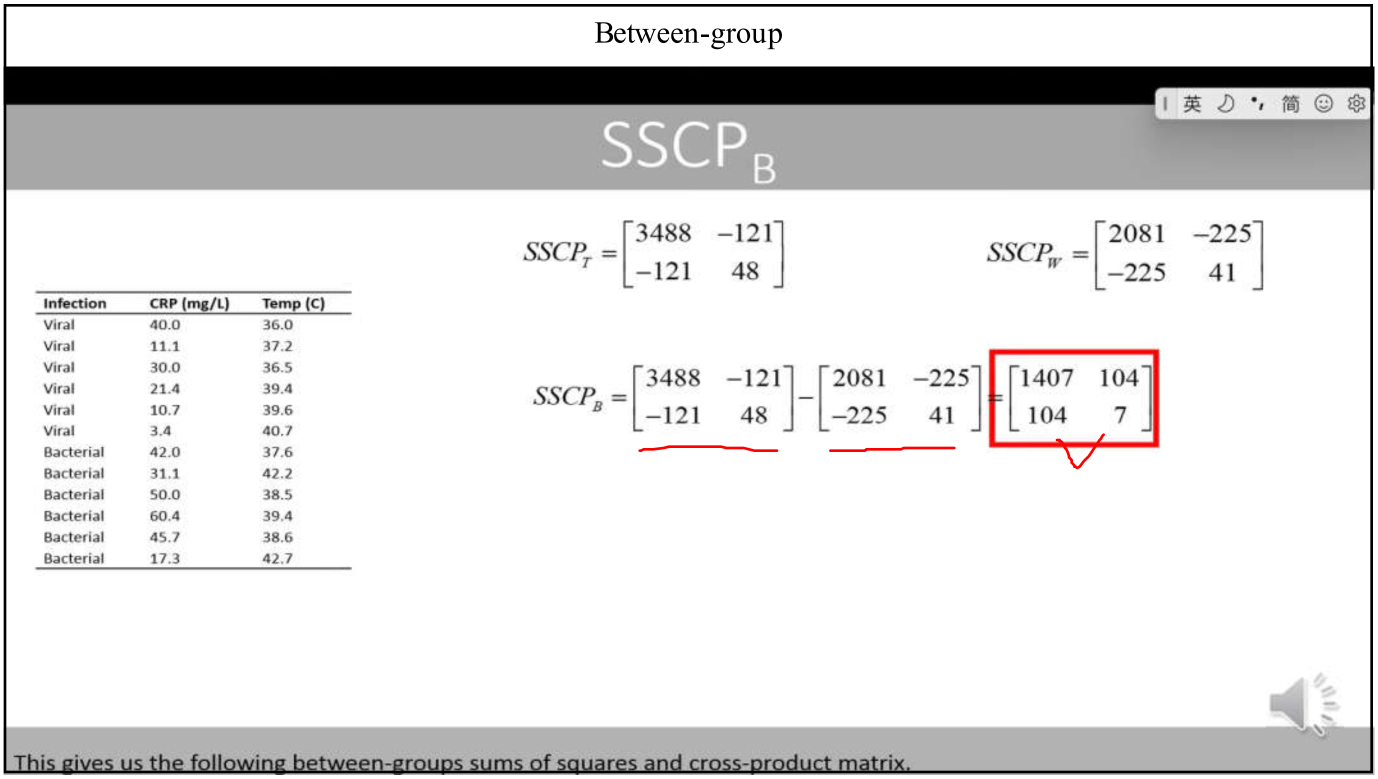
$$D^T_{\text{Viral}} D_{\text{Viral}} = \begin{bmatrix} 20.6 & -8.3 & 10.6 & 2.0 & -8.7 & -16.0 \\ -2.2 & -1.0 & -1.7 & 1.2 & 1.4 & 2.5 \end{bmatrix} \cdot \begin{bmatrix} 20.6 & -2.2 \\ -8.3 & -1.0 \\ 10.6 & -1.7 \\ 2.0 & 1.2 \\ -8.7 & 1.4 \\ -16.0 & 2.5 \end{bmatrix} = \begin{bmatrix} 941.3 & -104.8 \\ -104.8 & 18.4 \end{bmatrix}$$

$$D^T_{\text{Bacterial}} D_{\text{Bacterial}} = \begin{bmatrix} 0.9 & -10.0 & 8.9 & 19.3 & 4.6 & -23.8 \\ -2.2 & 2.4 & -1.3 & -0.4 & -1.2 & 2.9 \end{bmatrix} \cdot \begin{bmatrix} 0.9 & -2.2 \\ -10.0 & 2.4 \\ 8.9 & -1.3 \\ 19.3 & -0.4 \\ 4.6 & -1.2 \\ -23.8 & 2.9 \end{bmatrix} = \begin{bmatrix} 1140.1 & -119.8 \\ -119.8 & 22.3 \end{bmatrix}$$

$$SSCP_W = \begin{bmatrix} 941.3 & -104.8 \\ -104.8 & 18.4 \end{bmatrix} + \begin{bmatrix} 1140.1 & -119.8 \\ -119.8 & 22.3 \end{bmatrix} = \boxed{\begin{bmatrix} 2081.4 & -224.6 \\ -224.6 & 40.7 \end{bmatrix}}$$

so that we get the following within-groups sums of squares and cross-product matrix.

Between-group



SSCP$_B$

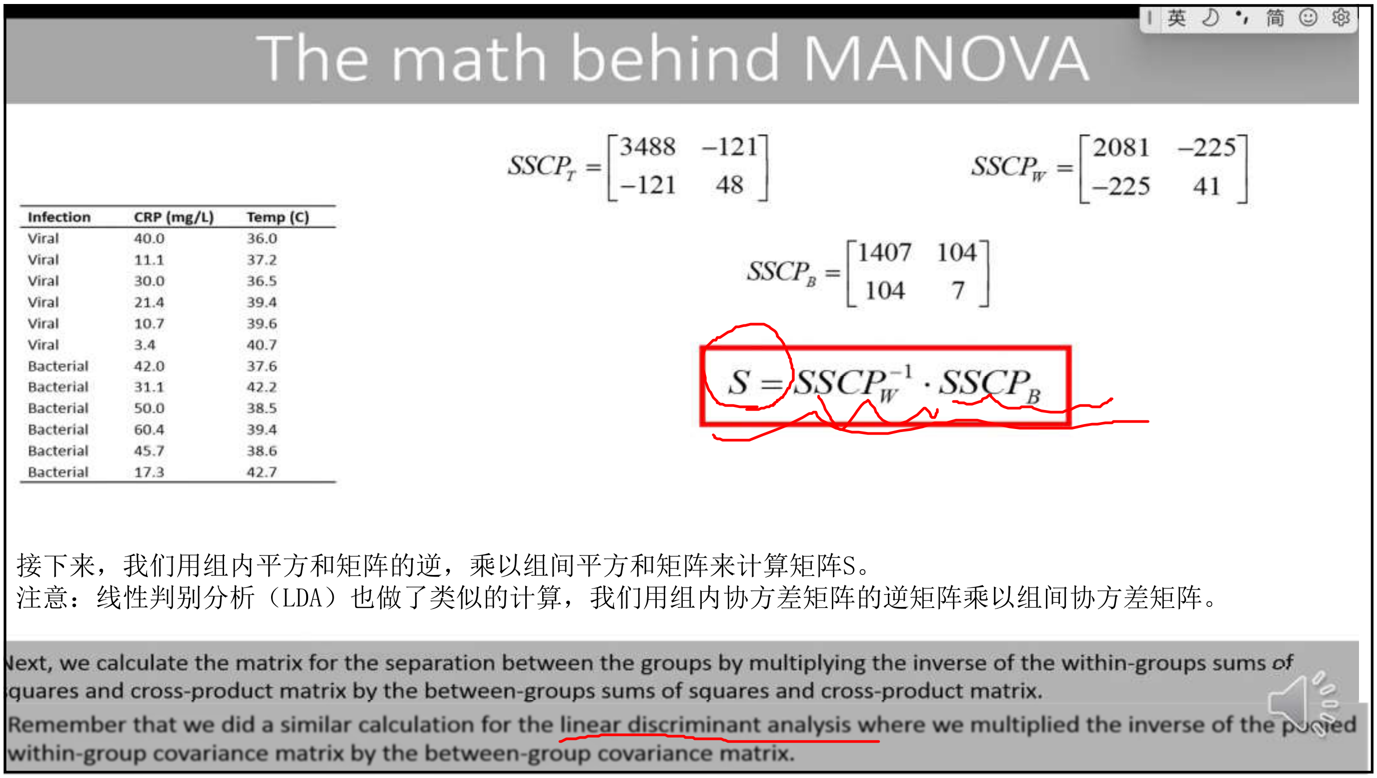
$$SSCP_T = \begin{bmatrix} 3488 & -121 \\ -121 & 48 \end{bmatrix} \qquad SSCP_W = \begin{bmatrix} 2081 & -225 \\ -225 & 41 \end{bmatrix}$$

Infection	CRP (mg/L)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

$$SSCP_B = \underline{\begin{bmatrix} 3488 & -121 \\ -121 & 48 \end{bmatrix}} - \underline{\begin{bmatrix} 2081 & -225 \\ -225 & 41 \end{bmatrix}} = \boxed{\begin{bmatrix} 1407 & 104 \\ 104 & 7 \end{bmatrix}}$$

✓

This gives us the following between-groups sums of squares and cross-product matrix.



The math behind MANOVA

$$SSCP_T = \begin{bmatrix} 3488 & -121 \\ -121 & 48 \end{bmatrix} \qquad SSCP_W = \begin{bmatrix} 2081 & -225 \\ -225 & 41 \end{bmatrix}$$

Infection	CRP (mg/L)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

$$SSCP_B = \begin{bmatrix} 1407 & 104 \\ 104 & 7 \end{bmatrix}$$

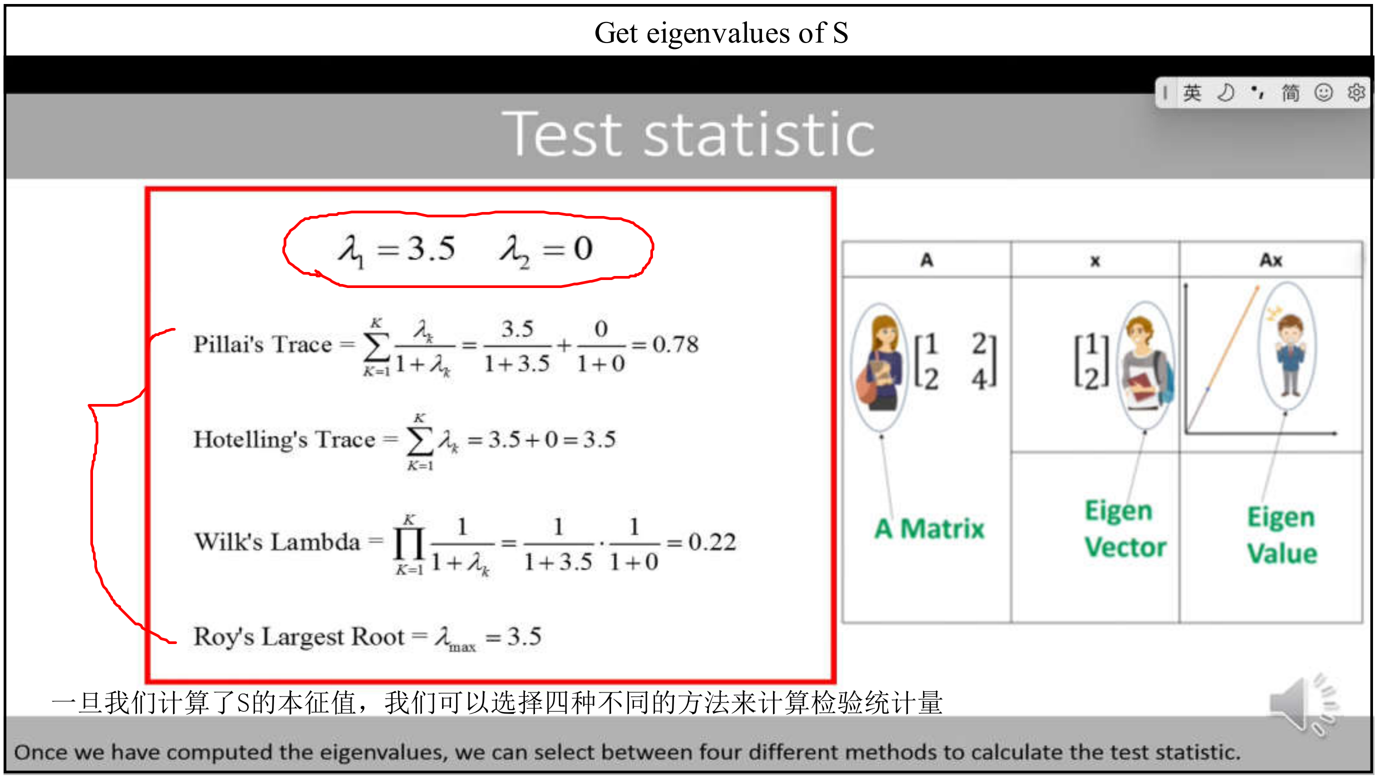
$$\boxed{S = SSCP_W^{-1} \cdot SSCP_B}$$

接下来，我们用组内平方和矩阵的逆，乘以组间平方和矩阵来计算矩阵S。
注意：线性判别分析（LDA）也做了类似的计算，我们用组内协方差矩阵的逆矩阵乘以组间协方差矩阵。

Next, we calculate the matrix for the separation between the groups by multiplying the inverse of the within-groups sums of squares and cross-product matrix by the between-groups sums of squares and cross-product matrix.

Remember that we did a similar calculation for the linear discriminant analysis where we multiplied the inverse of the pooled within-group covariance matrix by the between-group covariance matrix.

Get eigenvalues of S



Test statistic

$$\lambda_1 = 3.5 \quad \lambda_2 = 0$$

$$\text{Pillai's Trace} = \sum_{K=1}^{K} \frac{\lambda_k}{1+\lambda_k} = \frac{3.5}{1+3.5} + \frac{0}{1+0} = 0.78$$

$$\text{Hotelling's Trace} = \sum_{K=1}^{K} \lambda_k = 3.5 + 0 = 3.5$$

$$\text{Wilk's Lambda} = \prod_{K=1}^{K} \frac{1}{1+\lambda_k} = \frac{1}{1+3.5} \cdot \frac{1}{1+0} = 0.22$$

$$\text{Roy's Largest Root} = \lambda_{\max} = 3.5$$

A	x	Ax
$\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 2 \end{bmatrix}$	

一旦我们计算了S的本征值，我们可以选择四种不同的方法来计算检验统计量

Once we have computed the eigenvalues, we can select between four different methods to calculate the test statistic.

Hotelling’s T-square

Infection	CRP (mg/L)	Temp (C)
Viral	40.0	36.0
Viral	11.1	37.2
Viral	30.0	36.5
Viral	21.4	39.4
Viral	10.7	39.6
Viral	3.4	40.7
Bacterial	42.0	37.6
Bacterial	31.1	42.2
Bacterial	50.0	38.5
Bacterial	60.4	39.4
Bacterial	45.7	38.6
Bacterial	17.3	42.7

$$\text{cov}_{\text{virus}} = \begin{bmatrix} 188 & -21 \\ -21 & 4 \end{bmatrix}$$

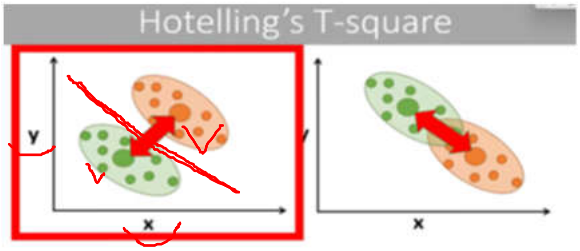
$$\text{cov}_{\text{Bacteria}} = \begin{bmatrix} 228 & -24 \\ -24 & 4 \end{bmatrix}$$

$$S = \frac{\begin{bmatrix} 188 & -21 \\ -21 & 4 \end{bmatrix} + \begin{bmatrix} 228 & -24 \\ -24 & 4 \end{bmatrix}}{2} = \begin{bmatrix} 208.1 & -22.5 \\ -22.5 & 4.1 \end{bmatrix}$$

Pooled

$$S = \frac{(n_1 - 1)\text{cov}(A) + (n_2 - 1)\text{cov}(B)}{n_1 + n_2 - 2}$$

马氏距离(Mahalanobis Distance): $MD = \sqrt{(\vec{x} - \vec{y})^T S^{-1} (\vec{x} - \vec{y})}$

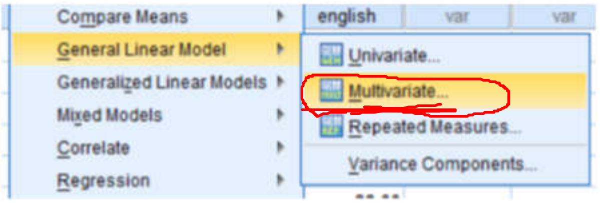


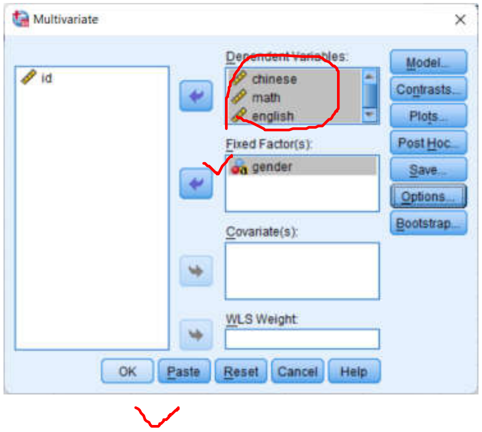
Hotelling’s T-square

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} MD^2$$

$$F = \frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} T^2$$

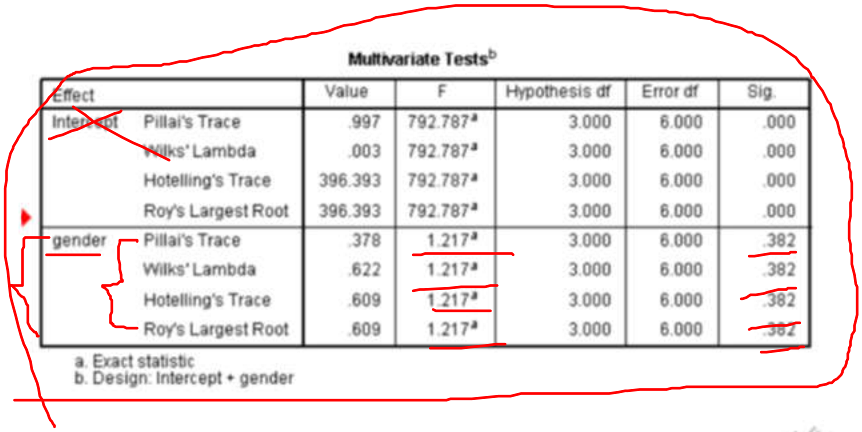
算得的p值与MANOVA是一致的





Dependent Variables: chinese, math, english

Fixed Factor(s): gender



Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.997	792.787[a]	3.000	6.000	.000
	Wilks' Lambda	.003	792.787[a]	3.000	6.000	.000
	Hotelling's Trace	396.393	792.787[a]	3.000	6.000	.000
	Roy's Largest Root	396.393	792.787[a]	3.000	6.000	.000
gender	Pillai's Trace	.378	1.217[a]	3.000	6.000	.382
	Wilks' Lambda	.622	1.217[a]	3.000	6.000	.382
	Hotelling's Trace	.609	1.217[a]	3.000	6.000	.382
	Roy's Largest Root	.609	1.217[a]	3.000	6.000	.382

a. Exact statistic

b. Design: Intercept + gender

本页很重要

小结：常见检验的原假设H0（status quo）

均值比较 原假设：均值相等。

单样本t检验：
H0： μ = μ0

独立样本t检验：
H0: μ1 = μ2

ANOVA
H0: μ1 = μ2 = μ3 = ...

方差齐次性检验：
H0: 方差/标准差相等
$\sigma_1 = \sigma_2$

相关分析：
H0:不相关 ρ = 0

回归分析 Y = a + bx
H0: b = 0

分布拟合检验|拟合优度检验（包括卡方，K-S，秩和检验）
H0: 服从某种分布|分布相同 PMF1 = PMF2, CDF1 = CDF2

REMEMBER

All null hypotheses include an equal sign in them.