补充：概率论与数理统计基础

经济学院：熊维勤

mada55@163.com

QQ：404304805

目标

- 了解概率论和数理统计学中的基本概念
- 掌握计量经济学中常用的几个分布及其性质和特征：正态分布、$t-$分布、$F-$分布、χ^2-分布
- 掌握随机变量的数字特征（数学期望、方差、协方差）及其性质
- 重点：掌握参数估计和假设检验的基本方法

第一节 概率论基础

一、正态分布及其特征

二、随机变量的数字特征

一、随机变量与确定性变量

随机变量：取值不定，随试验（或抽样）的变化而变化。

例：掷骰子的点数X（取值随试验的变化而变化）

二班同学的身高X（取值随抽样的变化而变化）

确定性变量：取值固定（在重复或抽样之中取固定值）。

例：张三同学的身高X（在重复抽样之中取固定值）

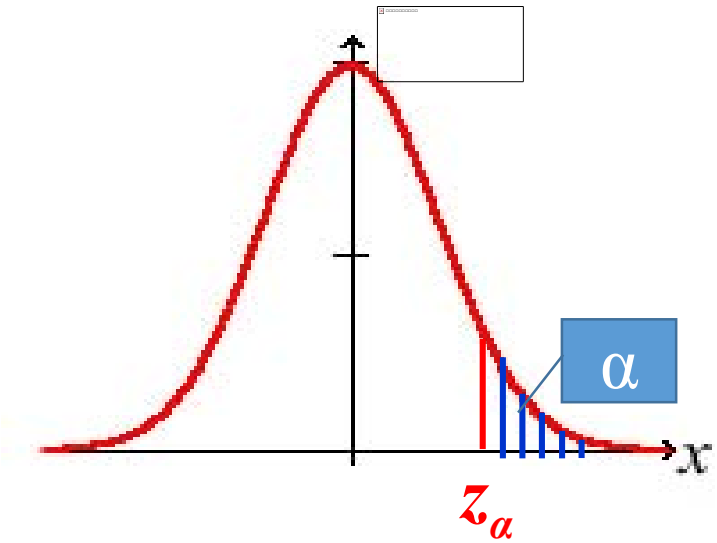
二、标准正态分布的α分位数

定理1 若 $X \sim N(\mu,\sigma^2)$, 则 $Z=\dfrac{X-\mu}{\sigma} \sim N(0,1)$.

1、标准正态分布的上α分位数

设 $X \sim N(0,1)$, 若数 z_α 满足条件 $P\{X > z_\alpha\}=\alpha$, $0<\alpha<1$

则称 z_α 为 标准正态分布的上 α 分位数.



例5：查表试求 $z_{0.05}$

解：由 $P\{X > z_{0.05}\}=0.05$

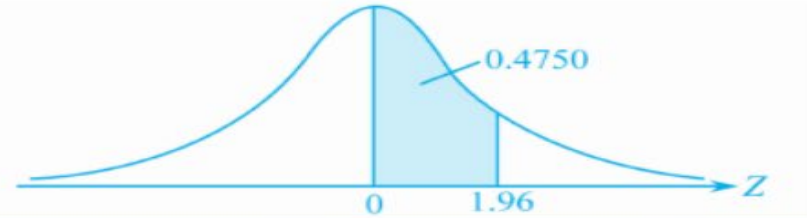
$$\Rightarrow 1-P\{X \le z_{0.05}\}=0.05 \Rightarrow P\{X \le z_{0.05}\}=0.95$$

$$\Rightarrow \Phi(z_{0.05})=0.95 \Rightarrow z_{0.05}=1.65$$

例：$p(0 \leqslant Z \leqslant 1.96) = 0.4750$

$p(Z \geqslant 1.96) = 0.5 - 0.4750$

$= 0.025$



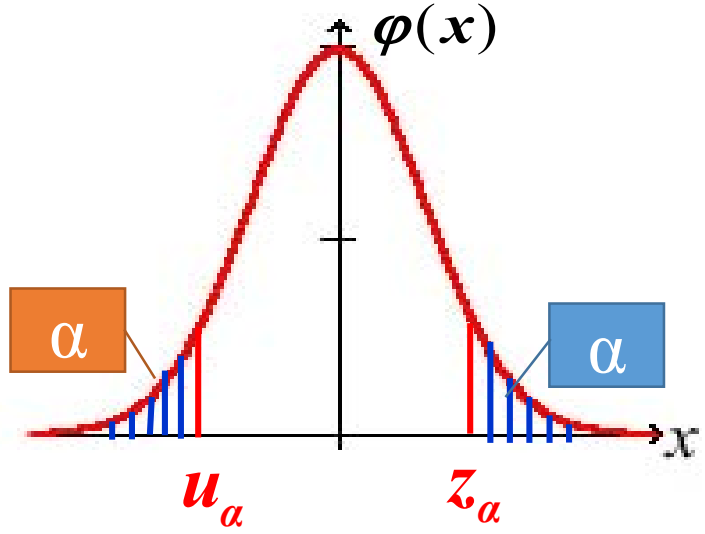
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4454	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	[illegible]	0.4803	0.4808	0.4812	0.4817

二、标准正态分布的α分位数

2、标准正态分布的下α分位数

设 $X \sim N(0,1)$，若数 u_α 满足条件 $P\{X < u_\alpha\} = \alpha, 0 < \alpha < 1$

则称 u_α 为标准正态分布的下 α 分位数.



根据对称性，显然有 $u_\alpha = -z_\alpha$，即：

标准正态分布的上、下α分位数互为相反数。

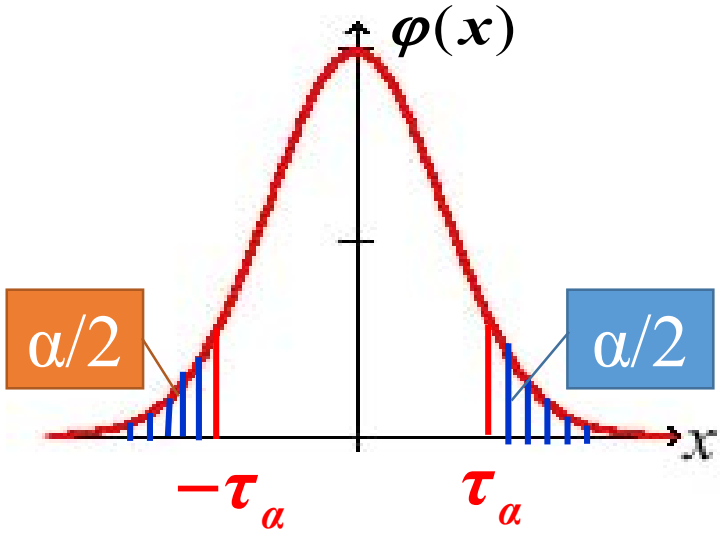
二、标准正态分布的α分位数

3、标准正态分布的双侧α分位数

设 $X \sim N(0,1)$，若数 τ_α 满足条件 $P\{|X| > \tau_\alpha\} = \alpha$，$0 < \alpha < 1$

则称 τ_α 为标准正态分布的双侧 α 分位数.

由于 $P\{|X| > \tau_\alpha\} = P\{X > \tau_\alpha\} + P\{X < -\tau_\alpha\} = \alpha$



故 $\tau_\alpha = z_{\alpha/2}$，即：

标准正态分布的双侧α分位数

等于其上α /2分位数。

例：$p(0 \leqslant Z \leqslant 1.96) = 0.4750$

$p(Z \geqslant 1.96) = 0.5 - 0.4750$

$= 0.025$

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4454	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817

二、随机变量的数字特征

1、数学期望

2、方差

3、协方差和相关系数

（一）数学期望的性质

1. 设C是常数，则$E(C)=C$;

2. 若k是常数，则$E(kX)=kE(X)$;

3. $E(X+Y)=E(X)+E(Y)$;

推广：$E[\sum_{i=1}^{n}X_i]=\sum_{i=1}^{n}E(X_i)$

请注意：
由$E(XY)=E(X)E(Y)$不一定能推出X,Y独立

4. 设X、Y相互独立，则 $E(XY)=E(X)E(Y)$;

推广：$E[\prod_{i=1}^{n}X_i]=\prod_{i=1}^{n}E(X_i)$ （诸X_i相互独立）

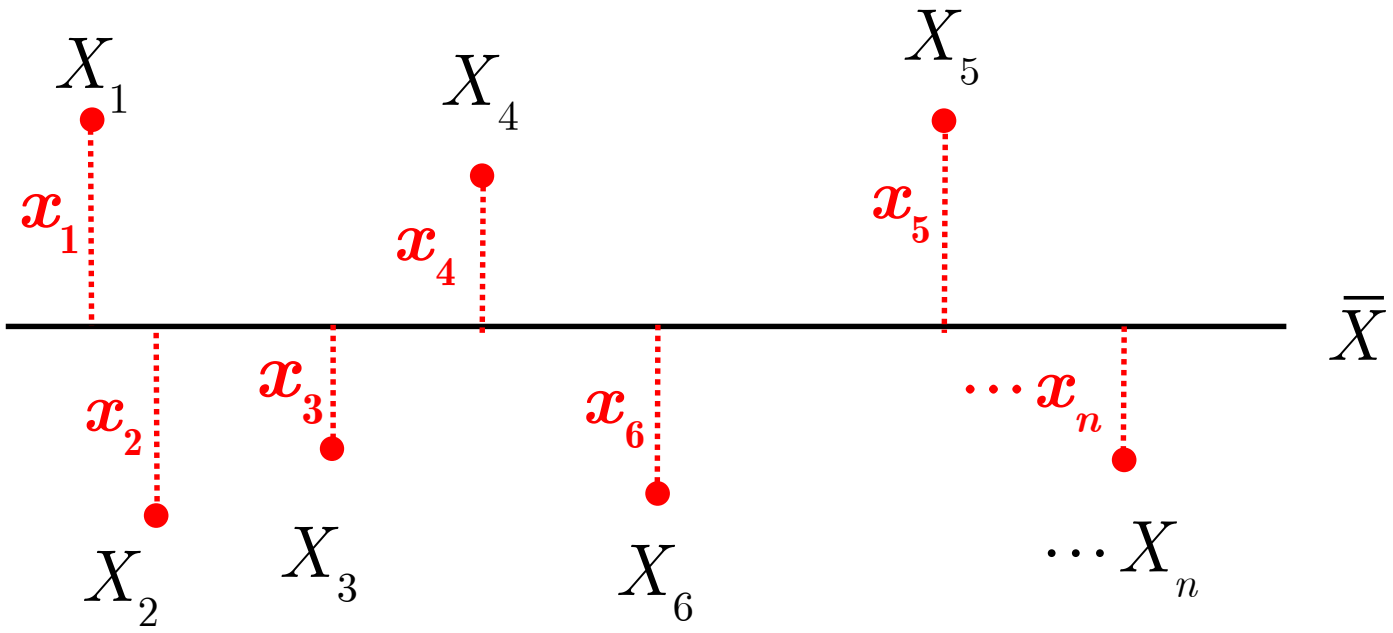
（二）随机变量的方差

1、方差的定义

设X是一个随机变量，若$E[(X-E(X)]^2$存在，称$E[(X-E(X)]^2$为 X 的方差. 记为$D(X)$或$Var(X)$，即

$$D(X)=Var(X)=E[X-E(X)]^2$$

方差的算术平方根$\sqrt{D(X)}$称为X的标准差或均方差记为$\sigma(X)$，它与X具有相同的量纲。



$x_i = X_i - \bar{X}$称为离差，它衡量了随机变量X的某个取值X_i与其均值$\bar{X}$的偏离程度

离差有正有负，且$\sum_{i=1}^{n} x_i = 0$

令$S = \dfrac{\sum_{i=1}^{n} |x_i|}{n} = \dfrac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n}$，则$S$衡量了随机变量$X$的取值偏离其均值$\bar{X}$的平均偏离程度(或围绕均值的平均波动程度)。

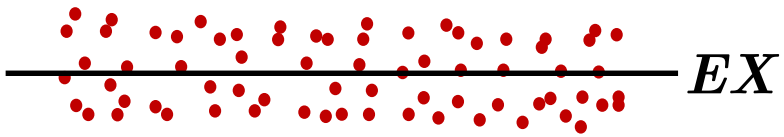
由于绝对值在数学上处理困难，因此改用平方消除x_i 的负号

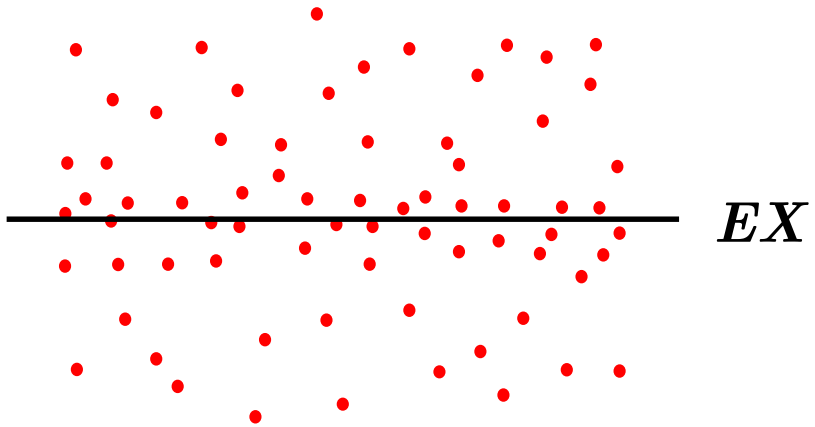
令$D(X)=\frac{\sum_{i=1}^{n}x_i^2}{n}=\frac{\sum_{i=1}^{n}(X_i-X)^2}{n}=E[X-E(X)]^2$，则$D(X)$就是方差的定义式

故方差的含义是：

度量了随机变量的取值偏离其均值的平均偏离程度(离中趋势)

或：随机变量的取值围绕其均值的波动性





2、计算方差的一个简化公式

$$D(X)=E(X^2)-[E(X)]^2$$

证明：由方差的定义式 $D(X)=E\left[X-E(X)\right]^2$

$$\Rightarrow D(X)=E\left\{X^2-2XE(X)+\left[E(X)\right]^2\right\}$$

$$=E(X^2)-2E(X)E(X)+\left[E(X)\right]^2$$

$$=E(X^2)-\left[E(X)\right]^2$$

3、方差的性质

1）设C是常数, 则 $D(C)=0$;

2）若 C 是常数, 则 $D(CX)=C^2\,D(X)$;

3）设 X 与 Y 是两个随机变量，则

$$D(X\pm Y)= D(X)+D(Y)\pm 2E\{[X-E(X)][Y-E(Y)]\}$$

$$= D(X)+D(Y) \pm 2Cov(X,Y)$$

（三）协方差和相关系数

1.定义： $E\{[X-E(X)][Y-E(Y)]\}$称为随机变量X和Y的协方差,记为$Cov(X,Y)$，即：$Cov(X,Y)=E\{[X-E(X)][Y-E(Y)]\}$

2.性质：

(1) $Cov(X,Y)=Cov(Y,X)$

(2) $Cov(aX,bY)=ab\,Cov(X,Y)$　a,b 是常数

(3) $Cov(X_1+X_2,Y)=Cov(X_1,Y)+Cov(X_2,Y)$

计算协方差的一个简单公式

由协方差的定义及期望的性质，可得

$$Cov(X,Y)=E(XY)-E(X)E(Y)$$

可见，若X与Y独立，$Cov(X,Y)=0$.

特别地

$$Cov(X,X)=E(X^2)-[E(X)]^2=D(X)$$

$$D(X\pm Y)=D(X)+D(Y)\pm 2Cov(X,Y)$$

二、相关系数

定义： 设$D(X)>0, D(Y)>0,$ 称

$$\rho_{XY}=\frac{Cov(X,Y)}{\sqrt{D(X)D(Y)}}$$

为随机变量 X 和 Y 的相关系数 .

在不致引起混淆时，记 ρ_{XY} 为 ρ

相关系数的性质：

1. $|\rho| \le 1$

2. X和Y独立时，$\rho = 0$，但其逆不真.

3. $|\rho| = 1 \Longleftrightarrow$ 存在常数 $a,b(b \neq 0)$，

使 $P\{Y = a + bX\} = 1$,

即 X 和 Y 以概率 1 线性相关.

正态分布的重要性质

定理： N个独立正态随机变量之和仍然服从正态分布

即若$X_i \sim N(\mu_i, \sigma_i^2) \quad i=1,2,\cdots n,$ 且X_i相互独立，

则：$\sum_{i=1}^{n} X_i \sim N(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2)$

推广： N个独立正态随机变量的任意线性组合仍然服从正态分布

即若$X_i \sim N(\mu_i, \sigma_i^2) \quad i=1,2,\cdots n,$ 且X_i相互独立，

则：$\sum_{i=1}^{n} \alpha_i X_i \sim N(\sum_{i=1}^{n} \alpha_i \mu_i, \sum_{i=1}^{n} \alpha_i^2 \sigma_i^2)$

第二节 数理统计基础

一、抽样分布

二、参数估计

三、假设检验

一、样本及其抽样分布

总体：$X\sim N(\mu,\sigma^2)$

抽样

独立性（等概原则）

代表性：$X_i\sim N(\mu,\sigma^2)$

样本：$X_1,\ X_2,\ \cdots,\ X_n$ ⇨ 统计量：$f(X_1, X_2,\cdots, X_n)$

测量

代入

样本观察值：$x_1,\ x_2,\ \cdots,\ x_n$　统计量值：$f(x_1, x_2,\cdots, x_n)$

一、样本及其抽样分布

Ⅰ、χ^2 分布

定义: 设$X_1, X_2, \cdots, X_n$ 相互独立，都服从标准正态分布$N(0,1)$, 则称随机变量：

$$\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$$

所服从的分布为自由度为 n 的 χ^2 分布

记为 $\chi^2 \sim \chi^2(n)$

自由度（degree of freedom, df）：是指在计算样本的统计量值时，样本中独立或能自由变化的数据的个数称为该统计量的自由度

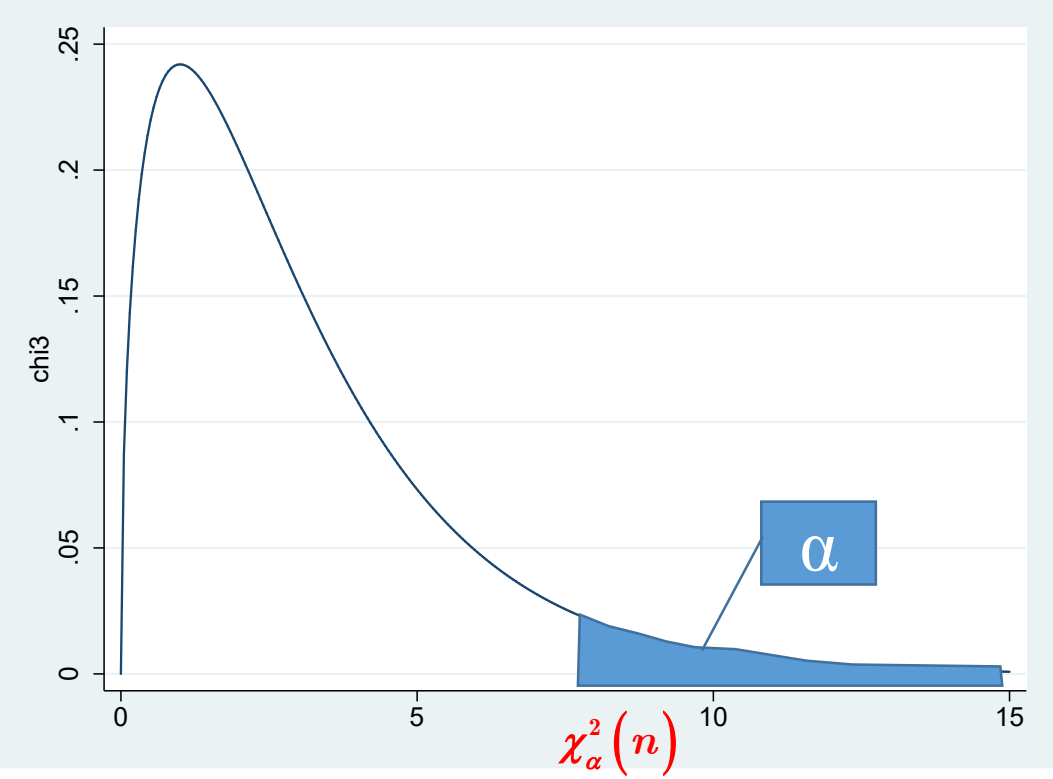
例如在计算样本均值统计量$\bar{X}=\frac{\sum_{i=1}^{n}X_i}{n}$时，变量$X$可以自由赋值的个数为$n$，因此均值统计量$\bar{X}$的自由度即为$n$

而在计算样本方差统计量$S^2=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)}{n-1}$时，因为$\bar{X}$在计算$S^2$ 时是给定的，故此时X可以自由赋值的个数为$n-1$，即方差统计量S^2的自由度即为$n-1$

在计算样本统计量时，若加入了k个约束条件，则其自由度就相应减少k

χ^2分布的上α分位数

对给定的正数$\alpha \in (0,1)$，称满足$P\{\chi^2 > \chi_\alpha^2(n)\} = \alpha$的数$\chi_\alpha^2(n)$为$\chi^2(n)$分布的上$\alpha$分位数



例如：

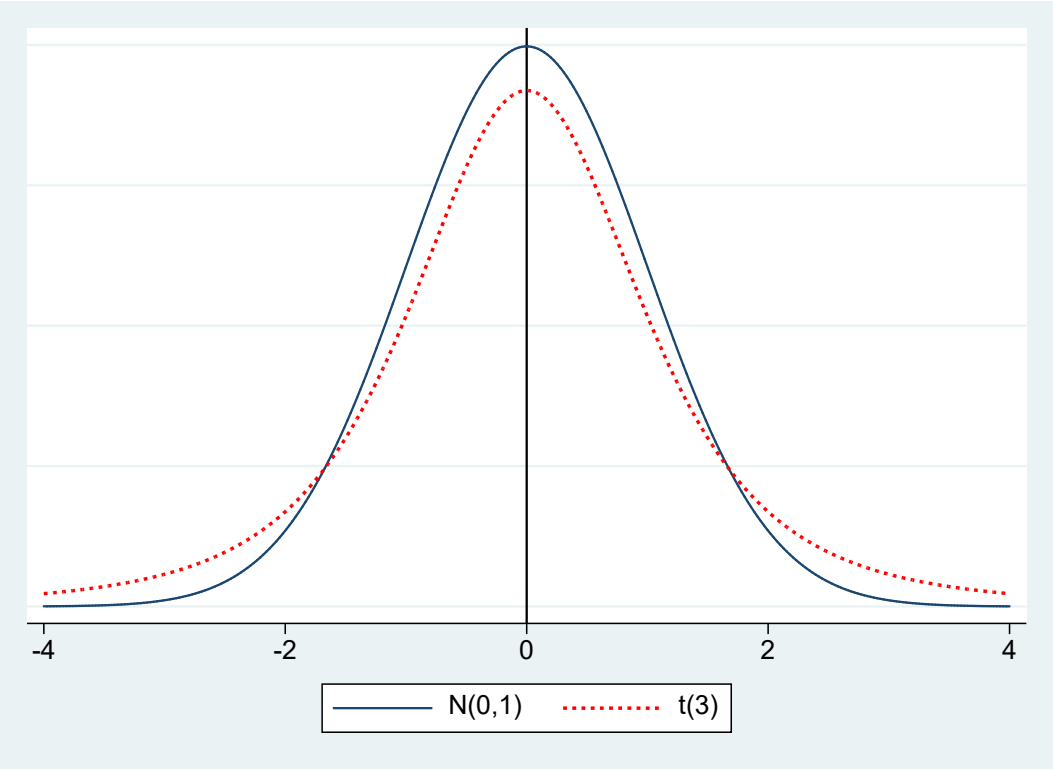
$$\chi_{0.05}^2(3) = 7.815$$

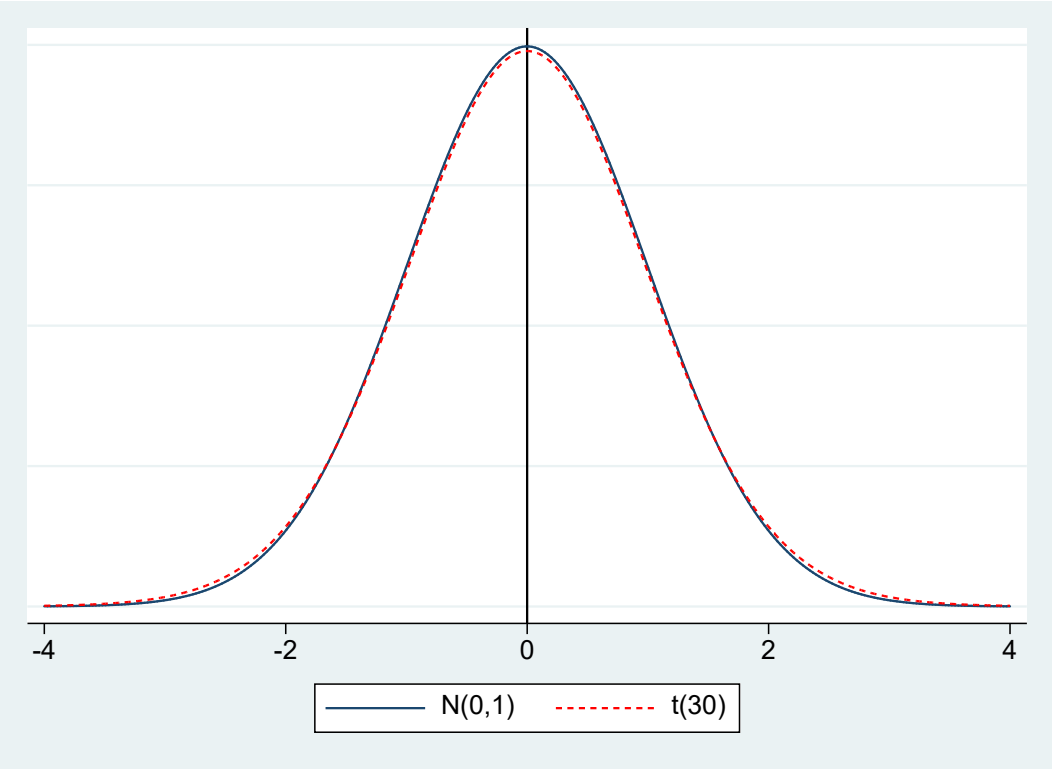
n	0.995	0.99	0.975	0.95	0.90	0.75	0.25	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	0.102	1.323	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	0.575	2.773	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	1.213	4.108	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	1.923	5.385	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	2.675	6.626	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	3.455	7.841	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	4.255	9.037	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	5.071	10.219	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	5.899	11.389	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	6.737	12.549	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	7.584	13.701	17275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	8.438	14.845	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	9.299	15.984	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	10.165	17.117	21.064	23.685	16.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	11.037	18.245	22.307	24.966	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	11.912	19.369	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	12.792	20.489	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	13.675	21.605	25.989	28.869	31.526	34.805	37.156

Ⅱ、t 分布

定义：设$X \sim N(0,1)$，$Y \sim \chi^2(n)$，且X与Y相互独立，则称变量

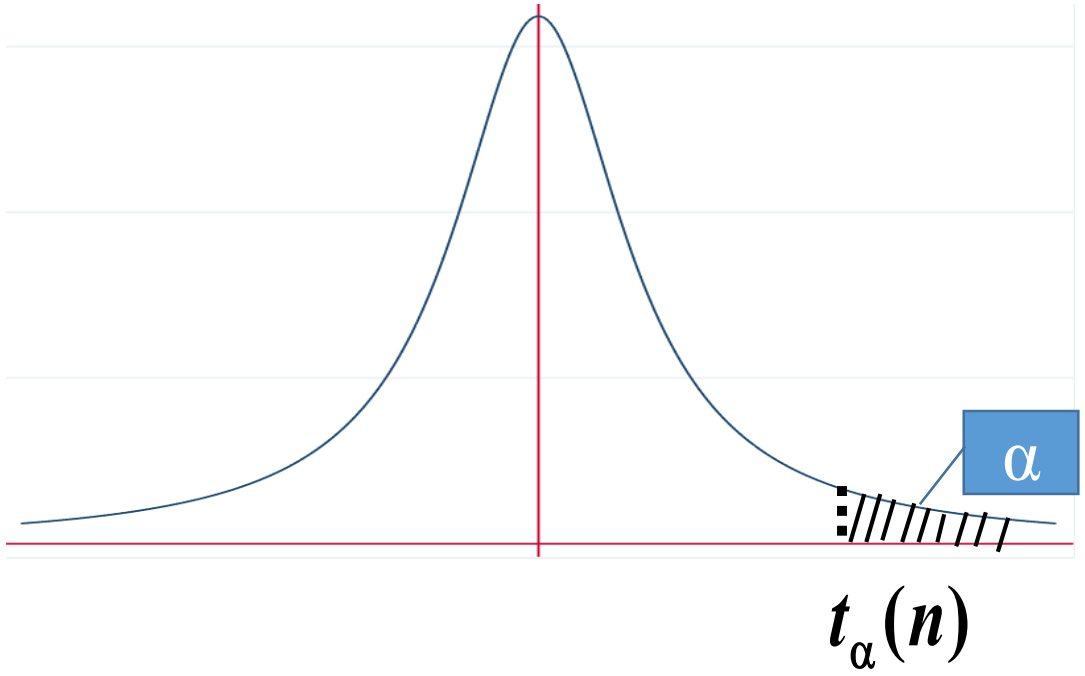
$t = \dfrac{X}{\sqrt{Y/n}}$服从自由度为$n$的$t$分布，记为$t \sim t(n)$





t分布的上α分位数：

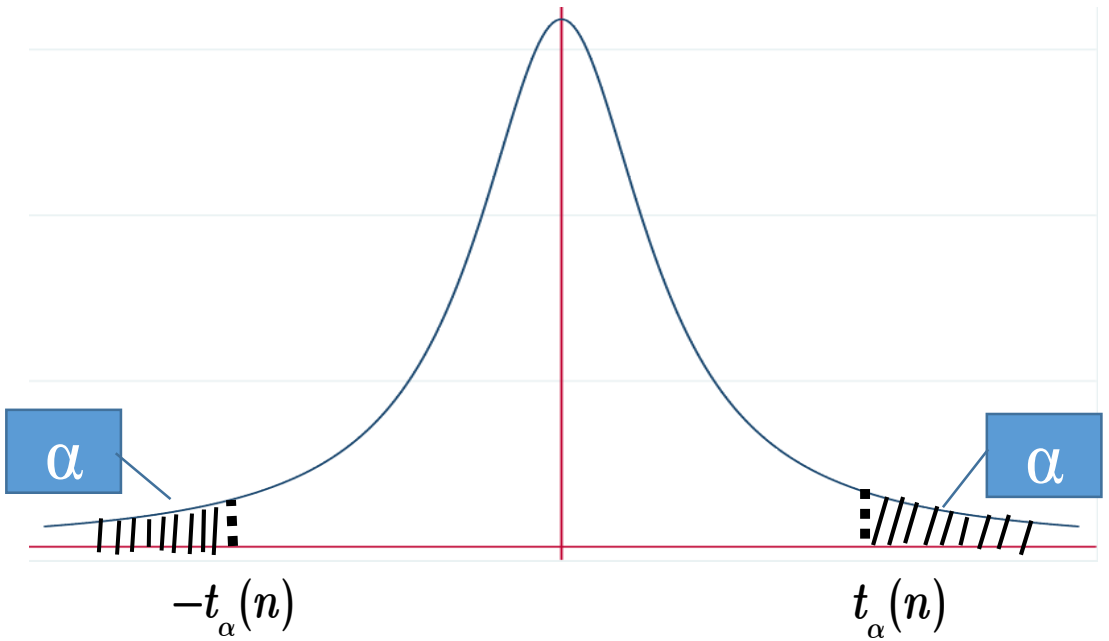
对给定的正数$\alpha \in (0,1)$，称满足$P\{t > t_\alpha(n)\} = \alpha$的数$t_\alpha(n)$为$t(n)$分布的上$\alpha$分位数

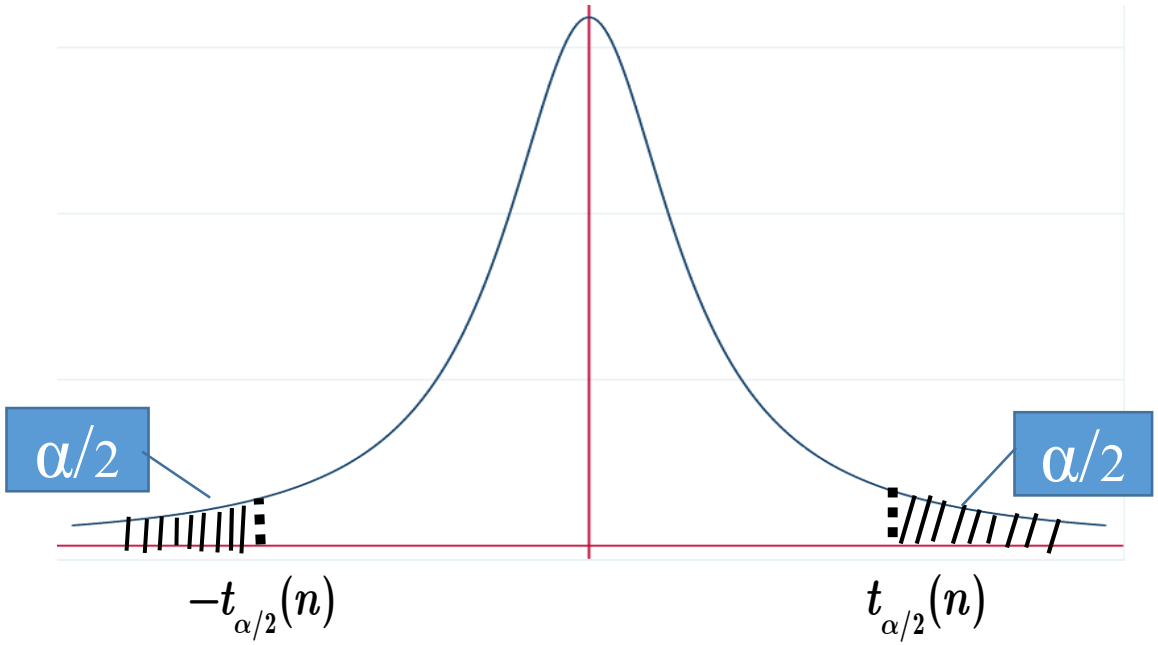


由t分布的对称性可知：

1）t分布上、下α分位数互为相反数

2）t分布双侧α分位数等于其上$\alpha/2$分位数





例$t_{0.025}(10)=2.228$

p_r / df	0.25	0.10	0.05	0.025	0.01	0.005	0.001
	0.50	0.20	0.10	0.05	0.02	0.010	0.002
1	1.000	3.078	6.314	12.706	31.821	63.657	318.31
2	0.816	1.886	2.920	4.303	6.965	9.925	22.327
3	0.765	1.638	2.353	3.182	4.541	5.841	10.214
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	0.718	1.440	1.943	2.447	3.143	3.707	5.208
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785
8	0.706	1.397	1.860	2.306	2.896	3.355	4.501
9	0.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.700	1.372	1.812	2.228	2.764	3.169	4.144
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025
12	0.695	1.356	1.782	2.179	2.681	3.055	3.930
13	0.694	1.350	1.771	2.160	2.650	3.012	3.852
14	0.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	0.690	1.337	1.746	2.120	2.583	2.921	3.686
17	0.689	1.333	1.740	2.110	2.567	2.898	3.646
18	0.688	1.330	1.734	2.101	2.552	2.878	3.610
19	0.688	1.328	1.729	2.093	2.[illegible]	2.861	3.579
20	0.687	1.325	1.725	2.086	2.528	2.845	3.552

当$n>45$时，对于常用的α 的值，可用正态近似$t_\alpha(n)\approx z_\alpha$

Ⅲ、F分布

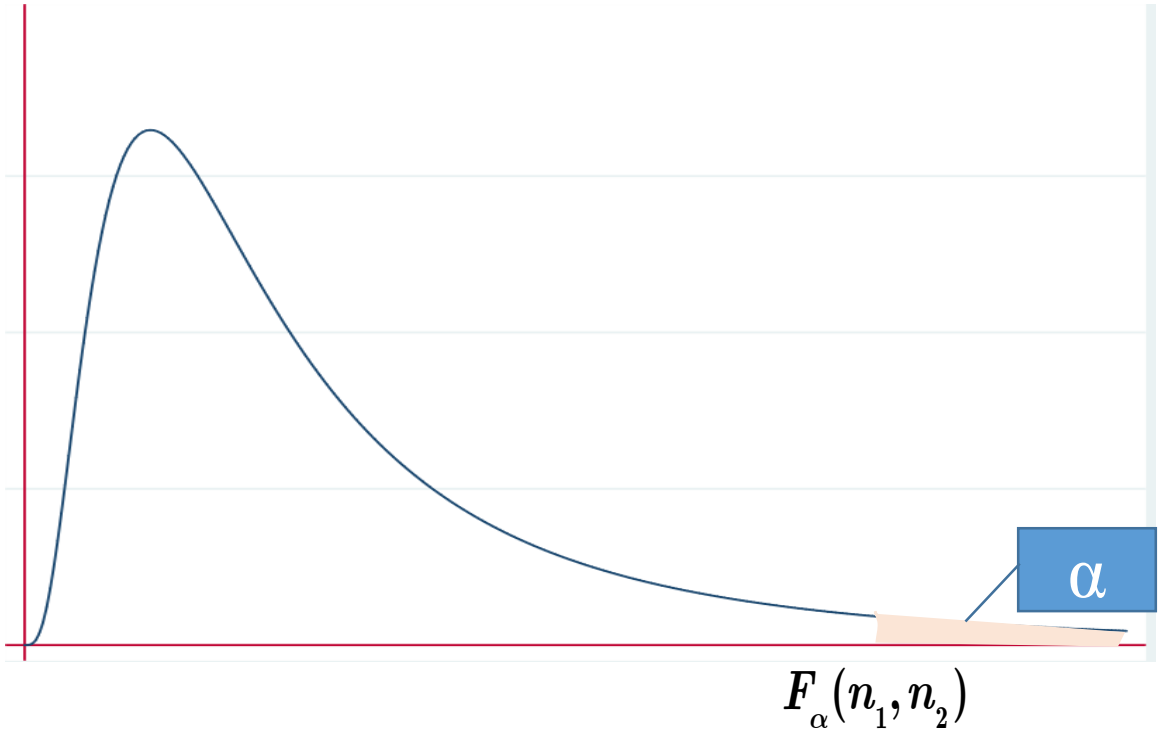
定义：设$U \sim \chi^2(n_1)$，$V \sim \chi^2(n_2)$且U、V相互独立，则称随机变量$F = \dfrac{U/n_1}{V/n_2}$服从第一自由度为n_1，第二自由度为n_2的F分布，记为$F \sim F(n_1, n_2)$

1、$\dfrac{1}{F} = \dfrac{V/n_2}{U/n_1} \sim F(n_2, n_1)$

2、$t^2 = \dfrac{X^2}{Y/n} \sim F(1, n)$

F分布的上α分位数

对给定的正数$\alpha \in (0,1)$，称满足$P\{F > F_\alpha(n_1,n_2)\} = \alpha$的数$F_\alpha(n_1,n_2)$为$F(n_1,n_2)$分布的上$\alpha$分位数



例：$F_{0.05}(9,18)=2.46$

分母自由度 N_2	p_r	分子自										
		1	2	3	4	5	6	7	8	9	10	11
17	0.25	1.42	1.51	1.50	1.49	1.47	1.46	1.45	1.44	1.43	1.43	1.42
	0.10	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.98
	0.05	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41
	0.01	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52
18	0.25	1.41	1.50	1.49	1.48	1.46	1.45	1.44	1.43	1.42	1.42	1.41
	0.10	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.96
	0.05	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37
	0.01	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43
19	0.25	1.41	1.49	1.49	1.47	1.46	1.44	1.43	1.42	1.41	1.41	1.40
	0.10	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.94
	0.05	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34
	0.01	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36

3、几个重要的抽样分布定理

设总体X的均值为μ，方差为σ^2，$X_1, X_2, \cdots, X_n$是来自总体的一个样本，则样本均值$\bar{X}$和样本方差S^2有：

$$E(\bar{X}) = \mu, \quad D(\bar{X}) = \frac{\sigma^2}{n}, \quad E(S^2) = \sigma^2 \left| \text{注：} S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \right.$$

证明：$\bar{X} = \dfrac{\sum_{i=1}^{n} X_i}{n} \Rightarrow E\bar{X} = E\left(\dfrac{\sum_{i=1}^{n} X_i}{n}\right) = \dfrac{E\left(\sum_{i=1}^{n} X_i\right)}{n} = \dfrac{\sum_{i=1}^{n} EX_i}{n} = \dfrac{\sum_{i=1}^{n} \mu}{n} = \mu$

$\bar{X} = \dfrac{\sum_{i=1}^{n} X_i}{n} \Rightarrow D\bar{X} = D\left(\dfrac{\sum_{i=1}^{n} X_i}{n}\right) = \dfrac{D\left(\sum_{i=1}^{n} X_i\right)}{n^2} = \dfrac{\sum_{i=1}^{n} DX_i}{n^2} = \dfrac{\sum_{i=1}^{n} \sigma^2}{n^2} = \dfrac{\sigma^2}{n}$

因$D\bar{X}=E\left(\bar{X}^2\right)-\left(E\bar{X}\right)^2 \Rightarrow E\left(\bar{X}^2\right)=D\bar{X}+\left(E\bar{X}\right)^2=\frac{\sigma^2}{n}+\mu^2$

同理：因$DX_i=E\left(X_i^2\right)-\left(EX_i\right)^2 \Rightarrow E\left(X_i^2\right)=DX_i+\left(EX_i\right)^2=\sigma^2+\mu^2$

$$\text{又}S^2=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1}=\frac{\sum_{i=1}^{n}\left(X_i^2-2X_i\bar{X}+\bar{X}^2\right)}{n-1}=\frac{\sum_{i=1}^{n}X_i^2-2\bar{X}\sum_{i=1}^{n}X_i+\sum_{i=1}^{n}\bar{X}^2}{n-1}$$

$$=\frac{\sum_{i=1}^{n}X_i^2-2n\bar{X}^2+n\bar{X}^2}{n-1}=\frac{\sum_{i=1}^{n}X_i^2-n\bar{X}^2}{n-1}$$

$$\Rightarrow ES^2=E\left(\frac{\sum_{i=1}^{n}X_i^2-n\bar{X}^2}{n-1}\right)=\frac{E\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)}{n-1}=\frac{\sum_{i=1}^{n}EX_i^2-nE\bar{X}^2}{n-1}$$

$$=\frac{\sum_{i=1}^{n}\left(\sigma^2+\mu^2\right)-n\left(\frac{\sigma^2}{n}+\mu^2\right)}{n-1}=\frac{n\left(\sigma^2+\mu^2\right)-\left(\sigma^2+n\mu^2\right)}{n-1}=\sigma^2$$

特别地，当总体为正态分布时，以下将给出几个重要的抽样分布定理

定理 1 (样本均值的分布)

设 $X_1, X_2, \ldots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$

的样本，$\overline{X}$是样本均值，则有

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

即 $$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

定理 2 (样本方差的分布)

设$X_1,X_2,\ldots,X_n$是来自正态总体 $N(\mu,\sigma^2)$ 的样本，$\overline{X}$和S^2分别为样本均值和样本方差，则有

(1) $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

(2) $\overline{X}$与S^2独立.

由于$\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1), \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$且$\bar{X}$与$S^2$独立，故由$t$分布的定义知：

$$\frac{\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\bar{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}\sqrt{\dfrac{S^2}{\sigma^2}}} = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

定理 3： 设$X_1, X_2, \ldots, X_n$是取自正态总体 $N(\mu, \sigma^2)$ 的样本，$\bar{X}$和S^2分别为样本均值和样本方差，则有：

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

总结：对正态总体$X \sim N(\mu, \sigma^2)$，其均值$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

(1)若总体方差σ^2已知，则$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

(2)若总体方差σ^2未知，则用样本方差S^2代替总体方差后有则$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

二、参数估计

- 点估计
- 估计量的评选标准
- 区间估计

1、点估计

例：已知全班同学身高$X \sim N\left(\mu, \sigma^2\right)$，总体参数$\mu, \sigma^2$未知，如何估计？

经验做法：

从中抽取n个同学作为样本，测得其身高，并利用样本均值$\bar{X} = \dfrac{\sum_{i=1}^{n} X_i}{n}$

估计总体均值μ，用样本方差$S^2 = \dfrac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}$估计总体方差$\sigma^2$。

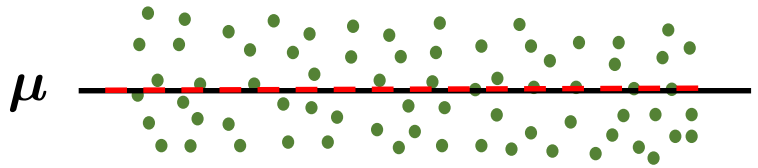
问题：$\bar{X}$、S^2是不是μ、σ^2的“好”估计量？

对一个点估计量进行评价的标准有三个：无偏性、有效性、一致性

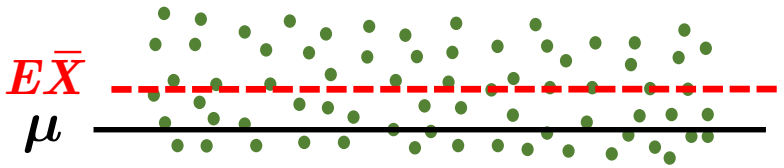
1）、无偏性

设$\bar{X}$是未知参数μ的估计量，若$E(\bar{X})=\mu$，则称$\bar{X}$是μ的无偏估计量。

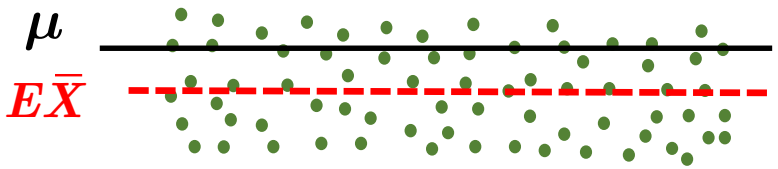
无偏估计量的含义是：当用$\bar{X}$来估计μ时，不会出现系统性偏高或偏低的情况。



$E\bar{X}=\mu$(无偏估计)



$E\bar{X}>\mu$(系统性偏高)



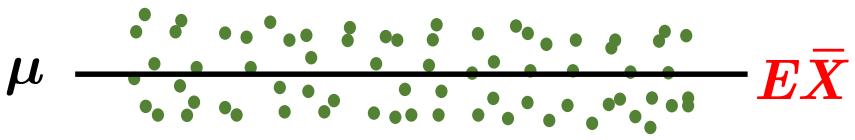
$E\bar{X}<\mu$(系统性偏低)

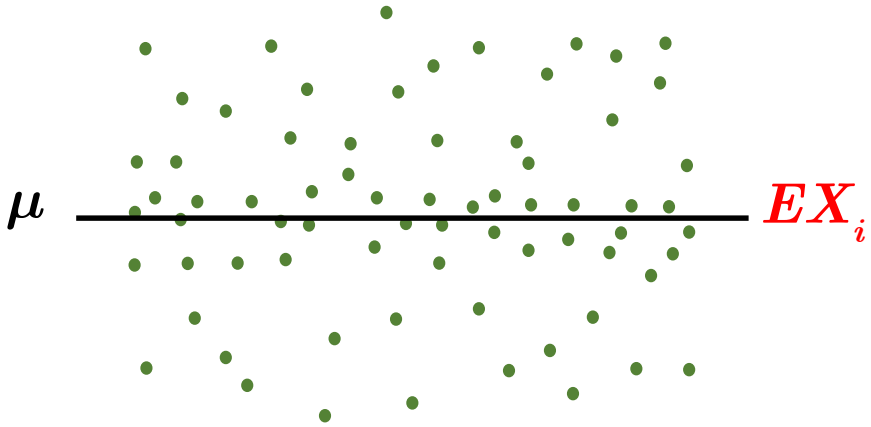
2）、有效性

若参数θ有两个无偏估计量$\tilde{\theta}_1(X_1 \cdots X_n)$、$\tilde{\theta}_2(X_1 \cdots X_n)$，且$D(\tilde{\theta}_1) < D(\tilde{\theta}_2)$，则称估计量$\tilde{\theta}_1$较$\tilde{\theta}_2$更有效。

在身高例子中，随机抽取一名同学，其身高X_i也是总体均值μ 的无偏估计量。因为$X_i \sim N(\mu, \sigma^2)$(样本与总体服从相同的分布)，故$EX_i = \mu$

但在μ 的两个无偏估计量X_i和$\bar{X}$中，因为$DX_i = \sigma^2$，而$D\bar{X} = \frac{\sigma^2}{n} < DX_i$ ，因此$\bar{X}$比X_i更有效。其含义是：用$\bar{X}$估计μ比用X_i估计μ的平均误差更小（精度更高）。





3）、一致性

无偏性和有效性都是在样本容量n固定的前提下提出来的，我们自然希望随着样本容量的增大，一个估计量的值能稳定于待估参数的真实值，因此对估计量又有下述一致性的要求：

设 $\tilde{\theta}(X_1,\ldots,X_n)$ 是参数 θ 的估计量，若对于任意$\theta \in \Theta$，当$n \to \infty$ $\tilde{\theta}(X_1,\ldots,X_n)$依概率收敛于 θ ，则称 $\hat{\theta}$ 为 θ 的一致估计量。

3、区间估计

点估计难以判断其精度和可行性。在全班同学身高例子中，我们放松要求，希望给出一个估计区间$(\underline{\theta},\overline{\theta})$，使得总体均值$\mu$ 落在该区间的概率足够大，即$P\{\underline{\theta}<\mu<\overline{\theta}\}=1-\alpha$。

区间$(\underline{\theta},\overline{\theta})$称为参数$\mu$ 的置信度为$1-\alpha$的置信区间。该区间的长度$\overline{\theta}-\underline{\theta}$刻画了估计的精度，概率$1-\alpha$刻画了估计的可靠性

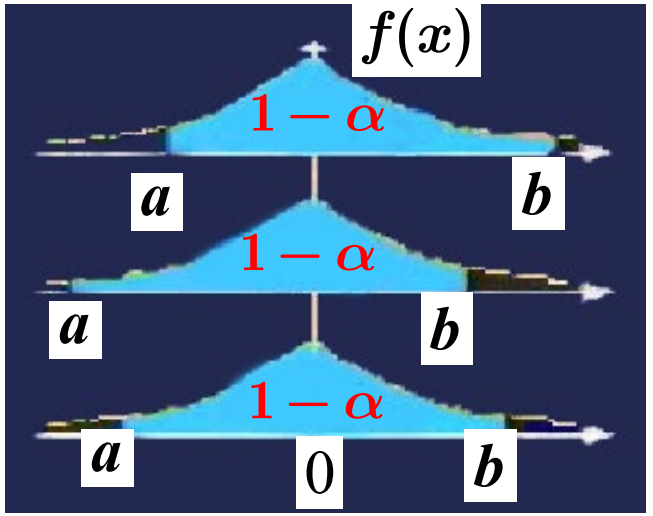
可靠性与精度是一对矛盾，一般是在保证可靠性的前提下尽可能提高精度

置信区间的求法

置信区间的求解通常始于如下概率问题：

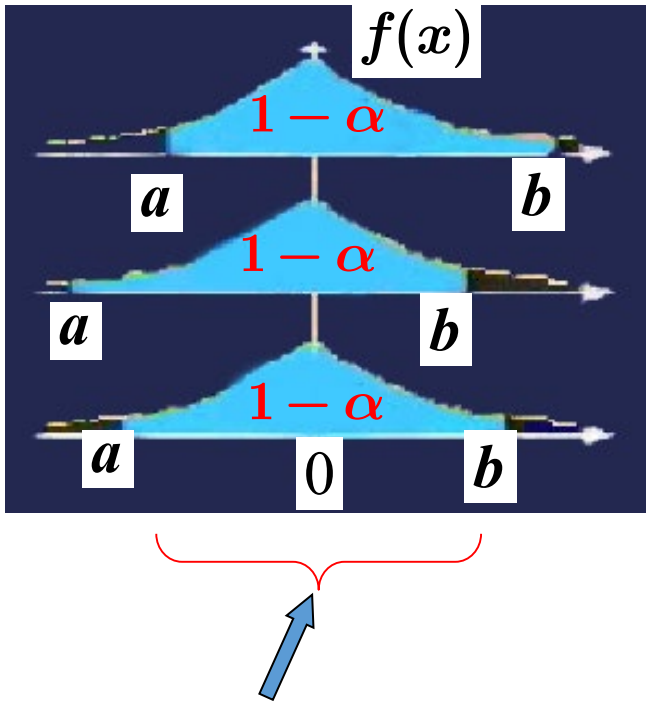
X为随机变量，给定概率$1-\alpha$，求解区间(a,b)，使得$P\{a<X<b\}=1-\alpha$

设X的概率密度函数为$f(x)$，则对任意两个数a和b，只要它们包含了$f(x)$下$1-\alpha$的面积，就确定一个解区间(a,b)，因此原问题有无穷多解。



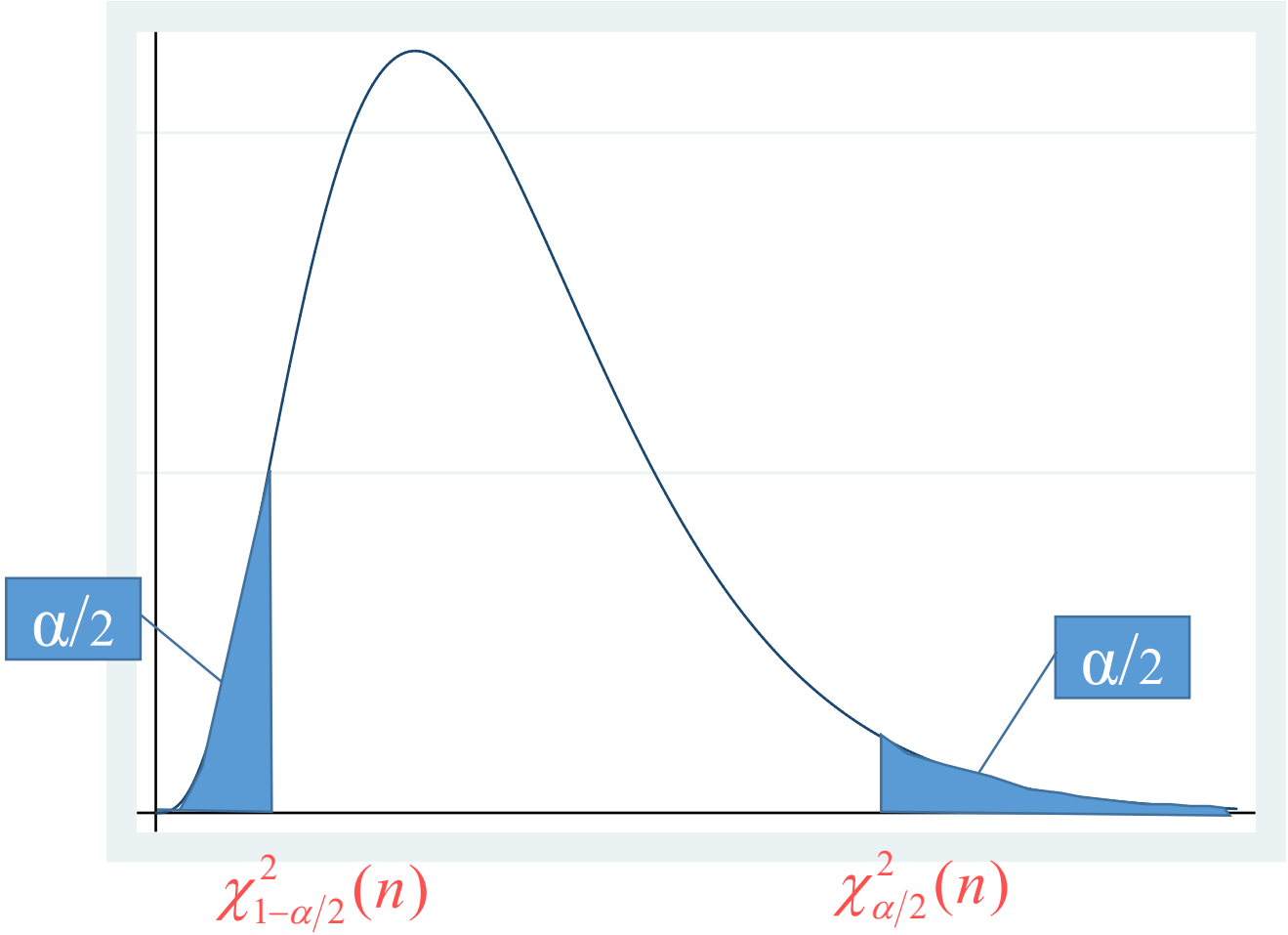
有所有可能的解中，我们希望得到精度最高（即区间长度最短）的解。

在概率密度函数为单峰且对称的情形下，当$a=-b$时求得的区间长度最短



显然，若$X\sim N(0,1)$，则$a=-z_{\alpha/2}$，$b=z_{\alpha/2}$

强制对称准则：若概率密度函数不对称，如χ^2分布，F分布，习惯上仍取其上、下$\alpha/2$分位数来计算置信区间.



单正态总体均值μ的置信区间求解

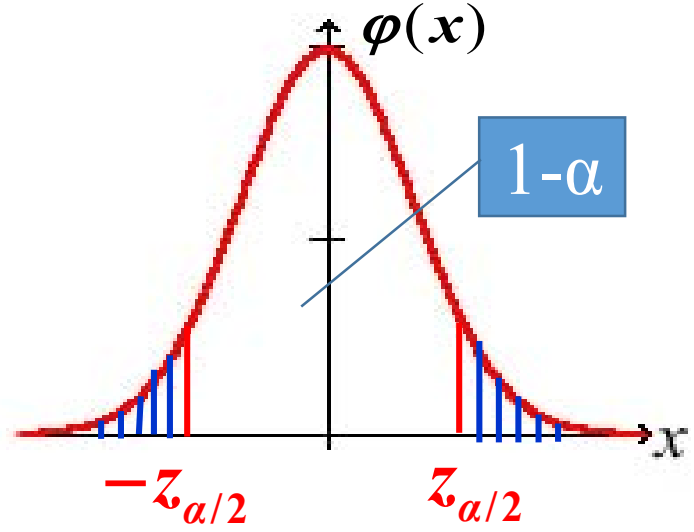
设总体$X \sim N\left(\mu,\sigma^2\right)$，$X_1, X_2,\cdots, X_n$为其一个样本，试求均值$\mu$ 的置信度为$1-\alpha$ 的置信区间。

1、若σ^2已知：

因$\bar{X} \sim N\left(\mu,\sigma^2/n\right)$，则$Z=\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$\Rightarrow P\left\{-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right\} = 1-\alpha$$

$$\Rightarrow P\left\{\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha$$



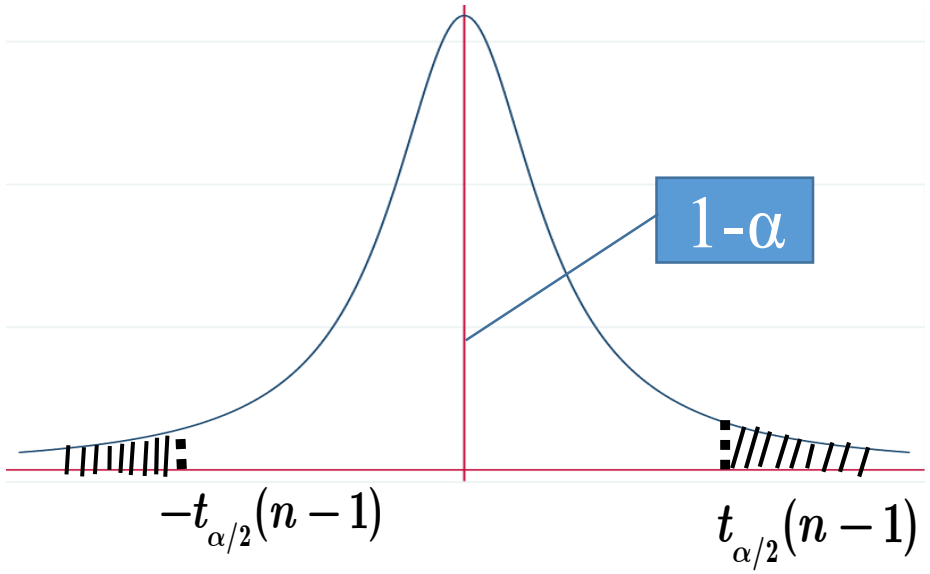
即μ的置信度为$1-\alpha$的置信区间为$\left[\bar{X} - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}, \bar{X}+z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right]$，简记为$\bar{X} \mp z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$

单正态总体均值μ的置信区间求解

2、若σ^2未知：

因$\bar{X} \sim N\left(\mu, \sigma^2/n\right)$，则$t=\dfrac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

$$\Rightarrow P\left\{-t_{\alpha/2} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2}\right\} = 1-\alpha$$



$$\Rightarrow P\left\{\bar{X}-t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X}+t_{\alpha/2}\frac{S}{\sqrt{n}}\right\} = 1-\alpha$$

即μ的置信度为$1-\alpha$的置信区间为$\left[\bar{X}-t_{\alpha/2}\dfrac{S}{\sqrt{n}}, \bar{X}+t_{\alpha/2}\dfrac{S}{\sqrt{n}}\right]$，简记为$\bar{X} \mp t_{\alpha/2}(n-1)\dfrac{\sigma}{\sqrt{n}}$

小结：单正态总体均值的置信区间

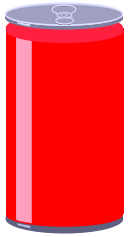
$X \sim N(\mu,\sigma^2)$，　并设 $X_1,\ldots,X_n$为来自总体的样本，$\bar{X},S^2$ 分别为样本均值和样本方差 .

均值μ的置信度为1-α的置信区间为：

	统计量	置信区间
σ^2已知	$\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$	$\overline{X} \pm z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$
σ^2未知	$\dfrac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$	$\overline{X} \pm t_{\alpha/2}(n-1)\dfrac{S}{\sqrt{n}}$

四、假设检验

例：罐装可乐生产线是否正常？



正常生产条件下，由于种种随机因素的影响，每罐可乐的容量应在355毫升上下波动，这些因素中没有哪一个占有特殊重要的地位，因此根据中心极限定理，假定总体X服从正态分布是合理的，不妨设为$X \sim N\left(\mu, \sigma^2\right)$。其中$\mu$=355且$\sigma^2$已知。

现在的问题是：如何判断罐装生产线是否正常？

一、假设检验的基本思想和方法

随机抽取n罐可乐，即$X_1,\ldots,X_n$是取自正态$N(\mu,\sigma^2)$的一个样本，样本均值$\bar{X}\sim N(\mu,\sigma^2/n)$。生产线是否正常的问题可以归结为检验如下假设：

$$H_0:\mu=\mu_0\ (=355)$$

在实际工作中，往往把不轻易否定的命题作为原假设.

它的对立假设是：

$$H_1:\ \mu\neq\mu_0$$

称H_0为原假设（或零假设）；

称H_1为备择假设（或对立假设）.

一、假设检验的基本思想和方法

那么，如何判断原假设H_0是否成立呢？

测量并计算出n罐样本可乐的平均容量$\bar{X}$并与标准容积$\mu=355$进行比较，如果$|\bar{X}-\mu|$的值足够大，我们通常会认为生产线工作不正常。

值得思考的问题是：$|\bar{X}-\mu|$足够大这一事件的发生是否必然意味着生产线工作不正常？或者说，$|\bar{X}-\mu|$足够大这一事件是不是一定是生产线工作不正常带来的？

答案是否定的。因为抽样的偶然性也可能导致这一事件发生。

一、假设检验的基本思想和方法

但常识告诉我们，因抽样原因带来$|\bar{X}-\mu|$足够大这一事件的概率是非常低的。这一思想的正式表述是：$P_{H_0}\{|\bar{X}-\mu|\geq M\}=\alpha$。其中$M$为一足够大的数，$\alpha$称为显著性水平，通常取1%、5%或10%，也表示拒绝H_0犯错误的概率。

概率论的一个基本结论是：小概率事件在一次抽样中不会发生。因此如果只进行了一次抽样，但$|\bar{X}-\mu|$足够大这一事件就发生了，那就意味着这一事件的发生绝对不是抽样原因带来的，只能是生产线工作不正常带来的。故此时我们必须拒绝机器设备工作正常的原假设。（当然，我们在作出这一断定时在可能出错，出错的概率就是事前给定的显著性水平α 。

剩下的问题是，足够大究竟是多大？

一、假设检验的基本思想和方法

因$P_{H_0}\left\{\left|\bar{X}-\mu\right|\geq M\right\}=\alpha\Leftrightarrow P_{H_0}\left\{\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right|\geq\frac{M}{\sigma/\sqrt{n}}\right\}=\alpha$。不妨令$\frac{M}{\sigma/\sqrt{n}}=k$,

则上式等价于$P_{H_0}\left\{\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right|\geq k\right\}=\alpha$。

由于$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$, 因此$k$就是标准正态分布的双侧$\alpha$分位数，即$k=z_{\alpha/2}$,

而$\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right|\geq z_{\alpha/2}$则称之为拒绝域。

拒绝域的含义是：如果检验统计量$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$的绝对值大于$z_{\alpha/2}$，则表明样本平均容积与标称容积的差异足够大。在一次抽样中，这种差异不可能是抽样原因带来的，只能是生产线工作不正常带来的，此时我们必须拒绝生产线工作正常的原假设。

二、单正态总体均值假设检验的标准步骤

一、提出假设：

H_0：$\mu=\mu_0$；；　H_1　$\mu \neq \mu_0$

二、在H_0成立的前提下构造检验统计量

三、给定显著性水平α，查表求出临界值$z_{\alpha/2}$

注：若检验统计量的概率密度函数对称，则通常做双侧假设检验，临界值为其双侧α分位数，拒绝域一定有绝对值符号；
若概率密度函数非对称，则做单侧假设检验，临界值为其上α分位数，拒绝域没有绝对值符号；

四、代入样本数据，计算出检验统计量值$Z=\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$并与临界值$z_{\alpha/2}$进行比较：

1、若$|Z| \geq z_{\alpha/2}$，落入拒绝域，拒绝原假设

2、若$|Z| < z_{\alpha/2}$，未落入拒绝域，不能拒绝原假设

若总体方差未知：

一、提出假设：

H_0：$\mu=\mu_0$；：　H_1　$\mu \neq \mu_0$

二、在H_0成立的前提下构造检验统计量：$t=\dfrac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

三、给定显著性水平α，查表求出临界值$t_{\alpha/2}(n-1)$，确定拒绝域$|t| \geq t_{\alpha/2}(n-1)$

四、代入样本数据，计算出检验统计量值$t=\dfrac{\bar{X}-\mu}{S/\sqrt{n}}$并与临界值$t_{\alpha/2}(n-1)$比较：

1、若$|t| \geq t_{\alpha/2}(n-1)$，落入拒绝域，拒绝原假设

2、若$|t| < t_{\alpha/2}(n-1)$，未落入拒绝域，不能拒绝原假设

例1：方差已知的双侧假设检验

某车间用一台包装机包装葡萄糖。当机器工作正常时，袋装葡萄糖的重量服从均值为0.5公斤，标准差为0.015公斤的正态分布。某日随机抽取了9袋，测得其重量分别为（公斤）：

0.497 , 0.506 , 0.518 , 0.524 , 0.498 , 0.511 , 0.520 , 0.515 , 0.512

问机器是否工作正常？（显著性水平$\alpha=0.05$）

一、提出假设：

H_0：$\mu=0.5$；：　H_1 0.5 $\neq$

二、在H_0成立的前提下构造检验统计量：$Z=\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$

三、给定$\alpha=5\%$，查表求出临界值$z_{0.025}=1.96$，确定拒绝域$|Z|\geq 1.96$

四、代入样本数据$\bar{X}=0.511$，$\sigma=0.015$，$n=9$可算出检验统计量值

$|Z|=\left|\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right|=\left|\dfrac{0.511-0.5}{0.015/3}\right|=2.2>1.96$。故落入拒绝域，拒绝原假设，即包装机工作不正常。

例2：方差未知的双侧假设检验

某种电子元件的寿命 X（以小时计）服从正态分布，μ, σ^2均未知。现测得16只元件的寿命如下：

159　280　101　212　224　379　179　264

222　362　168　250　149　260　485　170

问是否有理由在5%的显著性水平下认为元件的平均寿命等于225（小时）？

一、提出假设：

H_0：$\mu=225$；；　H_1 225 ≠

二、在H_0成立的前提下构造检验统计量：$t=\dfrac{\bar{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$

三、给定$\alpha=5\%$，查表求出临界值$t_{0.025}(15)=2.13$，确定拒绝域$|t|\geq 2.13$

四、由样本数据可计算出$\bar{X}=241.5$，$S=98.7259$，$n=16$。由此可计算出

检验统计量值$|t|=\left|\dfrac{\bar{X}-\mu}{S/\sqrt{n}}\right|=\left|\dfrac{241.5-225}{98.7259/4}\right|=0.6685<2.13$。未落入拒绝域，

不能拒绝原假设，即在5%的显著性水平下可以认为元件的平均寿命等于225小时。