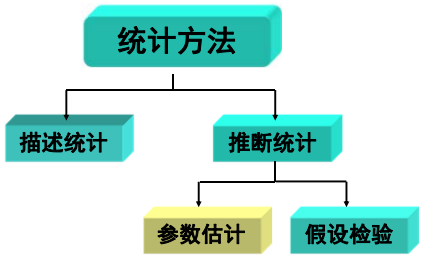
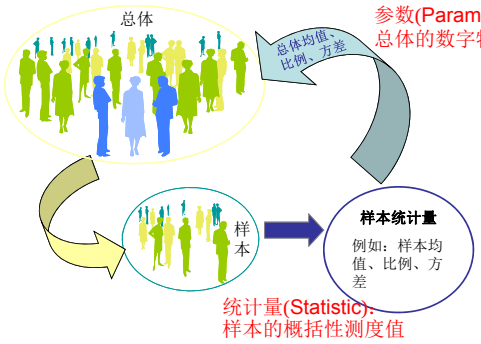
第四章　抽样分布与参数估计

参数估计在统计方法中的地位



统计推断的过程



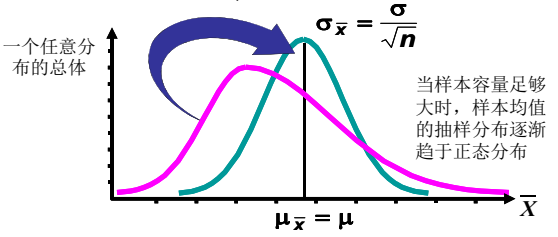
第一节　抽样分布

1. 所有样本指标（如均值、比例、方差等）所形成的分布称为抽样分布
2. 是一种理论概率分布
3. 随机变量是 **样本统计量**
   - 样本均值, 样本比例等
4. 结果来自容量相同的所有可能样本
5. 随机等概率抽样、放回式抽样

样本均值的抽样分布

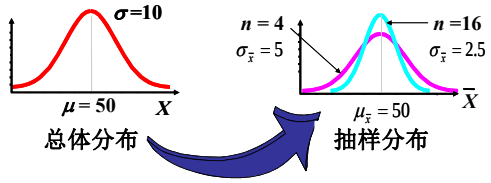
中心极限定理

中心极限定理：设从均值为μ，方差为σ^2的一个任意总体中抽取容量为n的样本，当n充分大时，样本均值的抽样分布近似服从均值为μ、方差为σ^2/n的正态分布



样本均值的抽样分布
（来自方差已知的正态总体）

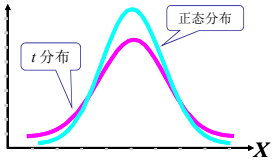
当总体服从正态分布$N\sim(\mu,\sigma^2)$时，来自该总体的所有容量为 n 的样本的均值 $\bar{X}$ 也服从正态分布，$\bar{X}$ 的数学期望为 μ，方差为 σ^2/n。即 $\bar{X}\sim N(\mu,\sigma^2/n)$



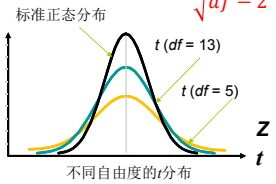
样本均值的抽样分布
（标准差未知的正态总体）

- 可以用样本标准差代替总体标准差
- 统计量 $T=\dfrac{\bar{X}-\mu}{s/\sqrt{n}}\sim t(n-1)$

$\sigma=\sqrt{\dfrac{df}{df-2}}$

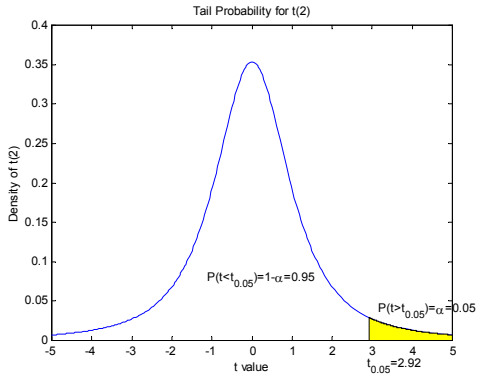


t 分布与正态分布的比较



不同自由度的t分布

t(2)分布右侧尾概率$P(t>t_\alpha)=\alpha$的示意图



n	P(2):双侧 P(1):单侧	0.5 0.25	0.2 0.1	0.1 0.05	0.05 0.025	0.02 0.01	0.01 0.005	0.005 0.0025	0.002 0.001
1		1	3.078	6.314	12.706	31.821	63.657	127.321	318.309
2		0.816	1.886	2.92	4.303	6.965	9.925	14.089	22.327
3		0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215
4		0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173
5		0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893
6		0.718	1.44	1.943	2.447	3.143	3.707	4.317	5.208
7		0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785
8		0.706	1.397	1.86	2.306	2.896	3.355	3.833	4.501
9		0.703	1.383	1.833	2.262	2.821	3.25	3.69	4.297
10		0.7	1.372	1.812	2.228	2.764	3.169	3.581	4.144
11		0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025
12		0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.93
13		0.694	1.35	1.771	2.16	2.65	3.012	3.372	3.852
14		0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787
15		0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733
16		0.69	1.337	1.746	2.12	2.583	2.921	3.252	3.686
17		0.689	1.333	1.74	2.11	2.567	2.898	3.222	3.646
18		0.688	1.33	1.734	2.101	2.552	2.878	3.197	3.61
19		0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579
20		0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552
21		0.686	1.323	1.721	2.08	2.518	2.831	3.135	3.527
22		0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505
23		0.685	1.319	1.714	2.069	2.5	2.807	3.104	3.485
24		0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467
25		0.684	1.316	1.708	2.06	2.485	2.787	3.078	3.45
26		0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435
27		0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421
28		0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408
29		0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396
30		0.683	1.31	1.697	2.042	2.457	2.75	3.03	3.385
31		0.682	1.309	1.696	2.04	2.453	2.744	3.022	3.375
32		0.682	1.309	1.694	2.037	2.449	2.738	3.015	3.365

样本方差的抽样分布

样本方差的分布

➡设总体服从正态分布$N\sim(\mu,\sigma^2)$，X_1，X_2，…，X_n为来自该正态总体的样本，则样本方差 s^2 的分布为

$$\frac{(n-1)s^2}{\sigma^2}\sim\chi^2(n-1)\qquad S^2=\frac{\sum_i(X_i-\mu)^2}{n-1}$$

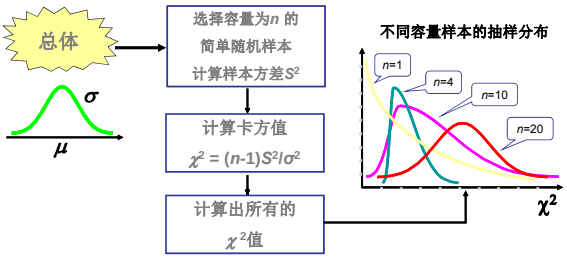
将$\chi^2(n-1)$称为自由度为$(n-1)$的卡方分布

χ^2-分布

$$\frac{(n-1)S^2}{\sigma^2}=\frac{(n-1)}{\sigma^2}\frac{\sum_i(X_i-\mu)^2}{n-1}=\sum_i(\frac{X_i-\mu}{\sigma})^2$$

- ***n*个独立正态变量平方和称为有 *n* 个自由度的 χ^2-分布 (chi-square distribution)，记为$\chi^2(n)$。**
- **χ^2-分布的可加性：独立的卡方变量之和仍然服从χ^2-分布，其自由度等于原来分布的自由度之和。**

卡方 (χ^2) 分布



卡方检验临界值表

自由度	显著性水平（α）					
	0.50	**0.25**	**0.10**	**0.05**	**0.03**	**0.01**
1	0.455	1.323	2.706	3.841	5.024	6.635
2	1.386	2.773	4.605	5.991	7.378	9.210
3	2.366	4.108	6.251	7.815	9.348	11.345
4	3.357	5.385	7.779	9.488	11.143	13.277
5	4.351	6.626	9.236	11.070	12.833	15.086
6	5.348	7.841	10.645	12.592	14.449	16.812
7	6.346	9.037	12.017	14.067	16.013	18.475
8	7.344	10.219	13.362	15.507	17.535	20.090
9	8.343	11.389	14.684	16.919	19.023	21.666
10	9.342	12.549	15.987	18.307	20.483	23.209
11	10.341	13.701	17.275	19.675	21.920	24.725
12	11.340	14.845	18.549	21.026	23.337	26.217
13	12.340	15.984	19.812	22.362	24.736	27.688
14	13.339	17.117	21.064	23.685	26.119	29.141
15	14.339	18.245	22.307	24.996	27.488	30.578
16	15.338	19.369	23.542	26.296	28.845	32.000
17	16.338	20.489	24.769	27.587	30.191	33.409
18	17.338	21.605	25.989	28.869	31.526	34.805
19	18.338	22.718	27.204	30.144	32.852	36.191
20	19.337	23.828	28.412	31.410	34.170	37.566
21	20.337	24.935	29.615	32.671	35.479	38.932
22	21.337	26.039	30.813	33.924	36.781	40.289
23	22.337	27.141	32.007	35.172	38.076	41.638
24	23.337	28.241	33.196	36.415	39.364	42.980

样本标准差的分布

$E(S^2)=\sigma^2$　　$E(S)\overset{?}{=}\sigma$

若 $E(S)=\sigma$，

$$\sigma_S^2=E(S^2)-[E(s)]^2=\sigma^2-[E(s)]^2=0$$

实际上，$E(S)=c\cdot\sigma,\quad c<1$

$$\sigma_S=\sqrt{1-c^2}\cdot\sigma$$

来自两个总体的样本统计量分布

两个样本均值的和与差的分布（总体方差已知）

- 设X_1，X_2，…，X_{n1}是来自正态总体$N(\mu_1,\sigma_1^2)$的一个样本，Y_1，Y_2，…，Y_{n2}是来自正态总体$N(\mu_2,\sigma_2^2)$的一个样本，且$X_i(i=1,2,\ldots,n_1)$，$Y_i(i=1,2,\ldots,n_2)$相互独立，则

$$\bar{X}\pm\bar{Y}\sim N\left((\mu_1\pm\mu_2),\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_1^2}{n_1}\right)\right)$$

两个样本方差比的抽样分布

➡设X_1，X_2，…，X_{n1}是来自正态总体$N\sim(\mu_1,\sigma_1^2)$的一个样本，Y_1，Y_2，…，Y_{n2}是来自正态总体$N\sim(\mu_2,\sigma_2^2)$的一个样本，且$X_i(i=1,2,\ldots, n_1)$，$Y_i(i=1,2,\ldots, n_2)$相互独立，则

$$\frac{S_x^2/S_y^2}{\sigma_1^2/\sigma_2^2}=\frac{S_x^2/\sigma_1^2}{S_y^2/\sigma_2^2}\sim F(n_1-1,n_2-1)$$

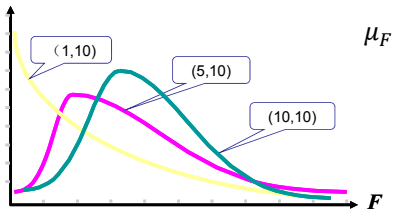
将$F(n_1\text{-}1, n_2\text{-}1)$称为第一自由度为$(n_1\text{-}1)$，第二自由度为$(n_2\text{-}1)$的$F$分布

F-分布

- F-分布变量为两个χ^2-分布变量（在除以它们各自自由度之后）的比；
- 而两个χ^2-分布的自由度则为F-分布的自由度，因此，F-分布有两个自由度；第一个自由度等于在分子上的χ^2-分布的自由度，第二个自由度等于在分母的χ^2-分布的自由度。

两个样本方差比的抽样分布

➡不同样本容量的抽样分布



$$\mu_F=\frac{df_2}{df_2-2}$$

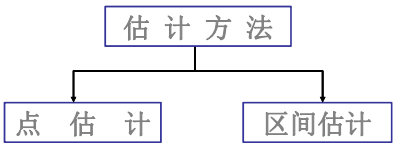
两个样本均值的和与差的分布（总体方差未知但相等）

➡设X_1，X_2，…，X_{n1}和Y_1，Y_2，…，Y_{n2}来自相互独立且具有相同方差的两个正态总体，那么

$$t=\frac{(\bar{X}\pm\bar{Y})-(\mu_1\pm\mu_2)}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1-1)+(n_2-1)}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

$$\sim t(n_1+n_2-2)$$

第二节　参数估计基本方法



点 估 计

- 点估计：用由样本数据所计算出来的单个数值，对总体参数所做的估计。
  - 点估计没有给出估计值接近总体未知参数程度的信息
- 估计量（estimator）：用于估计总体参数的统计量。
  - 求估计量的方法：矩估计法、极大似然法等
  - 优良性准则：无偏性、有效性和一致性

矩估计法

用样本的k阶矩作为总体的k阶矩的估计量，建立含待估计参数的方程，从而解出待估计参数。

例：设总体X的概率密度为

$$f(x,\theta)=\begin{cases}\theta x^{\theta-1}, & 0<x<1\\ 0, & others\end{cases}$$

$x_1, x_2, \ldots, x_n$为来自于总体X的样本，求参数θ的矩估计。

解：$E(X)=\int_0^1 x\cdot\theta x^{\theta-1}dx=\theta\int_0^1 x^{\theta}dx=\frac{\theta}{\theta+1}$

$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ 是 $E(X)$的估计量

$\Rightarrow$ $\hat{\theta}=\frac{\bar{x}}{1-\bar{x}}$

- 优点：
  - 不依赖总体分布，简便易行
  - 只要n充分大，精确度较高
- 缺点：
  - 矩估计的有效性较差
  - 要求总体的某个k阶矩存在
  - 要求未知参数能写成k阶矩的函数形式
- 注意：
  - 对相同的参数存在多个矩估计，不唯一

极大似然函数法

- 设随机变量X的概率密度函数为$f(x;\theta)$，其中θ为未知参数，$x_1,\ldots,x_n$为样本观察值，称

$$L(\theta)=L(x_1,\ldots,x_n;\theta)=\prod_{i=1}^{n}f(x_i;\theta)$$

为X关于样本观察值的似然函数。

- 如果似然函数在$\theta=\hat{\theta}$时达到最大值，则称$\hat{\theta}$是参数θ的极大似然估计。

例：设X服从（0-1）分布，P{X=1}=p未知，$x_1, x_2, \ldots, x_n$为来自于总体的样本值，求参数p的极大似然估计。

解：$L(x_1,\ldots,x_n;p)=\prod_{i=1}^{n}P\{X=x_i\}$

$P\{X=0\}=1-p,\ P\{X=1\}=p$

$\Leftrightarrow$ $P\{X=x_i\}=(1-p)^{1-x_i}p^{x_i}$

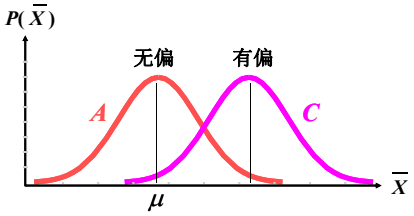
$\frac{d\,lnL}{dp}=\frac{1}{p}\sum x_i-\frac{1}{1-p}\sum(1-x_i)=0$

$\Rightarrow$ $\hat{p}=\bar{x}$

- 优点：
  - 利用了分布函数形式，得到的估计量的精度一般较高
  - 数学上可证明，在一定条件下，只要样本量足够大，极大似然估计和未知参数真值的差可任意小
- 缺点：
  - 要求必须知道总体的分布函数形式

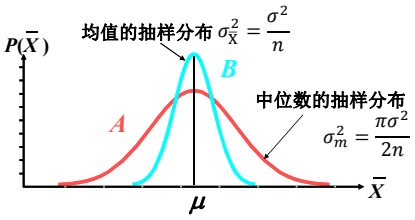
估计量的优良性准则

无偏性：估计量的数学期望等于被估计的总体参数



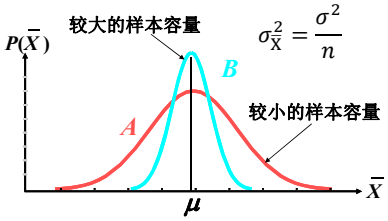
估计量的优良性准则

有效性：样本含量相同的情况下，一个方差较小的无偏估计量称为一个更有效的估计量。



估计量的优良性准则

一致性：随着样本容量的增大，估计量越来越接近被估计的总体参数。也叫相容性。



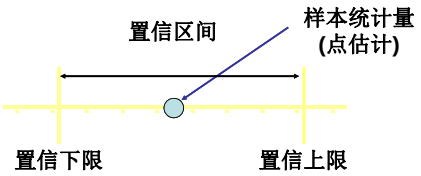
被估计的总体参数

	总体参数	符号表示	用于估计的样本统计量
一个总体	均值	μ	$\bar{x}$
	比例	P	$\hat{p}$
	方差	σ^2	s^2
两个总体	均值之差	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$
	比例之差	$P_1 - P_2$	$\hat{p}_1 - \hat{p}_2$
	方差比	σ_1^2/σ_2^2	s_1^2/s_2^2

区间估计

区间估计

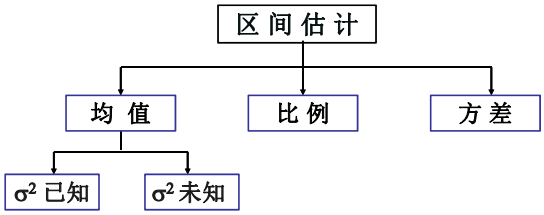
1. 根据一个样本的观察值给出总体参数的估计范围
2. 给出以一定概率包含总体参数的区间



抽样分布

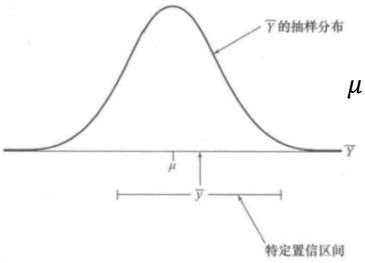
数据来源	统计量	总体方差已知	总体方差未知
一个总体	样本均值		
	样本方差		
两个总体	均值之和或差		
	方差之比		

置信区间估计



区间估计的一般方法

由中心极限定理，$\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} \sim N(0,1)$，其中$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$



$$\mu = \bar{X} \pm Z_{\alpha/2} \cdot \sigma_{\bar{X}}$$

第三节　总体均值和总体方差的区间估计

一. 总体均值的区间估计
二. 总体比例的区间估计
三. 样本容量的确定

总体均值的区间估计
(σ^2 已知)

总体均值的置信区间 (σ^2 已知)

1. 假定条件
   - 总体服从正态分布, 且总体方差（σ^2）已知
   - 如果不是正态分布，可以由正态分布来近似 ($n \geq 30$)
2. 使用正态分布统计量 Z

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

3. 总体均值 μ 在1-α置信水平下的置信区间为

$$\left(\bar{x} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \quad \bar{x} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

总体均值的区间估计

【例】某大学从该校学生中随机抽取100人，调查到他们平均每天参加体育锻炼的时间为26分钟。试以95%的置信水平估计该大学全体学生平均每天参加体育锻炼的时间（已知总体方差为36小时）。

解： 已知 $\bar{x}$=26, σ=6，n=100,
1-α = 0.95， $Z_{\alpha/2}$=1.96

$$\left(\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

$$= \left(26 - 1.96\frac{6}{\sqrt{100}}, 26 + 1.96\frac{6}{\sqrt{100}}\right)$$

$$= (24.824, 27.176)$$

我们可以95%的概率保证平均每天参加锻炼的时间在24.824～27.176分钟之间

总体均值的置信区间 (σ^2 未知)

1. 假定条件
   - 总体方差（σ^2）未知
   - 总体必须服从**正态分布**
2. 使用 t 分布统计量

$$t = \frac{\bar{X} - \mu}{S_{n-1}/\sqrt{n}} \sim t(n-1)$$

3. 总体均值 μ 在1-α置信水平下的置信区间为

$$\left(\bar{X} - t_{\frac{\alpha}{2}}\frac{S_{n-1}}{\sqrt{n}}, \quad \bar{X} + t_{\frac{\alpha}{2}}\frac{S_{n-1}}{\sqrt{n}}\right)$$

总体均值的区间估计

【例】从一个正态总体中抽取一个随机样本，n = 25，其均值 $\bar{x}$ = 50，标准差 s = 8。建立总体均值μ 的95%的置信区间。

解： 已知 $X \sim N(\mu, \sigma^2)$， $\bar{x}$=50, s=8，
n=25, 1-α = 0.95， $t_{\alpha/2}$=2.0639。

$$\left(\bar{x} - t_{\alpha/2}\frac{s_{n-1}}{\sqrt{n}}, \bar{x} + t_{\alpha/2}\frac{s_{n-1}}{\sqrt{n}}\right)$$

$$= \left(50 - 2.0639\frac{8}{\sqrt{25}}, 50 + 2.0639\frac{8}{\sqrt{25}}\right)$$

$$= (46.69, 53.3)$$

我们可以95%的概率保证总体均值在46.69～53.30 之间

正态总体方差的区间估计

1. 假设总体方差 σ^2 的点估计量为S^2,且

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

$$P\left(\chi^2_{1-\frac{\alpha}{2}} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

2. 总体方差在1-α置信水平下的置信区间为

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}\right]$$

总体比例的区间估计

总体比例的置信区间

1. 假定条件
   - 两类结果
   - 总体服从二项分布
   - 可以由正态分布来近似
2. 使用正态分布统计量 Z

$$Z=\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}\sim N(0,1)$$

3. 总体比例 P 的置信区间为

$$\hat{p}\pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

总体比例的置信区间

【例】某企业在一项关于职工流动原因的研究中，从该企业前职工的总体中随机选取了200人组成一个样本。在对其进行访问时，有140人说他们离开该企业是由于同管理人员不能融洽相处。试对由于这种原因而离开该企业的人员的真正比例构造95%的置信区间。

解： 已知 n=200 ，$\hat{p}$ =0.7 ，$n\hat{p}$ =140>5, $n(1-\hat{p})$=60>5，α= 0.95， $Z_{\alpha/2}$=1.96

$$\hat{p}\pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$=0.7\pm 1.96\sqrt{\frac{0.7(1-0.7)}{200}}$$

$$(0.636,0.764)$$

我们可以95%的概率保证该企业职工由于同管理人员不能融洽相处而离开的比例在63.6%~76.4%之间

样本容量的确定

估计总体均值时样本容量的确定

$$\left(\bar{x}-Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\quad \bar{x}+Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

若样本均值允许误差为 Δ，

也就是说 $\Delta=Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$

所以样本容量 $n=\frac{Z_{\alpha/2}^2\sigma^2}{\Delta^2}$

样本容量的确定

【例】一家广告公想估计某类商店去年所花的平均广告费用有多少。经验表明，总体方差约为1800000元。如置信度取95%，并要使估计处在总体平均值附近500元的范围内，这家广告公司应抽多大的样本？

解：已知 σ^2=1800000， α =0.05，$Z_{\alpha/2}$=1.96，Δ=500

应抽取的样本容量为

$$n=\frac{Z_{\alpha/2}^2\sigma^2}{\Delta^2}$$

$$=\frac{(1.96)^2(1800000)}{500^2}$$

$$=27.65\approx 28$$

估计总体比例时样本容量的确定

1. 根据比例区间估计公式可得样本容量n为

$$n=\frac{Z_{\alpha/2}^2p(1-p)}{\Delta^2}$$

其中：$\Delta=Z_{\alpha/2}\sqrt{\frac{n}{p(1-p)}}$

2. 若总体比例P未知时，可用样本比例 $\hat{p}$ 来代替

样本容量的确定

【例】一家市场调研公司想估计某地区有彩色电视机的家庭所占的比例。该公司希望对比例p的估计误差不超过0.5，要求的可靠程度为95%，应抽多大容量的样本（没有可利用的p估计值）。

解: 已知$\Delta=0.05$，$\alpha=0.05$，$Z_{\alpha/2}=1.96$，当$\hat{p}$未知时用最大方差0.25代替

应抽取的样本容量为

$$n=\frac{Z_{\alpha/2}^2 p(1-p)}{\Delta^2}$$

$$=\frac{(1.96)^2(0.5)(1-0.5)}{(0.5)^2}$$

$$\approx 385$$

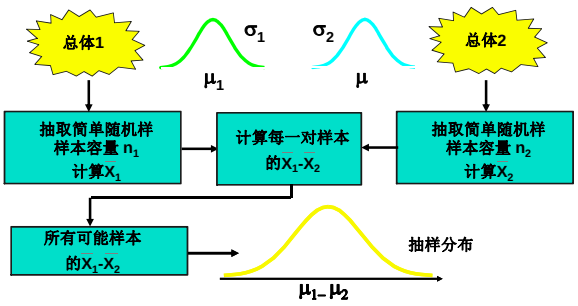
第四节　两个总体均值及两个总体比例之差估计

一．两个总体均值之差估计

二．两个总体比例之差估计

两个总体均值之差的估计

两个样本均值之差的抽样分布



两个总体均值之差的估计
(σ_1^2、σ_2^2 已知)

1. 假定条件
   - 两个样本是独立的随机样本
   - 两个总体都服从正态分布
   - 若不是正态分布，可以用正态分布来近似($n_1\geq30$和$n_2\geq30$)
2. 两个独立样本均值之差的抽样分布服从正态分布，其期望值为

$$E(\bar{x}_1-\bar{x}_2)=\mu_1-\mu_2$$

其标准误差为

$$\sigma_{(\bar{x}-\bar{x}_2)}=\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

两个总体均值之差的估计
(σ_1^2、σ_2^2 已知)

3. 使用正态分布统计量Z

$$Z=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0,1)$$

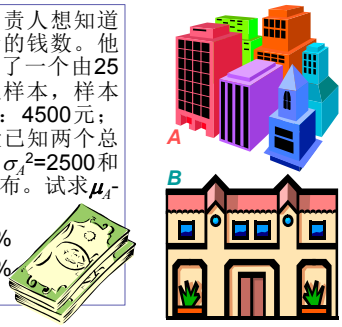
4. 两个总体均值之差$\mu_1-\mu_2$在$1-\alpha$ 置信水平下的置信区间为

$$(\bar{x}_1-\bar{x}_2)\pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

两个总体均值之差的估计

【例】一个银行负责人想知道储户存入两家银行的钱数。他从两家银行各抽取了一个由25个储户组成的随机样本，样本均值如下：银行A：4500元；银行B:3250元。设已知两个总体服从方差分别为σ_A^2=2500和σ_B^2=3600的正态分布。试求μ_A-μ_B的区间估计

- （1）置信度为95%
- （2）置信度为99%



两个总体均值之差的估计

解:已知

$X_A \sim N(\mu_A, 2500)$

$X_B \sim N(\mu_B, 3600)$

$\bar{x}_A$=4500，

$\bar{x}_B$=3250，

σ_A^2 =2500

σ_B^2 =3600

$n_A = n_B$ =25

(1) μ_A- μ_B置信度为95%的置信区间为

$$(\bar{x}_1 - \bar{x}_2) \pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(4500-3250) \pm 1.96\sqrt{\frac{2500}{25} + \frac{3600}{25}}$$

$$(1219.78, 1280.62)$$

两个总体均值之差的估计
(σ_1^2、σ_2^2未知，但相等)

1. 假定条件
   - 两个总体都服从正态分布
   - σ_1^2、σ_2^2未知，但$\sigma_1^2=\sigma_2^2$
2. 总体方差σ^2的联合估计量为

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

3. 估计量 $\bar{x}_1 - \bar{x}_2$的标准差为

$$\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

两个总体均值之差的估计
(σ_1^2、σ_2^2未知，但相等)

- 使用 t 分布统计量

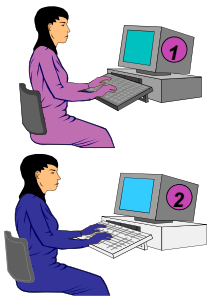
$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1+n_2-2)$$

- 两个总体均值之差μ_1-μ_2在1-α 置信水平下的置信区间为

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

两个总体均值之差的估计

【例】为比较两位银行职员为新顾客办理个人结算账目的平均时间长度，分别给两位职员随机安排了10位顾客，并记录下为每位顾客办理账单所需的时间（单位：分钟），相应的样本均值和方差分别为： $\bar{x}_1$=22.2，s_1^2=16.63，$\bar{x}_2$=28.5，s_2^2=18.92。假定每位职员办理账单所需时间均服从正态分布，且方差相等。试求两位职员办理账单的服务时间之差的95%的区间估计。



两个总体均值之差的估计

解:已知

$X_1 \sim N(\mu_1, \sigma^2)$

$X_2 \sim N(\mu_2, \sigma^2)$

$\bar{x}_1$=22.2，

$\bar{x}_2$=28.5，

s_1^2=16.63

s_2^2=18.92

$n_1 = n_2$=10

$\sigma_1^2 = \sigma_1^2$

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

$$= \sqrt{\frac{(10-1)16.36 + (10-1)18.92}{10+10-2}} = 4.2$$

μ_1- μ_2置信度为95%的置信区间为

$$(22.2 - 28.5) \pm 2.1 \times 4.2 \times \sqrt{1/10 + 1/10}$$

$$= (-10.2, -2.4)$$

两个总体均值之差的估计
(σ_1^2 、σ_2^2未知，且不相等)

1. 假定条件
   - 两个总体都服从正态分布
   - σ_1^2、σ_2^2未知，且$\sigma_1^2 \neq \sigma_2^2$
2. 使用的统计量为

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t(df)$$

$$df = \frac{1}{\frac{k^2}{df_1} + \frac{(1-k)^2}{df_2}} \quad \text{其中 } k = \frac{s_1^2/n_1}{s_1^2/n_1 + s_2^2/n_2}$$

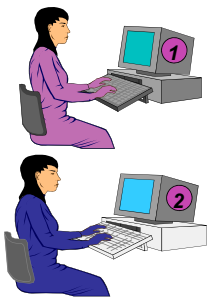
两个总体均值之差的估计
(σ_1^2、σ_2^2未知，且不相等)

- 两个总体均值之差μ_1-μ_2在1-α 置信水平下的置信区间为

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}(df)\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

两个总体均值之差的估计

【例】为比较两位银行职员为新顾客办理个人结算账目的平均时间长度，分别给两位职员随机安排了10位顾客，并记录下了为每位顾客办理账单所需的时间（单位：分钟），相应的样本均值和方差分别为：x_1=22.2，s_1^2=16.63，x_2=28.5，s_2^2=18.92。假定每位职员办理账单所需时间均服从正态分布，但**方差不相等**。试求两位职员办理账单的服务时间之差的95%的区间估计。



两个总体均值之差的估计

解:已知

$X_1 \sim N(\mu_1, \sigma^2)$

$X_2 \sim N(\mu_2, \sigma^2)$

$\bar{x}_1$=22.2，

$\bar{x}_2$=28.5，

s_1^2=16.63

s_2^2=18.92

$n_1 = n_2$=10

$\sigma_1^2 \neq \sigma_1^2$

自由度f为

$$f = \frac{\left(\frac{16.36}{10} + \frac{18.92}{10}\right)^2}{\left(\frac{16.36}{10}\right)^2 \frac{1}{9} + \left(\frac{18.92}{10}\right)^2 \frac{1}{9}} = 17.9 \approx 18$$

μ_1- μ_2置信度为95%的置信区间为

$$(22.2 - 28.5) \pm 2.1009\sqrt{\frac{16.36}{10} + \frac{18.92}{10}}$$

$$= (-10.25, -2.35)$$

配对数据

- 实验配对设计
- 例：mCPP被认为会影响人们对食物的吸收。在一项关于mCPP减肥效果的研究中，9名适度肥胖的女性以双盲和安慰剂对照实验法给予mCPP。一些受试者先服用mCPP两周，随后的两周不服用任何有关药物，其后再服用安慰剂两周。另外一些受试者刚好相反。记录每一名受试者在每一种处理条件下减轻的体重。

mCPP减肥效果的实验结果

受检者序号	减轻的体重		
	x_1(mCPP)	x_2(安慰剂)	d=$x_1 - x_2$
1	1.1	0.0	1.1
2	1.3	-0.3	1.6
3	1.0	0.6	0.4
4	1.7	0.3	1.4
5	1.4	-0.7	2.1
6	0.1	-0.2	0.3
7	0.5	0.6	-0.1
8	1.6	0.9	0.7
9	-0.5	-2.0	1.5

$$\left(\bar{d} - t_{\frac{\alpha}{2}}\frac{S_d}{\sqrt{n}}, \quad \bar{d} + t_{\alpha/2}\frac{S_d}{\sqrt{n}}\right)$$

两个总体均值之差

【例】研究人员研究了温室植物肥料在萝卜苗生长中的效果。随机选择了一些萝卜种子作为对照组，另外一些种植在添加了颗粒肥料的铝盘中，两组的其他条件完全一致。出芽两周后植株高度如下表。

温室植物肥料对萝卜苗生长效果的数据表

对照组	3.4	4.4	3.5	2.9	2.7	2.6	3.7	2.7	2.3	2.0	1.8	2.3	2.4	2.5
(1)	1.6	2.9	2.3	2.8	2.5	2.3	1.6	1.6	3.0	2.3	3.2	2.0	2.6	2.4
肥料组	2.8	1.9	3.6	1.2	2.4	2.2	3.6	1.2	0.9	1.5	2.4	1.7	1.4	1.8
(2)	1.9	2.7	2.3	1.8	2.7	2.6	1.3	3.0	1.4	1.2	2.6	1.8	1.7	1.5

已知求肥料使得苗株增长的高度的0.95置信区间。

- 独立样本（Independent sample）

  从两个总体中，分别独立的抽取一个随机样本进行比较和研究。

- 配对样本（paired sample）

  只有一个样本，但其中的个体都要先后观测两次。

两个总体比例之差的区间估计

1. 假定条件
   - 两个总体是独立的
   - 两个总体服从二项分布
   - 可以用正态分布来近似
2. 两个总体比例之差P_1-P_2在1-α置信水平下的置信区间为

$$(\hat{p}_1-\hat{p}_2)\pm Z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$$

两个正态总体方差比的区间估计

两个正态总体方差比的区间估计

1. 用两个样本的方差比来判断
   - 如果S_1^2/S_2^2接近于1,说明两个总体方差很接近
   - 如果S_1^2/S_2^2远离1,说明两个总体方差之间存在差异

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

2. 总体方差比在1-α置信水平下的置信区间为

$$\left[\frac{s_1^2/s_2^2}{F_{1-\alpha/2}(n_1-1,n_2-1)},\frac{s_1^2/s_2^2}{F_{\alpha/2}(n_1-1,n_2-1)}\right]$$

σ_1^2/σ_2^2 的置信区间为

$$\left[\frac{s_1^2/s_2^2}{F_{1-\alpha/2}(n_1-1,n_2-1)},\frac{s_1^2/s_2^2}{F_{\alpha/2}(n_1-1,n_2-1)}\right]$$

σ_2^2/σ_1^2 的置信区间为

$$\left[\frac{s_2^2/s_1^2}{F_{1-\alpha/2}(n_2-1,n_1-1)},\frac{s_2^2/s_1^2}{F_{\alpha/2}(n_2-1,n_1-1)}\right]$$

$$\left[\frac{F_{\alpha/2}(n_1-1,n_2-1)}{s_1^2/s_2^2},\frac{F_{1-\alpha/2}(n_1-1,n_2-1)}{s_1^2/s_2^2}\right]$$

$$F_{1-\alpha/2}(n_1-1,n_2-1)=\frac{1}{F_{\alpha/2}(n_2-1,n_1-1)}$$

两个正态总体方差比的区间估计

【例】用某一特定工序生产的一批化工产品中的杂质含量的变异依赖于操作过程中处理的时间长度。某生产商拥有两条生产线，为了降低产品中杂质平均数量的同时降低杂质的变异，对两条生产线进行了很小的调整，研究这种调整是否确能达到目的。为此从两条生产线生产的两批产品中各随机抽取了25个样品，它们的均值和方差为

- $\bar{x}_1=3.2$，$S_1^2=1.04$
- $\bar{x}_2=3.0$，$S_2^2=0.51$
- 试确定两总体方差比σ_1^2/σ_2^2的90%的置信区间。

两个正态总体方差比的区间估计

解:已知

$\bar{x}_1=3.2$，$S_1^2=1.04$

$\bar{x}_2=3.0$，$S_2^2=0.51$

$F_{1-\alpha/2}(24, 24)$
$=F_{0.95\,2}(24, 24)$
$=1.98$

$F_{\alpha/2}(24, 24)$
$=F_{0.052}(24, 24)$
$=0.51$

σ_1^2/σ_2^2置信度为90%的置信区间为

$$\frac{1.04}{0.51}\cdot\frac{1}{1.98}\le\frac{\sigma_1^2}{\sigma_2^2}\le\frac{1.04}{0.51}(1.98)$$

$$1.03\le\frac{\sigma_1^2}{\sigma_2^2}\le 4.04$$

小结

- **置信度95%仅仅描述用来构造该区间上下界的(随机)统计量覆盖总体参数的概率；**
- **置信区间的论述是由区间和置信度两部分组成。在公布调查结果时应给出样本数。**
- **区间估计的方法来自于相应的抽样分布**

估计方法有效性的条件

- 样本必须近似来自于总体中的随机样本
- 使用$\frac{\sigma}{\sqrt{n}}$的两个条件：
  - 总体容量必须比样本容量大（有时使用“有限总体校正系数”）
  - 观察值彼此之间相互独立
- 置信区间有效性的条件：总体分布的非正态性的程度

给定条件下，检查假定条件是否合理

结 束