大数据的统计学基础——第10周

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

关注炼数成金企业微信

■提供全面的数据价值资讯，涵盖商业智能与数据分析、大数据、企业信息化、数字化技术等，各种高性价比课程信息，赶紧掏出您的手机关注吧！

假设检验

- 假设检验(hypothesis test/test of significance)
- 零假设(null hypothesis)
- 备择假设(alternative hypothesis)
- 检验统计量(test statistic)
- 拒绝域(critical region)
- 临界值(critical value)
- p-值(p-value)
- 双侧假设(two-tailed test)
- 单侧假设(left-tailed test/right-tailed test)
- 第一类错误(type Ⅰ error)
- 第二类错误(type Ⅱ error)

假设检验

- 假设:
- 一份报纸头条声称大部分求职者通过网络找到工作
- 医学调查工作者说健康成年人的平均体温并不等于98.6 °F
- 某工厂引进了一台新的设备，称新设备生产的零件误差比旧设备小
- 某航空公司说一位飞机乘客（包括他的行李）的平均重量超过185lb已经是20年前的数据了，与现今数据不符。

- 如何证明上述论断是否正确？

统计推断中的罕见事件规则

- Rare event rule of inferential statistics
- If, under a given assumption, the probability of a particular observed event is exceptionally small, we conclude that the assumption is probably not correct.

- 在一个已知的假设下，如果一个特定事件发生的概率格外小，那么我们认为，这个假设可能不对。

- 实际推断原理：概率很小的事件在一次试验中实际上几乎是不发生的

Gender Selection

- 某药厂曾经推出过一款名为“Gender Choice”的药，该要的宣传广告中声称这药可以让夫妻怀上女孩的几率上升到80%。假设我们现在有100对想要生女儿的夫妻，他们都服用了这款“Gender Choice”。如果我们假设这款药是没有效果的，吃了这个药对他们生男孩还是女孩没有影响。那么基于以上的假设，这100对夫妻生了多少名女孩才符合假设内容（假设一对夫妻只生一名孩子）？
- a.52个女孩
- b.97个女孩
- 一般情况下，生女孩的概率为50%。100对夫妻中理论上会生50个女孩。如果是52个的话，与50比较接近，发生的概率还是比较大的。但是如果是97个的话，这种情况发生的概率很小，如果它真的发生了，那么我们就不得不考虑前提假设是否有误。

Gender Selection and Probability

- 要靠数据说话
- 假设服用“Gender Choice”没有效果，也就是生女孩的概率还是p=0.5。我们来计算100对夫妻生了52名女孩的概率是大还是小。
- 我们的假设是p=0.5；该药厂的假设是p>0.5
- 设X为100对夫妻中生育女孩的个数，则在我们的假设下，X~B(100,0.5)
- 现在我们考虑P(X>=52)=$C_{100}^{52}0.5^{52}0.5^{48} = 0.3821$（用正态分布近似计算）

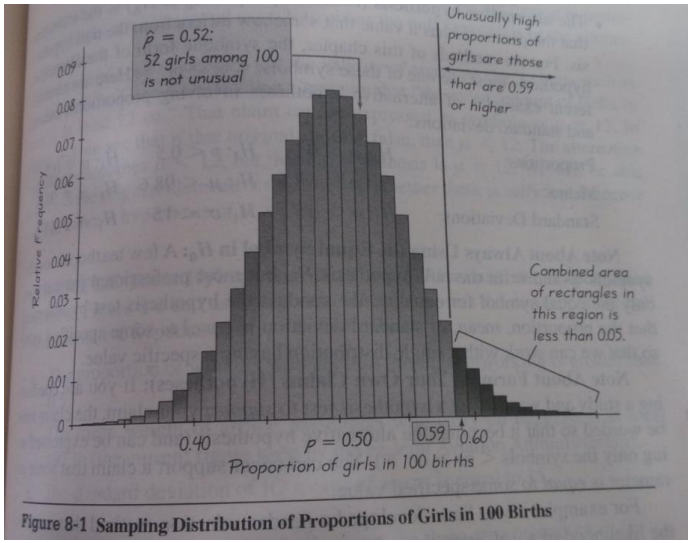


Figure 8-1 Sampling Distribution of Proportions of Girls in 100 Births

零假设与备择假设

- null hypothesis is a statement that the value of a population parameter is equal to some claimed value.
- 零假设是假定一个总体参数等于某个特定值的一个声明，用H_0表示。如$H_0: p = 0.5$; H_0:μ=98.6; H_0:σ=15
- 备择假设是假定该总体参数为零假设中假设的值除外的值，用H_1表示。如$H_1: p > 0.5; H_1: p < 0.5; H_1: p \neq 0.5$

- 零假设与备择假设的选择：
- 若希望假设的论断成立，设为备择假设；若希望假设的论断不成立，设为零假设

- 例：可口可乐公司声称：每罐可乐的平均重量最小为12安士

零假设与备择假设

- 对于以下假设：
- 1.通过网站找到工作的求职者比例超过0.5
- 2.飞机乘客的平均重量（包括乘客的手提行李）最多为195磅
- 3.演员的IQ成绩的标准差为15
- 分别写出它们的零假设与备择假设

检验统计量

- The test statistic is a value used in making a decision about the null hypothesis , and it is found by converting the sample statistic to a score with the assumption that the null hypothesis is true.
- 检验统计量是一个用于确定零假设是否为真的一个值，这个值在假定零假设为真时由样本数据计算得到的。

- 如： 检验总体比例的检验统计量：$z = \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}}$

拒绝域与显著性水平

- 拒绝域，也称否定域，是指检验统计量所有可以拒绝零假设的取值所构成的集合。

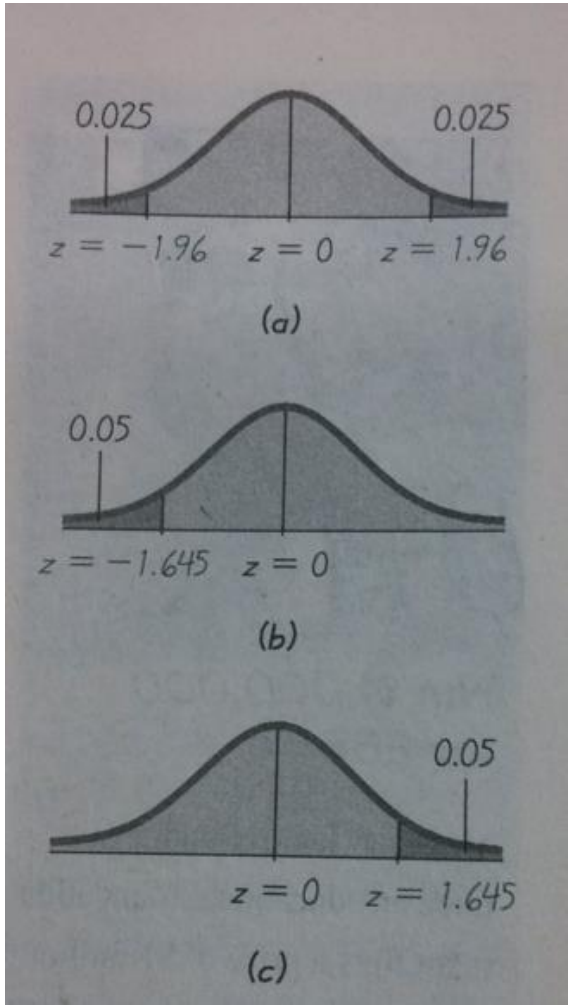
- 显著性水平，α，指当零假设正确的时候，检验统计量落在拒绝域的概率。也就是当零假设为真而我们却拒绝零假设这种错误发生的概率。与置信区间中的显著性水平α意义一致。常用取值：0.1,0.05,0.01

- 临界值，拒绝域与非拒绝域的分界线

双侧检验与单侧检验

- 双侧检验：拒绝域落在分布两边的尾部
- 左侧检验：拒绝域落在分布左边的尾部
- 右侧检验：拒绝域落在分布右边的尾部

- The p-value is the probability of getting a value of the test statistic that is at least as extreme as the one representing the sample data,assuming that the null hypothesis is true.
- P-值：当原假设为真时所得到的样本观察结果或更极端结果出现的概率
- 当p-值足够小时，即小于置信水平时，我们可以拒绝零假设。



决定规则

方法	拒绝零假设	不拒绝零假设
临界值法	检验统计量落在拒绝域	检验统计量没有落在拒绝域
P-value法	P-value<=α	P-value>α
另一个选择	不采用具体的α值，写出p-value留给读者自己判断	
置信区间	置信区间中没有包含零假设的参数值	置信区间中有包含零假设的参数值

第一类错误与第二类错误

	真实情况	
	零假设为真	零假设为假
拒绝零假设	第一类错误，α	正确决定
不拒绝零假设	正确决定	第二类错误，β

- 第一类错误，零假设正确的情况下拒绝了零假设，犯错概率：α
- 第二类错误，零假设错误的情况下没有拒绝零假设，犯错概率：β

总体比例的假设检验

- 适用条件：
- 1. 样本是一个简单随机样本
- 2. 满足二项分布的条件
- 3. 满足np≥5，nq≥5（满足这样的条件的话，二项分布可以使用$N(\mu = np, \sigma^2 = npq)$来近似）
- n：试验次数
- $\hat{p} = \frac{x}{n}$：样本比例
- p：总体比例（在零假设中使用）
- q：1-p

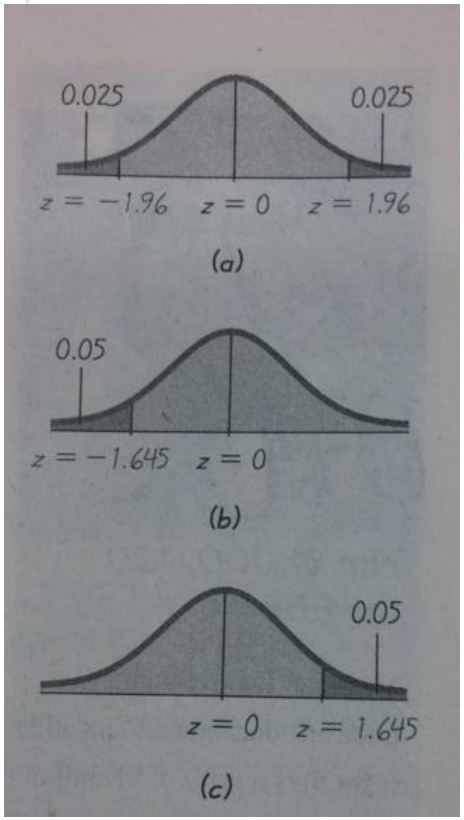
- 检验统计量：$z = \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}} \sim N(0,1)$

总体比例的假设检验

定理三(棣莫弗－拉普拉斯(De Moivre-Laplace)定理) 设随机变量 η_n ($n=1,2,\cdots$)服从参数为 n,p ($0<p<1$)的二项分布,则对于任意 x,有

$$\lim_{n\to\infty} P\left\{\frac{\eta_n - np}{\sqrt{np(1-p)}} \leqslant x\right\} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \Phi(x). \qquad (2.5)$$

- 根据中心极限定理，$z = \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}} \sim N(0,1)$
- 当检验统计量z的值落在正态分布的尾部时，该样本发生的概率非常小，这种情况下，我们认为零假设有误，从而拒绝零假设



例子——找工作

- 根据某项调查的数据，随机抽取的703名工作者中，有61%的人通过网络得到他们的工作。根据样本数据，设置信水平α=0.05，判断以下论断：超过50%的人通过网络找到他们的工作。
- 1. 先确立是否适合总体比例的假设检验的条件：
- a. 样本是一个简单随机样本
- b. 满足二项分布的条件
- c. 满足np≥5，nq≥5（满足这样的条件的话，二项分布可以使用$N(\mu = np, \sigma^2 = npq)$来近似）
- 3个条件都满足

例子——找工作

- 2. 设定零假设与备择假设

$$H_0：p = 0.5，H_1：p > 0.5$$

- 3. 计算检验统计量的值

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.61 - 0.5}{\sqrt{\frac{0.5 * 0.5}{703}}} = 5.83$$

- 4. 作出判断
- a.临界值法：找出拒绝域，先找出临界值：$z_{0.95} = 1.645$，从而拒绝域为(1.645, ∞)。5.83落在拒绝域中，故拒绝零假设
- b. p-值法：P{Z>=z=5.83}=0.0001<0.05，故拒绝零假设
- 5. 最终结论
- 根据上述结果，我们拒绝零假设，即认为超过50%的人通过网络找到他们的工作。

总体均值的假设检验（σ 已知）

- 适用条件：
- 1. 样本是简单随机样本
- 2. 总体标准差σ已知
- 3. 总体是正态分布或n>30

- 检验统计量：$z = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{\sigma}{\sqrt{n}}}$

- 根据中心极限定理，我们知道$z = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$

例子——M&M豆

- 下表显示的是随机从一包有465颗M&M豆中抽取的13颗红色M&M豆的重量。

13颗红色M&M豆的重量（单位：g）						
0.751	0.841	0.856	0.799	0.966	0.859	0.857
0.942	0.873	0.809	0.890	0.878	0.905	

- 所有M&M豆重量的标准差为σ=0.0565g。上述13颗红色M&M豆的平均重量$\bar{x} = 0.8635g$。而M&M的包装上写着净含量：396.9g。即平均每颗M&M豆重396.9/465=0.8535g。设置信水平α=0.05，根据样本数据，检验下面的说法是否正确：M&M豆每颗的平均重量大于0.8535g，也就是说消费者拿到的比包装上所写的多。假设M&M豆的重量服从正态分布。

例子——M&M豆

- 1. 适用条件都满足
- 2. 确定零假设与备择假设
- $H_0：\mu = 0.8535$ v.s $H_1：\mu > 0.8535$
- 3. 计算检验统计量的值

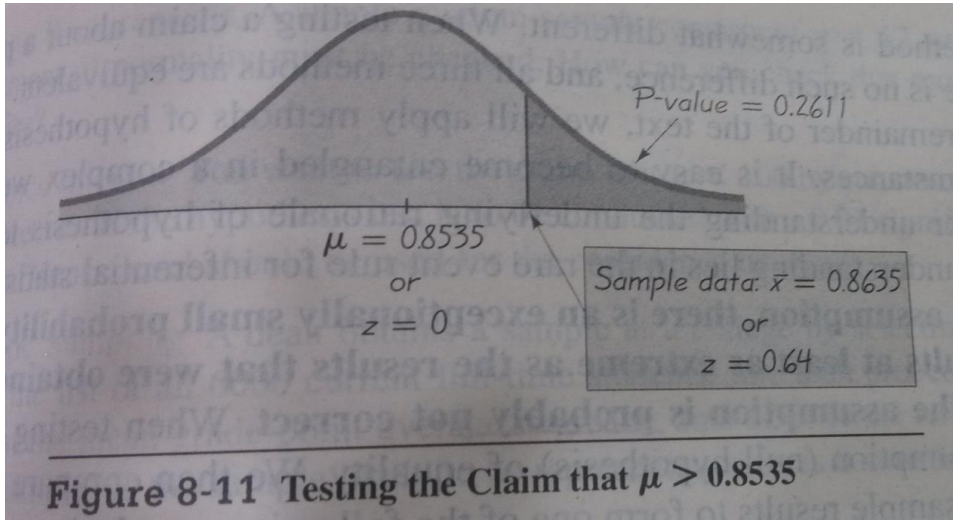


Figure 8-11 Testing the Claim that $\mu > 0.8535$

- $z = \frac{\bar{x}-\mu_{\bar{x}}}{\frac{\sigma}{\sqrt{n}}} = \frac{0.8635-0.8535}{\frac{0.0565}{\sqrt{13}}} = 0.64$
- 4. 作出判断
- a. 临界值法：右侧检验，临界值为$z_{0.95} = 1.645 > 0.64$，即检验统计量没有落入拒绝域，从而不能拒绝零假设
- b. p-值法：P{Z>z=0.64}=1-0.7389=0.2611>α=0.05，故不能拒绝零假设
- 5. 最终结论
- 不能拒绝零假设，也就是我们不接受“M&M豆每颗的平均重量大于0.8535g”这个说法

总体均值的假设检验（σ 未知）

- 适用条件：
- 1. 样本是简单随机样本
- 2. 总体标准差σ 未知
- 3. 总体为正态分或n>30

- 检验统计量：$t = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{s}{\sqrt{n}}}$

- 根据t分布的定义，我们知道$t = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{s}{\sqrt{n}}} \sim t(n-1)$

例子——M&M豆

- 下表显示的是随机从一包有465颗M&M豆中抽取的13颗红色M&M豆的重量。

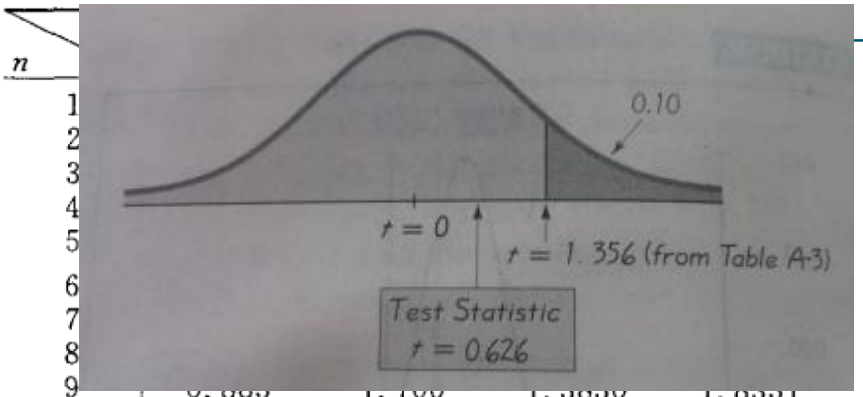
13颗红色M&M豆的重量（单位：g）						
0.751	0.841	0.856	0.799	0.966	0.859	0.857
0.942	0.873	0.809	0.890	0.878	0.905	

- 上述13颗红色M&M豆的平均重量$\bar{x} = 0.8635g$。而M&M的包装上写着净含量：396.9g。即平均每颗M&M豆重396.9/465=0.8535g。设置信水平α=0.05，根据样本数据，检验下面的说法是否正确：M&M豆每颗的平均重量大于0.8535g，也就是说消费者拿到的比包装上所写的多。假设M&M豆的重量服从正态分布。
- 与上题最大的分别在于这题没有给总体标准差。

例子——M&M豆

- 1. 适用条件都满足
- 2. 确定零假设与备择假设
- $H_0：\mu = 0.8535$ v.s $H_1：\mu > 0.8535$
- 3. 计算检验统计量的值
- $t = \frac{\bar{x}-\mu_{\bar{x}}}{\frac{s}{\sqrt{n}}} = \frac{0.8635-0.8535}{\frac{0.0576}{\sqrt{13}}} = 0.626$
- 4. 作出判断

$P\{t(n) > t_\alpha(n)\} = \alpha$



n				
10	0.879	1.093	1.3722	1.8125
11	0.876	1.088	1.3634	1.7959
12	0.873	1.083	1.3562	1.7823
13	0.870	1.079	1.3502	1.7709
14	0.868	1.076	1.3450	1.7613
15	0.866	1.074	1.3406	1.7531

- a. 临界值法：右侧检验，临界值为$t_{0.05,12} = 1.782 > 0.626$，即检验统计量没有落入拒绝域，从而不能拒绝零假设
- b. p-值法：P{Z>z=0.64}=1-0.7285=0.2715>α=0.05，故不能拒绝零假设
- 5. 最终结论
- 不能拒绝零假设，也就是我们不接受“M&M豆每颗的平均重量大于0.8535g”这个说法

总体方差的假设检验

- 适用条件：
- 1. 样本是简单随机样本
- 2. 总体必须是正态分布

- 检验统计量：$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

例子——零部件长度

- 在正常生产条件下，某种零部件的长度服从正态分布，标准差不得超过 0.13 厘米。现从一批准备出厂的零部件中，随机抽取20件，测得标准差为 0.16 厘米。试问：在 α=0.05的显著性水平下，能否得出这批零部件不合格的结论。
- 1. 条件符合
- 2. 确定零假设和备择假设

$$H_0：\sigma^2 = 0.13^2 \text{ v.s } H_1: \sigma^2 > 0.13^2$$

- 3. 计算检验统计量的值

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{19 * 0.16^2}{0.13^2} = 28.78$$

例子——零部件长度

- 4. 作出判断
- a. 临界值法：临界值为30.143>28.78，所以检验统计量的值没有落入拒绝域中，从而不能拒绝零假设
- b. p-值法：P{χ^2>28.78}=1-0.9305=0.0695>0.05，故不能拒绝零假设
- 5.最终结论
- 不能拒绝零假设，也就是我们不能得出“这批零部件不合格”的结论

n \ α	0.995	0.99	0.975	0.95	0.90	0.10	0.05
1	0.000	0.000	0.001	0.004	0.016	2.706	3.843
2	0.010	0.020	0.051	0.103	0.211	4.605	5.992
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685
15	4.600	5.229	6.262	7.261	8.547	22.307	24.996
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296
17	5.697	6.407	7.564	8.682	10.085	24.769	27.587
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869
19	6.843	7.632	8.906	10.117	11.651	27.203	30.143
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410

两个独立总体比例的比较

- 适用条件：
- 1. 两个样本为相互独立的简单随机样本
- 2. 对于每个样本，成功个数和失败个数都要大于等于5

- 混合样本比例/聚合估计值(pooled sample proportion)
- $\bar{p} = \frac{x_1+x_2}{n_1+n_2}$
- 检验两个总体比例是否相等

- 检验统计量：$z = \frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\frac{\bar{p}\bar{q}}{n_1}+\frac{\bar{p}\bar{q}}{n_2}}} \sim N(0,1)$

例子——腕隧道症候群

- 对于腕隧道症候群，现有两种治疗方法可供选择：使用夹板或进行手术。

	手术	夹板
经治疗一年后痊愈人数	67	60
总治疗人数	73	83
成功率	92%	72%

- 在 α=0.05的显著性水平下，利用以上数据能否说明进行手术治疗的成功率比使用夹板治疗高。
- 1. 条件符合
- 2. 确定零假设和备择假设

$$H_0: p_1 = p_2 \ \ v.s \ \ \ H_1: p_1 > p_2$$

- 3.计算检验统计量的值

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{67 + 60}{73 + 83} = 0.81410256$$

例子——腕隧道症候群

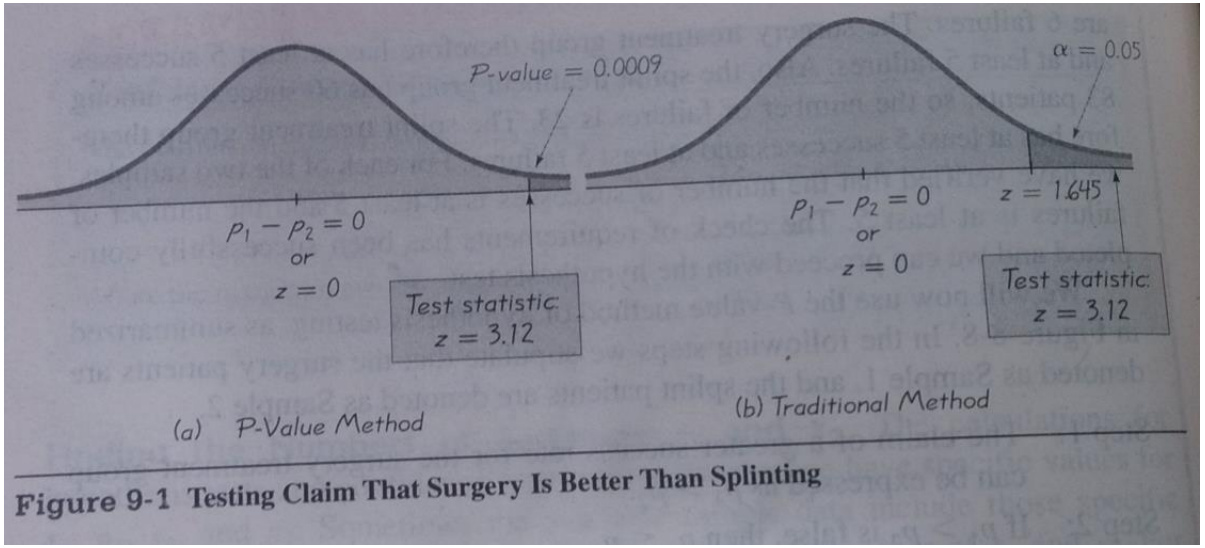


Figure 9-1 Testing Claim That Surgery Is Better Than Splinting

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}} = \frac{\frac{67}{73} - \frac{60}{83} - 0}{\sqrt{\frac{0.81410256 * 0.18589744}{73} + \frac{0.81410256 * 0.18589744}{83}}}$$

$$= 3.12$$

- 4. 作出判断
- a. 临界值法：$z_{0.95} = 1.645 < 3.12$，故检验统计量落在拒绝域中，从而拒绝零假设
- b. p-值法：P{Z>z=3.12}=0.0009<0.05，故拒绝零假设

两个独立总体均值的比较

- 情况一：两个独立样本的总体标准差σ_1、σ_2已知
- 适用条件：
- 1. 两个独立样本的总体标准差σ_1、σ_2已知
- 2. 两个样本相互独立
- 3. 两个样本都是简单随机样本
- 4. 两个样本的总体服从正态分布或是样本容量$n_1 > 30$，$n_2 > 30$

- 检验统计量：$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}}} \sim N(0,1)$

例子——城市农村成绩差异

- 某地区历年高考成绩的统计资料显示，城市考生的标准差为50分，农村考生的标准差为55分。现从城市考生中随机抽取32人组成一个样本，测得平均成绩为515分；从农村考生中随机抽取40人成一个样本，测得平均成绩为545分。假设高考成绩服从正态分布，在显著性水平α=0.05下，试问：城市考生与农村考生的高考成绩是否有显著差别。
- 设X_1、X_2分别表示城市考生的成绩和农村考生的成绩，根据题意有
- $\bar{x}_1 = 515$，$n_1 = 32$，$\sigma_1 = 50; \bar{x}_2 = 545$，$n_1 = 40$，$\sigma_1 = 55$
- 1. 条件符合情况一
- 2. 确定零假设和备择假设

$$H_0: \mu_1 = \mu_2; H_1: \mu_1 \neq \mu_2$$

例子——城市农村成绩差异

- 3.计算检验统计量的值

$$z=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_1^2}{n_2}}}=\frac{515-545-0}{\sqrt{\frac{50^2}{32}+\frac{55^2}{40}}}=-2.42$$

- 4. 作出判断
- a. 临界值法： α=0.05，对应的临界值$z_{0.975}=1.96<|-2.42|$,故检验统计量落在拒绝域中，从而拒绝零假设
- b. p-值法：P{Z<-2.42}=P{Z>2.42}=1-P{Z<=2.42}=1-0.9922=0.0078<0.025
- 或P{|Z|>|-2.42|}= 2*P{Z>2.42}=2(1-P{Z<=2.42})=2（1-0.9922）=0.0156<0.025 ，故拒绝零假设
- 5.最终结论
- 表明城市考生与农村考生的高考成绩有显著差别。

两个独立总体均值的比较

- 情况二：两个独立样本的总体标准差σ_1、σ_2未知，但认为$\sigma_1 = \sigma_2$
- 适用条件：
- 1.两个独立样本的总体标准差σ_1、σ_2未知，但认为$\sigma_1 = \sigma_2$
- 2. 两个样本相互独立
- 3. 两个样本都是简单随机样本
- 4. 两个样本的总体服从正态分布或是样本容量$n_1 > 30$，$n_2 > 30$

- 检验统计量：$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \sim t(n_1 + n_2 - 2)$，其中$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$
- S_p^2称为合并方差（pooled variance）

例子——汽车耗油量

- 为测试两种型号载重汽车每百公里的油耗，分别随机抽取10辆车进行检测，记录其数据如下(单位：公升/每百公里)：
- A型：11.9 12.3 11.7 12.0 12.8 11.6 12.5 11.9 12.8 12.4
- B型：11.6 12.0 12.4 11.8 12.4 12.8 11.6 12.6 11.9 11.7
- 假设，汽车每百公里的油耗服从正态分布且方差大致相等，在显著性水平α=0.05下，试检验两种型号载重汽车的油耗是否存在显著差别。
- 设X_1、X_2分别表示A型汽车耗油量和B型汽车耗油量，根据题意有
- $\bar{x}_1 = 12.19$，$n_1 = 10$，$s_1^2 = 0.1877; \bar{x}_2 = 12.08$，$n_1 = 10$，$s_2^2 = 0.1907$
- 1. 条件符合情况二
- 2. 确定零假设和备择假设

$$H_0: \mu_1 = \mu_2; H_1: \mu_1 \neq \mu_2$$

例子——汽车耗油量

- 3.计算检验统计量的值

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)} = 0.1892$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{12.19 - 12.08 - 0}{\sqrt{\frac{0.1892}{10} + \frac{0.1892}{10}}} = 0.5655$$

- 4. 作出判断
- a. 临界值法： α=0.05，对应的临界值$t_{0.025,18} = 2.1009 > |0.5655|$,故检验统计量没有落在拒绝域中，从而不能拒绝零假设
- b. p-值法：P{t_{18}>0.5655}=0.2893577>0.025，或P{|t_{18}|>0.5655}=2*0.2893577>0.05故不能拒绝零假设
- 5.最终结论
- 认为两种型号载重汽车的油耗不存在显著差别

两个独立总体均值的比较

- 情况三：两个独立样本的总体标准差σ_1、σ_2未知且不相等。
- 适用条件：
- 1. 两个样本的总体标准差σ_1、σ_2未知且不相等。
- 2. 两个样本相互独立
- 3. 两个样本都是简单随机样本
- 4. 两个样本的总体服从正态分布或是样本容量$n_1 > 30$，$n_2 > 30$

- 检验统计量：$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(v), v = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(s_1^2/n_1)^2}{n_1} + \frac{(s_2^2/n_2)^2}{n_2}}$
- 简化计算，v可以近似取$\min(n_1 - 1, n_2 - 1)$

例子——年龄歧视

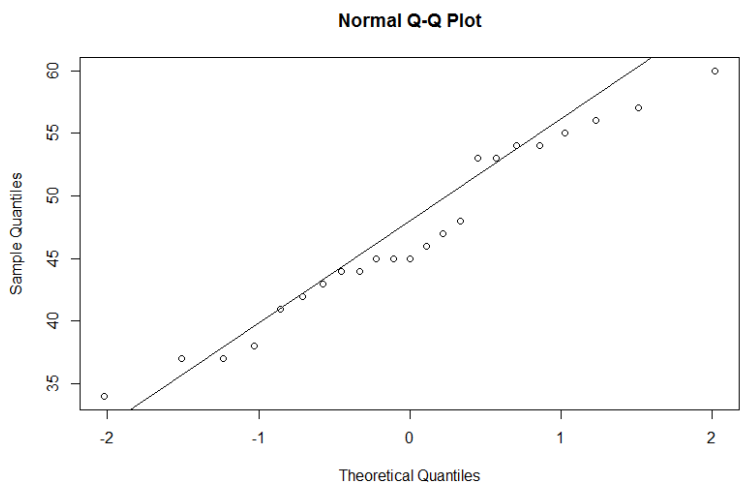
- 某公司的某一次内部提升中，不成功提升者与成功提升者的年龄分别如下：
- 不成功：34,37,37,38,41,42,43,44,44,45,45,45,46,48,47,53,53,54,54,55,56,57,60
- 成功：
  27,33,36,37,38,38,39,42,42,43,43,44,44,44,45,45,45,45,46,46,47,47,48,48,49,49,51,51,52,54
- 某些申请提升者投诉该公司对申请者有年龄歧视。把上述数据看做是来自总体容量较大总体的一个样本。设显著性水平α=0.05，请判断失败的申请者的平均年龄是否比成功的申请者平均年龄要大，并由此该公司是否有年龄歧视。
- 1. 条件检查
- （1）两个样本的总体标准差σ_1、σ_2未知，可以认为它们不相等
- （2）两个样本独立
- （3）两个样本都是简单随机样本

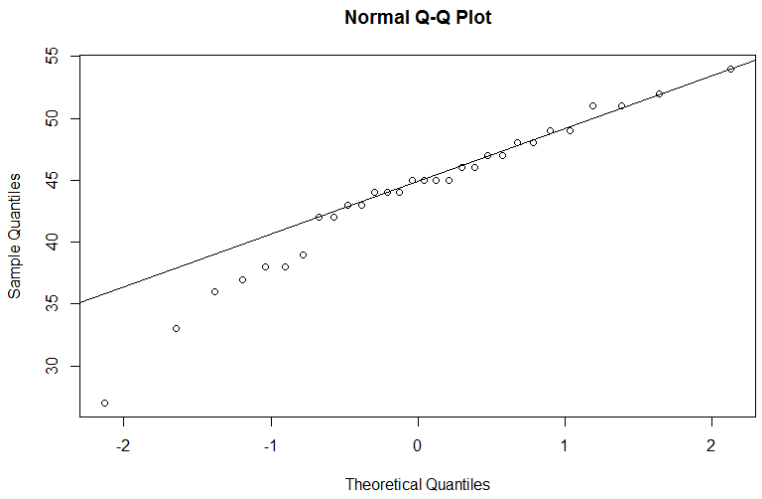
例子——年龄歧视

- （4）数据的正态性

Unsuccessful	Successful
$n = 23$	$n = 30$
$\bar{x} = 47.0$	$\bar{x} = 43.9$
$s = 7.2$	$s = 5.9$

Unsuccessful		Successful
	2	7
4	3	3
877	3	67889
44321	4	2233444
986555	4	555566778899
4433	5	1124
765	5	
0	6	





例子——年龄歧视

- 2. 确定零假设和备择假设

$$H_0: \mu_1 = \mu_2; H_1: \mu_1 > \mu_2$$

- 3.计算检验统计量的值

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(47.0 - 43.9) - 0}{\sqrt{\frac{7.2^2}{23} + \frac{5.9^2}{30}}} = 1.678$$

- 自由度计算：$v = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(s_1^2/n_1)^2}{n_1} + \frac{(s_2^2/n_2)^2}{n_2}} = 42.03 \approx 42$

例子——年龄歧视

- 4. 作出判断
- a. 临界值法： α=0.05，对应的临界值是$t_{0.05,42} = 1.6820 > 1.678$，故检验统计量没有落入拒绝域，不能拒绝零假设
- b. p-值法：P{$t_{42} > 1.678$}=0.05038718>0.05。故不能拒绝零假设。
- 5. 最终结论
- 从p-值法来看，当取α=0.05时，我们不能拒绝零假设，也就是认为失败的申请者的平均年龄是否比成功的申请者平均年龄没有显著性差异。
- 但是，如果我们将α取得更小一点，例如0.01，那么这时我们就可以拒绝零假设，认为失败的申请者的平均年龄是否比成功的申请者平均年龄大，也就是这家公司在进行内部提升的时候，是带有年龄歧视的。

配对样本（matched pairs）t检验

- 同源配对
- 自身配对

- 适用条件：
- 1. 两个样本数据时配对样本
- 2. 样本是简单随机样本
- 3. 配对样本数据量足够大（大于30对数据）或配对样本数据差值的总体服从或近似服从正态分布

配对样本（matched pairs）t检验

- d：配对数据的差值
- μ_d：差值的总体均值
- $\bar{d}$：样本差值的均值
- S_d：样本差值的标准差
- n：配对数据的对数

- 检验统计量：$t = \frac{\bar{d}-\mu_d}{\frac{s_d}{\sqrt{n}}} \sim t(n-1)$

例子

- 下表记录了某5天实际的最低温度与预测的最低温度

实际与预测的最低温度（华氏度）					
实际的最低温度	54	54	55	60	64
预测的最低温度	56	57	59	56	64
实际与预测的温度差d	-2	-3	-4	4	0

- 设α=0.05。根据上述数据，是否能充分说明实际的最低温度与预测的最低温度间有显著性差异。
- 属于配对样本检验，因为温度是同一天的数据。
- 设定假设：$H_0: \mu_d = 0$ v.s $H_1: \mu_d \neq 0$
- 检验统计量：$t = \frac{\bar{d}-\mu_d}{\frac{s_d}{\sqrt{n}}} = -0.699$；p-value：P(|t|>0.699)=2P(t<-0.699)>0.2>0.05
- 临界值：t=±2.776，不能拒绝零假设
- 故认为实际的最低温度与预测的最低温度间没有显著性差异

两个独立总体方差的比较

- 适用条件：
- 1. 两个总体相互独立
- 2. 两个总体服从正态分布
- 3. 样本是简单随机样本

- 假设两总体方差相等
- 检验统计量：$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$，其中，$S_1^2$是两样本方差中较大者
- 设X1，X2，……Xn1与Y1,Y2,……，Yn2分别来自正态总体$N(\mu_1, {\sigma_1}^2)$，$N(\mu_2, {\sigma_2}^2)$的样本，且这两个样本相互独立。其样本均值分别为$\bar{X}$，$\bar{Y}$，其方差分别为${S_1}^2$，${S_2}^2$,则有
- (1)$\frac{{S_1}^2/{S_2}^2}{{\sigma_1}^2/{\sigma_2}^2} \sim F(n_1 - 1, n_2 - 1)$

例子

- 教育考试中心进行了一项学生的性别对学生能力测试分数的方差是否存在显著差异的研究。研究人员随机抽取了72名学生的数据，其中41名女生测试分数的标准差为 15.3 分，31名男生测试分数的标准差为 9.6 分。假设学生能力测试成绩服从正态分布，在显著性水平 α=0.05下，试问：这些样本数据是否表明女生测试分数的标准差比男生大。
- 设X_1, X_2分别表示女生与男生能力测试的分数，则根据题意有
- $X_1 \sim N(\mu_1, {\sigma_1}^2), n_1 = 41, S_1 = 15.3$分
- $X_2 \sim N(\mu_2, {\sigma_2}^2), n_2 = 31, S_2 = 9.6$分
- 提出的假设是
- $H_0: {\sigma_1}^2 = {\sigma_2}^2; H_1: {\sigma_1}^2 > {\sigma_2}^2$
- 这是一个右侧检验问题。

例子

检验的统计量为$\mathrm{F}=\frac{{S_1}^2}{{S_2}^2}$

规定的显著性水平为 $\alpha=0.05$ ，查表得临界值$F_{0.05,40,30}=1.79$ ，原假设 H_0 的否定域为

$$V_2=\left\{F\geq F_\alpha(n_1-1,n_2-1)\right\}=\left\{F\geq 1.79\right\}$$

计算检验统计量 F 的值

$$F=\frac{S_1^2}{S_2^2}=\frac{15.3^2}{9.6^2}=2.54$$

因为 $F_{0.05,40,30}=1.79$，落在否定域里，所以否定 H_0 ，表明女生能力测试分数的标准差要显著地大于男生。

炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

Thanks

FAQ时间