数学实验

Experiments in Mathematics

实验9　数据的统计描述和分析

Dept. of Mathematical Sciences
Tsinghua University
Beijing 100084, China

对数据进行统计分析是工程技术，经济管理，科学研究，社会调查等领域中经常要做的事情.

区分两种不同意义的统计

对产品全部检验,废品数除以产品总数得到废品率

对产品进行抽检,废品数除以检验数得到废品率

通过人口普查数据得到甲乙两地生育率的差距

两地各抽1000名妇女判断生育率有无差别

普通意义的统计

数理统计

数理统计的对象------受随机因素影响的数据

数据的统计描述和分析

1、统计的基本概念

2、参数估计

3、假设检验

4、**MATLAB**统计工具箱**(Statistics Toolbox)**的使用

数据的统计分析示例1　学生的身高和体重

学校随机抽取 100 名学生，测量他们的身高(cm)和体重(kg)

身高	体重	身高	体重	身高	体重	身高	体重	身高	体重	身高	体重
172	75	169	55	169	64	171	65	167	47	177	64
171	62	168	67	165	52	169	62	168	65	184	70
166	62	168	65	164	59	170	58	165	64	166	49
160	55	175	67	173	74	172	64	168	57	171	71
155	57	176	64	172	69	169	58	176	57	170	59
173	58	168	50	169	52	167	72	170	57	171	59
166	55	161	49	173	57	175	76	158	51	175	68
170	63	169	63	173	61	164	59	165	62	165	56
167	53	171	61	166	70	166	63	172	53	169	65
173	60	178	64	163	57	169	54	169	66	168	62
178	60	177	66	170	56	167	54	169	58	182	63
173	73	170	58	160	65	179	62	172	50	176	68
163	47	173	67	165	58	176	63	162	52	172	56
165	66	172	59	177	66	182	69	175	75	173	59
170	60	170	62	169	63	186	77	174	66	174	64
163	50	172	59	176	60	166	76	167	63		
172	57	177	58	177	67	169	72	166	50		

问题

1）对这些数据作初步整理并给出图形描述；

2）根据这些数据对全校学生的平均身高和体重作出估计，并给出估计的误差范围;

3）学校10年前作过普查，学生的平均身高为167.5厘米，平均体重为60.2公斤，试根据这次抽查的数据，对学生的平均身高和体重有无明显变化作出结论。

示例2　胃溃疡病人的溶菌酶含量

患胃溃疡的病人组与无胃溃疡的对照组各取30人，两组人胃液中溶菌酶含量如下

病人	0.2	10.4	0.3	0.4	10.9	11.3	1.1	2.0	12.4	16.2
	2.1	17.6	18.9	3.3	3.8	20.7	4.5	4.8	24.0	25.4
	4.9	40.0	5.0	42.2	5.3	50.0	60.0	7.5	9.8	45.0
正常	0.2	5.4	0.3	5.7	0.4	5.8	0.7	7.5	1.2	8.7
	1.5	8.8	1.5	9.1	1.9	10.3	2.0	15.6	2.4	16.1
	2.5	16.5	2.8	16.7	3.6	20.0	4.8	20.7	4.8	33.0

1）判断患病人的溶菌酶含量与“正常人”有无显著差别;

2）若表中病人组最后5个数据有误，去掉后再作判断。

统计的基本概念

样本——统计研究的主要对象

- 总体--研究对象的全体。如学校全体学生的身高。
- 个体--总体中一个基本单位。如一个学生的身高。
- 样本--若干个体的集合。如**100**个学生的身高。
- 样本容量--样本中个体数。如**100**。

整个学校学生身高~随机变量X，概率分布$F(x)$;

n个学生的身高$\{x_i, i=1,\ldots n\}$ (样本)~相互独立的、分布均为$F(x)$的一组随机变量。

样本：随机取值的一组数据；

一组相互独立的、同分布的随机变量。

频数和直方图

将数据取值范围划分为若干区间，统计这组数据在每个区间中出现的次数——频数

示例**1**-----学生身高与体重的数据处理

100名学生身高、体重频数表

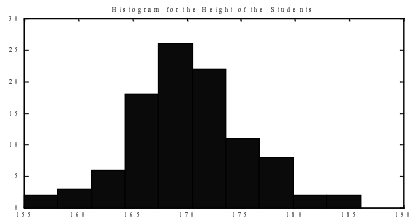
student.m

身高N	2	3	6	18	26	22	11	8	2	2
身高X	156.55	159.65	162.75	165.85	168.95	172.05	175.15	178.25	181.35	184.45
体重N	8	6	8	21	13	19	11	5	4	5
体重X	48.50	51.50	54.50	57.50	60.50	63.50	66.50	69.50	72.50	75.50

[N,X]=hist(data,k) ~ 数组data的频数。将由数组data的最小、最大值给出的区间k等分（缺省时k=10），N返回小区间频数， X返回小区间中点。

对数据作直观的图形描述——直方图

100名学生身高的直方图



hist(data,k) ~ 数组data的直方图 (**k**同前）

统计的基本概念----统计量

$x=(x_1, \ldots x_n)$ ~ 容量为n的样本

统计量——由样本加工的、反映样本的数量特征的函数

表示位置的统计量

平均值（均值）：数据取值的平均位置 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ mean(x)

中位数：将数据由小到大排序后位于中间位置的那个数值 median(x)

表示变异程度的统计量

标准差 $s=[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2]^{1/2}$ std(x)

方差 标准差的平方 s^2 var(x)

极差 x 的最大值与最小值之差 range(x)

表示分布形态的统计量

偏度 $g_1=\frac{1}{s^3}\sum_{i=1}^{n}(x_i-\bar{x})^3$ 对称性的度量 skewness(x)

峰度 $g_2=\frac{1}{s^4}\sum_{i=1}^{n}(x_i-\bar{x})^4$ 尾部形态的度量 kurtosis(x)

（正态分布g_2=3）

统计中常用的几个概率分布

分布函数$F(x)$与密度函数$p(x)$

$F(x)=P\{X\le x\}\quad F(-\infty)=0, F(\infty)=1$

$p(x)=dF/dx\quad F(x)=\int_{-\infty}^{x}p(x)dx$

$EX=\mu=\int_{-\infty}^{\infty}xp(x)dx,\ DX=\sigma^2=\int_{-\infty}^{\infty}(x-\mu)^2p(x)dx$

若对某分布函数F(x)和实数α(0<α<1)有$F(x_\alpha)=\alpha$, 称x_α为该分布的α分位数，或

$\int_{-\infty}^{x_\alpha}p(x)dx=\alpha\qquad x_\alpha=F^{-1}(\alpha)$

1）正态分布 $N(\mu,\sigma^2)$

(1) 密度函数 $p(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$

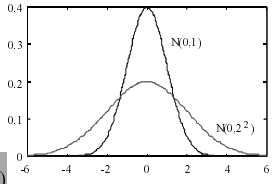
⑵均值$EX=\mu$, 方差$DX=\sigma^2$

(3)标准正态分布

$N(0,1)$的分布函数

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp(-x^2/2)dx$$

$$(4)X\sim N(\mu,\sigma^2)\Rightarrow\frac{X-\mu}{\sigma}\sim N(0,1)$$

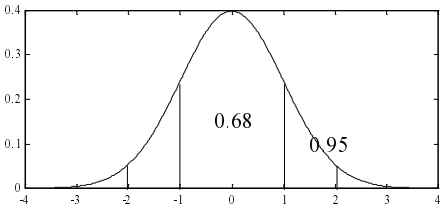


(5)均值和中位数相等，偏度为 0，峰度为 3

13

(6)常用概率

$$P(\mu-\sigma\le X\le\mu+\sigma)=68\%$$
$$P(\mu-2\sigma\le X\le\mu+2\sigma)=95\%$$
$$P(\mu-3\sigma\le X\le\mu+3\sigma)=99.7\%$$

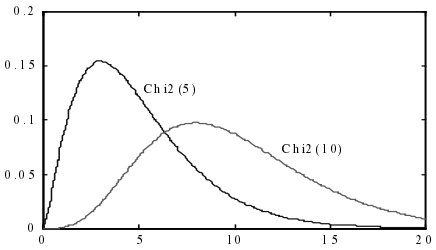


14

2）χ^2分布（Chi square）

条件：$X_1,X_2,\cdots X_n$相互独立、服从标准正态分布$N(0,1)$的随机变量，

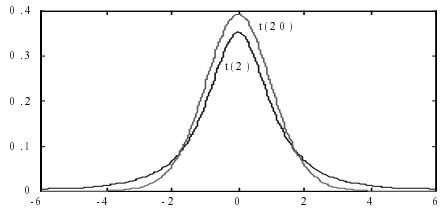
χ^2分布：$Y=\sum_{i=1}^{n}X_i^2\sim\chi^2(n)$ (n自由度)，$EY=n$，$DY=2n$



3）t分布(student)

条件：$X\sim N(0,1)$，$Y\sim\chi^2(n)$，且相互独立

t 分布：$T=\frac{X}{\sqrt{Y/n}}\sim t(n)$（n 称自由度）

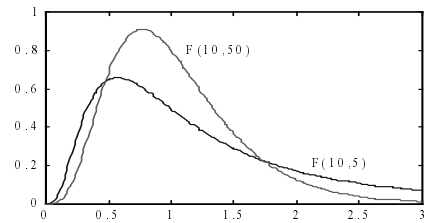


理论上$n\to\infty$时$T\sim t(n)\to N(0,1)$，实际上n≥30时即近似$N(0,1)$

4）F分布

条件：$X\sim\chi^2(n_1)$，$Y\sim\chi^2(n_2)$，且相互独立

F分布：$F=\frac{X/n_1}{Y/n_2}\sim F(n_1,n_2)$，$(n_1,n_2)$称自由度。



17

MATLAB统计工具箱(toolbox\stats)中的概率分布

norm	chi2	t	f
正态分布	χ^2分布	t 分布	F 分布

pdf	cdf	inv	stat	rnd
概率密度	概率分布	逆概率分布	均值与方差	随机数生成

使用方法：**将分布命令字符与函数命令字符接起来，并输入自变量（可以是标量、数组或矩阵）和参数**

p=normpdf(x,mu,sigma):

均值mu、标准差sigma的正态分布在x的密度函数

p=p(x)（mu=0，sigma=1时可缺省）

18

P=tcdf(x,n):
t分布(自由度n) 在x的分布函数P=F(x)

x=chi2inv(P,n):
χ^2 分布(自由度n) 使分布函数F(x)=P的x(即P分位数)

```
x=norminv(0.9,0,2)
x =
   2.5631
```

```
p=normcdf(2.5631,0,2)
p =
   0.9000
```

[m,v]=fstat(n1,n2):
F分布(自由度n1,n2)的均值m和方差v

example0901.m
example0902.m

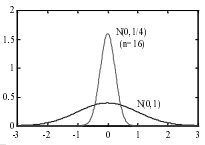
19

正态总体的样本的统计量的分布

• 最常用的统计量——样本均值，样本方差

均值 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，方差 $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$

仅当总体为正态分布$X\sim N(\mu,\sigma^2)$，$\bar{x},s^2$的分布才有便于使用的结果。



$\bar{x}\sim N(\mu,\sigma^2/n)$ 即 $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$

$\frac{\bar{x}-\mu}{s/\sqrt{n}}\sim t(n-1)$　　$\frac{(n-1)s^2}{\sigma^2}\sim\chi^2(n-1)$

20

参数估计-----点估计与区间估计

参数估计：利用样本统计量对总体参数进行估计，分点估计和区间估计两种。

点估计　• 用样本统计量确定总体参数的一个数值

• 评价估计优劣的标准:无偏性，有效性等
• 估计的方法有矩法、极大似然法等。

对总体均值μ，方差σ²的点估计：

$\hat{\mu}=\bar{x},\quad \hat{\sigma}^2=s^2,\quad \hat{\sigma}=s$

点估计未给出估计值的精度和可信程度

21

区间估计

总体的待估参数θ，估计量 $\hat{\theta}$，求区间 $[\hat{\theta}_1,\hat{\theta}_2]$，使θ满足

$P(\hat{\theta}_1\le\theta\le\hat{\theta}_2)=1-\alpha$　（$0<\alpha<1$,给定）

$[\hat{\theta}_1,\hat{\theta}_2]$：θ 的置信区间　　$\hat{\theta}_1,\hat{\theta}_2$:置信下限和置信上限

$1-\alpha$：置信概率或置信水平　　α:显著性水平

$\alpha=0.05$ ⇨ 由样本得到的置信区间以 0.95 的概率包含了待估参数θ

置信区间越小，估计精度越高　⇔　二者矛盾　置信水平越大，可信程度越高

在一定置信水平下使置信区间尽量小

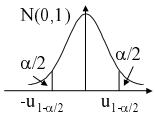
22

对总体均值μ，方差σ²的区间估计

• 假设总体服从正态分布N(μ,σ²)

1) 总体方差σ^2已知，估计均值μ，置信水平 $1-\alpha$

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



$N(0,1)$的$1-\alpha/2$分位数$u_{1-\alpha/2}$满足 $P(|z|\le u_{1-\alpha/2})=1-\alpha$

$\Rightarrow P(\bar{x}-u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\le\mu\le\bar{x}+u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}})=1-\alpha$

置信区间　$[\bar{x}-u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x}+u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}]$

23

2) 总体方差σ^2未知，估计均值μ，置信水平 $1-\alpha$

样本均值 $\bar{x}$
样本均方差s

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

$t(n-1)$的$1-\alpha/2$分位数$t_{1-\alpha/2}$满足：

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

置信区间　$[\bar{x}-t_{1-\alpha/2}\frac{s}{\sqrt{n}},\ \bar{x}+t_{1-\alpha/2}\frac{s}{\sqrt{n}}]$

与σ²已知比较　$[\bar{x}-u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x}+u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}]$

$\alpha=0.05,\ u_{1-\alpha/2}=1.96,\quad t_{1-\alpha/2}=2.06\ (n=25)$

24

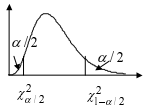
估计总体方差σ^2，置信水平 $1-\alpha$

样本方差s^2 $\quad \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$

χ^2分布的$\alpha/2$分位数 $\chi^2_{\alpha/2}$和 $1-\alpha/2$分位数 $\chi^2_{1-\alpha/2}$ 满足

$$P(\chi^2_{\alpha/2} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{1-\alpha/2}) = 1-\alpha$$

σ^2的置信区间 $[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2}}]$



对总体均值μ，方差σ^2的区间估计

μ的置信区间(σ^2已知) $[\bar{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}]$

μ的置信区间(σ^2未知) $[\bar{x} - t_{1-\alpha/2}\frac{s}{\sqrt{n}},\ \bar{x} + t_{1-\alpha/2}\frac{s}{\sqrt{n}}]$

σ^2的置信区间 $[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}},\ \frac{(n-1)s^2}{\chi^2_{\alpha/2}}]$

μ的置信区间长度 $L_\mu = 2t_{1-\frac{\alpha}{2}}s/\sqrt{n} \approx 2u_{1-\frac{\alpha}{2}}s/\sqrt{n}$ (n大)

σ^2的置信区间长度 $L_{\sigma^2} = (n-1)s^2(1/\chi^2_{\alpha/2} - 1/\chi^2_{1-\alpha/2})$

给定α，n越大，L_μ越小，估计精度越高；L_{σ^2} 呢？

参数估计-----MATLAB实现

```
[mu sigma muci sigmaci]=normfit(x,alpha)
```

其中：x为样本　　alpha为显著性水平（缺省时为0.05）

返回值：mu--　均值μ的点估计

　　sigma---标准差σ的点估计

　　muci--均值μ的区间估计

　　sigmaci---标准差σ的区间估计。

示例1-----学生身高与体重数据处理（续）

根据采集数据对全校学生的平均身高和体重作出估计，并给出估计的误差范围.

studentcg.m

假设检验

问题

学生的身高和体重

学校随机抽取 100 名学生，测量他们的身高和体重

3）学校10年前作过普查，学生的平均身高为167.5厘米，平均体重为60.2公斤，试根据这次抽查的数据，对学生的平均身高和体重有无明显变化作出结论。

假设：学生身高(总体)均值μ=167.5，

根据样本对该假设进行检验.

答案只有两种：接受；拒绝。

总体均值的假设检验

例　甲方：$x \sim N(50,1)$, 批量供给乙方。

对于每批产品μ=50是否成立，双方商定检验方案。

- 每批抽取25件测量，计算均值 $\bar{x}$
- 制订数量标准δ，若 $|\bar{x}-50| \le \delta$, 认为 μ=50成立，接受该批产品（合格品）；否则，拒绝。
- 商定水平α，使合格品被错误拒绝的概率不超过α （通常α =0.05)。

解　$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$　$P(|z| \le 2) = 0.95$　$P(|\bar{x}-\mu| \le 2\sigma/\sqrt{n}) = 0.95$

$\Rightarrow \delta = 2\sigma/\sqrt{n} = 0.4$　当 $|\bar{x}-50| \le 0.4$, 接受；否则，拒绝。

总体均值的假设检验

已有样本(容量n，均值 $\bar{x}$，标准差s)，要对总体均值μ是否等于给定值μ_0进行检验(假定总体服从正态分布)

假设　$H_0: \mu = \mu_0;\quad H_1: \mu \ne \mu_0$

称H_0为原假设（或零假设），H_1为备选假设，二者择其一：接受H_0；拒绝H_0，即接受H_1。

显著性水平α～ H_0成立时被错误拒绝的概率。

总体方差σ^2已知　$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$

$P(|z| \le u_{1-\alpha/2}) = 1-\alpha$

$\Rightarrow$ $|z| \le u_{1-\alpha/2}$ 时接受H_0; 否则拒绝H_0(接受H_1).

称z检验或u检验

总体均值的假设检验

总体方差 σ^2未知 $t=\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}\sim t(n-1)$ ⇨ $|t|\le t_{1-\alpha/2}$ 时接受H_0; 否则拒绝H_0(接受H_1).

$P(|t|\le t_{1-\alpha/2})=1-\alpha$ 称 t检验

常用: $\alpha=0.05\rightarrow u_{1-\alpha/2}=1.96$; $\alpha=0.01\rightarrow u_{1-\alpha/2}=2.575$

当n较大时(n>30) $t_{1-\alpha/2}$ 与$u_{1-\alpha/2}$相近.

思考 设从一个样本得到 z=2.2, 那么若取α=0.05，将拒绝 H_0; 若取α=0.01，将接受H_0。你怎样评价这两个不同的结果.

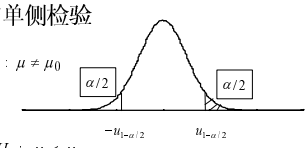
α是错误地拒绝H_0的概率，α不是越小越好吗?

31

总体均值的假设检验

双侧检验与单侧检验

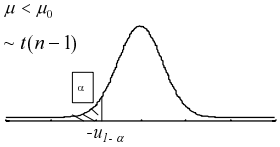
双侧检验 $H_0:\mu=\mu_0;\quad H_1:\mu\neq\mu_0$



单侧检验 $H_0:\mu\ge\mu_0;\quad H_1:\mu<\mu_0$

$z=\dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\sim N(0,1)$ $t=\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}\sim t(n-1)$

$z\ge u_\alpha(=-u_{1-\alpha})$时接受$H_0$; 否则拒绝$H_0$(接受$H_1$).



单侧检验 $H_0:\mu\le\mu_0;\quad H_1:\mu>\mu_0$

32

总体均值假设检验的MATLAB实现

总体方差σ^2已知 $H_0:\mu=\mu_0;\quad H_1:\mu\neq\mu_0$ $H_0:\mu\ge\mu_0;\quad H_1:\mu<\mu_0$

```
[h,p,ci]=ztest(x,mu,sigma,alpha,tail)
```

输入：x~样本(数组), mu~μ_0, sigma~ σ, alpha~ α (缺省时α=0.05), tail~H_1: $\mu\neq\mu_0$, tail=0(可缺省); H_1: μ<μ_0, tail=-1; H_1: μ>μ_0, tail=1

输出：h=0~接受H_0, h=1~拒绝H_0，p~ H_0下样本均值出现的概率，ci~(由样本均值估计的)μ的置信区间.

总体方差σ^2未知

```
[h,p,ci]=ttest(x,mu,alpha,tail)
```

除不需sigma外，与ztest相同。

Hypo1.m

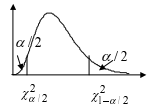
Hypo2.m

33

总体方差的假设检验

双侧检验 $H_0:\sigma^2=\sigma_0^2,\quad H_1:\sigma^2\neq\sigma_0^2$

$\chi^2=\dfrac{(n-1)s^2}{\sigma_0^2}\sim\chi^2(n-1)$



$\chi^2_{\alpha/2}\le\chi^2\le\chi^2_{1-\alpha/2}$ 时接受H_0;否则拒绝H_0.

单侧检验 $H_0:\sigma^2\ge\sigma_0^2,\quad H_1:\sigma^2<\sigma_0^2$

$H_0:\sigma^2\le\sigma_0^2,\quad H_1:\sigma^2>\sigma_0^2$

34

问题

粮食加工厂的一台自动装包机的设定值为：每包装50公斤，标准差0.3。现抽查了20包，得到如下数据

49.8 50.1 50.5 49.7 49.0 50.0 50.3 50.0 49.9 49.9

50.5 49.2 49.7 49.8 50.1 50.0 50.3 50.2 50.4 50.1

问该包装机运转是否正常?

单侧检验 $H_0:\sigma^2\le0.09,\quad H_1:\sigma^2>0.09$ 取α=0.05

计算样本方差 $s^2=0.1483$

$\chi^2=\dfrac{(n-1)s^2}{\sigma_0^2}=31.3056>\chi^2_{1-\alpha}=30.1435$ 拒绝$\mathbf{H_0}$

取α=0.02 试试看!

hypotestsigma.m

35

两个总体均值的假设检验

问题

胃溃疡病人的溶菌酶含量

患胃溃疡的病人组与无胃溃疡的对照组各取30人，测得两组人胃液中溶菌酶含量，

判断患病人的溶菌酶含量与"正常人"有无显著差别。

已有样本1(容量n_1, 均值 $\bar{x}$ 标准差s_1), 样本2(容量n_2，均值 $\bar{y}$标准差s_2), 要对2个总体均值μ_1, μ_2是否相等进行检验(假定总体均服从正态分布)

假设 $H_0:\mu_1=\mu_2,\quad H_1:\mu_1\neq\mu_2$

36

总体方差σ_1^2, σ_2^2已知　$H_0:\mu_1=\mu_2,\quad H_1:\mu_1\neq\mu_2$

$X\sim N(\mu_1,\sigma_1^2)$

$Y\sim N(\mu_2,\sigma_2^2)$

$$\Rightarrow\quad \frac{(\bar{x}-\mu_1)-(\bar{y}-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0,1)$$

若H_0成立 $\Rightarrow$ $$z=\frac{\bar{x}-\bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0,1)$$

给定显著性水平α

取$N(0,1)$的$1-\alpha/2$分位数$u_{1-\alpha/2}$ $\Rightarrow$ $|z|\le u_{1-\alpha/2}$时接受H_0; 否则拒绝H_0(接受H_1).

37

总体方差σ_1^2, σ_2^2未知，但可假定$\sigma_1^2=\sigma_2^2$　$H_0:\mu_1=\mu_2,\quad H_1:\mu_1\neq\mu_2$

$$\frac{(\bar{x}-\mu_1)-(\bar{y}-\mu_2)}{\sqrt{\frac{s^2}{n_1}+\frac{s^2}{n_2}}}\sim t(n_1+n_2-2),\quad s^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

H_0成立 $\Rightarrow$ $$t=\frac{\bar{x}-\bar{y}}{\sqrt{\frac{s^2}{n_1}+\frac{s^2}{n_2}}}\sim t(n_1+n_2-2)$$

取分位数$t_{1-\alpha/2}$, $|t|\le t_{1-\alpha/2}$时接受H_0; 否则拒绝H_0(接受H_1).

MATLAB实现

x，y~2个样本(数组，长度可不同), 其余用法与ttest相同

ill.m

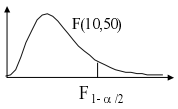
两个总体方差的假设检验

$X\sim N(\mu_1,\sigma_1^2)$　$Y\sim N(\mu_2,\sigma_2^2)$ $\Rightarrow$ $$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}\sim F(n_1-1,n_2-1)$$

已知2个样本n_1,n_2,s_1^2,s_2^2

双侧假设检验　$H_0:\sigma_1^2=\sigma_2^2,\quad H_1:\sigma_1^2\neq\sigma_2^2$

若H_0成立 $\frac{s_1^2}{s_2^2}\sim F(n_1-1,n_2-1)$



取$s_1^2\ge s_2^2$，记$F=\frac{s_1^2}{s_2^2}$

给定α，取$F(n_1-1,n_2-1)$的$1-\alpha/2$分位数$F_{1-\alpha/2}$

当$F\le F_{1-\alpha/2}$时接受H_0; 否则拒绝H_0（接受H_1）

置信水平仍为1-α

39

假设检验的进一步讨论

某炼油厂(甲方)向用户(乙方)成批(若干桶)供货，现只考虑汽油合格的一项指标：含硫量不超过1%。若双方商定每批抽检10桶，试以下面数据为例，在以下情况下讨论乙方是否应接受该批汽油。

1.08 0.93 1.08 0.94 0.97 1.19 1.17 1.27 0.99 1.10 (%)

A. 显著性水平α=0.05, (1)甲方提供(总体)标准差σ=0.1, (2)甲方提供σ=0.15, (3)甲方未提供σ;

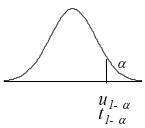
B. 甲方一向信誉很好，将显著性水平改为α=0.01;

C. 现乙方与一新炼油厂(丙方)谈判，如沿用与甲方订的合同(α由0.05改为0.01), 会有什么后果（风闻丙方有用含硫量1.08% 的汽油顶替合格品的前科）。

解　单侧检验 $H_0:\mu\le\mu_0, H_1:\mu>\mu_0$ $(\mu_0=1)$　Hypo3.m

A. 显著性水平α=0.05

甲方提供σ　$u=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$



$u\le u_{1-\alpha}$时接受H_0; 否则, 拒绝H_0.

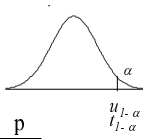
甲方未提供σ　$t=\frac{\bar{x}-\mu}{s/\sqrt{n}}\sim t(n-1)$　$t\le t_{1-\alpha}$时接受H_0; 否则, 拒绝H_0.

σ	u	$u_{1-\alpha}$	h	p
0.10	2.2768	1.6449	1	0.0114
0.15	1.5179	1.6449	0	0.0645
未知	1.9854	1.8331	1	0.0392
s=0.1147	t	$t_{1-\alpha}$		

问： σ=0.15时接受H_0合理吗，为什么会被接受?

单侧检验 $H_0:\mu\le\mu_0, H_1:\mu>\mu_0$ $(\mu_0=1)$

B. 显著性水平α=0.01



σ	u	$u_{1-\alpha}$	h	p
0.10	2.2768	2.3263	0	0.0114
0.15	1.5179	2.3263	0	0.0645
未知	1.9854	2.8214	0	0.0392
s=0.1147	t	$t_{1-\alpha}$		

由于甲方一向信誉很好，H_0不宜轻易拒绝，将显著性水平α (H_0被错误否定的概率)减小是合适的;

42

单侧检验 $H_0:\mu\le\mu_0, H_1:\mu>\mu_0\ (\mu_0=1)$

C. 考察丙方一旦用含硫量1.08% 的汽油顶替合格品，在α分别为0.05和0.01时, 错误接受H_1的概率(β)有多大

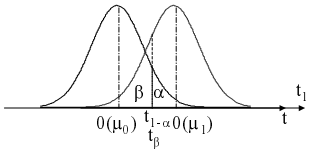
$t=\frac{\bar{x}-\mu}{s/\sqrt{n}}\sim t(n-1)$ $P(t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}<t_{1-\alpha}|H_0)\ge 1-\alpha$ 若$t\le t_{1-\alpha}$, 接受H_0

当一样本来自μ= μ_1=1.08> μ_0=1的总体时 $t_1=\frac{\bar{x}-\mu_1}{s/\sqrt{n}}\sim t(n-1)$

$\beta=P(t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}<t_{1-\alpha}|H_1)$ ⇨ $\beta=P(t_1<t_{1-\alpha}-\frac{\mu_1-\mu_0}{s/\sqrt{n}}|H_1)$

$t=t_1+\frac{\mu_1-\mu_0}{s/\sqrt{n}}$ $=F_{t(n-1)}(t_{1-\alpha}-\frac{\mu_1-\mu_0}{s/\sqrt{n}})$

当μ= μ_1=1.08> μ_0=1时 $\beta=F_{t(n-1)}(t_{1-\alpha}-\frac{\mu_1-\mu_0}{s/\sqrt{n}})=F_{t(n-1)}(t_\beta)$



α由0.05降为0.01,错误接受H_1的概率β太大，对于一向信誉很好的甲方合适，对于可疑的新厂(丙方)则不合适。

α	$t_{1-\alpha}$	t_β	β
0.05	1.8331	-0.3729	0.3589
0.01	2.8214	0.6154	0.7232

Hypo3.m

$\alpha\downarrow\rightarrow t_{1-\alpha}\uparrow, t_\beta\uparrow\rightarrow\beta\uparrow$

假设检验中的两类错误

- 第一类错误“弃真”——本应接受的H_0被拒绝, 概率α
- 第二类错误“取伪”——本应拒绝的H_0被接受, 概率β

当样本容量一定时，二者矛盾：α减小导致β增加.

通常α选得较小(0.05,0.01), β则较大(具体数值取决于μ_1).

原假设H_0和备选假设H_1是不平等的：人们保护、偏爱H_0；“歧视”H_1。

实际问题中选择什么样H_0的是重要的

0—1分布总体均值的假设检验

问题 甲方向乙方成批供货，双方商定废品率不超过 3%。今从一批中抽取 100 件，发现有 5 件废品，问乙方是否应接受这批产品。（设$\alpha=0.05$）

分析 总体X服从0—1分布：X=0—合格品; X=1—废品

X 的均值$\mu=p$ (废品率）, X 的方差$\sigma^2=p(1-p)$

样本容量n，均值 $\bar{x}$ (平均废品率)

n充分大时近似地有 $z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$

总体废品率的假设检验(双侧)

$H_0: p=p_0,\quad H_1: p\ne p_0$

H_0成立时 $z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}=\frac{\bar{x}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}\sim N(0,1)$

取$N(0,1)$的1-α/2分位数$u_{1-\alpha/2}$ $P(|z|\le u_{1-\alpha/2})=1-\alpha$

$|z|\le u_{1-\alpha/2}$ 时接受 H_0; 否则拒绝 H_0(接受 H_1).

假设检验(单侧) $H_0: p\le p_0,\quad H_1: p>p_0$

$|z|\le u_{1-\alpha}$ 时接受 H_0; 否则拒绝 H_0(接受 H_1).

问题 $\bar{x}=5/100, p_0=0.03, n=100$ $z=1.17<u_{1-\alpha}=u_{0.95}=1.65$

H_0成立，乙方应接受那批产品.

问：如果将α提高，会有什么结果？

总体分布的正态性检验

Q—Q图检验: `normplot(x)`, x ~样本

将样本从小到大排序得$x_1\le x_2\le\cdots\le x_n$，经验分布函数为 $F_n(x)=\begin{cases}0, & x<x_1\\ k/n, & x_k\le x<x_{k+1}\ (k=1,2,\cdots n-1)\\ 1, & x\ge x_n\end{cases}$

若样本来自正态分布总体，则$F_n(x)\approx F(x)$，$F_n(x)\approx\Phi(\frac{x-\mu}{\sigma})$。

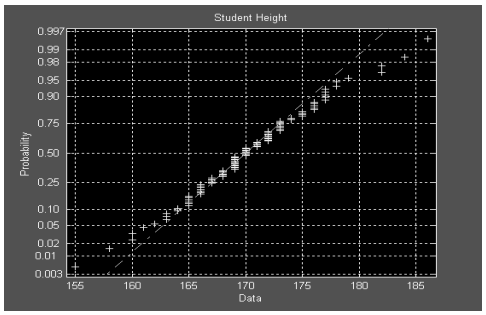
令$u=\Phi^{-1}(F_n(x))$，则 $x\approx\sigma u+\mu$，在$x\sim u$平面上是一条直线。

实际上取$F_n(x_k)=\frac{k-1/2}{n}$，相应的$u_k=\Phi^{-1}(\frac{k-1/2}{n})$

如果由样本计算出的 n 个点(x_k,u_k)近似在直线$x=\sigma u+\mu$上，则可认为它来自正态分布。

学生身高分布的正态性检验

qqtest.m



49

目的

1、掌握数据统计描述和分析的基本方法

2、根据问题的要求提出模型

3、对已经确定的模型，确定参数、使用MATLAB

作业

3)，6)(问：两机床的精度是否一样)，8)

50