大数据的统计学基础——第11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！

本周内容

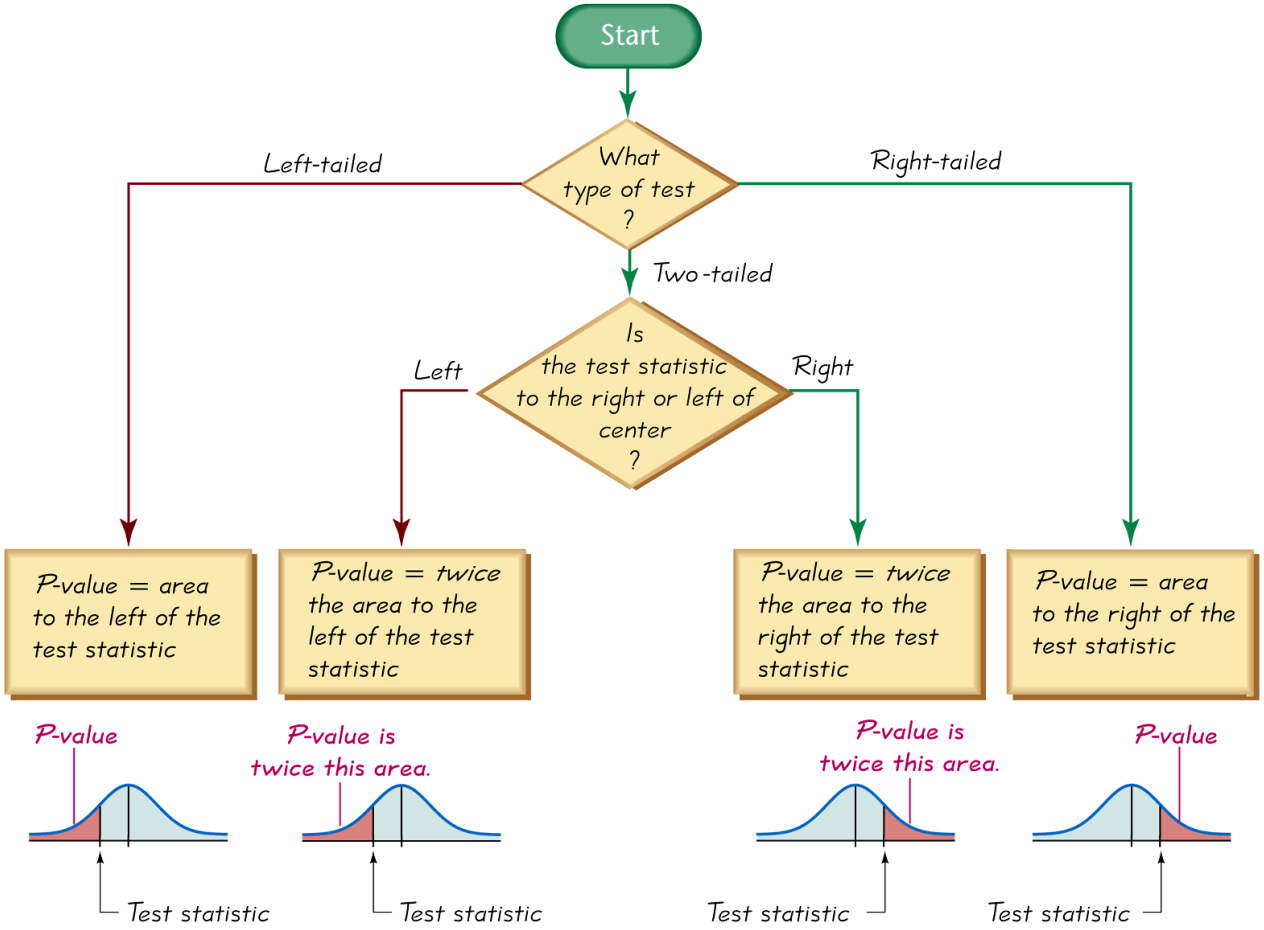
- 假设检验：
- 假设检验与置信区间
- 假设检验的检验能力(power)

- 非参数方法(Nonparametric methods)：
- 符号检验（sign test）
- Wilcoxon符号检验(Wilcoxon signed-rank test)
- Wilcoxon秩和检验(Wilcoxon rank Sum test)

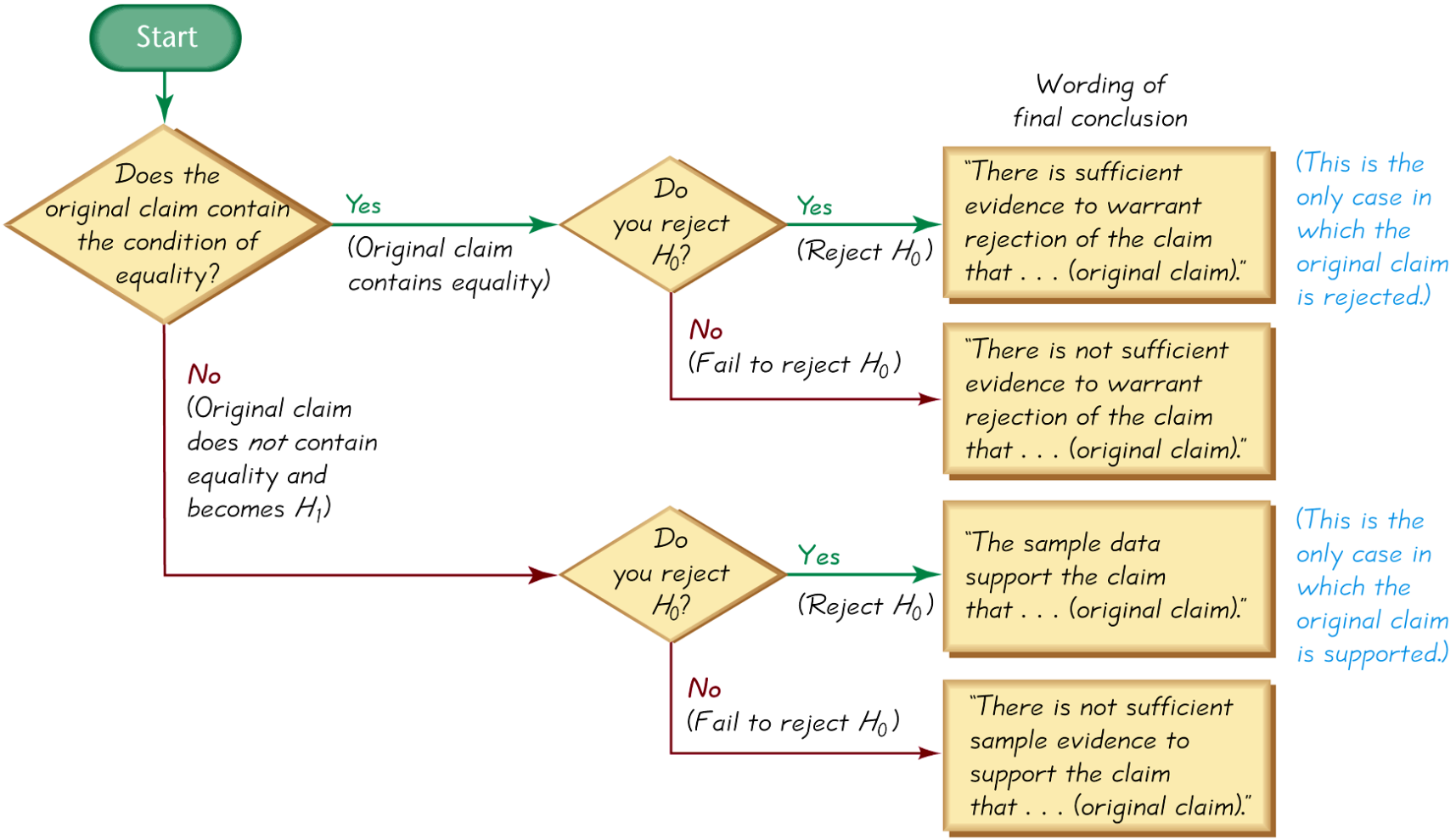
- 补充内容：
- 正态分布近似的连续项修正（Continuity Correction）

复习

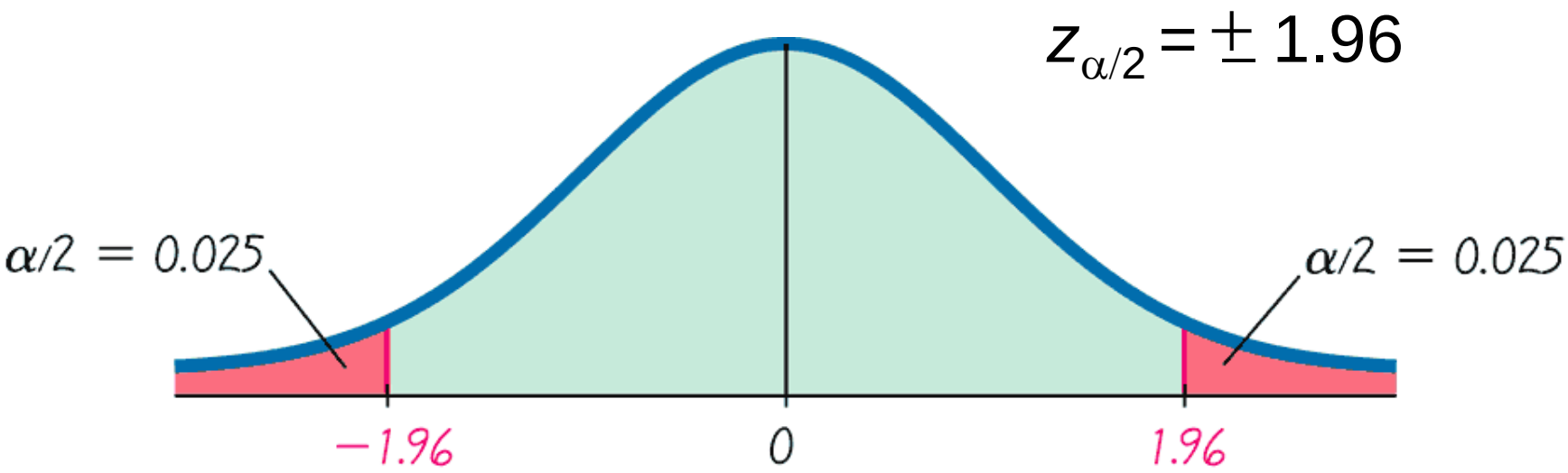
- 假设检验


Start
What type of test ?
Left-tailed
Right-tailed
Two-tailed
Is the test statistic to the right or left of center ?
Left
Right
P-value = area to the left of the test statistic
P-value = twice the area to the left of the test statistic
P-value = twice the area to the right of the test statistic
P-value = area to the right of the test statistic
P-value
Test statistic
P-value is twice this area.
Test statistic
P-value is twice this area.
Test statistic
P-value
Test statistic

复习


Start
Does the original claim contain the condition of equality?
Yes
(Original claim contains equality)
Do you reject H_0?
Yes
(Reject H_0)
No
(Fail to reject H_0)
No
(Original claim does not contain equality and becomes H_1)
Do you reject H_0?
Yes
(Reject H_0)
No
(Fail to reject H_0)
Wording of final conclusion
"There is sufficient evidence to warrant rejection of the claim that . . . (original claim)."
(This is the only case in which the original claim is rejected.)
"There is not sufficient evidence to warrant rejection of the claim that . . . (original claim)."
"The sample data support the claim that . . . (original claim)."
(This is the only case in which the original claim is supported.)
"There is not sufficient sample evidence to support the claim that . . . (original claim)."

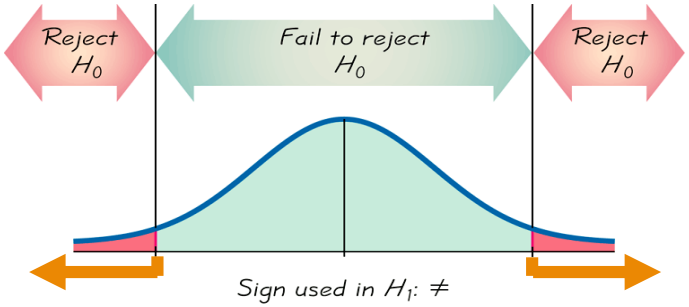
复习



假设检验与置信区间

- 关于总体均值的假设检验与置信区间， σ 未知的情况
- 假设检验：$H_0：\mu = \mu_0$ v.s $H_1：\mu \neq \mu_0$

  检验统计量：$t = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$，临界值：$t_{\alpha/2,n-1}$



  拒绝域：$\left(\left|t = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}\right| > t_{\frac{\alpha}{2},n-1}\right) = \left(|\bar{x}-\mu_0| > t_{\frac{\alpha}{2},n-1} \times \frac{S}{\sqrt{n}}\right) = (\mu_0 > \bar{x} + t_{\frac{\alpha}{2},n-1} \times \frac{S}{\sqrt{n}}$或$\mu_0 < \bar{x} - t_{\frac{\alpha}{2},n-1} \times \frac{S}{\sqrt{n}})$

- 置信区间：$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$
- $1-\alpha = P\left\{\left|\frac{\bar{x}-\mu}{S/\sqrt{n}}\right| < t_{\alpha/2,n-1}\right\} = P\left\{\bar{x} - t_{\alpha/2,n-1} * \frac{S}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2,n-1} * \frac{S}{\sqrt{n}}\right\} = P\{\bar{x} - E < \mu < \bar{x} + E\}$,此时，$E = t_{\alpha/2,n-1} * \frac{S}{\sqrt{n}}$
- 故μ的$1-\alpha$的置信区间为：$(\bar{x} - t_{\alpha/2,n-1} * \frac{S}{\sqrt{n}}$ ，$\bar{x} + t_{\alpha/2,n-1} * \frac{S}{\sqrt{n}})$

假设检验与置信区间

- 其他情况
- 1. 单个总体
- （1）总体均值（ σ 已知）
- （2）总体比例
- （3）总体方差

- 2. 两个总体
- （1）两个独立总体的均值之差（三种情况）
- （2）两个配对样本的均值之差
- （3）两个独立总体比例之差
- （4）两个独立正态总体的方差之比

配对样本的置信区间

- 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

- $t = \frac{\bar{d}-\mu_d}{\frac{s_d}{\sqrt{n}}} \sim t(n-1)$

- $1-\alpha = P\left(\left|\frac{\bar{d}-\mu_d}{\frac{s_d}{\sqrt{n}}}\right| < t_{\alpha/2,n-1}\right) = P\left(\left|\bar{d}-\mu_d\right| < t_{\alpha/2,n-1} \times \frac{s_d}{\sqrt{n}}\right) = P\left(\bar{d} - t_{\alpha/2,n-1} \times \frac{s_d}{\sqrt{n}} <\right.$

例子——温度差

- 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，构造温度差的置信区间。
- $\bar{d} = -1.0$，$S_d = 3.2$，$n = 5$，$t_{\alpha/2,n-1} = 2.776$

总体比例的假设检验与置信空间

- 双侧置信区间
- $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$，所以 $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$
- $P\left\{\left|\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}\right| < z_{\frac{\alpha}{2}}\right\} = 1-\alpha$，故 $P\left\{\hat{p} - z_{\frac{\alpha}{2}} * \sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + z_{\frac{\alpha}{2}} * \sqrt{\frac{p(1-p)}{n}}\right\} = 1-\alpha$
- 由于p值的真实值不知道，一般采用 $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ 去代替 $\sqrt{\frac{p(1-p)}{n}}$
- 所有总体比例p的1-α置信区间为

$$(\hat{p}-E\ ,\ \hat{p}+E) = \left(\hat{p} - z_{\frac{\alpha}{2}} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\ \hat{p} + z_{\frac{\alpha}{2}} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

- 其中，$E = z_{\frac{\alpha}{2}} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

总体比例的假设检验与置信空间

- $H_0：p = p_0$ v.s $H_1：p \neq p_0$
- 检验统计量：$z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0 q_0}{n}}} \sim N(0,1)$；临界值：$z_{\frac{\alpha}{2}}$
- 接受域：$\left(\left|\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}\right| < z_{\frac{\alpha}{2}}\right) = \left(|\hat{p}-p_0| < z_{\frac{\alpha}{2}} \times \sqrt{\frac{p_0(1-p_0)}{n}}\right) = \left(\hat{p} - z_{\frac{\alpha}{2}} * \sqrt{\frac{p_0(1-p_0)}{n}} < p_0 <\right.$

例子

- 某项调查想分析手机号码中的最后一位数字，随机抽取了1000个手机号码，发现其中有119个号码最后一位数字为0.设请使用临界值法、p-值法、置信空间法去检验下面的说法是否正确：手机最后一位数字为0的概率是0.1.


```
> (0.119-0.1)/sqrt(0.1*0.9/1000)
[1] 2.002776
> 2*(1-pnorm(2.002776))
[1] 0.04520134
> 0.119-1.96*sqrt(0.119*0.881/1000)
[1] 0.09893138
> 0.119+1.96*sqrt(0.119*0.881/1000)
[1] 0.1390686
```

例子

- 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。

检验能力

- 第二类错误：β
- 检验能力：1-β
- 计算方法
- 例：某个假设检验的零假设、备择假设、显著性水平，样本容量如下
- $H_0: p = 0.5\ v.s\ H_1: p \neq 0.5\ ;\ \alpha = 0.05\ ; n = 100\ ;$样本比例：$\hat{p} = 0.57$

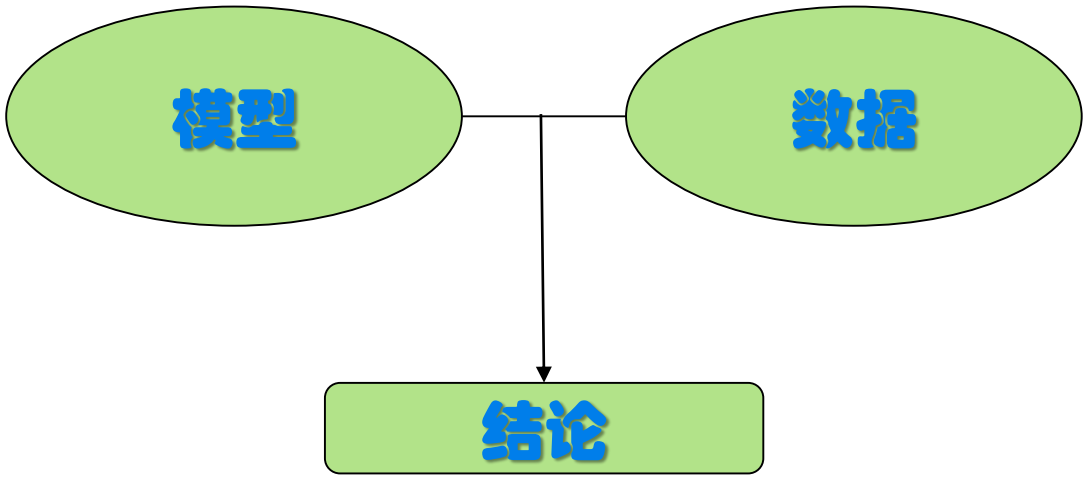
Specific Alternative Value of p	β	Power of Test $(1 - \beta)$
0.3	0.013	0.987
0.4	0.484	0.516
0.6	0.484	0.516
0.7	0.013	0.987

检验能力

- 提高检验能力的方法：
- 1. 增大样本容量
- 2. 提高显著性水平α
- 3. 采用与零假设相差更远的总体参数估计值
- 4. 降低标准差

非参数方法

- 什么是非参数方法？
- 非参数方法 v.s 任意分布检验（distribution-free test）
- 非参数方法 v.s 参数方法
- 非参数方法的好处与坏处



秩（Rank）

- 对n个数据按照从小到大的顺序排序，第一个数字秩为1，第二个数字秩为2，……，最大的数字秩为n
- 例：

7.3	6.1	2.4	12.3
3	2	1	4

结（Ties）

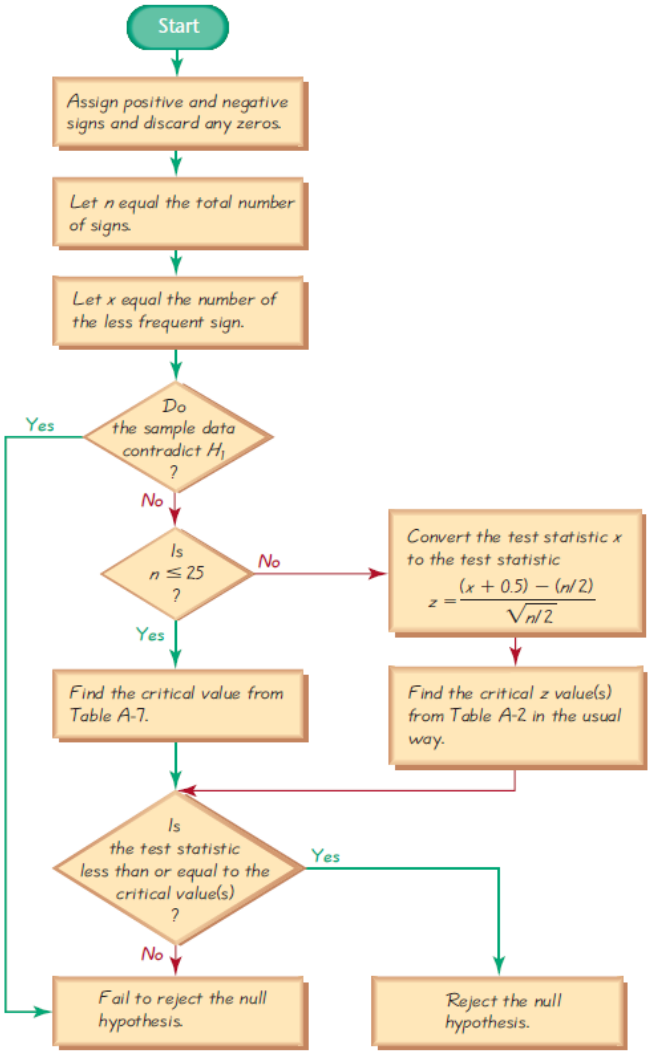
- **Handling ties in ranks:** If a tie in ranks occurs, the usual procedure is to find
- the mean of the ranks involved and then assign this mean rank to each of the tied
- items, as in the following example.

EXAMPLE The numbers 4, 5, 5, 5, 10, 11, 12, and 12 are given ranks of 1, 3, 3, 3, 5, 6, 7.5, and 7.5, respectively. See the table below and note the procedure for handling ties.

Sorted Data	Preliminary Ranking		Rank
4	1		**1**
5	2	Mean is 3.	**3**
5	3		**3**
5	4		**3**
10	5		**5**
11	6		**6**
12	7	Mean is 7.5.	**7.5**
12	8		**7.5**

符号检验

- 符号：+、 -
- 符号检验可以检验以下论断：
- 1. 关于配对样本数据的论断
- 2. 关于名词性数据的论断
- 3. 关于一个总体的中位数的论断



符号检验

- 适用条件：
- 1. 样本是简单随机样本
- 2. 对总体分布形式不作任何要求

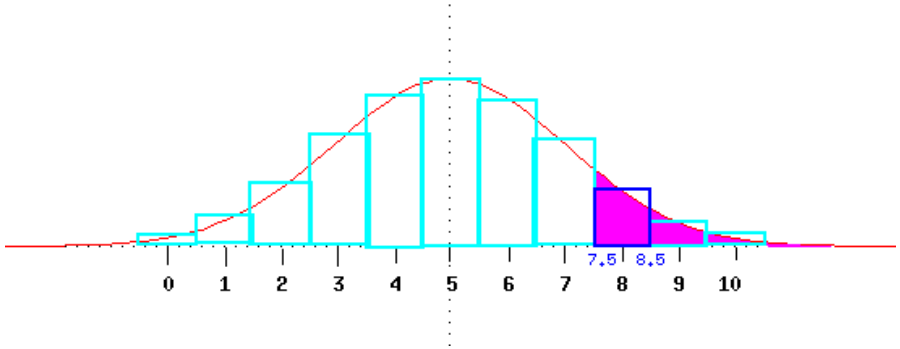
- B: min(+,-)
- n：样本数据去掉0后的总个数

- $H_0：v=0$
- 检验统计量：
- $n \leq 25$时：$B \sim B(n, 0.5)$

- $n > 25$时：$Z = \frac{(B+0.5)-(\frac{n}{2})}{\sqrt{n}/2} \sim N(0,1)$(近似)或$B$

连续项修正

- 使用连续型分布去近似离散型分布时，为了取得更好的近似值，常常会做连续型修正
- 例：正态分布近似二项分布
- P(X>7) =P(X>7.5)



- 例： X~B(20,0.25),求P(X>=8)
- 正常：$P(X \geq 8) = P(\frac{X-20*0.25}{\sqrt{20*0.258*0.75}} \geq \frac{8-20*0.25}{\sqrt{20*0.258*0.75}}) \approx \Phi(\frac{8-20*0.25}{\sqrt{20*0.258*0.75}})$
- 使用连续项修正:$P(X \geq 8) = P(X \geq 7.5) = P(\frac{X-20*0.25}{\sqrt{20*0.258*0.75}} \geq \frac{7.5-20*0.25}{\sqrt{20*0.258*0.75}}) \approx$

  $\Phi(\frac{7.5-20*0.25}{\sqrt{20*0.258*0.75}})$

符号检验——单个总体的中位数

- 设X_1，X_2，……，X_n是来自总体X的一个样本。
- 设X服从一个连续分布，具体是什么分布不作要求，总体中位数为v

H_0	H_1	检验统计量B
$v = 0$	$v \neq 0$	+或-的个数
	$v > 0$	
	$v < 0$	

- n = 12, +的数目为9。假设检验 $H_0 : v = 0$ v:s: $H_1 : v \neq 0$
- n = 12, +的数目为9 。假设检验 $H_0 : v = 0$ v:s: $H_1 : v > 0$
- n = 12, +的数目为9 。假设检验 $H_0 : v = 0$ v:s: $H_1 : v < 0$

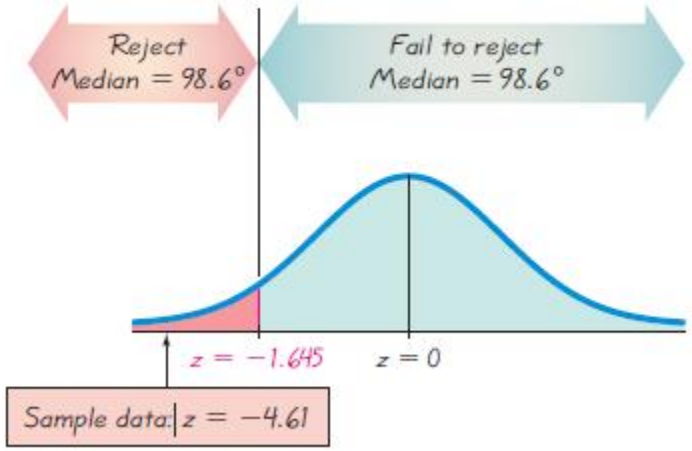
符号检验——单个总体的中位数

- 例子——体温
- 根据106个样本数据，我们要检验总体体温的中位数是否小于98.6华氏度。在106个样本数据中，有68个观察数据是低于98.6华氏度，23个观察数据是高于98.6华氏度，15个观察数据是等于98.6华氏度

$$H_0 : v = 98.6°\mathrm{F} \quad v.s \quad H_1 : v < 98.6°\mathrm{F}$$

```
> pbinom(23,91,0.5)
[1] 1.261077e-06

> pnorm(-4.61)
[1] 2.013345e-06
```



符号检验——配对样本

- 记配对样本中的观察值分别为Y_i，$Z_i(i = 1,2,\dots\dots,n)$
- 令$X_i = Y_i - Z_i(i = 1,2,\dots\dots,n)$
- 对于X_i的总体中位数v，作假设检验$H_0: v = 0$

Y (diet A)	Z (diet B)	X= Y-Z
85	83	2
69	78	-9
81	70	11
112	72	40
77	67	10
86	68	18
113	113	0

- 例子：双胞胎的体重
- 为研究不同食谱对孩子发育的影响，随机抽取了7对双胞胎进行试验，给双胞胎中的其中一位喂食食谱A，另一位喂食食谱B。判断食谱A是否比食谱B对孩子发育好。
- $H_0: v = 0$ v.s $H_1: v > 0$

```
> 1-pbinom(4,6,0.5)
[1] 0.109375
```

符号检验——名词性数据（Nominal Data）

- $H_0: p = 0.5$

- 例子：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%。随机抽取了325个曾服用过该药的夫妻所生下来的孩子，其中有295个女孩。设显著性水平为0.05，检验“这个药是没有效果”的这个说法是否正确。
- $H_0: p = 0.5$ v.s $H_1: p > 0.5$

```
> 1-pbinom(294,325,0.5)
[1] 0
> 1-pnorm((294.5-325*0.5)/sqrt(325*0.25))
[1] 0
```

符号检验

- 符号检验的不足
- 对比下面两组数据
- A : 1,2,3,4,-10
- B : 11,12,13,14,-1
- 做符号检验：$H_0：v=0$ v.s $H_1：v>0$ 得到相同的结果

Wilcoxon符号秩检验

- 设X_1，X_2，……，X_n是来自总体X的一个样本。
- 要求：X服从一个连续并且对称的分布

- 假设检验：H_0：$v = 0$

- 步骤：
- 1. 把0去掉，并重新计算n
- 2. 对$|X_1|$，$|X_2|$，……，$|X_n|$排序，得到其对应的秩R_1，R_2，……，R_n
- 3. 根据X_1，X_2，……，X_n的正负情况，给予其对应的秩R_1，R_2，……，R_n相应的+，-号，得到符号秩
- 4. 令T=sum（R_i），得到检验统计量
- 5. 在零假设的条件下，查看检验统计量T的值是否为一个极端值

Wilcoxon符号秩检验

- 例：H_0：$v=0$ v.s H_1：$v>0$ ，显著性水平0.05

Y (diet A)	Z (diet B)	$X=Y-Z$
85	83	2
69	78	-9
81	70	11
112	72	40
77	67	10
86	68	18
113	113	0

- 符号秩：

X_i	2	-9	11	40	10	18	0
rank	1	2	4	6	3	5	omit
Signed- rank	1	-2	4	6	3	5	

- T=1+4+6+3+5=19
- P(T>=19)=0.047<0.05

Wilcoxon符号秩检验

- T在零假设成立下的分布：
- 1. 没有ties的情况：例如n=3

R_1	R_2	R_3	T
1	2	3	6
-1	2	3	5
1	-2	3	4
1	2	-3	3
-1	-2	3	3
-1	2	-3	2
1	-2	-3	1
-1	-2	-3	0

- P(T>=3) = 5=8 = 0.625;P(T >= 4) = 3=8 = 0.375;P(T >= 5) = 2=8 = 0.25;
- P(T >= 6) = 1=8 = 0.125

Wilcoxon符号秩检验

- T在零假设成立下的分布：
- 1. 没有ties的情况：
- $E(T)=\frac{n(n+1)}{4}; Var(T)=\frac{n(n+1)(2n+1)}{24}$
- 2. 有ties的情况：
- 精确分布比较难求，查表不再适用
- $E(T)=\frac{n(n+1)}{4}; Var(T)=\frac{n(n+1)(2n+1)}{24}-\frac{1}{48}\sum_{i=1}^{r}t_i(t_i^2-1)$
- ti表示第i个结中相等数据的数目

$$1 \quad \underbrace{2.5 \quad 2.5}_{t_1=2} \quad 4 \quad \underbrace{6 \quad 6 \quad 6}_{t_2=3} \quad 8 \quad 9$$

- 近似：$\frac{T-E(T)}{Var(T)}\sim N(0,1)$（$n$充分大时可以近似）

Wilcoxon秩和检验

- 适用条件：
- 1. 两个独立的简单随机样本
- 2. 每个样本至少有10个样本值

- 设X_1，X_2，……，X_m和Y_1，Y_2，……，Y_n（n<=m,N=n+m）分别来说总体X与总体Y，是两个独立的样本

- 假设：H_0：X与Y具有相同的中位数 ；H_1：X与Y具有不同的中位数

Wilcoxon秩和检验

- 步骤：
- 1. 混合两个样本的数据
- 2. 将所有样本值按照从小到大的顺序排列，得到所有样本值的秩
- 3. 将样本容量较小的样本值所对应的秩加起来，得到检验统计量

$$W = \sum_{i=1}^{n} R_i$$

其中R_i是Y_i在所有样本值中的秩

- 例： X 1.2,8.3,3.6,5.4　　Y 4.1，9.2
- 1.2 3.6 4.1 5.4 8.3 9.2
- 1　2　3　4　5　6

Wilcoxon秩和检验

- W的精确分布：

$$P(R_1 = r_1, \cdots, R_n = r_n) = \frac{1}{\binom{N}{n}}.$$

- 例：n=2,m=4,N=6的情况

$$P(W \geq 11) = \frac{1}{15} = 0.067.$$

$$P(W \geq 10) = \frac{2}{15} = 0.133.$$

$$P(W \geq 19) = \frac{4}{15} = 0.267.$$

$$\Delta = Y - X$$

$$H_0: \ \Delta = 0 \quad v.s. \quad H_1: \ \Delta > 0.$$

- P(W>=9)=4/15=0.267

Rank of Y_i	W
1, 2	3
1, 3	4
1, 4	5
1, 5	6
1, 6	7
2, 3	5
2, 4	6
2, 5	7
2, 6	8
3, 4	7
3, 5	8
3, 6	9
4, 5	9
4, 6	10
5, 6	11

Wilcoxon秩和检验

- W的近似分布
- 1. 没有ties

$$E(W) = n(N+1)/2.$$

$$\mathrm{Var}(W) = mn(N+1)/12.$$

- 2. 有ties

$$E(W) = \frac{n(N+1)}{2}.$$

$$\mathrm{Var}(W) = \frac{mn(N+1)}{12} - \frac{mn}{12N(N-1)} \sum_{j=1}^{g} t_j(t_j^2 - 1),$$

- ti表示混合样本中第i个结中相等数据的数目
- 当n和m足够大时（n、m>=8），$\frac{W - E(W)}{\sqrt{\mathrm{Var}(W)}} \sim N(0,1).$

Wilcoxon秩和检验

- ties的情况：

X	12	16	16	12	10
Y	30	12	24	32	24

数据	10	12	12	12	16	16	24	24	30	32
秩	1	3	3	3	5.5	5.5	7.5	7.5	9	10

$\Delta = Y - X$

$H_0: \Delta = 0 \quad v.s. \quad H_1: \Delta > 0.$

- W=3+7.5+7.5+9+10=37
- P(W>=37)≈0.028（查表）

Wilcoxon秩和检验

- ties的情况（精确计算）：
- 1 3 3 3 5.5 5.5 7.5 7.5 9 10

$$\text{P(ranks : 3; 7:5; 7:5; 9; 10)} = \frac{\binom{3}{1}\binom{2}{2}\binom{1}{1}\binom{1}{1}}{\binom{10}{5}} = \frac{3\times 1\times 1\times 1}{\binom{10}{5}} = \frac{3}{\binom{10}{5}}$$

$$P(\text{ranks : } 5.5, 5.5, 7.5, 9, 10) = \frac{2}{\binom{10}{5}}.$$

Ranks	3 7.5 7.5 9 10	5.5 5.5 7.5 9 10	5.5 7.5 7.5 9 10
W	37	37.5	39
Probability	$\frac{3}{\binom{10}{5}}$	$\frac{2}{\binom{10}{5}}$	$\frac{2}{\binom{10}{5}}$

$$p\text{-value} = P(W \geq 37) = \frac{3+2+2}{\binom{10}{5}} = 0.0278 \approx 0.028.$$

Wilcoxon秩和检验

- 使用正态分布近似

$$\sum_{j=1}^{g} t_j(t_j^2-1) = 3\times(3^2-1)+2\times(2^2-1)+2\times(2^2-1) = 36.$$

$$E(W) = \frac{n(N+1)}{2} = \frac{5\times 11}{2} = 27.5.$$

$$\begin{aligned}\mathrm{Var}(W) &= \frac{mn(N+1)}{12} - \frac{mn}{12N(N-1)}\sum_{j=1}^{g} t_j(t_j^2-1)\\ &= \frac{5\times 5\times 11}{12} - \frac{5\times 5}{12\times 10\times 9}\times 36\\ &= 22.916 - 0.83 = 22.083.\end{aligned}$$

$$P(W\geq 37) = P\left(N(0,1)\geq \frac{37-27.5}{\sqrt{22.083}}\right) = 0.020.$$

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