

生物医学统计基础笔记

0 Thinking Statistically

1. 从样本推断总体
2. 统计推断的依据:
 1. 随机采样
 2. 样本足够大
 3. 样本统计量的分布特征已知
3. 数据可视化提早统计分析
4. p值与 H_0 / H_1 成立的关系: 没有关系, 是拒绝用的, 不能用于成立
5. 统计显著性 (α) 不代表实际显著性 (样本反映的)

报告格式: APA格式

单样本t-检验:

根据单样本t-检验的结果, 样本均值 ($M = 3.20$, $SD = 0.80$) 显著高于总体均值 ($t(29) = 2.60$, $p < .05$, $95\%CI=[xxx,xxx]$)。

独立样本t-检验:

根据独立样本t-检验的结果, 组A的平均值 ($M = 5.10$, $SD = 1.20$) 显著高于组B的平均值 ($M = 4.20$, $SD = 0.90$), $t(58) = 2.90$, $p < .05$, cohen's d=0.8, $95\%CI=[xxx,xxx]$ 。

配对样本t-检验:

根据配对样本t-检验的结果, 前后两次测量的平均值存在显著差异 ($M1 = 3.20$, $SD1 = 0.80$; $M2 = 4.10$, $SD2 = 0.90$), $t(29) = 2.60$, $p < .05$, cohen's d=0.8, $95\%CI=[xxx,xxx]$ 。

$$\bullet \text{ cohen's } d \text{ 是一个效应量, } d = \frac{\bar{x}_1 - \bar{x}_2}{s}, s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

1-way ANOVA:

三个组的成绩分别是12.3($n=12, SD=4.1$), 7.4($n=9, SD=2.3$), 6.6($n=8, SD=3.1$), 显示方案对成绩有显著效应($F(2,26)=8.76, p=.012, \eta^2=0.1$).

进一步组间多重比较Tukey HSD 检验发现, anxifree 与 joyzepam 对情绪改善差异 ($d=0.77$) 达到统计显著性 ($p=.0015$), joyzepam 与 placebo 对情绪改善差异 ($d=-1.49$) 也达到了统计显著性 ($p=.0001$)

卡方检验

卡方检验表面剂量与治疗效果具有显著的弱关联性 ($\chi^2(2)=14.23, p<.001, \text{Cramer's } V=0.18$)

GoF卡方检验

通过XXX检验, 感染人数的分布显著服从/不服从XX分布, $\chi^2(df, N) = xx, p=.xxx$
其中N是样本观测总数

相关系数

Pearson相关分析表明, 两个变量之间存在很强/中等/弱/没有线性相关性 ($r(N)=.30, p<.001, 95\%CI=[0.15, 0.35]$)

N是样本观测量

回归模型

Social support significantly predicted/explained depression scores ($b_i = -.34$, $t(225) = 6.53$, $p < .001$)

Social support also explained a significant proportion of variance in depression scores ($R^2 = .12$, $F(1, 225) = 42.64$, $p < .001$)

F分布的自由度分别是自变量个数，样本观测量-1-自变量个数

1 预备知识

1.1 常用代码

► click here

1.2 生物医学数据类型

两个大类，四个小类：

可数的：

离散型：病毒感染人数，人的心率

连续性：人的血压值

可分类的：

有序(ordinal)：学生成绩 (A/B/C/D)

无序(nominal)：病毒诊断结果 (阳性/阴性)

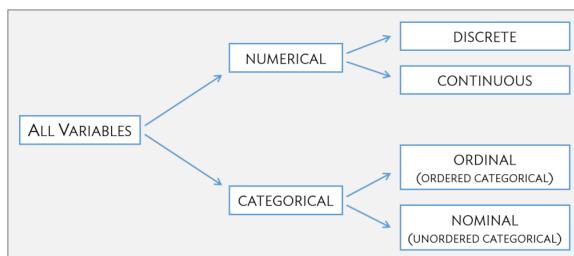


Figure 1.8: Breakdown of variables into their respective types.

1.3 常用术语

总体、样本、自变量、因变量、概率

样本的叫**统计量**，总体的叫**参数**

样本 (Sample) 和总体 (Population) 的例子：

对照组和实验组各10只卒中小鼠，研究某药物对卒中的干预作用，这个例子中有：

总体1：药物干预的任何卒中小鼠 (大小不确定)

总体2：没有用药物干预的任何卒中小鼠 (大小也不确定)

样本1：本实验中的药物干预组 ($n=10$)

样本2：本实验中的对照组 ($n=10$)

本实验有两个总体，两个样本

另一个例子：



1.2 Week 1 / 课堂复习

问题：一项虚构的新降血压药物研究实验设计，研究它是否比目前的一个常规药物（药效8小时）具有更长的药效时间，招募了一批高血压病人（200人），随机分成两组（暂不考虑伦理问题）：

- 组A：常规降压药物；
- 组B：新型降压药

统计方法：两组病人在用药8小时后，分别检测血压值(SBP, mmHg)，比较新药组是否比常规药物组平均血压值更低，并推断8小时后，病人的血压是否显著高于正常人的SBP值（130mmHG）

问题：这个案例中

1. 这个实验设计中的总体，样本分别是什么，各有多少个？
2. 这个研究设计中，哪些是统计量，哪些值是参数？
3. 组A的病人数n=100是（样本大小，样本数量，样本量）
4. 如果从自变量，因变量角度来分析，这个实验设计的因变量，自变量分别是什么？



- 1、总体有3个，分别是正常人群和两个服用不同药的人群；样本两个，分别100人
- 2、统计量：8h后两组血糖平均值；参数：总体数量，正常人SBP
- 3、**样本大小**
- 4、自变量是吃哪种药,因变量是8h后**每个个体**的血压

推断性统计：用样本来推测总体

描述性统计：用全体的数字特征来描述总体

1.4 大数定律和中心极限定理

注意，这与概统中的概念有所区别

大数定律(LLN)：样本观测值越大，样本均值观测结果越接近总体均值

中心极限定理(CLT)：样本均值的分布服从正态分布，该分布均值等于总体均值

► 用中心极限定理求均值click

1.5 常用分布和函数

正态分布：stats.norm() Z分布就是标准正态分布

T分布：stats.t()

二项分布：stats.binom()

卡方分布：stats.chi2()

pdf:概率密度 stats.norm.pdf(x,μ,σ)

pmf:概率质量 stats.binom.pmf(k,n,p)

cdf:分布函数，用法同上

rvs:生成随机变量 stats.norm.rvs(size,μ,σ)

ppf:下侧分位数 stats.norm.ppf(p,μ,σ)

isf:上侧分位数 stats.norm.isf(p,μ,σ)

2 数据可视化

2.1 一般建议

看变化趋势: lineplot

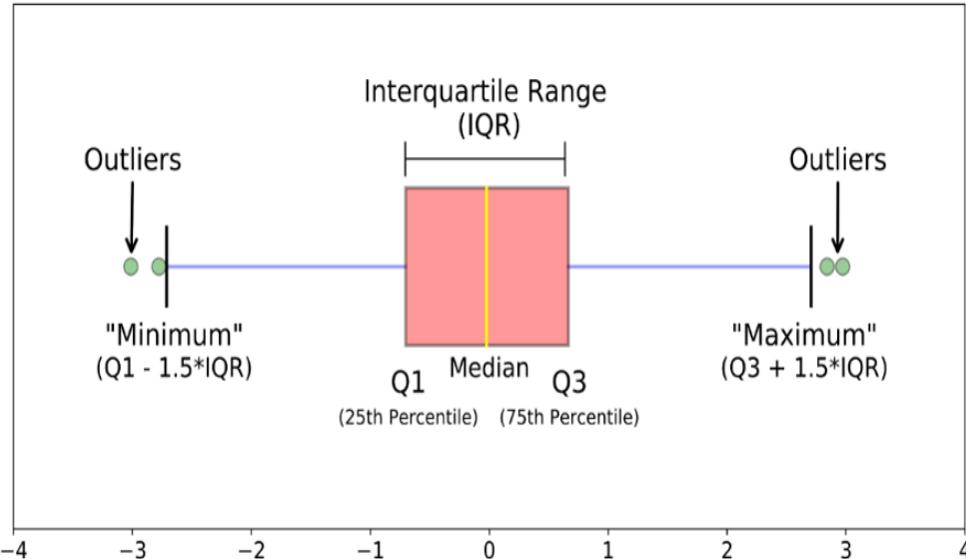
看分布特征: boxplot, scatterplot, 直方图, pdf

看关系: 回归线, scatterplot+回归线, 热图

plt.subplot(2, 2)用于生成一个 2×2 的画布, 可以存放四张图

2.2 箱体图认读

箱体图认读:



Q1, Median, Q3 分别表示 25, 50, 75 分位点 (非均值)

Minimum = $Q1 - 1.5 * IQR$, Maximum = $Q3 + 1.5 * IQR$

IQR = $Q3 - Q1$ (箱体长度)

3 描述性统计

3.1 分位 Quantiles

几个常见的分位数:

1. median 二分位 $1/2$
2. tercile 三分位 $1/3$
3. quartile 四分位 $1/4$
4. quintile 五分位 $1/5$
5. decile 十分位 $1/10$
6. percentile 百分位 $1/100$

▶ python 代码 click

3.2 数据的集中趋势 Central Tendency

平均值 mean

众数 mode

中位数 median

Mean

有n个观测值的样本 $\{x_i | i = 1, 2, \dots, n\}$ 定义

算术平均: $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$

几何平均: $\bar{x}_g = \sqrt[n]{x_1 x_2 \dots x_n}$

调和平均: $\bar{x}_H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}$

Mode

可以是数据型的，也可以是类别型的

Median

若有偶个数，则任意取两个数平均，取低，取高都行

► python代码click

3.3 描述数据的分散性

极差 Range $Range = x_{(n)} - x_{(1)}$

变异系数 Coefficient of variation $cv = \frac{\sigma}{\mu} = \frac{s}{m}$

方差 Variation $\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$, $s^2 = \frac{\sum(x_i - m)^2}{n-1}$

标准差 Standard Deviation σ, s

► python代码click

3.4 描述数据的分布

峰度 kurtosis

偏度 skewness

变异系数 coefficient of variation

峰度

Pearson's Kurtosis:

$Kurtosis(x) = E[(\frac{x-\mu}{\sigma})^4] = \beta_2 = \frac{\mu_4}{\sigma^4}$

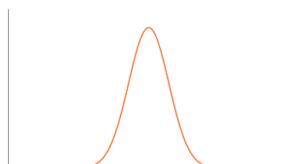
对于正态分布，上式恒等于3

Fisher's Kurtosis/Excess of Kurtosis = Pearson's Kurtosis - 3

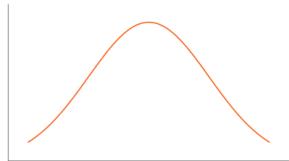
FK=0时，为高斯分布



FK>0时，为超高斯分布

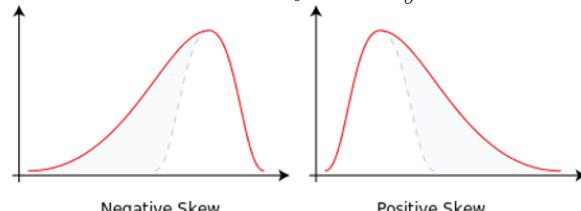


$FK < 0$ 时，为欠高斯分布



偏度

$$Skewness = \tilde{\mu}_3 = E[(\frac{X-\mu}{\sigma})^3] = \frac{\mu_3}{\sigma^3}$$



上图中，左边是向左偏、偏度<0，右边是向右偏、偏度>0（看尾巴）

注意，当偏度=0时仅能说明对称分布，不能说明正态性

变异系数

$$CoV = \frac{\sigma}{\mu} \times 100\% = \frac{s}{m} \times 100\%$$

用途：

1. 比较不同数据集的离散程度
2. 无量纲

缺点：当均值接近0时不好用

► python代码click

4 均值的比较

4.1 置信区间 confidence interval(CI)

4.1.1 单样本

4.1.1.1 单样本总体方差已知，用Z分布

对正态总体或近似正态总体，有

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

若要求置信度为 $1 - \alpha$ ，则

$$\begin{aligned} \therefore P(-u_{\frac{\alpha}{2}} < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq u_{\frac{\alpha}{2}}) &= 1 - \alpha \\ \Rightarrow CI_{1-\alpha} &= (\bar{x} - u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) \end{aligned}$$

4.1.1.2 单样本总体方差未知，用t分布

$$\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t(n-1)$$

$$\therefore P(-t_{\frac{\alpha}{2}} < \frac{\bar{x}-\mu}{s/\sqrt{n}} \leq t_{\frac{\alpha}{2}}) = 1 - \alpha$$

$$\Rightarrow CI_{1-\alpha} = (\bar{x} - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}})$$

4.1.2 双样本

4.1.2.1 双样本是配对的（非独立样本）

何为配对：

同一批人的两次成绩差的均值、双胞胎之间的身高差均值等，可以找到配对的双方的量等价于单样本的分布，参见上方4.1.1

4.1.2.2 两个独立样本，总体方差各自已知

已知 σ_1^2, σ_2^2 ，要求 $\bar{x}_1 - \bar{x}_2$ 的置信区间

$$\therefore \bar{x}_1 - \bar{x}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

记 $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \sigma^2 = sem^2$ ，称其为 $\bar{x}_1 - \bar{x}_2$ 的标准方差，也称标准误差的平方

$$\therefore \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma} \sim N(0, 1)$$

$$\Rightarrow CI_{1-\alpha} = (\bar{x}_1 - \bar{x}_2) \pm u_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

4.1.2.3 两个独立样本，总体方差未知但相等

已知 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ ，但是多少不知道

使用t分布

$$\therefore \bar{x}_1 - \bar{x}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2})$$

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

$$\therefore \frac{(n_1-1)s_1^2}{\sigma^2} \sim \chi^2(n_1-1), \frac{(n_2-1)s_2^2}{\sigma^2} \sim \chi^2(n_2-1)$$

$$\therefore \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}} \sim t(n_1 + n_2 - 2)$$

$$\text{def } \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} = s_p, \bar{x}_1 - \bar{x}_2 \text{的标准误差为} sem = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\therefore \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

$$\Rightarrow CI_{1-\alpha} = (\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}}(n_1 + n_2 - 2) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

4.1.2.4 两个独立样本，总体方差未知且不相等

使用Welch's t分布

推导过于复杂，直接给出结论

$$CI_{1-\alpha} = (\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}}(\nu) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

其中 ν 为t分布的自由度，通常不是整数，用最接近的整数计算，有

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{(n_1-1)}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{(n_2-1)}\left(\frac{s_2^2}{n_2}\right)^2}$$

4.2 零假设检验 Null Hypothesis Significance Test(NHST)

NHST注意事项

含义

NHST的含义仅仅是 $P(x > \text{Data} | H_0)$

1. 只反应数据和假设的兼容性
2. 不能推断 H_0, H_1 成立的概率
3. 不能直接回答真正关心的问题： $P(H_0 | \text{Data})$

统计显著性 α 不等于实际显著性

1. n足够大时，尽管差距很小，p仍然可以很小
2. p很小不代表实际差距很大

常见错误：

1. p值之间的比较：两组实验组，分别和对照组求p，p=0.04的结果要比p=0.05的结果更显著吗？不
2. 只做了一次实验就不能讲概率

p值到底反映了什么？

1. $p < 0.05$ ，说明两个样本均值差异有显著性（在 H_0 假设成立的情况下）
2. $p < 0.05$ ，说明两个总体均值存在差异的显著性很大？**错误**
3. “认为 H_0 假设错误”这个**结论错误的概率**小于p值

4.2.1 NHST与CI的关系

1. NHST的回答比CI**简洁**，只回答Yes or No
2. NSHT没有CI准确，CI还提供大/小多少的信息
3. NSHT要回答的问题**都可以用CI回答**

4.2.2 NHST的推理逻辑

详见概统笔记

1. 先做零假设和备择假设
 - 什么样的假设可以作为零假设
 - 假设完可以确定枢轴量的分布
 - 用想要拒绝的做假设
 - 不能用p不够小来确定其成立
2. 选择枢轴量
3. 根据枢轴量在 H_0 成立时的分布，根据检验方式算概率（p值）
 1. 双边检验： $p = P(|x| \geq |X|)$
 2. 右侧检验： $p = P(x \geq X)$
 3. 左侧检验： $p = P(x \leq X)$
4. 检验p的显著性水平，如果 $p < \alpha$ ，则拒绝原假设， H_1 成立

4.2.3 单样本t检验

总体 σ^2 未知，使用t分布
拒绝域是同 α 下CI的**补集**

► python代码click

4.2.4 双样本t检验

一个错误例子：现有A($mean=12$)，B($mean=8$)，得到 $t(29)=xxx$, $p < .05$ ，认为A的均值显著大于B
错误：只能认为 $mean(A) \neq mean(B)$ ，因为做的是双侧检验，无法得出单边结论。**存疑**

4.2.4.1 配对的样本

配对做差，和单样本相同

► python代码click

4.2.4.2 独立的样本

和单样本类似的，拒绝域都是相应的 α 下的CI的补集

► python代码click

4.3 对方差分析 (ANOVA)

4.3.1 从t检验到ANOVA检验

有n组数据，两两之间做t检验（多重检验），重复m次，假设每两组、每次检验的假阳性（去真错误，第一类）概率为 α ，那么所有检验结后有假阳性的概率是：

$$P = 1 - (1 - \alpha)^{C_n^2 \cdot m}$$

这是非常大的一个数，所以抛弃多重检验

4.3.2 ANOVA原理

假定分为k组，共N个对象，每组 N/k 个

ANOVA做F检验，其核心是 $SST = SSB + SSW$ (Total=Between+Within)，注意自由度，证明见下

$\{x_{ij}\}$: The j^{th} observation in the i^{th} group \bar{x} : The mean over all groups n_i : The number of observations within the i^{th} group $SST = \sum_{i,j} (x_{ij} - \bar{x})^2$ $= \sum_{i,j} (x_{ij}^2 - 2x_{ij}\bar{x} + \bar{x}^2)$ $SSW = \sum_{i,j} (x_{ij} - \bar{x}_i)^2$ $= \sum_{i,j} (x_{ij}^2 - 2x_{ij}\bar{x}_i + \bar{x}_i^2)$ $SSB = \sum_i n_i \cdot (\bar{x}_i - \bar{x})^2$	$\bar{x} = \frac{\sum_{i,j} x_{ij}}{n}$ $\bar{x}_i = \frac{\sum_j x_{ij}}{n_i}$ $n = \sum_i n_i$ $\sum_i n_i \bar{x}_i = n\bar{x}$ $SST = \sum_i \sum_j (x_{ij} - \bar{x})^2 = \sum_i \sum_j (x_{ij}^2 - 2x_{ij}\bar{x} + \bar{x}^2)$ $= \sum_i \sum_j x_{ij}^2 - 2\bar{x} \sum_i \sum_j x_{ij} + \sum_i \sum_j \bar{x}^2$ $= \sum_i \sum_j x_{ij}^2 - 2\bar{x} \cdot n\bar{x} + n\bar{x}^2$ $= \sum_i \sum_j x_{ij}^2 - n\bar{x}^2$	$SSB = \sum_i n_i (\bar{x}_i - \bar{x})^2$ $= \sum_i n_i \bar{x}_i^2 - 2 \sum_i n_i \bar{x}_i \bar{x} + \sum_i n_i \bar{x}^2$ $= \sum_i n_i \bar{x}_i^2 - 2\bar{x} \sum_i n_i \bar{x}_i + n\bar{x}^2$ $= \sum_i n_i \bar{x}_i^2 - 2n\bar{x}^2 + n\bar{x}^2$ $= \sum_i n_i \bar{x}_i^2 - n\bar{x}^2$ $SSW = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 = \sum_i \sum_j x_{ij}^2 - 2 \sum_i \sum_j x_{ij} \bar{x}_i + \sum_i \sum_j \bar{x}_i^2$ $= \sum_i \sum_j x_{ij}^2 - 2 \sum_i n_i \bar{x}_i \cdot \bar{x}_i + \sum_i n_i \cdot \bar{x}_i^2$ $= \sum_i \sum_j x_{ij}^2 - \sum_i n_i \bar{x}_i^2$ $\therefore SST = SSW + SSB$
---	---	--

得到ANOVA的报告表

Source	SS	df	MS	F	p-value	eta2
Total	SSB+SSW	N-1	/	MSB/MSW	F(k-1,N-k)对应的p	SSB/SST
Between	SSB	k-1	MSB	/	/	/
Within	SSW	N-K	MSW	/	/	/

其中 $MS = SS/df$, $F(df_B, df_W) = \frac{MSB}{MSW} = \frac{SSB/df_B}{SSW/df_W}$
 $\eta^2 = \frac{SSB}{SST}$, 表示SST中有多少是SSB贡献的，也叫effect size

4.3.3 单因素对方差分析 1-Way ANOVA

类似于下面的分组，每组的受试样本都是不一样的，是1-Way ANOVA

Drug1	Drug2	Drug3
sub1,sub2,sub3	sub4,sub5,sub6	sub7,sub8,sub9

类似下面的，每组的受试样本一样的，是1-Way Repeated Measure ANOVA

Drug1	Drug2	Drug3
sub1,sub2,sub3	sub1,sub2,sub3	sub1,sub2,sub3

► python代码click

1-way ANOVA, 1-way RM ANOVA条件

1. 独立观测
2. 独立样本 (1-way) , 重复/配对样本 (1-way RM)
3. 正态分布
4. 方差齐性 (1-way) , 球性 (1-way RM)
5. 连续变量
6. 组内随机采样
7. 组间样本量平衡

4.3.4 两两配对比较

当通过1-way ANOVA**拒绝零假设后**, 要想确定是哪几组之间的存在显著差异, 就要使用Post Hoc多重检验 (事后检验) (不使用多重t检验的原因同上)

► python代码

最后得到的结果是

A	B	mean(A)	mean(B)	diff	se	T	p-tukey
hedges							
0	anxitfree	joyzepam	0.7167	1.4833	-0.7667	0.1759	-4.3596
	-2.3234						0.0015
1	anxitfree	placebo	0.7167	0.4500	0.2667	0.1759	1.5164
	0.8081						0.3117
2	joyzepam	placebo	1.4833	0.4500	1.0333	0.1759	5.8760
	3.1315						0.0010

其中, diff=mean(A)-mean(B), se是枢轴量的standard error, T是枢轴量观测值, p-tukey是T对应的p值的修正值, hedges不管

4.3.5 2-way ANOVA

原理

$$SST = SSB_A + SSB_B + SS_{AB} + SSW$$

Source	SS	DF	MS	F
$A_{(Rows)}$	$n_r \sum_{w=1}^r (M_w - G)^2$	$r - 1$	$\frac{SS_r}{df_r}$	$\frac{MS_r}{MS_W}$
$B_{(Col)}$	$n_c \sum_{u=1}^c (M_u - G)^2$	$c - 1$	$\frac{SS_c}{df_c}$	$\frac{MS_c}{MS_W}$
AxB	$SS_B - SS_r - SS_c$	$(r - 1)(c - 1)$	$\frac{SS_{rxc}}{df_{rxc}}$	$\frac{MS_{rxc}}{MS_W}$
Within	$\sum_{w=1}^r \sum_{j=1}^n (X_{wj} - M_w)^2 + \sum_{u=1}^c \sum_{j=1}^n (X_{uj} - M_u)^2$	$nrc - rc$	$\frac{SS_w}{df_w}$	
Total	$\sum_{j=1}^N (X_j - G)^2$	$nrc - 1$		

Gender\DRug	A	B
male	sub1,sub2,sub3	sub4,sub5,sub6
female	sub7,sub8,sub9	sub10,sub11,sub12

当出现上面这种有两个变量时，使用2-way ANOVA，假设是

- H01: Factor1不同水平的均值没有差异
- H02: Factor2不同水平的均值没有差异
- H03: Factor1不同水平的均值差异与Factor2无关，即不存在交互作用

► python代码

其报告表如下

Source	SS	DF	MS	F	p-unc	np2
0 Drug	3.453333	2	1.726667	31.714286	0.000016	0.840909
1 Therapy	0.467222	1	0.467222	8.581633	0.012617	0.416956
2 Drug * Therapy	0.271111	2	0.135556	2.489796	0.124602	0.293269
3 Residual	0.653333	12	0.054444	NaN	NaN	NaN

意义和1-way相同，但是算法不同

2-way ANOVA比两个独立的ANOVA具有更高的统计功效 (statistic power) ,即更容易检测出主效应 (很小的p值)

故即使不关心两个变量间的相互作用，也使用2-way ANOVA进行检验

5 类别的比较

5.1 单个样本比例的置信区间

5.1.1 例子引入

例：某社区有12000人，但只有200个试剂盒，要如何检测社区的感染率？即：如何从样本推断总体比例一共有200次检测机会，记n=200, N=12000

在一次检测中，测出了m个阳性，那么这组的感染率就是 $p_1 = \frac{m}{n}$

假设12000中一共有M个阳性，那么很显然 $E(p_i) = \frac{M}{12000} \triangleq p$ 其中p就是每个人的感染率，即要求的东西

p_i 的方差满足 $D(p_i) = D(m/n) = \frac{p(1-p)}{n} \Rightarrow \sigma_p = \sqrt{\frac{p(1-p)}{n}}$

根据大数定律和中心极限定理，当满足下面的条件时，就有

$p_i \sim N(p, \frac{p(1-p)}{n})$, $p \approx p_i \Rightarrow \sigma_p \approx \sigma_{p_i}$

故用 $p_i \sim N(p, \frac{p_i(1-p_i)}{n})$ ，作为枢轴量计算置信区间，得到p的置信区间 $CI_{1-\alpha} = p_i \pm Z_{\frac{\alpha}{2}} \cdot \sigma_{p_i}$

要满足的条件是：

1. $n < 5\%N$ ，即每次抽样的样本数少于总体的样本数的5%（为了保证 p_i 的数量满足中心极限定理）

2. 阳性和阴性数目都不少($np_i > 10, n(1 - p_i) > 10$)

实际上，这个想法和概统中所说的中心极限是一致的。

假定个体 x_i ，其得病概率p，那么总体中所有患病人数满足 $\Sigma x \sim B(N, p)$

根据中心极限定理，有 $\Sigma x \stackrel{\text{近似}}{\sim} N(Np, Np(1 - p)) \approx N(Np, Np_i(1 - p_i))$

$\therefore p \stackrel{\text{近似}}{\sim} N(p, \frac{p_i(1-p_i)}{N})$

此时只需要满足条件2即可

5.1.2 求单样本CI的两种方法

- Simple Asymptotic: 即上面的方法
- Simple Asymptotic with **continuity correction**: $CI_{1-\alpha} = p_i \pm (Z_{\frac{\alpha}{2}} \cdot \sigma_{p_i} + \frac{1}{2n})$

► python代码

5.2 多个独立样本的比例差异的统计

注意，以下的样本必须满足5.1的条件

5.2.1 差异的置信区间

原理

两个独立的正态总体 $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$ ，枢轴量如前文得到的独立样本的t检验

现在就是将 X_1, X_2 换成 \hat{p}_1, \hat{p}_2 ，其均值是要求的，方差是用已知估计的

有 $\hat{p}_1 - \hat{p}_2 \sim N(p_1 - p_2, \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2})$

得到枢轴量

几种算法

- Pearson's: $CI = \hat{p}_1 - \hat{p}_2 \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
- Yate's: $CI = \hat{p}_1 - \hat{p}_2 \pm (z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} + \frac{1}{2}(\frac{1}{n_1} + \frac{1}{n_2}))$

5.2.2 差异的假设检验

5.2.2.1 两个样本用NHST(zTest)

$H_0 : p_1 = p_2$, 枢轴量如5.2.1所示, 做NHST

► python代码

5.2.2.2 多个样本用RC联表的 χ^2 检验

例：药物剂量对阳性率的影响

症状	剂量1	剂量2	剂量3
阳性	10	40	5
阴性	90	160	95
总计	100	200	100

即 $p_1 = p_2 = p_3$? (H_0)

上表格是观测值(Observation), 下面给出当 H_0 成立时的期望人数(Expectation)

症状	剂量1	剂量2	剂量3	总计
阳性	10	40	5	55
阴性	90	160	95	345
总计	100	200	100	400

由总计一栏计算如果相等各个计量的人数分布

症状	剂量1	剂量2	剂量3
阳性	13.75	27.5	13.75
阴性	86.25	172.5	86.25
总计	100	200	100

得到 $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$, 其自由度 $df = rc - 1 - r - c = (r - 1)(c - 1)$, 其中 r, c 分别是行数、列数(不包括总计), 上例中 $df = (3-1)(2-1) = 2$

由于仅当 H_0 成立时上式才服从卡方分布, 故可以做检验

效应量Cramer's V(φ_c)

$$V = \sqrt{\frac{\chi^2}{N \cdot \min(r-1, c-1)}}, \text{ 其中 } N \text{ 是样本量, } V \text{ 表示变量对因变量的关联性强弱}$$

Rules of thumb:

$V \in [0.1, 0.2]$: 弱关联, $V \in [0.2, 0.5]$: 中等关联, $V > 0.5$: 强关联

RC联表卡方检验的条件

- 各个单元格独立观测

- 2x2表:

- $E_i \geq 10$

- $5 \leq E_i < 10$, 使用Yate's correction (仍是卡方检验)

3. $E_i < 5$, 使用Fisher's Exact Test (不是卡方检验)

- 大于等于2x3表:

1. $E < 5$ 的单元格小于20%

2. 可以合并几列再做

► python代码

5.3 计数数据的卡方检验: Goodness of Fit

例: 五个班级的流感人数分布是否均匀

班级	1	2	3	4	5
观测值(Obs)	8	10	12	7	13

先列出预期值

班级	1	2	3	4	5
观测值(Obs)	8	10	12	7	13
预期值(Exp)	10	10	10	10	10

类似于RC联表的卡方检验, 就是验证Obs和Exp是否同分布

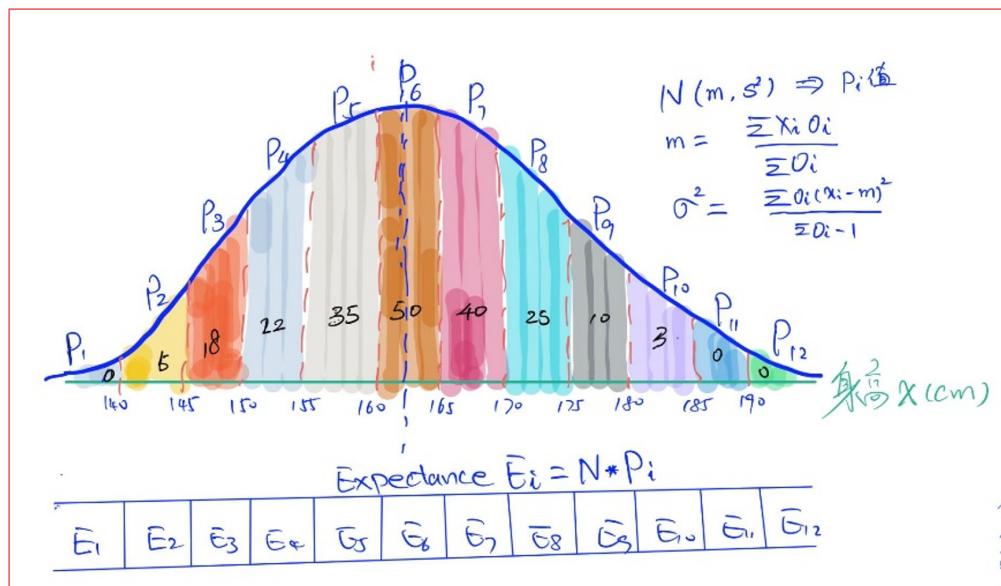
原理是 $\chi^2 = \sum_{i=1}^c \frac{(O_i - E_i)^2}{E_i}$, $df=c-q-1$, 其中c是样本数, q是列出预期值所需的由样本推出的未知量, 1是为了保持总体不变

本例中 $df=5-1=4$, $q=0$, 本例只用到了总体不变

APA报告格式: 通过XXX检验, 感染人数的分布显著服从/不服从XX分布, $\chi^2(df, N) = xx, p=.xxx$
N是总人数 (不是班级数)

► python代码

正态分布的例子



选取每个区间的中间值作为代表值，构造观测值数组count

如果认为应该服从 $N(\mu, \sigma^2)$ ，通过观测值计算，即 $\mu = \text{mean}(\text{Count})$, $\sigma = s$, 此时q=2

根据每组差值计算此时正态分布的 E_i , 即 $E_i = (\Phi(u_i) - \Phi(u_{i-1})) \times N$, 再加上两头的用 O_i 和 E_i 构造 χ^2

如果认为服从 $N(1, 8)$, 那么此时q=0, 因为没用样本观测值, $\mu = 1, \sigma = 2\sqrt{2}$

例：如下图

2. 如果要调查不同城市中80岁以上人口总糖尿病人数的**比例分布**情况，随机挑选了50个城市，每个城市随机采样100个80岁以上的老人，统计其中的糖尿病人的数据，能否用卡方检验推断下面的问题？如果能，自由度是多少？

- (1) 不同城市的80岁以上的**老人糖尿病比例**是否存在差异？
- (2) 80岁以上的**糖尿病老人的数目**是否服从正态分布？
- (3) 50年后，再按相同的方法采样这50个城市，**80岁以上的糖尿病人的比例**是否发生变化？

(1) 能，使用GoF卡方检验，Exp是分布均匀，故df=50-1=49

或者，使用RC联表卡方检验，df= (50-1)(2-1)=49

(2) 能，划分一个区间，往里用计数的方法，如上例，分为n组，df=n-2-1

(3) 能，此时认为50年前测得的数据为Exp，50年后测得的数据为Obs

$$\chi^2 = \sum_{i=1}^{50} \frac{(O_i - E_i)^2}{E_i}, \text{ 由于此时的Exp和Obs总数无关, df=50, 不需要-1}$$

5.4 判断服从正态分布的方法

```
print(stats.skew(data))      %偏度
print(stats.kurtosis(data))   %峰度
print(stats.shapiro(data))    %NHST
fig = sm.qqplot(data, stats.norm,line='s')    %QQplot
以及卡方检验（最准确）
```

6 两个、多个变量之间的关系

6.1 相关性分析

6.1.1 协方差

某个样本有两个变量，记为 X, Y , 其观测值记为 x_i, y_i , 则 x, y 的协方差为

$$\text{cov}(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}, n-1 \text{是无偏估计}$$

注意在概统中曾经提过的协方差 $\text{cov}(X, Y) = E(\sum_i (X_i - \bar{X})(Y_i - \bar{Y}))$ 是总体的，所以不除 $n - 1$

► python代码

很显然 $\text{cov}(2x, 2y) = 4\text{cov}(x, y)$, 即协方差受量纲的影响，为了解决这个问题，使用相关系数

6.1.2 相关系数

以下三个相关系数，只需且只能算一个

6.1.2.1 Pearson's linear(Pearson 相关系数)

$$r_{x,y} = \frac{cov(x,y)}{s_x \cdot s_y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

其中 σ 是各自的样本方差，除以样本方差后变为无量纲的量

经验上将 $|r|$ 分为以下三类

1. 0.1~0.3, Small
2. 0.3~0.5, Medium
3. >0.5, large

r对outlier很敏感，r是数据分布是否线性的判断；r=0仅能说明线性关系弱，不是没有关系

► python代码

6.1.2.2 Spearman's ρ

使用情况：

1. X,Y不是数值，而是一个排序的变量， or
2. X,Y没有明显的线性关系

首先将原始的 (x_i, y_i) 排序，用排序后的顺序作为新的值，带入 $r_{x,y}$ 的公式计算
用于评估单调关系

例：X[5,2,8,1,4];Y[10,4,12,3,8]

首先排序，得到新的：R[3,2,5,1,4];S[4,2,5,1,3]，对应的 (r_i, s_i) 不变

如果有相同的，就并列，如R[1,2,2,4,5]

而平均值就是排序后正常计算的平均值，用新的数组带入上述公式

def $d_i = R_i - S_i$ ，则有

$\rho = 1 - \frac{6 \sum_i d_i^2}{N(N^2 - 1)}$ ，其中N是样本观测量，上例中N=5，和上式计算结果一致

► python代码

6.1.2.3 Kendall's τ

使用情况：

1. X,Y不是数值，而是一个排序的变量， or
2. X,Y没有明显的线性关系

变量X和Y的任何两次观测值 (x_i, y_i) vs (x_j, y_j) 间是否具有一致性大小关系，比如：

- 关系一致 if $(x_i > x_j \text{ and } y_i > y_j) \text{ or } (x_i < x_j \text{ and } y_i < y_j)$: 总计: n_c
- 关系不一致: if $(x_i > x_j \text{ and } y_i < y_j) \text{ or } (x_i < x_j \text{ and } y_i > y_j)$: 总计: n_d
- 关系不确定: if $x_i = x_j \text{ or } y_i = y_j$, t_i, t_j

$$\tau_A = \frac{n_c - n_d}{n_c + n_d} = \frac{n_c - n_d}{n(n-1)/2}$$

$$\tau_B = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}$$

$$n_0 = n(n-1)/2$$

$$n_1 = \sum_i t_i(t_i - 1)/2$$

$$n_2 = \sum_j u_j(u_j - 1)/2$$

n_c = Number of concordant pairs

n_d = Number of discordant pairs

t_i = Number of tied values in the i^{th} group of ties for the first quantity

u_j = Number of tied values in the j^{th} group of ties for the second quantity

`stats.kendalltau()`默认计算的是 τ_b

▶ python代码

6.1.3 相关系数的置信区间估计, NHST

用样本观测值得到的样本相关系数, 可以推测总体相关系数的区间

实际上为了得到总体的相关系数也只能这么做

使用Fisher-z变换修正偏度, 得到以下表格

	Fisher-z	CI of z	CI we need
Pearson's r	$z_r = \frac{1}{2} \ln \frac{1+r}{1-r}$	$z_r \pm z_{\alpha/2} \sqrt{\frac{1}{n-3}}$	$r_L = \frac{e^{2z_L} - 1}{e^{2z_L} + 1}, r_U = \frac{e^{2z_U} - 1}{e^{2z_U} + 1}$
Spearman's ρ	$z_\rho = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$	$z_\rho \pm z_{\alpha/2} \sqrt{\frac{1+\rho^2/2}{n-3}}$	$\rho_L = \frac{e^{2z_L} - 1}{e^{2z_L} + 1}, \rho_U = \frac{e^{2z_U} - 1}{e^{2z_U} + 1}$
Kendall's τ	$z_\tau = \frac{1}{2} \ln \frac{1+\tau}{1-\tau}$	$z_\tau \pm z_{\alpha/2} \sqrt{\frac{0.437}{n-4}}$	$\tau_L = \frac{e^{2z_L} - 1}{e^{2z_L} + 1}, \tau_U = \frac{e^{2z_U} - 1}{e^{2z_U} + 1}$

注意, 求得的置信区间是**非对称的**

6.1.4 Pearson's r相关分析假设条件

1. 两个连续变量 (或近似于连续)
2. 描述**线性**关系
3. 没有异常值outlier
4. **相关性不等于因果性**; p值不反应总体的相关性
5. 观测值相互独立

6.1.5 APA报告

Pearson相关分析表明, 两个变量之间存在很强/中等/弱/没有线性相关性 ($r(N)=.30, p<.001, 95\% CI=[0.15, 0.35]$)

N是样本观测量

6.2 简单线性回归分析

根据样本值, 建立一阶线性模型 $Y = \beta_0 + \beta_1 X$, 要求XY都是连续变量

实际上, 根据有限的样本值得到的样本观测模型为 $y_i = b_0 + b_1 x_i + \epsilon_i$, 其中 ϵ_i 为第i组观测值的误差, $(x_i, b_0 + b_1 x_i)$ 为第i组回归值/预测值

注意, 每一组样本值都应该独立

例: 建立物理成绩和数学成绩的简单回归模型, 下面的计算结果都是正确的, 问哪个说法是正确的

1. 某同学, 数学80分的, 建议推断其物理成绩是80分
2. **数学成绩为80分的同学, 物理平均分为82分**
3. 可以通过提高数学成绩, 来提高物理成绩
4. 一组学生经过训练, 数学成绩提高5分, 则平均物理成绩提高约4.5分
5. **数学成绩平均为90分的学生比数学成绩平均80分的学生, 物理成绩平均高9分**

正确的是标红的; 不能推断单个; 不能推断一个样本的变化

6.2.1 估计样本回归模型

为了得到 $y = b_0 + b_1 x$, 需要进行线性拟合, 这里使用最小二乘法(Ordinary Least Square)

预测值 $\hat{y}_i = b_0 + b_1 x_i$, 误差 $\epsilon_i = y_i - \hat{y}_i = (y_i - b_0 - b_1 x_i)$

def 样本整体预测误差(sum of squared errors)SSE = $\sum_i (y_i - \hat{y}_i)^2$

b_0, b_1 要使得SSE最小, 则 $\frac{\partial SSE}{\partial b_0} = \frac{\partial SSE}{\partial b_1} = 0$

$$\Rightarrow b_0 = \bar{y} - b_1 \bar{x}, b_1 = \frac{n \sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n \sum_i x_i^2 - (\sum_i x_i)^2} = \frac{\text{COV}(x,y)}{\text{D}(x)}$$

注意 b_1 的第二种算法使用的是计算总体的方法，即/ n 而非 $/(n-1)$ ，当然用样本的计算结果是一致的

得到样本回归模型 $y = b_0 + b_1 x$ ，注意这只能表示观测对象和自变量的关系，**对观测值不适用**

6.2.2 总体模型

由样本推总体无非是CI和NHST

假定残差项在任何 x_i 处服从正态分布： $\epsilon_i \sim N(0, \sigma^2)$

分布特征	总体参数 ($\beta 0$)	总体参数 ($\beta 1$)
点估计 (样本统计量-均值)	b_0	b_1
样本均值方差 σ^2 用样本MSE, or s^2 代替 $\sigma^2 \approx MSE = s^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}$ 则 b_0, b_1 为t分布	$\sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} \right]$ $sem(b0) = \sqrt{s^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} \right]}$	$\frac{\sigma^2}{\sum(x_i - \bar{x})^2}$ $sem(b1) = \sqrt{\frac{s^2}{\sum(x_i - \bar{x})^2}}$

得到 β_0 和 β_1 的CI，注意t分布的自由度是 $n-2$ ，因为使用样本估计了方差 (n-1-1)

► python代码

为了使 $X=0$ 时的Y有意义，即为了 $x=0$ 时截距有意义，一般会将 X 做随机变量中心化，此时截距的意义就是 X 为平均值时的Y值；做随机变量标准化是为了解决量纲的影响

6.2.3 模型误差估计

6.2.3.1 均值CI(CIB)

每个观测值 x 处，对应的 y （多个）的均值 \hat{y} 的CI，有

$$\hat{y}_{ci} = \hat{y} \pm t_{\alpha/2}(n-2) \cdot \sqrt{MSE} \cdot \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$$

$$\text{其中 } MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}$$

注意这是 \hat{y} 在 x 处的CI

6.2.3.2 观测值CI(PIB)

$$\text{每个观测值 } x \text{ 处，对应的观测值 } y \text{ 的CI，有 } y_{pi} = \hat{y} \pm t_{\alpha/2}(n-2) \cdot \sqrt{MSE} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$$

MSE见上

► python代码

模型概括											
Dep. Variable:	mpg	R-squared:	0.840								
Model:	OLS	Adj. R-squared:	0.823								
Method:	Least Squares	F-statistic:	48.96								
Date:	Sat, 28 May 2022	Prob (F-statistic):	2.91e-11								
Time:	14:18:08	Log-Likelihood:	-73.067								
No. Observations:	32	AIC:	154.1								
Df Residuals:	28	BIC:	160.0								
Df Model:	3										
Covariance Type:	nonrobust										
模型系数											
	coef	std err	t	P> t	[0.025	0.975]					
Intercept	34.0029	2.643	12.867	0.000	28.590	39.416					
C(am) [T. 1]	2.0837	1.376	1.514	0.141	-0.736	4.903					
hp	-0.0375	0.010	-3.902	0.001	-0.057	-0.018					
wt	-2.8786	0.905	-3.181	0.004	-4.732	-1.025					
残差分析											
Omnibus:	2.810	Durbin-Watson:	1.433								
Prob(Omnibus):	0.245	Jarque-Bera (JB):	2.339								
Skew:	0.654	Prob (JB):	0.311								
Kurtosis:	2.790	Cond. No.	1.08e+03								

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.08e+03. This might indicate that there are strong multicollinearity or other numerical problems.

当共线性很强（残差线性相关）时，模型不能用来解释变量之间的关系（拟合地不好），但是仍然能用于预测数值

NHST的零假设是 $b_i=0$ ；当残差的方差过大时，模型不能用来预测，但是解释的效果可以

6.2.4 一阶线性回归模型的假设

1. X, Y都是连续变量
2. Y随X线性变化
3. 残差正态性, $\epsilon \sim N(0, \sigma^2)$
4. 所有数据i.i.d.
5. 自变量在同一个level上

6.2.5 方差分析

定义

$$SST = \sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2 \sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

当满足OLS时，第三项=0，定义 $SSR = \sum_i (\hat{y}_i - \bar{y})^2$, $SSE = \sum_i (y_i - \hat{y}_i)^2$

故 $SST = SSR + SSE$, SSR是自变量能解释的SS, SSE是自变量不能解释的SS(Error)

$$\text{定义} R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

代入定义式有 $R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$, 又 $\hat{y}_i = b_1 x_i + b_0$, 最终得到 $r(x, y) = \pm \sqrt{R^2}$

$$\text{又有} r(x, y) = \frac{\text{COV}(x, y)}{\sqrt{D(x)D(y)}}, b_1 = \frac{\text{COV}(x, y)}{D(x)}$$

$$\therefore b_1 = r(x, y) \cdot \frac{\sqrt{D(y)}}{\sqrt{D(x)}}$$

当然用样本方差也行，因为n都除掉了，实际上r使用样本的量来计算得到的结果也是一致的

故可以根据两个变量的均值，方差，R2来估计线性回归模型

R2越大，模型预测越准确

6.3 多元线性回归模型(MLR)

连续的观测量与多个因变量的关联: $SBP_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \dots + \epsilon_i$, 线性指的是系数 (bi) 是线性的，而非自变量，如 $y = b_0 + b_1 x + b_1 x^2$ 也是线性回归模型

而一般线性回归模型 (GLM) 通常还有一个类别变量

同样的，MLR也是用最小二乘法估计

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_N x_{Ni} + \epsilon_i$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1K} \\ 1 & x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{NK} \end{bmatrix} * \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

最优参数选择，最小化 $\epsilon' \epsilon$

$$\epsilon' \epsilon = [e_1 \ e_2 \ \cdots \ e_N] \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix} = \sum_{i=1}^N e_i^2$$

So minimizing $e' e$ gives us:

$$\min_b e' e = (y - Xb)'(y - Xb)$$

$$\min_b e' e = y'y - 2b'X'y + b'X'Xb$$

$$\frac{\partial(e'e)}{\partial b} = -2X'y + 2X'Xb \stackrel{!}{=} 0$$

$$X'Xb = X'y$$

$$b = (X'X)^{-1}X'y$$

这也导致outlier的干扰很大

► python代码

6.3.1 MLR系数解释

- 无交互作用, $y = b_0 + b_1 x_1 + \cdots + b_n x_n$

在解释系数时, 不能说单个样本的变化, 而是总体的差异

报告给出的t检验的 $H_0 : b_i = 0$, 即因变量和这个因素无关

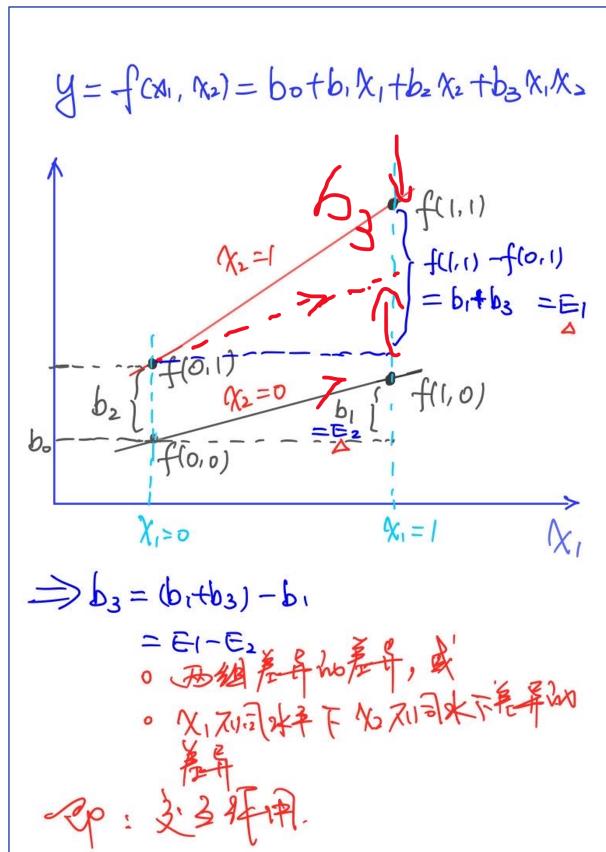
- 存在交互作用, $y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2$

b_1, b_2 的涵义变化为**其他自变量为0时**, 该变量的影响

b_3 可以这么看: $y = b_0 + b_1 x_1 + (b_2 + b_3 x_1)x_2$

即 x_2 的系数与 x_1 有关, 在不同的 x_1 水平下, x_2 对 y 的作用存在变化, x_1, x_2 存在交互作用

b_3 可以认为是两组差异的差异



如要说明某种新药的显著性，也是用上面的方法

即枢轴量不使用 $f(1, 1) - f(1, 0)$ ，而使用 $b_3 = [f(1, 1) - f(0, 1)] - [f(1, 0) - f(0, 0)]$ ，即差异的差异

这也说明，如果两条线不平行，那么说明可能存在交互作用

6.3.2 模型评价

6.3.2.1 残差特性

类似于6.2.3.2

好的模型，残差特性有

1. 正态分布性：

- Jarque-Bera检验的零假设是数据服从正态分布，备择假设是数据不服从正态分布
它基于样本的skewness和kurtosis来进行检验

2. 自相关性弱：残差的值是否独立，是否有 $\epsilon_i = f(\epsilon_{i-1})$ ，不是一般意义上的两个变量相关

- 当Durbin-Watson统计量接近2时(1.5~2.5)，表示残差不存在自相关性（即残差之间相互独立）
- 当Durbin-Watson统计量接近0或4时，表示存在正向或负向的自相关性

3. 共线性弱：即残差之间线性无关；当残差共线性强时，说明模型漏了某些解释变量

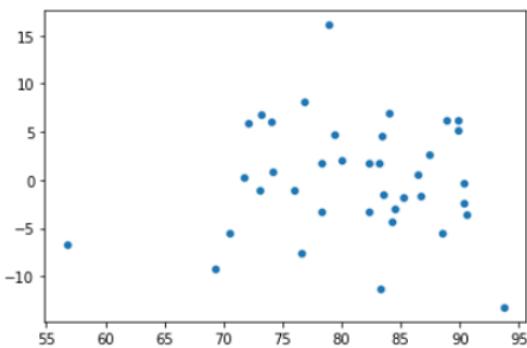
4. 方差齐性：数据不是喇叭口状的，在x不同水平下残差的范围比较一致

```

residuals = results.resid
fitted_value = results.fittedvalues
sns.scatterplot(x=fitted_value, y=residuals)

```

<AxesSubplot:>



好的残差分布，已经归一化

6.3.2.2 总体评价

包括 R^2 , Adjusted R^2 , AIC, BIC, F-value, P-value

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}, R_{adj}^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

其中n为样本大小, p为自变量个数

当自变量数量变多, Adj R²会减小, 模型存在过拟合风险

模型总体的显著性为F-value

OLS Regression Results						
Dep. Variable:	MathScores	R-squared:	0.593			
Model:	OLS	Adj. R-squared:	0.582			
Method:	Least Squares	F-statistic:	53.97			
Date:	Mon, 27 May 2024	Prob (F-statistic):	9.83e-09			
Time:	16:30:34	Log-Likelihood:	-125.48			
No. Observations:	39	AIC:	255.0			
Df Residuals:	37	BIC:	258.3			
Df Model:	1					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
Intercept	32.1283	6.740	4.767	0.000	18.472	45.784
PhyScores	0.6097	0.083	7.347	0.000	0.442	0.778
=====						
Omnibus:		7.642	Durbin-Watson:			2.076
Prob(Omnibus):		0.022	Jarque-Bera (JB):			6.812
Skew:		0.725	Prob(JB):			0.0332
Kurtosis:		4.446	Cond. No.			551.
=====						

R²是服从F分布的, 自由度不用管, **F-value是优先于T-value的**, 当模型的F-value不足时, t的P值再显著, 也不太有效, t的P值是H₀: bi=0的NHST检验值

AIC与BIC

AIC从预测角度, 表明模型对**未知数据的预测程度**, **AIC越小**, 意味着模型更简洁和精确

BIC从拟合角度, 表明模型对**当前数据的拟合程度**, **BIC越小**, 意味着模型更简洁

6.3.3 模型汇报

最好用表格

文字的APA:

Social support significantly predicted depression scores ($bi = -.34$, $t(225) = 6.53$, $p < .001$)

也要汇报R², F, p

Social support also explained a significant proportion of variance in depression scores ($R^2 = .12$, $F(1, 225) = 42.64, p < .001$)

如果模型有显著的交互作用，主效应的解释要谨慎，因为主效应可能是交互作用引起的

模型描述的变量间的关联性，避免把结果解释成因果性

大的R²值不一定好的模型，同样小的R²不一定是差的模型

6.3.4 实际使用的优化方法

数据可视化: pairplot/scatterplot/jointplot, 发现存在相关性的变量 (predictors)

物理原理/实际问题: 根据实际问题选取存在相关性的变量

数据驱动的模型选择 (比较常用, 但需要谨慎)

- Stepwise 选择: 增、减变量, 观察AIC/BIC, R^2_{adj} 值的变化, 朝增加 R^2_{adj} 方向, AIC/BIC减小的方向选择

模型评价: 看残差特征

7 复习

7.1 总论

例: 以下说法全错

均值95% 置信区间包含大约95%的观测值

总体均值落入95%置信区间的概率是95% **正确的是有95%的信心, 即对这个方法, 不是对单个区间**

t检验的p<0.05,说明零假设成立的概率小于5% **仅仅一次不能说概率**

t检验的p<0.001,说明对照组和实验组来自的均值差异较大 **统计显著性不是实际显著性**

Pearson 相关分析, 得到r=0.8,p<0.05,说明两个变量存在强相关性

Pearson 相关分析, 得到r=0.05,p>0.5,说明两个变量关联性很弱, 或者不存在关联性 **没有线性关系**

简单线性模型y=1+1.5x,说明如果自变量x增加1个单位, 因变量y增加1.5

简单线性模型y=1+1.5x,说明如果自变量x增加1个单位, 因变量y均值增加1.5 **增加都是指同一个观测对象**

如果对照组和实验组在干预前不具有显著差异($p > 0.05$), 在干预后具有显著性差异($p < 0.01$), 可以推断干预有效 **要比差异的差异, 见6.3.1**

血压研究实验, 把50只动物随机分配到对照组和实验组, 每组25只动物, 实验完成后想知道对照组($n=25$)动物血压是否高于实验组动物($n=25$)的血压均值, 可以用单边t-检验 **这是检验总体是否高于的方法, 如果是对照组和实验组, 则是样本, 直接算就行**

实验分成三组(对照组, 低剂量组, 高剂量组), 统计检验发现对照组和低剂量组差异不具有统计显著性($p=0.50$), 对照组与高剂量组差异具有统计显著性($p=0.01$), 可以推断高剂量组比低剂量组干预有效 **直接比较p不可取**

t-test, ANOVA都可以用线性模型来完成

7.2 第六章

如果一个线性模型得到的R Squared 值为0.95, 说明该模型的正确的可能性很大
错误, 单单R2无法说明问题, 要综合R2和t检验和其他参数

简单线性回归模型的置信区间带 (CIB) 比预测区间带 (PIB) 要窄, 并且这两个带在**自变量均值处都最窄**

正确

多元线性回归模型, 某自变量的系数如果不显著 (比如 $p=0.4$), 说明该系数可以从模型中去掉
错误, 要综合考虑