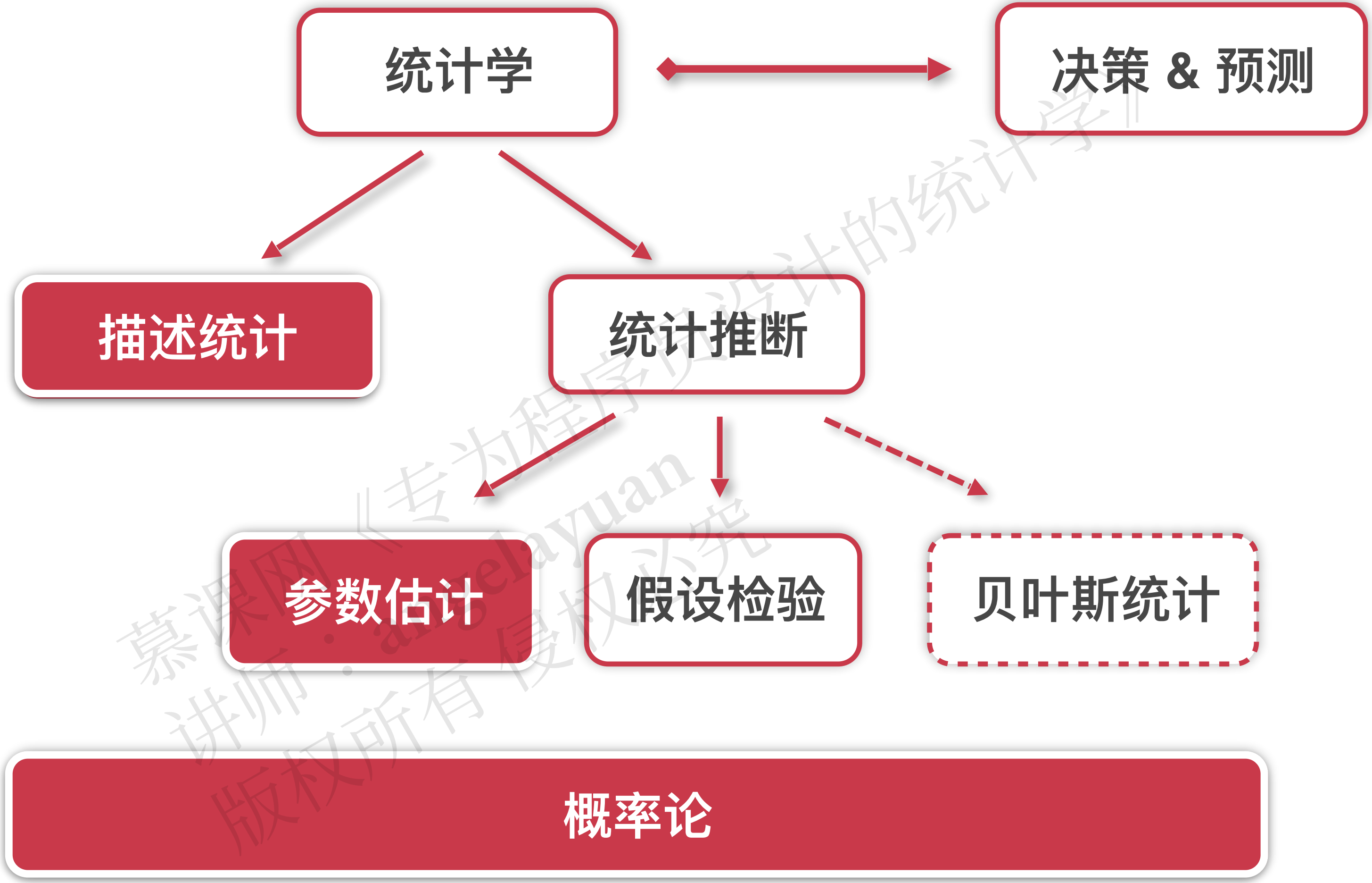

统计学
决策 & 预测
描述统计
统计推断
参数估计
假设检验
贝叶斯统计
概率论

点估计 (point estimate)

点估计

- 设总体X的分布函数的形式已知，但它的一个或多个参数未知，借助于总体X的一个样本来估计总体未知参数的值的问题称为参数的点估计问题

总体 → 样本 → 统计量 ≈ 总体未知参数

- 形式已知
- 一个或多个参数未知

- 样本均值
- 样本方差

- 总体均值
- 总体方差

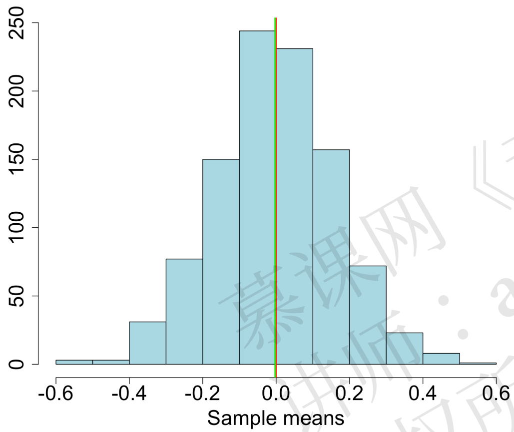
点估计

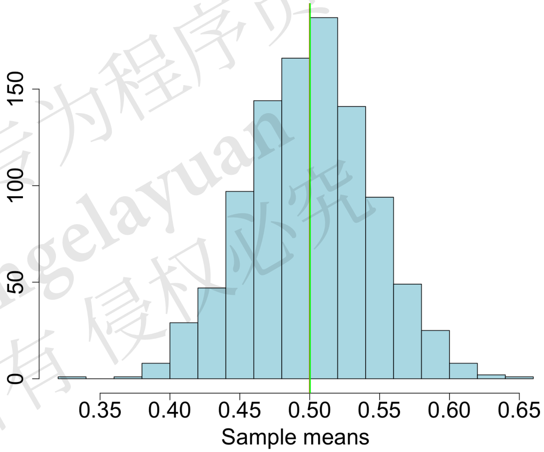
- 估计量、估计值

θ是待估参数，$X_1, X_2, \ldots X_n$ 是总体X的一个样本，$x_1, x_2, \ldots x_n$ 是一个样本值

点估计的问题就是要构造一个适当的统计量(估计量)，用它的观察值作为未知参数的近似值(估计值)

总体 → 1000个样本
样本容量= 40 → 计算1000个样本的样本均值 → 样本均值的分布





对于不同的样本值，
估计值一般是不同的

估计量的评选标准

- 对于同一参数，使用不同的估计方法求出的估计量可能不相同。采用哪一个估计量好呢?
  - 无偏性: 若估计量的数学期望存在，并且该期望等于总体参数，则称为无偏估计
  - 均值 vs 期望: 均值是一个统计量(基于样本构造的函数)；期望完全由随机变量的概率分布所确定(“上帝视角”)；两者常混用

估计量的评选标准

- 对于同一参数，使用不同的估计方法求出的估计量可能不相同。采用哪一个估计量好呢?
  - 对于某些样本值，由这一估计量得到的估计值相对于真值来说偏大或偏小，但是**反复将这一估计量使用多次，就“平均”来说其偏差为零**
  - **E(估计值) - 真值**称为系统误差；无偏估计的实际意义就是**无系统误差**

上一章，我们讲过....

设总体X的均值为μ, 方差为 σ^2

$X_1, X_2, \dots X_n$ 是来自总体X的一个样本

$\bar{X}, S^2$ 分别为样本均值和样本方差

则有 $E(\bar{X}) = \mu, \mathrm{Var}(\bar{X}) = \sigma^2/n$

$E(S^2) = \sigma^2$

说明不论总体服从什么分布，样本均值是总体均值的无偏估计；样本方差是总体方差的无偏估计

样本均值抽样分布的离散程度

标准误(standard error of mean; SE)

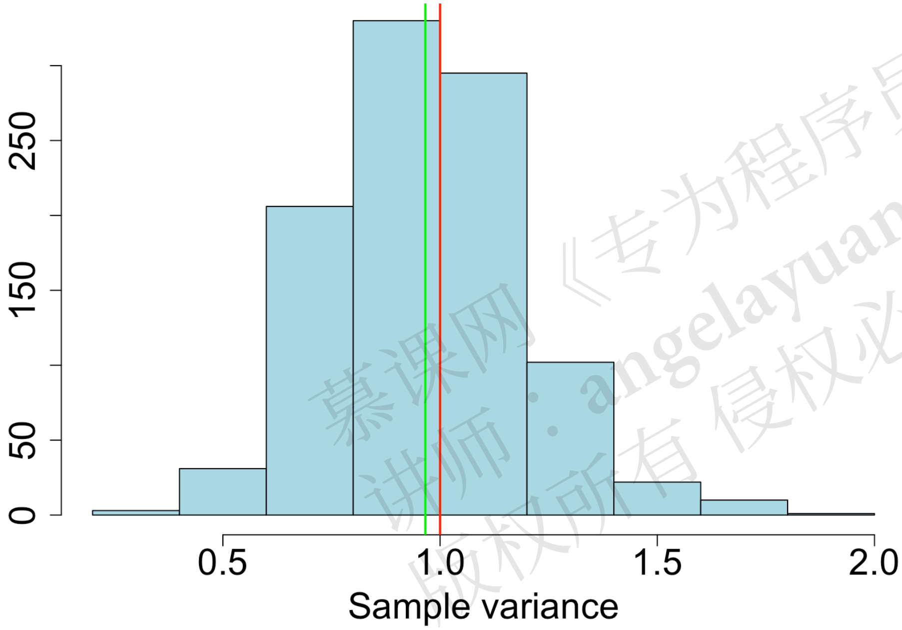
样本方差: 除以n vs 除以(n-1)

$$\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$ → 如果是基于样本计算的，则与总体方差有系统偏差

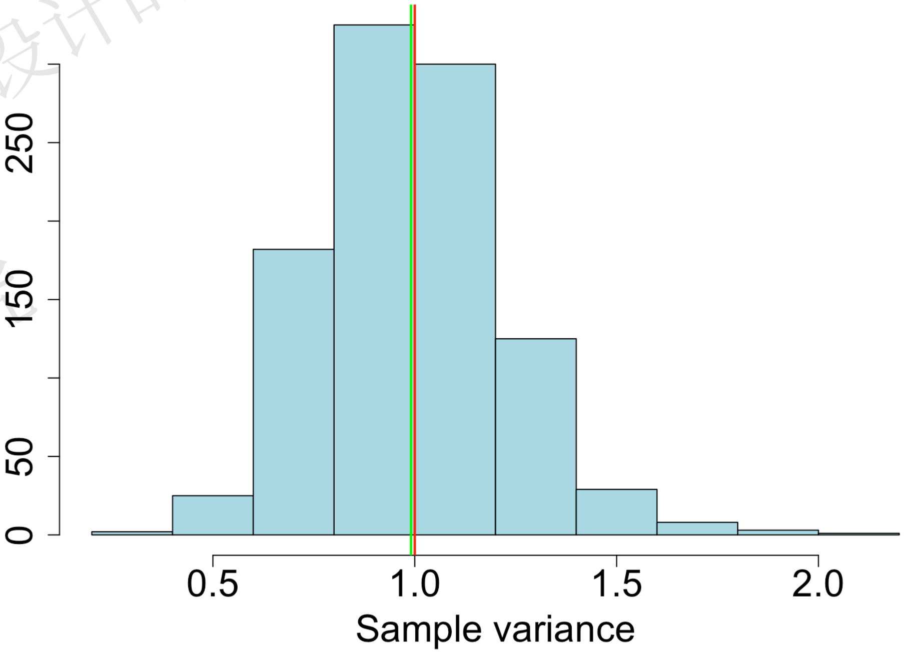
$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$ → 是总体方差的无偏估计

总体服从标准正态分布N(0,1)；样本容量 = 40；样本个数 = 1000

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

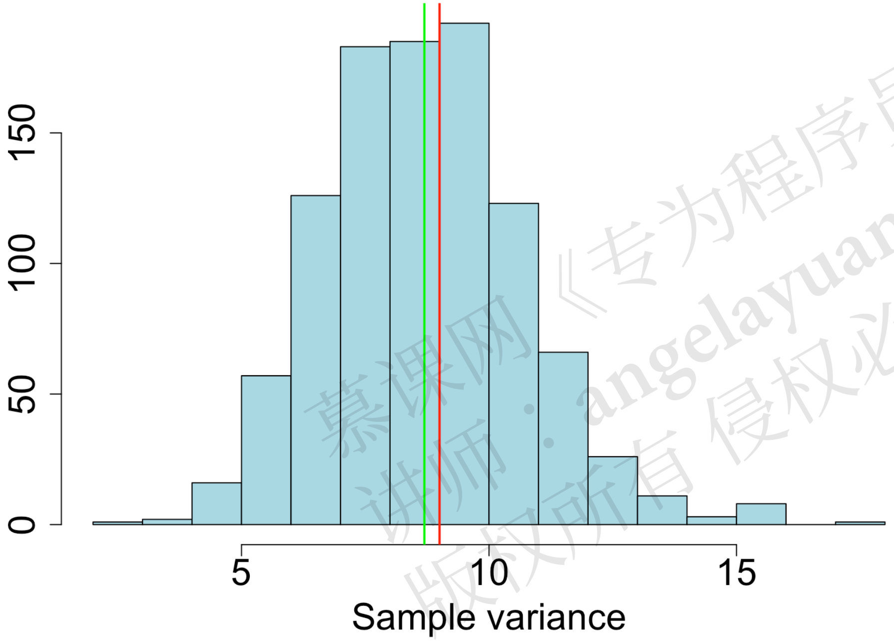


$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

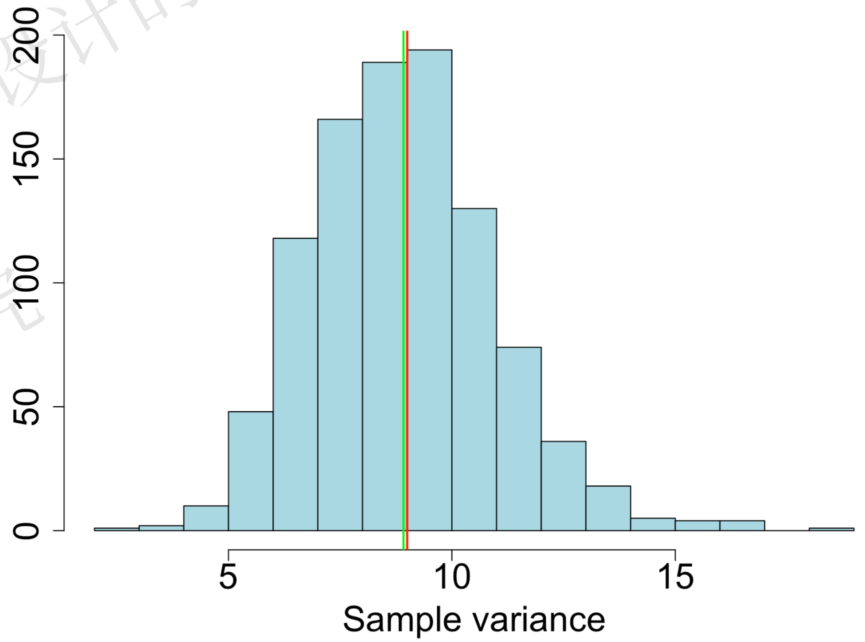


总体服从正态分布N(5,9)；样本容量 = 40；样本个数 = 1000

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

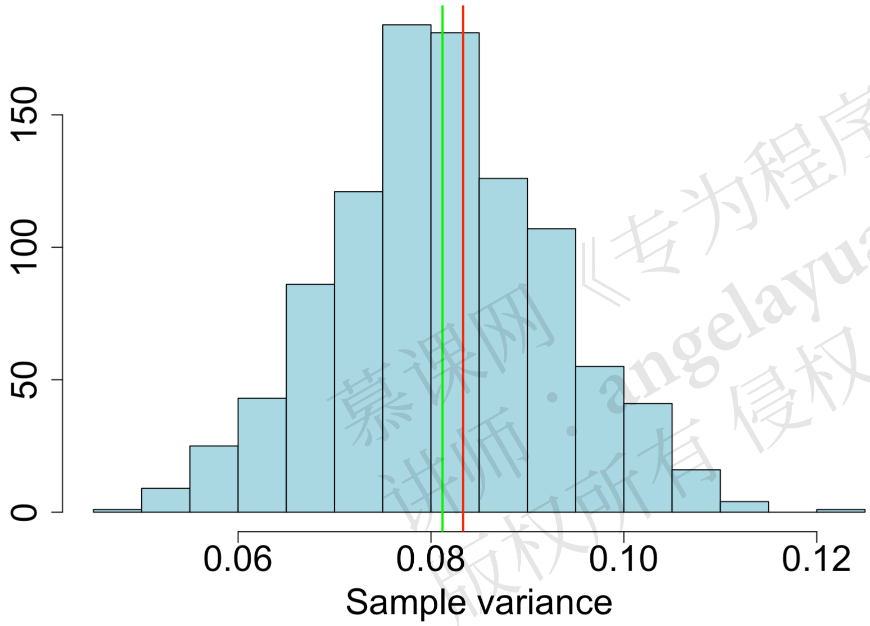


$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

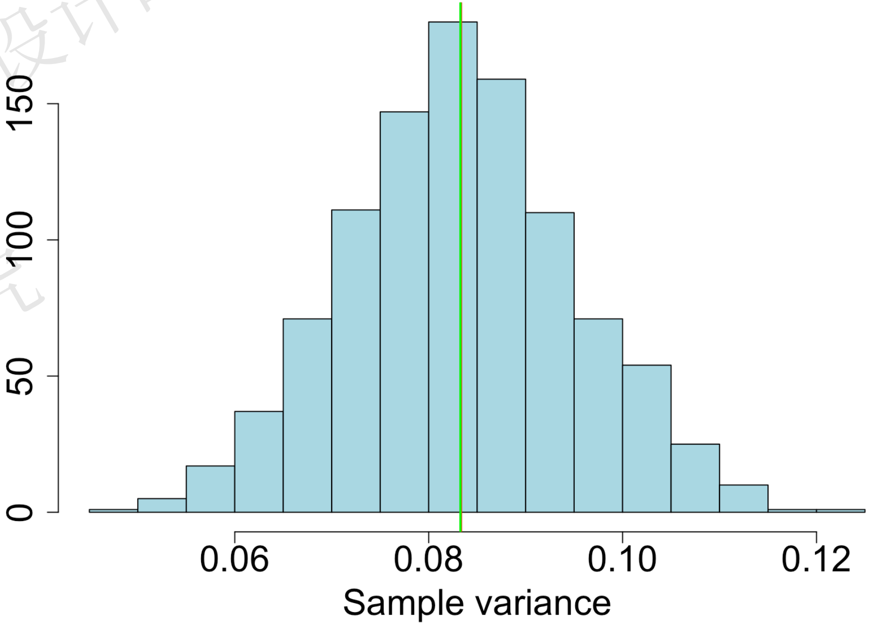


总体服从均匀分布(0,1)；样本容量 = 40；样本个数 = 1000

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



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



估计量的评选标准

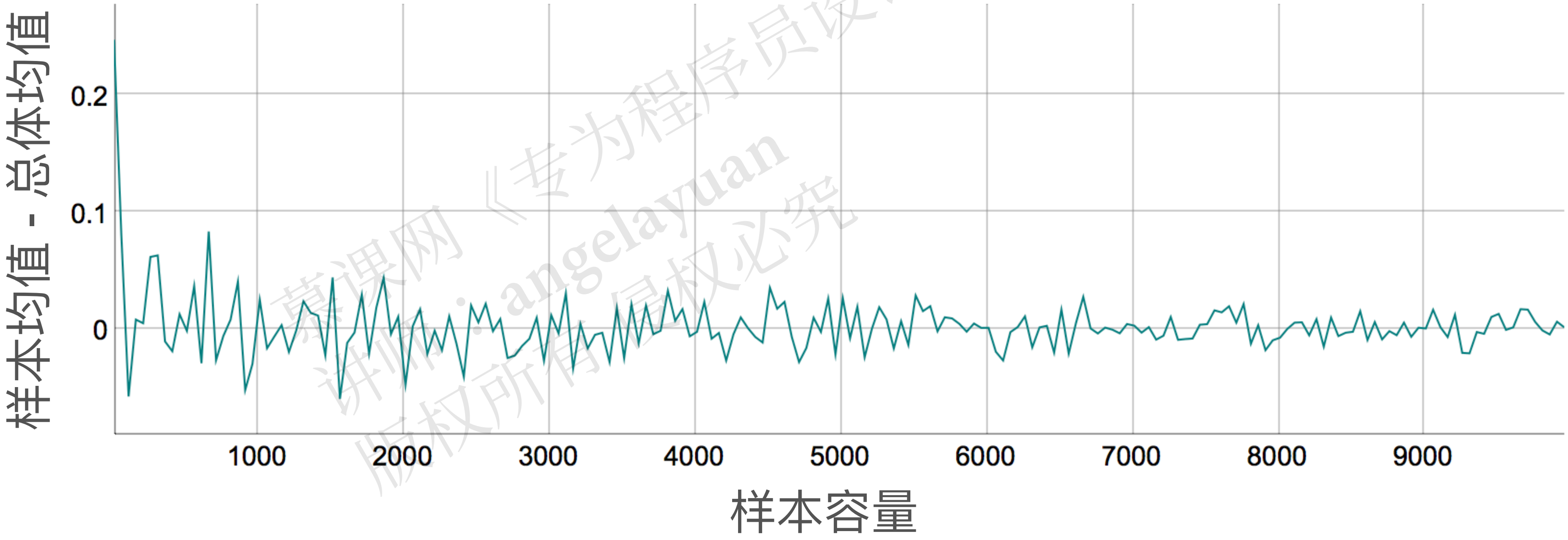
- 对于同一参数，使用不同的估计方法求出的估计量可能不相同。采用哪一个估计量好呢?
  - 有效性: 有两个无偏估计 θ_1和 θ_2，如果在样本容量n相同的情况下, θ_1 比 θ_2 更密集在真值附近，就认为θ_1比θ_2 更理想；
  - 由于方差是随机变量取值与其数学期望的偏离程度的测量，所以无偏估计以方差最小者为好

估计量的评选标准

- 对于同一参数，使用不同的估计方法求出的估计量可能不相同。采用哪一个估计量好呢?
  - 无偏性和有效性都是在样本容量n固定的前提下提出的
  - 相合性：我们希望随着样本容量的增大，一个估计量的值稳定于待估参数的真值。满足此条件的估计量为相合估计量

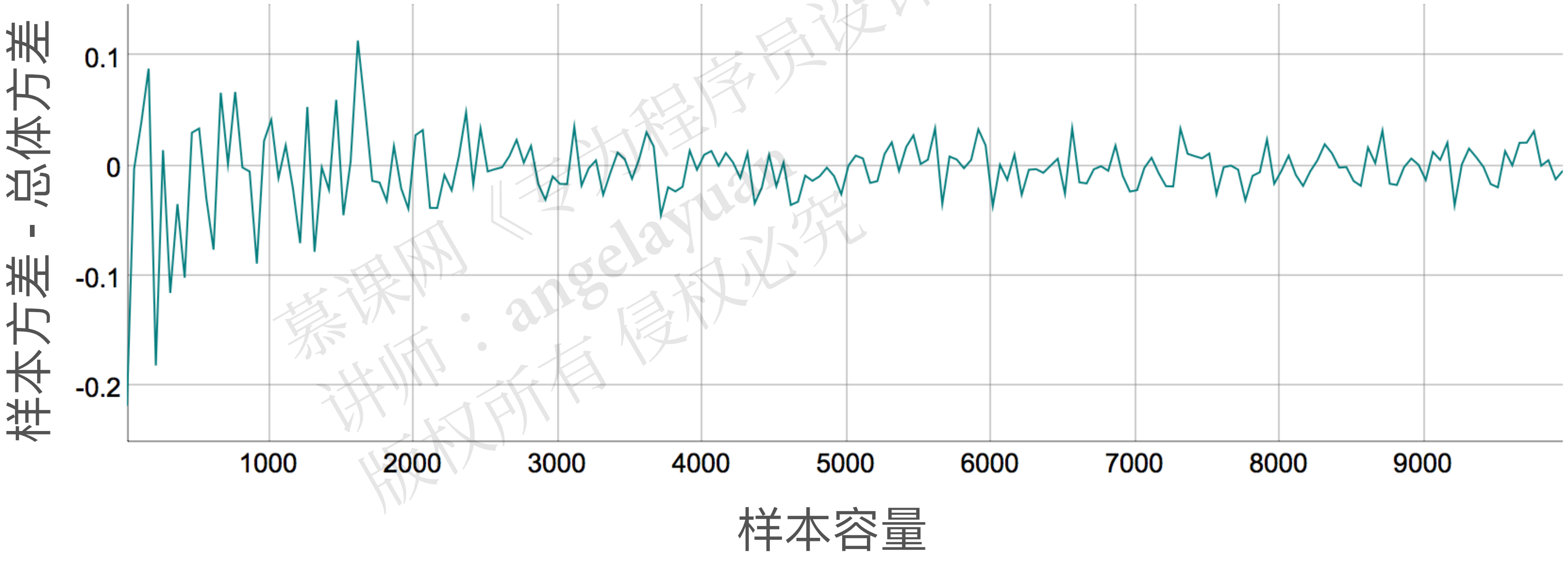
样本均值满足相合性

总体服从标准正态分布N(0,1)；样本容量 = 20, 70, 120, …, 10000



样本方差满足相合性

总体服从标准正态分布N(0,1)；样本容量 = 20, 70, 120, …, 10000



编程理解无偏性与相合性

区间估计 (interval estimate)

区间估计

- 对于一个未知量，我们在测量或计算时，不仅要得到近似值，还要估计误差，即近似值的精确程度/所求真值所在范围

- 对于未知参数，我们不仅要得到近似值(点估计)，还希望估计出一个范围(区间)，并希望知道这个范围包含参数真值的可信程度。这种形式的估计称为区间估计，这样的区间称为置信区间

置信区间

- 设总体的分布函数含有一个未知参数$\theta, \theta \in \Theta$($\Theta$是$\theta$可能的取值范围)；对于给定值$\alpha$ $(0 < \alpha < 1)$, 若由来自X的样本$X_1, X_2, \dots X_n$确定的两个统计量$\underline{\theta} = \underline{\theta}(X_1, X_2, \dots, X_n)$和$\overline{\theta} = \overline{\theta}(X_1, X_2, \dots, X_n)$ $(\underline{\theta} < \overline{\theta})$对于任意$\theta \in \Theta$满足

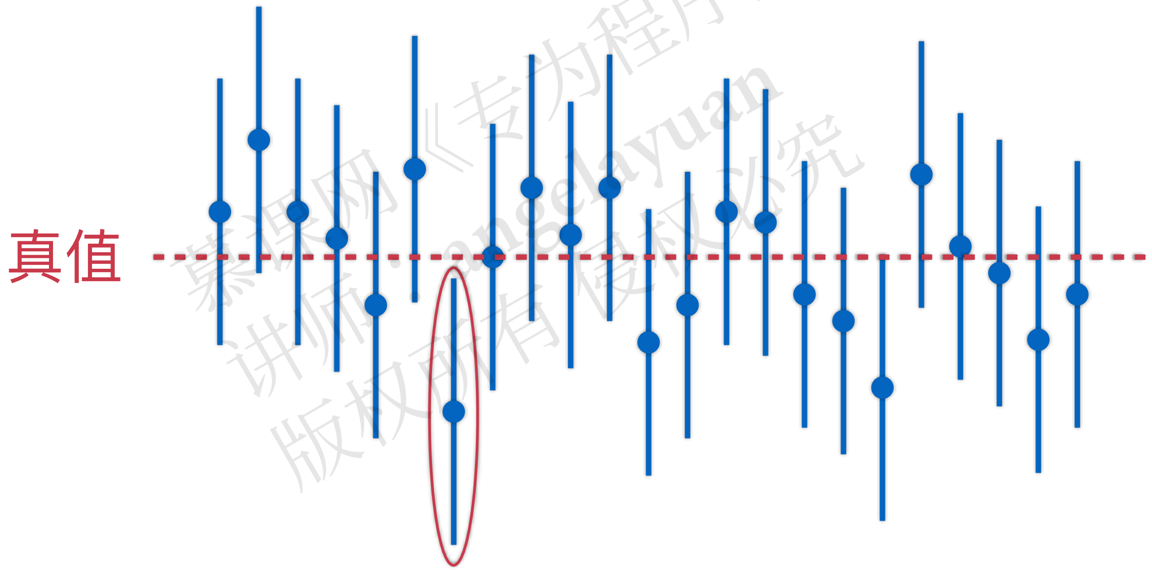
$$P\{\underline{\theta}(X_1, X_2, \dots, X_n) < \theta < \overline{\theta}(X_1, X_2, \dots, X_n)\} \geq 1 - \alpha$$

则称随机区间$(\underline{\theta}, \overline{\theta})$是$\theta$的置信水平为$1 - \alpha$的置信区间

置信下限　置信上限　　　　置信水平

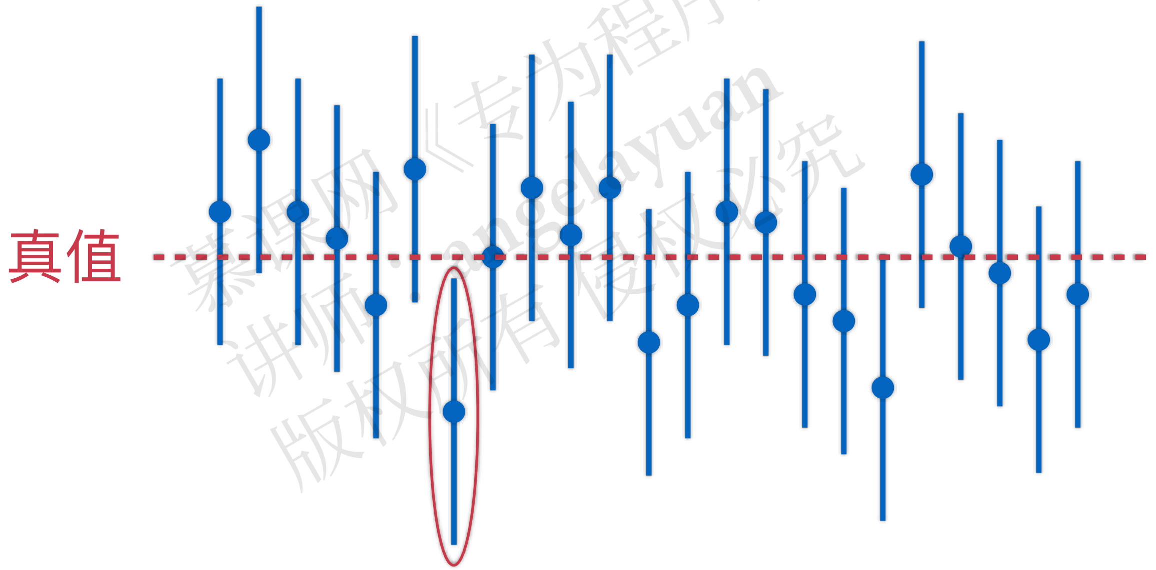
置信区间

- 固定样本容量n, 若反复抽样多次, 每个样本值确定一个区间, 每个这样的区间要么包含θ的真值, 要么不包含θ的真值



置信区间

- 按大数定律，在这么多区间中，包含真值的约占100*(1-α)%，不包含真值的占100*α%



1-α = 0.95

95%的区间包含真值

5%的区间不包含真值

求未知参数θ的置信区间的步骤

- 寻求一个样本 $X_1, X_2, \ldots X_n$，和一个统计量 $W(X_1, X_2, \ldots X_n; \theta)$
  使统计量W的分布不依赖于θ和其他未知参数
  统计量W的构造，通常可以从θ的点估计着手

- 对于给定的置信水平 $1-\alpha$，定出两个常数a和b，使得
  $P(a < W(X_1, X_2, \ldots X_n; \theta) < b) = 1-\alpha$
  若能从中得到θ的不等式 $\underline{\theta}(X_1, X_2, \ldots, X_n) < \theta < \overline{\theta}(X_1, X_2, \ldots, X_n)$
  则 $(\underline{\theta}, \overline{\theta})$ 为θ的一个置信水平为 $1-\alpha$ 的置信区间

置信区间

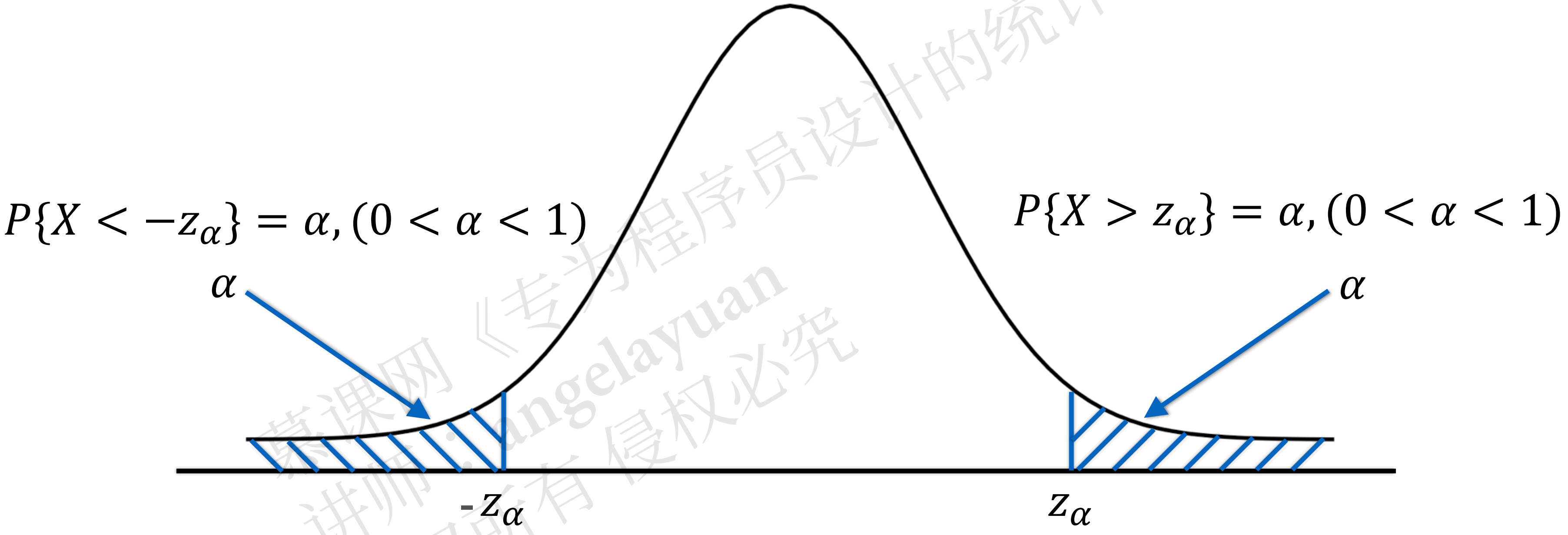
- 设总体$X \sim N(\mu, \sigma^2)$, σ^2已知, μ为未知, 设$X_1, X_2, \dots X_n$是来自X的样本, 求μ的置信水平为$1-\alpha$的置信区间

第1步 $W(X_1, X_2, \dots X_n; \theta)$

W的分布不依赖于θ和其他未知参数 → $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

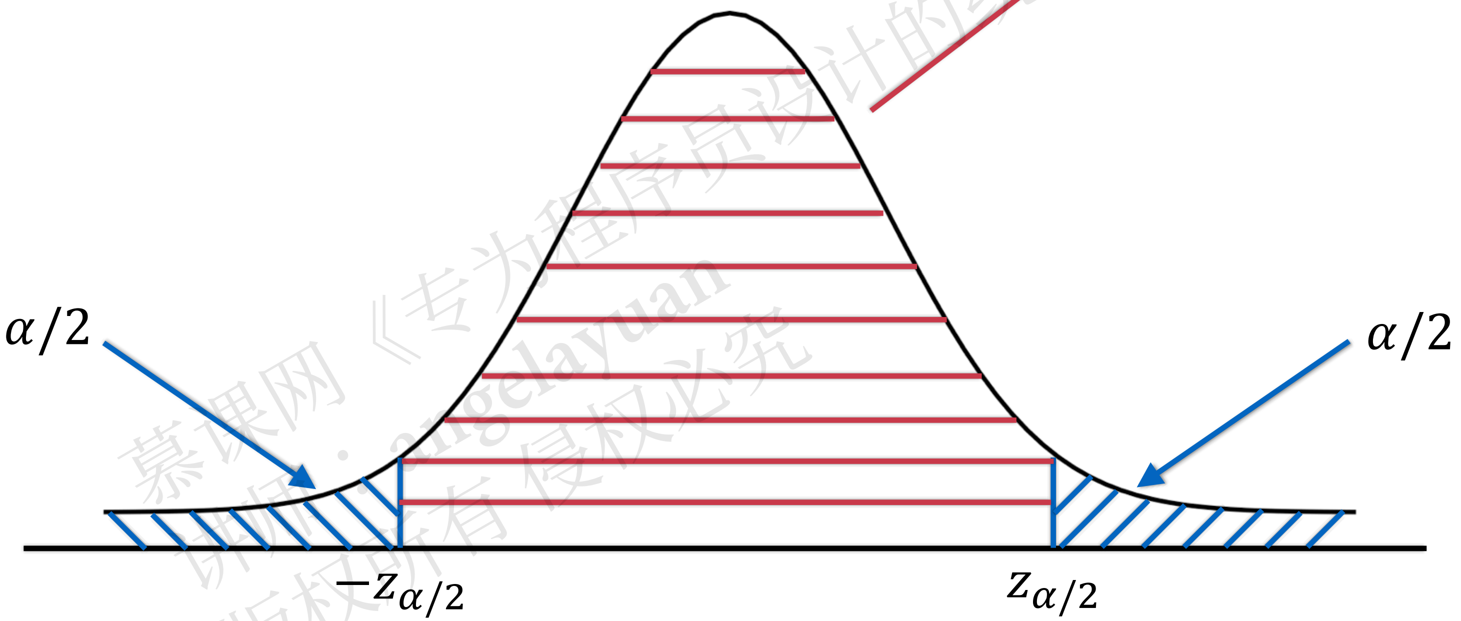
第2步 $P(a < W(X_1, X_2, \dots X_n; \theta) < b) = 1-\alpha$

标准正态分布



$P\{X < -z_\alpha\} = \alpha, (0 < \alpha < 1)$
$P\{X > z_\alpha\} = \alpha, (0 < \alpha < 1)$
α
α
$-z_\alpha$
z_α

$$P(a < W(X_1, X_2, \dots X_n; \theta) < b) = 1 - \alpha$$



置信区间

- 设总体 $X \sim N(\mu, \sigma^2)$, σ^2 已知, μ为未知, 设 $X_1, X_2, \dots X_n$是来自X的样本, 求 μ 的置信水平为 $1-\alpha$ 的置信区间

第1步

$W(X_1, X_2, \dots X_n; \theta)$

W的分布不依赖于θ和其他未知参数 → $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

第2步 $P(a < W(X_1, X_2, \dots X_n; \theta) < b) = 1-\alpha$

置信区间

- 设总体$X \sim N(\mu, \sigma^2)$, σ^2已知, μ为未知, 设$X_1, X_2, \ldots X_n$是来自X的样本, 求μ的置信水平为$1-\alpha$的置信区间

第2步 $P(a < W(X_1, X_2, \ldots X_n; \theta) < b) = 1 - \alpha$

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

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

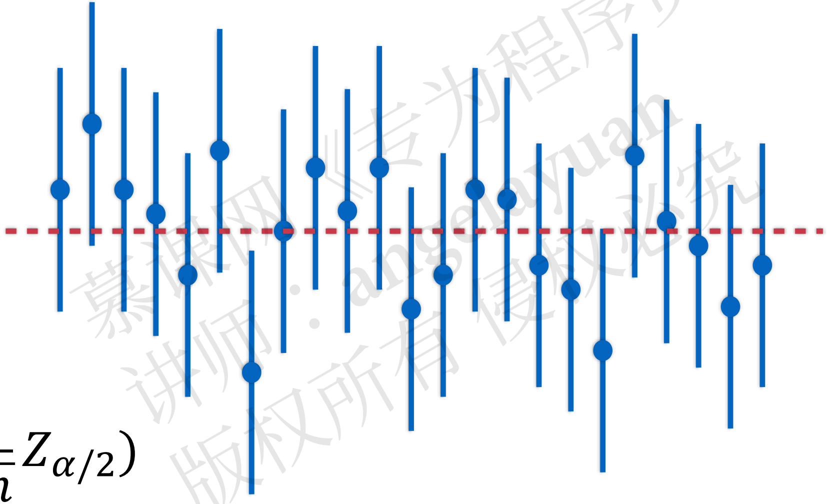
$$\rightarrow \left(\bar{X} - \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}\right) \quad \left(\bar{X} \pm \boxed{\frac{\sigma}{\sqrt{n}}}\boxed{Z_{\alpha/2}}\right)$$

误差范围(margin of error; ME)

置信区间

- 设总体 $X \sim N(\mu, \sigma^2)$, σ^2 已知, μ为未知, 设 $X_1, X_2, \ldots X_n$ 是来自X的样本, 求μ的置信水平为 $1-\alpha$ 的置信区间

真值

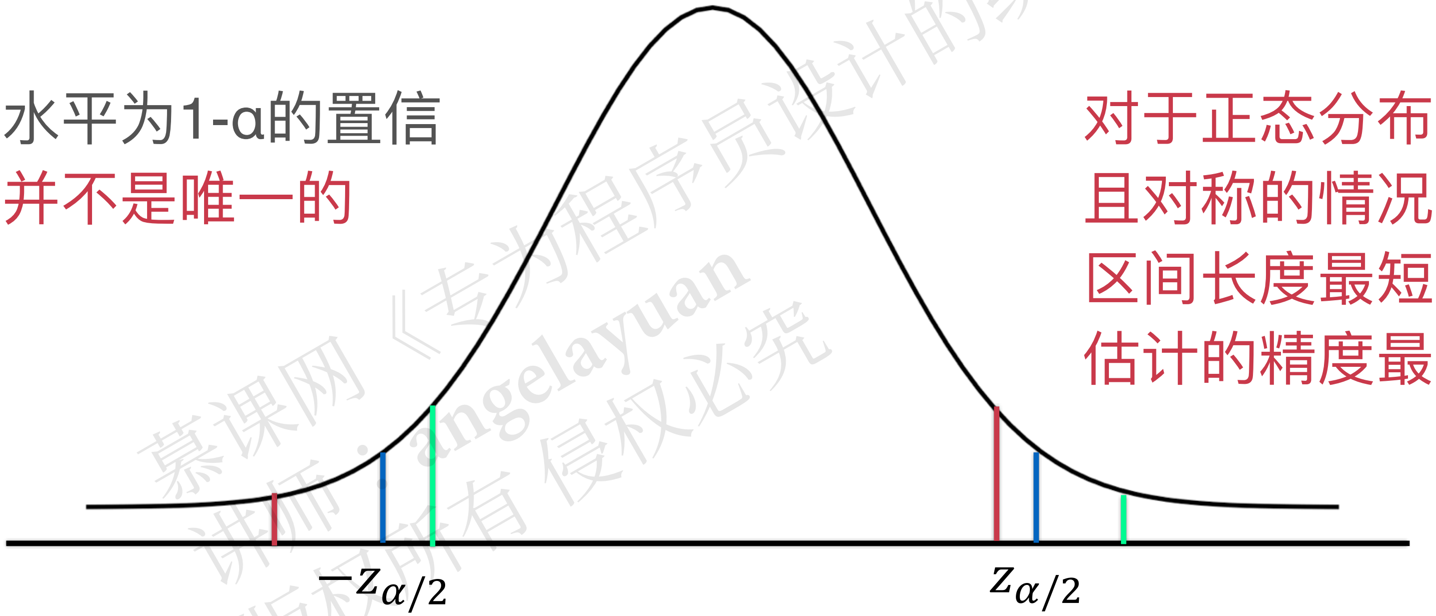


$$(\bar{X} \pm \frac{\sigma}{\sqrt{n}} Z_{\alpha/2})$$

- 现在得到的区间属于那些包含μ的区间的可信程度为100*(1-α)%, 或“该区间包含真值”这一陈述的可信程度为100*(1-α)%

$$P(a < W(X_1, X_2, \dots X_n; \theta) < b) = 1 - \alpha$$

置信水平为1-α的置信区间并不是唯一的

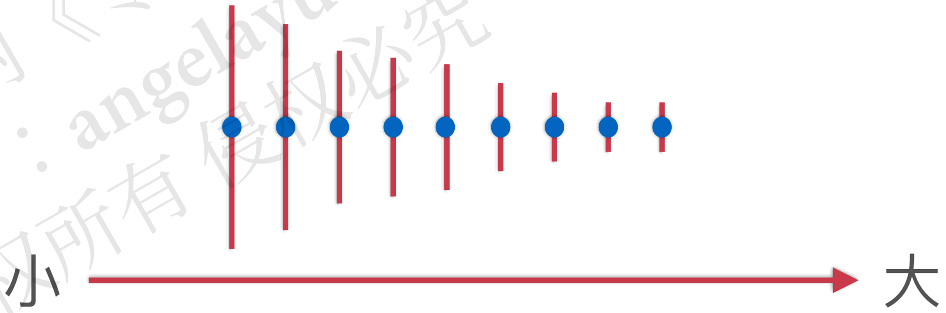


对于正态分布这种单峰且对称的情况, 对称的区间长度最短，意味着估计的精度最高

置信区间与样本容量

$$(\bar{X} \pm \frac{\sigma}{\sqrt{n}} Z_{\alpha/2})$$

- 标准误随着样本容量的增加而减小
- 误差范围随着样本容量的增加而减小



一个正态总体的情况

方差已知，求均值的置信区间

- 已知

$X \sim N(\mu, \sigma^2)$

$X_1, X_2, \dots X_n$

$1-\alpha$

$\bar{X}, S^2$

σ^2 已知

- 求均值μ的置信区间

$\bar{X} \sim N(\mu, \sigma^2/n) \qquad \dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

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

$$(\bar{X} \pm \frac{\sigma}{\sqrt{n}} Z_{\alpha/2})$$

例题

- 歌曲的时长服从正态分布$X\sim N(\mu,\sigma^2)$，$\sigma^2=1$；手机里有100首歌曲，这100首歌曲的平均时长是4分钟，求μ的置信水平为95%的置信区间

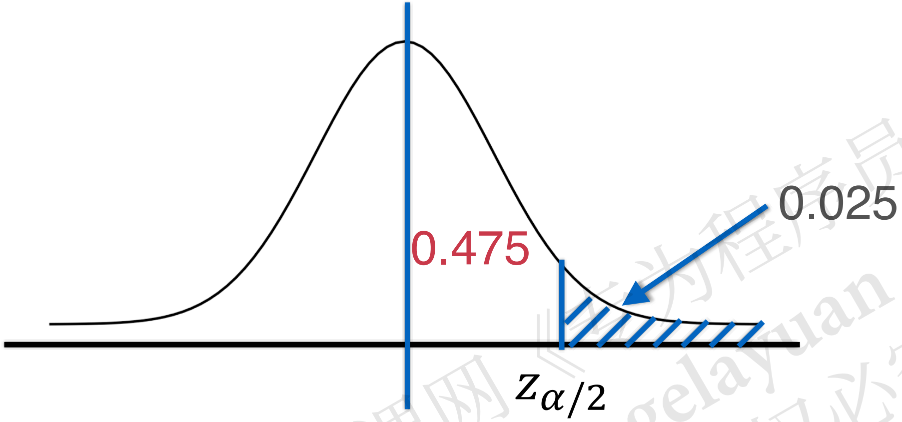
样本均值 = 4
样本容量 = 100
总体标准差 = 1
α = 0.05

$$\left(\bar{X} \pm \frac{\sigma}{\sqrt{n}} \boxed{Z_{\alpha/2}}\right)$$

查表

python函数

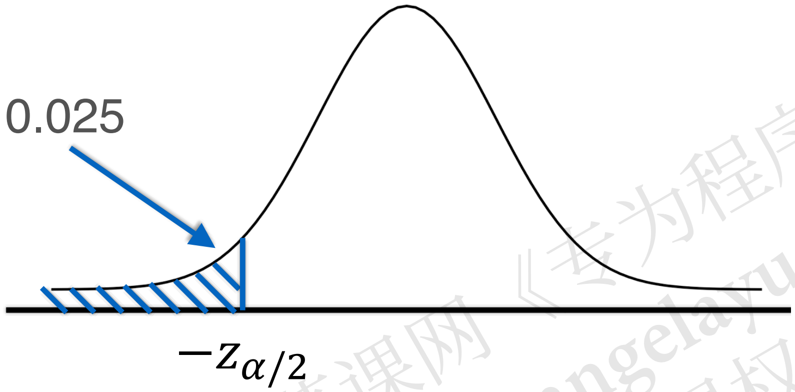
例题



95%置信水平对应的Z值为1.96

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	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	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	0.4554	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

例题



95%置信水平对应的Z值为1.96

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
–3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
–3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
–3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
–3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
–3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
–2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
–2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
–2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
–2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
–2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
–2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
–2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
–2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
–2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
–2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
–1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233

例题

- 歌曲的时长服从正态分布$X \sim N(\mu, \sigma^2)$，$\sigma^2 = 1$；
手机里有100首歌曲，这100首歌曲的平均时长是4分钟，求μ的置信水平为95%的置信区间

样本均值 = 4
样本容量 = 100
总体标准差 = 1
α = 0.05

$$\left(\bar{X} \pm \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}\right) = \left(4 \pm \frac{1}{\sqrt{100}} \times 1.96\right)$$

$$= (3.804, 4.196)$$

(3.804, 4.196)属于那些包含真值的区间的可信程度为95%；(3.804, 4.196)包含真值这一陈述的可信程度为95%

方差未知，求均值的置信区间

- 已知

$X \sim N(\mu, \sigma^2)$

$X_1, X_2, \ldots X_n$

$1-\alpha$

$\bar{X}, S^2$

σ^2 未知

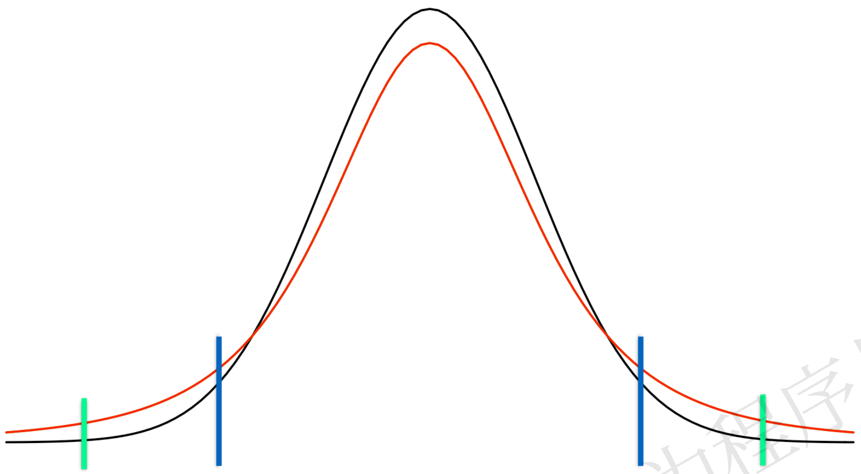
- 求均值μ的置信区间

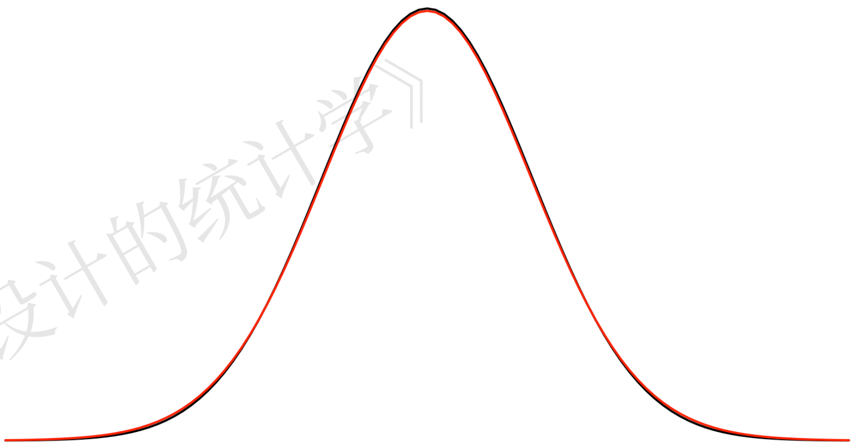
$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$(\bar{X} \pm \frac{\sigma}{\sqrt{n}} Z_{\alpha/2})$$

用样本方差代替总体方差

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

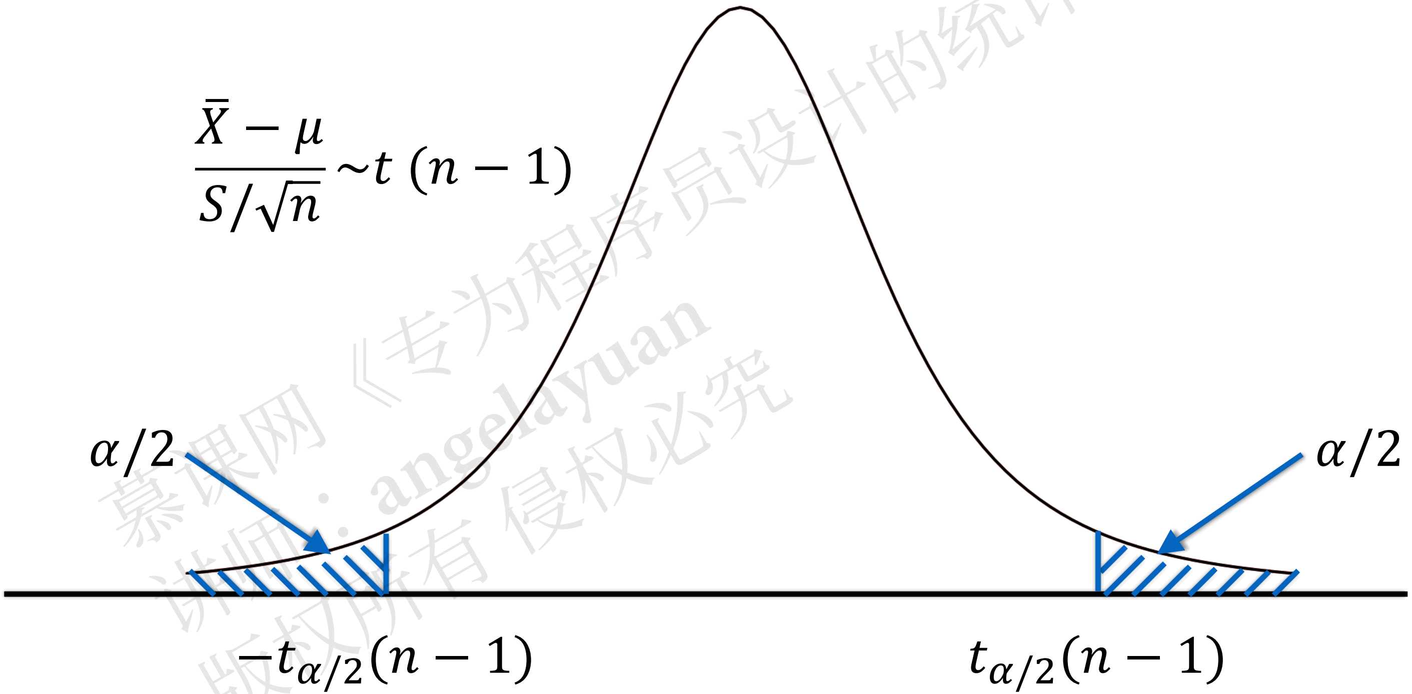




- 样本容量小时，t分布比正态分布略宽，置信区间略大；这反映了基于小样本得到的结论具有更大的不确定性
- 随着样本容量增大，t分布近似于标准正态分布；大样本也可以使用标准正态分布来计算置信区间

$$P(a < W(X_1, X_2, \dots X_n; \theta) < b) = 1 - \alpha$$

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



方差未知，求均值的置信区间

- 已知

$X \sim N(\mu, \sigma^2)$

$X_1, X_2, \ldots X_n$

$1-\alpha$

$\bar{X}, S^2$

σ^2 未知

- 求均值μ的置信区间

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

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

$$(\bar{X} \pm \boxed{\frac{S}{\sqrt{n}}}\boxed{t_{\alpha/2}(n-1)})$$

例题

- 歌曲的时长服从正态分布$X \sim N(\mu, \sigma^2)$，σ^2未知；
  手机里有100首歌曲，这100首歌曲的平均时长是4分钟，方差为1.44, 求μ的置信水平为95%的置信区间

样本均值 = 4
样本容量 = 100
样本标准差 = 1.2
α = 0.05

$$\left(\bar{X} \pm \frac{S}{\sqrt{n}} \boxed{t_{\alpha/2}(n-1)}\right)$$

95%置信水平
自由度 = 99
对应的t值为1.98

95%置信水平
自由度 = 19
对应的t值为2.09

例题

- 歌曲的时长服从正态分布$X \sim N(\mu, \sigma^2)$，σ^2未知；手机里有100首歌曲，这100首歌曲的平均时长是4分钟，方差为1.44, 求μ的置信水平为95%的置信区间

样本均值 = 4
样本容量 = 100
样本标准差 = 1.2
α = 0.05

$$\left(\bar{X} \pm \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1)\right)$$

$$= \left(4 \pm \frac{1.2}{\sqrt{100}} \times 1.98\right) = (3.762, 4.238)$$

(3.762, 4.238)属于那些包含真值的区间的可信程度为95%；(3.762, 4.238)包含真值这一陈述的可信程度为95%

均值未知，求方差的置信区间

- 已知

$X \sim N(\mu, \sigma^2)$

$X_1, X_2, \dots X_n$

$1-\alpha$

$\bar{X}, S^2$

μ 未知

- 求方差σ^2的置信区间

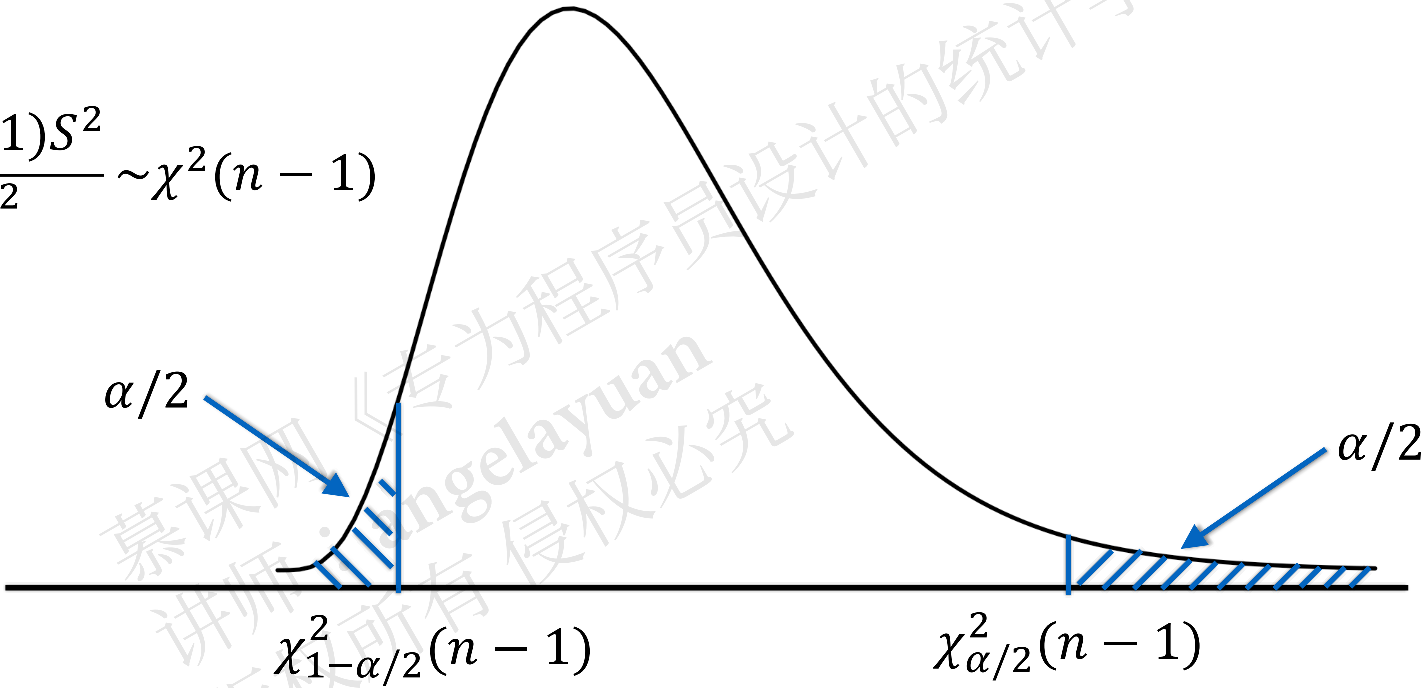
$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

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

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

$$P(a < W(X_1, X_2, \ldots X_n; \theta) < b) = 1 - \alpha$$

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



在密度函数不对称时，习惯上仍取对称的分位点

均值未知，求方差的置信区间

- 已知

$X \sim N(\mu, \sigma^2)$

$X_1, X_2, \ldots X_n$

$1-\alpha$

$\bar{X}, S^2$

μ 未知

- 求方差σ^2的置信区间

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

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

例题

- 歌曲的时长服从正态分布$X \sim N(\mu, \sigma^2)$，μ未知
  手机里有100首歌曲，这100首歌曲的时长的方差为1.44，求σ^2的置信水平为95%的置信区间

样本容量 = 100

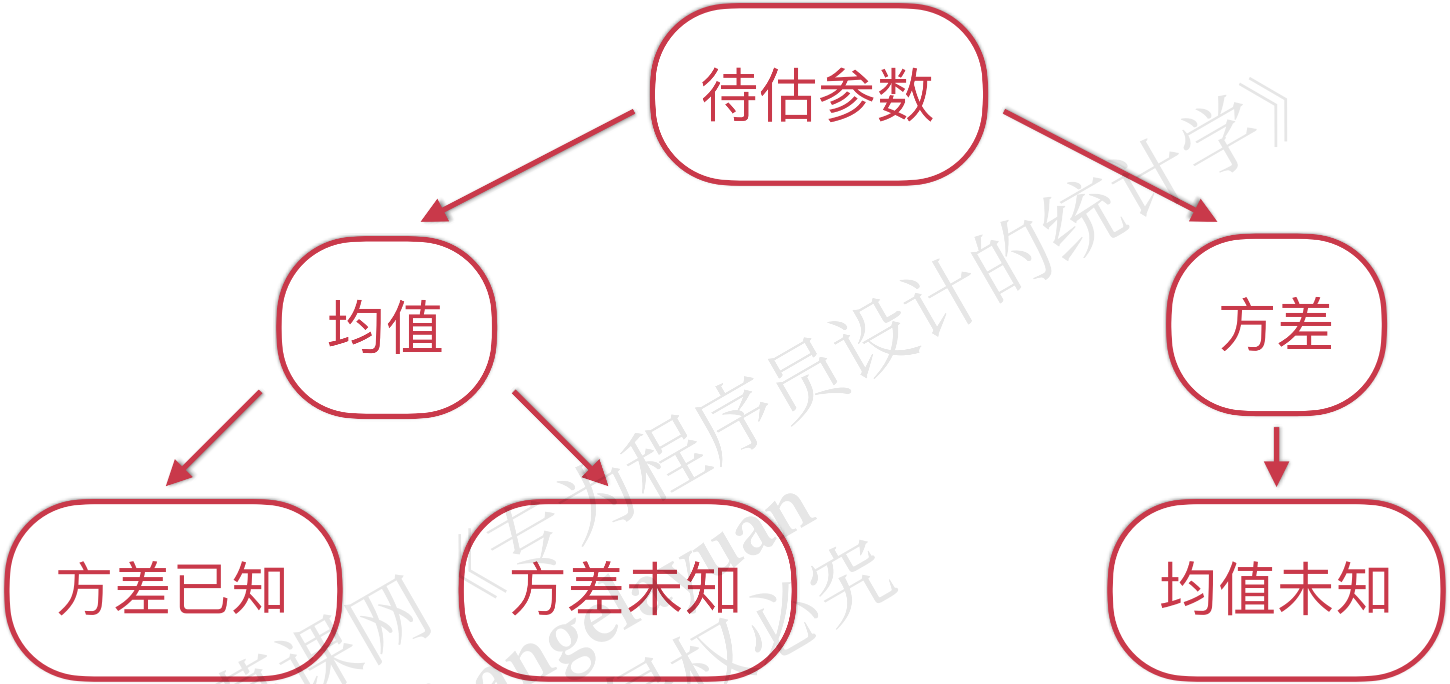
样本方差 = 1.44

α = 0.05

$$\left(\frac{(n-1)S^2}{\boxed{\chi^2_{\alpha/2}(n-1)}}, \frac{(n-1)S^2}{\boxed{\chi^2_{1-\alpha/2}(n-1)}}\right)$$

128.42　　73.36

$$= \left(\frac{99 * 1.44}{128.42}, \frac{99 * 1.44}{73.36}\right) = (1.11, 1.94)$$

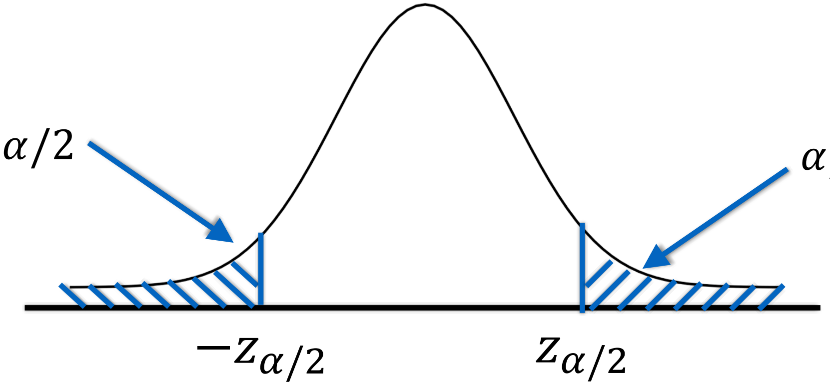


$$\left(\bar{X} \pm \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}\right) \qquad \left(\bar{X} \pm \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1)\right) \qquad \left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right)$$

编程求解置信区间
一个正态总体的情况

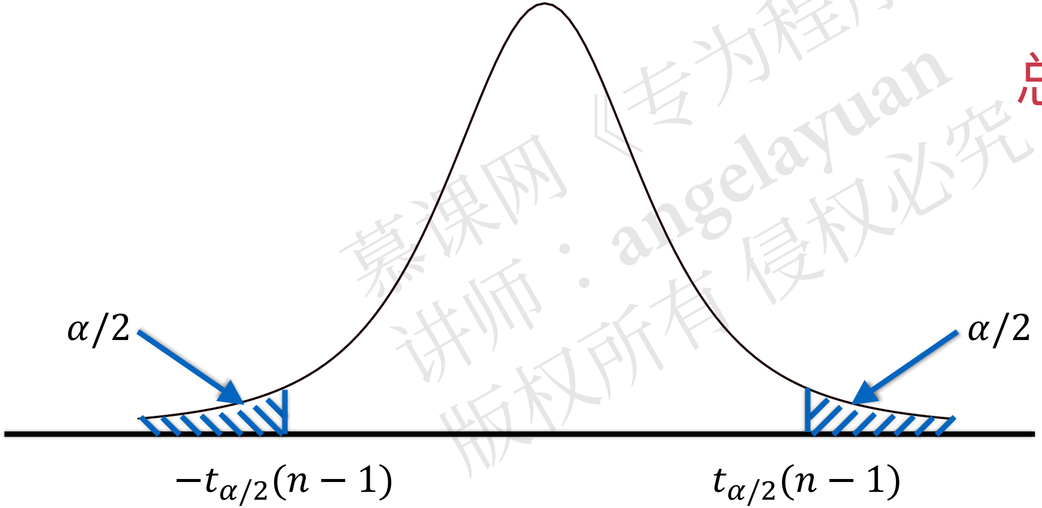
总体方差已知，求均值的置信区间

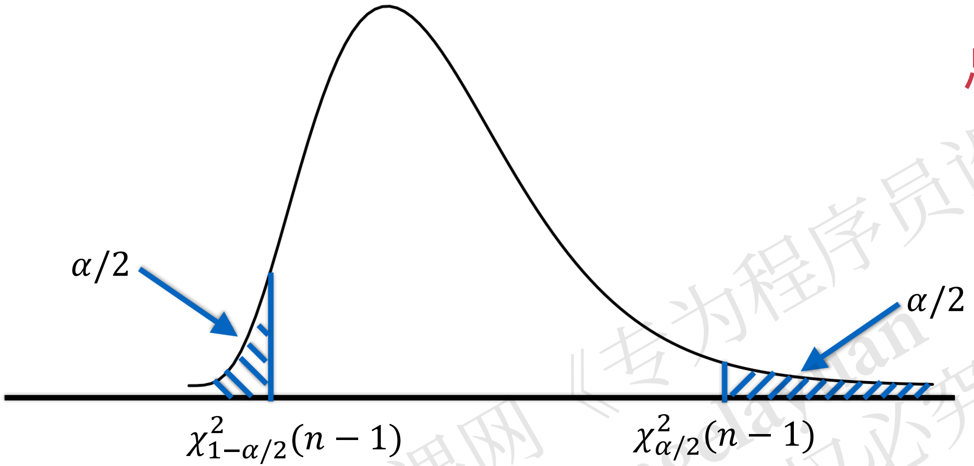
$$(\bar{X} \pm \frac{\sigma}{\sqrt{n}} Z_{\alpha/2})$$



总体方差未知，求均值的置信区间

$$(\bar{X} \pm \frac{S}{\sqrt{n}} t_{\alpha/2}(n-1))$$





总体均值未知，求方差的置信区间

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

两个正态总体的情况

两个方差已知，求均值差的置信区间

- 已知

$N(\mu_1, \sigma_1{}^2)$　　$N(\mu_2, \sigma_2{}^2)$

$X_1, X_2, \ldots X_{n_1}$　　$Y_1, Y_2, \ldots Y_{n_2}$

$\bar{X}, S_1{}^2$　　$\bar{Y}, S_2{}^2$

$1-\alpha$

$\sigma_1{}^2, \sigma_2{}^2$ 已知

- 求均值差μ_1-μ_2的置信区间

$$\bar{X}-\bar{Y} \sim N(\mu_1-\mu_2, \frac{\sigma_1{}^2}{n_1}+\frac{\sigma_2{}^2}{n_2})$$

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

两个方差已知，求均值差的置信区间

- 已知

$N(\mu_1, {\sigma_1}^2)$ $N(\mu_2, {\sigma_2}^2)$

$X_1, X_2, \dots X_{n_1}$ $Y_1, Y_2, \dots Y_{n_2}$

$\bar{X}, {S_1}^2$ $\bar{Y}, {S_2}^2$

$1-\alpha$

${\sigma_1}^2, {\sigma_2}^2$ 已知

- 求均值差μ_1-μ_2的置信区间

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \qquad (\bar{X} \pm \frac{\sigma}{\sqrt{n}} Z_{\alpha/2})$$

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

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

例题

- 25左右人群的月收入服从正态分布$N(\mu_1,{\sigma_1}^2)$，35左右人群的月收入服从正态分布$N(\mu_2,{\sigma_2}^2)$，σ_1,σ_2 分别为2000和8000；我们记录了30名25岁和40名35岁个体的月收入。这30名25岁个体平均收入为16,000元，这40名35岁个体平均收入为25,000元。求μ_1-μ_2置信水平为95%的置信区间

$$\left(\bar{X}-\bar{Y}\pm\sqrt{\frac{{\sigma_1}^2}{n_1}+\frac{{\sigma_2}^2}{n_2}}Z_{\alpha/2}\right)=\left(16000\ -25000\ \pm\sqrt{\frac{2000^2}{30}+\frac{8000^2}{40}}\times1.96\right)$$

$$=(-9000\ \pm 1316.561\times1.96)$$

$$=(\text{-}10316.56,\text{-}7683.44)$$

两个方差相等且未知，求均值差的置信区间

- 已知

$N(\mu_1, {\sigma_1}^2)$ $N(\mu_2, {\sigma_2}^2)$

$X_1, X_2, \dots X_{n_1}$ $Y_1, Y_2, \dots Y_{n_2}$

$\bar{X}, {S_1}^2$ $\bar{Y}, {S_2}^2$

$1-\alpha$

${\sigma_1}^2 = {\sigma_2}^2 = \sigma^2$ 未知

- 求均值差μ_1-μ_2的置信区间

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$$

$$S_w = \sqrt{\frac{(n_1-1){S_1}^2+(n_2-1){S_2}^2}{n_1+n_2-2}}$$

合并标准差

pooled standard deviation

两个方差相等且未知，求均值差的置信区间

- 已知

$N(\mu_1,{\sigma_1}^2)$ $\quad$ $N(\mu_2,{\sigma_2}^2)$

$X_1,X_2,\ldots X_{n_1}$ $\quad$ $Y_1,Y_2,\ldots Y_{n_2}$

$\bar{X},{S_1}^2$ $\quad$ $\bar{Y},{S_2}^2$

$1-\alpha$

${\sigma_1}^2={\sigma_2}^2=\sigma^2$ 未知

- 求均值差μ_1-μ_2的置信区间

$$\frac{\bar{X}-\mu}{S/\sqrt{n}}\sim t\,(n-1)\quad (\bar{X}\pm\frac{S}{\sqrt{n}}t_{\alpha/2}(n-1))$$

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2)$$

$$(\bar{X}-\bar{Y}\pm S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}t_{\alpha/2}(n_1+n_2-2))$$

例题

- 25左右人群的月收入服从正态分布$N(\mu_1, {\sigma_1}^2)$, 35左右人群的月收入服从正态分布$N(\mu_2, {\sigma_2}^2)$, σ_1, σ_2相等但未知；我们记录了30名25岁和40名35岁个体的月收入。这30名25岁个体平均收入为16,000元, 标准差为2500元, 这40名35岁个体平均收入为25,000元, 标准差为7000元。求μ_1-μ_2置信水平为95%的置信区间

$$S_w = \sqrt{\frac{(n_1 - 1){S_1}^2 + (n_2 - 1){S_2}^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(30-1)\times 2500^2 + (40-1)\times 7000^2}{30+40-2}}$$

$$= 5546.925$$

例题

- 25左右人群的月收入服从正态分布$N(\mu_1,{\sigma_1}^2)$, 35左右人群的月收入服从正态分布$N(\mu_2,{\sigma_2}^2)$, σ_1,σ_2相等但未知；我们记录了30名25岁和40名35岁个体的月收入。这30名25岁个体平均收入为16,000元, 标准差为2500元, 这40名35岁个体平均收入为25,000元, 标准差为7000元。求μ_1-μ_2置信水平为95%的置信区间

$$(\bar{X}-\bar{Y}\pm S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\underbrace{t_{\alpha/2}(n_1+n_2-2)}_{1.995}) = (-9000\ \pm 5546.925\times\sqrt{\frac{1}{30}+\frac{1}{40}}\times 1.995)$$

$$= (-11672.72, -6327.28)$$

两个方差不等且未知，求均值差的置信区间

- 已知

$N(\mu_1, {\sigma_1}^2)$ $N(\mu_2, {\sigma_2}^2)$

$X_1, X_2, \dots X_{n_1}$ $Y_1, Y_2, \dots Y_{n_2}$

$\bar{X}, {S_1}^2$ $\bar{Y}, {S_2}^2$

$1-\alpha$

${\sigma_1}^2, {\sigma_2}^2$ 未知且不等

- 求均值差μ_1-μ_2的置信区间

$$\boxed{\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}} \sim t$$

$\downarrow$

Student's t 统计量

William Sealy Gosset, who developed the "*t*-statistic" and published it under the pseudonym of "Student".

两个方差不等且未知，求均值差的置信区间

- 已知

$N(\mu_1, {\sigma_1}^2)$　　$N(\mu_2, {\sigma_2}^2)$

$X_1, X_2, \ldots X_{n_1}$　　$Y_1, Y_2, \ldots Y_{n_2}$

$\bar{X}, {S_1}^2$　　$\bar{Y}, {S_2}^2$

$1-\alpha$

${\sigma_1}^2, {\sigma_2}^2$ 未知且不等

- 求均值差μ_1-μ_2的置信区间

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{{S_1}^2}{n_1}+\frac{{S_2}^2}{n_2}}}$$

Welch's t
统计量

Bernard Lewis Welch

$$df=\frac{\left(\frac{{S_1}^2}{n_1}+\frac{{S_2}^2}{n_2}\right)^2}{\frac{({S_1}^2/n_1)^2}{n_1-1}+\frac{({S_2}^2/n_2)^2}{n_2-1}}$$

两个方差不等且未知，求均值差的置信区间

- 已知

$N(\mu_1, {\sigma_1}^2)$ $N(\mu_2, {\sigma_2}^2)$

$X_1, X_2, \ldots X_{n_1}$ $Y_1, Y_2, \ldots Y_{n_2}$

$\bar{X}, {S_1}^2$ $\bar{Y}, {S_2}^2$

$1-\alpha$

${\sigma_1}^2, {\sigma_2}^2$ 未知且不等

- 求均值差μ_1-μ_2的置信区间

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{{S_1}^2}{n_1}+\frac{{S_2}^2}{n_2}}} \longrightarrow$$

Welch's t 统计量

Bernard Lewis Welch

$$(\bar{X}-\bar{Y} \pm \sqrt{\frac{{S_1}^2}{n_1}+\frac{{S_2}^2}{n_2}}\, t_{\alpha/2})$$

例题

- 25左右人群的月收入服从正态分布$N(\mu_1, {\sigma_1}^2)$, 35左右人群的月收入服从正态分布$N(\mu_2, {\sigma_2}^2)$, σ_1, σ_2不等且未知；我们记录了30名25岁和40名35岁个体的月收入。这30名25岁个体平均收入为16,000元, 标准差为2500元, 这40名35岁个体平均收入为25,000元, 标准差为7000元。求μ_1-μ_2置信水平为95%的置信区间

$$df = \frac{\left(\frac{{S_1}^2}{n_1} + \frac{{S_2}^2}{n_2}\right)^2}{\frac{({S_1}^2/n_1)^2}{n_1 - 1} + \frac{({S_2}^2/n_2)^2}{n_2 - 1}} = \frac{\left(\frac{2500^2}{30} + \frac{7000^2}{40}\right)^2}{\frac{(2500^2/30)^2}{30 - 1} + \frac{(7000^2/40)^2}{40 - 1}} = 51.394$$

例题

- 25左右人群的月收入服从正态分布$N(\mu_1, {\sigma_1}^2)$, 35左右人群的月收入服从正态分布$N(\mu_2, {\sigma_2}^2)$, σ_1, σ_2不等且未知；我们记录了30名25岁和40名35岁个体的月收入。这30名25岁个体平均收入为16,000元, 标准差为2500元, 这40名35岁个体平均收入为25,000元, 标准差为7000元。求μ_1-μ_2置信水平为95%的置信区间

$$\left(\bar{X} - \bar{Y} \pm \sqrt{\frac{{S_1}^2}{n_1} + \frac{{S_2}^2}{n_2}}\, \underbrace{t_{\alpha/2}}_{2.007}\right) = \left(-9000 \pm \sqrt{\frac{2500^2}{30} + \frac{7000^2}{40}} \times 2.007\right)$$

$$= (\text{-}11402.82, \text{-}6597.18)$$

两个均值未知，求两个方差比的置信区间

- 已知

$N(\mu_1, {\sigma_1}^2)$ $N(\mu_2, {\sigma_2}^2)$

$X_1, X_2, \ldots X_{n_1}$ $Y_1, Y_2, \ldots Y_{n_2}$

$\bar{X}, {S_1}^2$ $\bar{Y}, {S_2}^2$

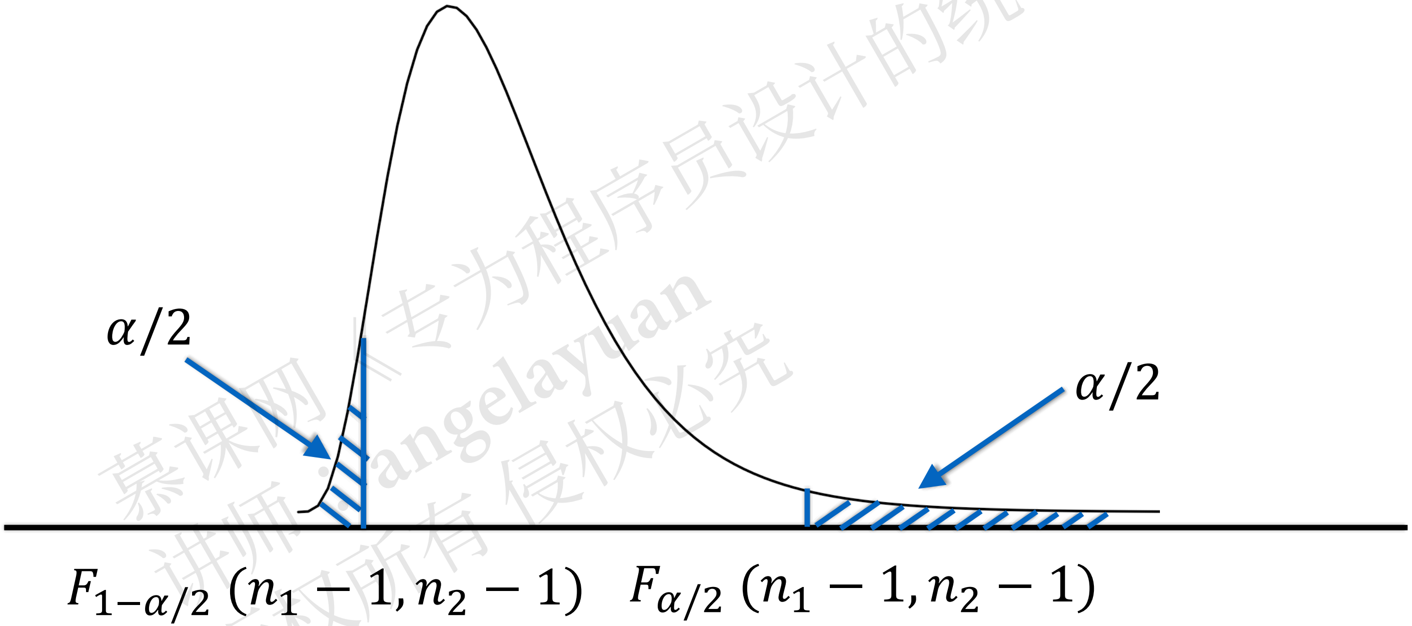
$1-\alpha$

μ_1, μ_2 未知

- 求方差比 ${\sigma_1}^2/{\sigma_2}^2$ 的置信区间

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

$$P\left\{F_{1-\alpha/2}(n_1-1,n_2-1)<\frac{S_1{}^2/S_2{}^2}{\sigma_1{}^2/\sigma_2{}^2}<F_{\alpha/2}(n_1-1,n_2-1)\right\}=1-\alpha$$



在密度函数不对称时，习惯上仍取对称的分位点

两个均值未知，求两个方差比的置信区间

- 已知

$N(\mu_1, {\sigma_1}^2)$ $N(\mu_2, {\sigma_2}^2)$

$X_1, X_2, \dots X_{n_1}$ $Y_1, Y_2, \dots Y_{n_2}$

$\bar{X}, {S_1}^2$ $\bar{Y}, {S_2}^2$

$1-\alpha$

μ_1, μ_2 未知

- 求方差比 ${\sigma_1}^2/{\sigma_2}^2$ 的置信区间

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

$$\left(\frac{{S_1}^2}{{S_2}^2}\frac{1}{F_{\alpha/2}\,(n_1-1, n_2-1)}, \frac{{S_1}^2}{{S_2}^2}\frac{1}{F_{1-\alpha/2}\,(n_1-1, n_2-1)}\right)$$

例题

- 25左右人群的月收入服从正态分布$N(\mu_1,{\sigma_1}^2)$, 35左右人群的月收入服从正态分布$N(\mu_2,{\sigma_2}^2)$, μ_1,μ_2未知；我们记录了$\boxed{30}$名25岁和$\boxed{40}$名35岁个体的月收入。这30名25岁个体收入的标准差为$\boxed{2500元}$, 这40名35岁个体收入的标准差为$\boxed{7000元}$。求${\sigma_1}^2/{\sigma_2}^2$ 置信水平为95%的置信区间

$$\left(\frac{{S_1}^2}{{S_2}^2}\underbrace{\frac{1}{F_{\alpha/2}\left(n_1-1,n_2-1\right)}}_{1.962},\frac{{S_1}^2}{{S_2}^2}\underbrace{\frac{1}{F_{1-\alpha/2}\left(n_1-1,n_2-1\right)}}_{0.492}\right)=\left(\frac{2500^2}{7000^2}\times\frac{1}{1.962},\frac{2500^2}{7000^2}\times\frac{1}{0.492}\right)$$

$$=(0.065,0.259)$$

- 待估参数
  - 均值差
    - 两个方差已知

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

    - 两个方差未知

      $$(\bar{X} - \bar{Y} \pm S_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{\alpha/2}(n_1 + n_2 - 2))$$

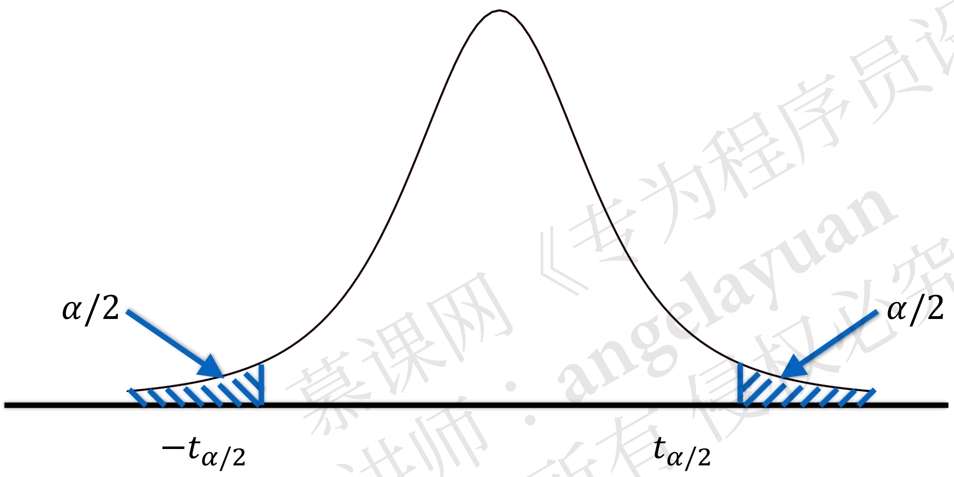
      $$(\bar{X} - \bar{Y} \pm \sqrt{\frac{S_1{}^2}{n_1} + \frac{S_2{}^2}{n_2}} t_{\alpha/2})$$

  - 方差比
    - 两个均值未知

      $$\left(\frac{S_1{}^2}{S_2{}^2} \frac{1}{F_{\alpha/2}\,(n_1 - 1, n_2 - 1)}, \frac{S_1{}^2}{S_2{}^2} \frac{1}{F_{1-\alpha/2}\,(n_1 - 1, n_2 - 1)}\right)$$

编程求解置信区间
两个正态总体的情况

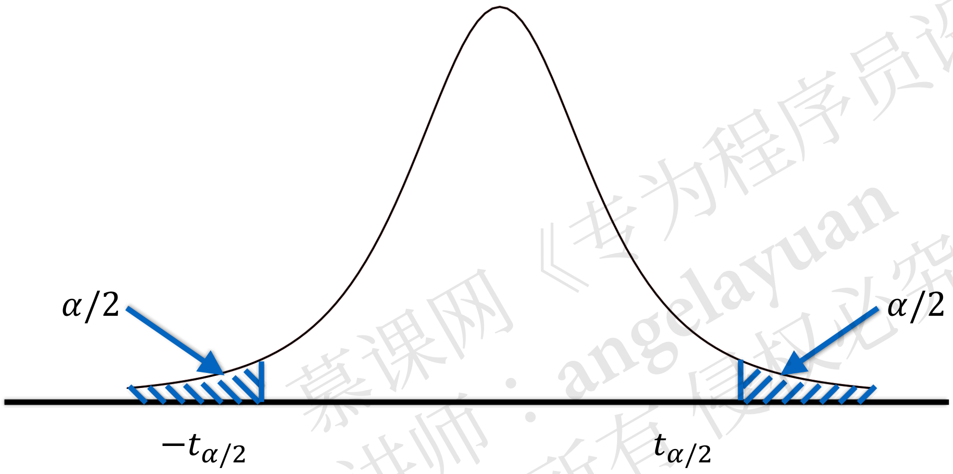
两个总体方差未知且相等，求均值差的置信区间



$$\left(\bar{X}-\bar{Y} \pm S_w \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\, t_{\alpha/2}(n_1+n_2-2)\right)$$

$$S_w=\sqrt{\frac{(n_1-1){S_1}^2+(n_2-1){S_2}^2}{n_1+n_2-2}}$$

两个总体方差未知且不等，求均值差的置信区间

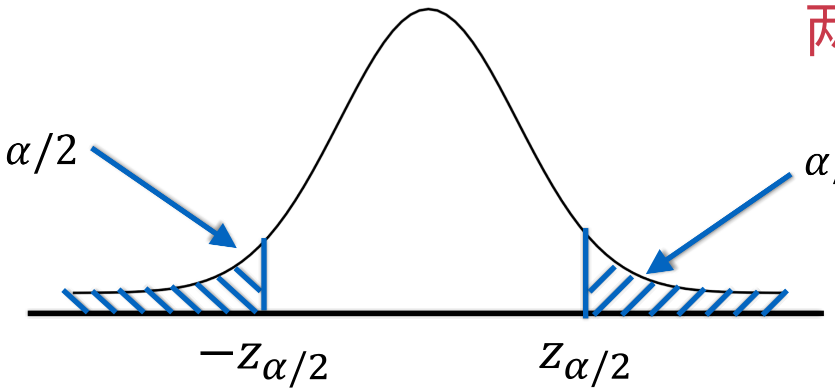


$$(\bar{X}-\bar{Y} \pm \sqrt{\frac{S_1{}^2}{n_1}+\frac{S_2{}^2}{n_2}}\, t_{\alpha/2})$$

$$df = \frac{\left(\frac{S_1{}^2}{n_1}+\frac{S_2{}^2}{n_2}\right)^2}{\frac{(S_1{}^2/n_1)^2}{n_1-1}+\frac{(S_2{}^2/n_2)^2}{n_2-1}}$$

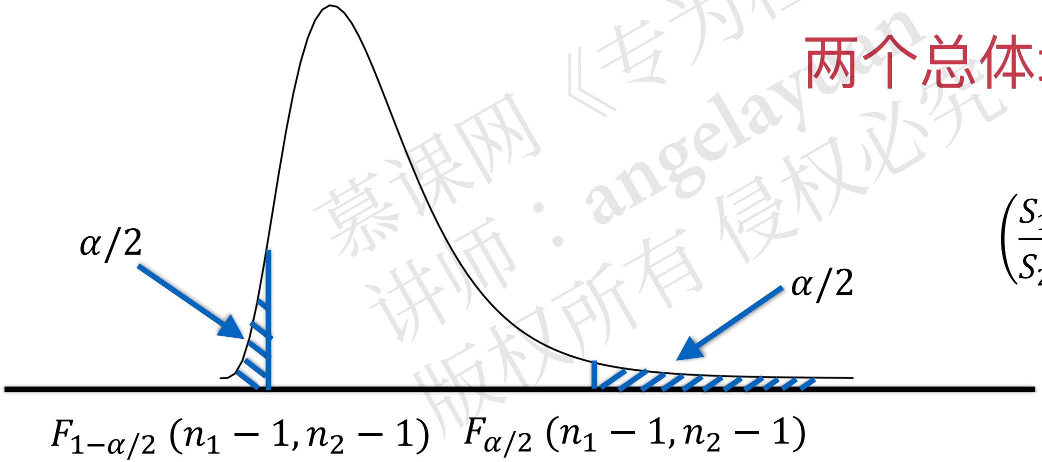
两个总体方差已知，求均值差的置信区间

$$\left(\bar{X} - \bar{Y} \pm \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}\, Z_{\alpha/2}\right)$$



两个总体均值未知，求方差比的置信区间

$$\left(\frac{{S_1}^2}{{S_2}^2}\frac{1}{F_{\alpha/2}\,(n_1-1, n_2-1)}, \frac{{S_1}^2}{{S_2}^2}\frac{1}{F_{1-\alpha/2}\,(n_1-1, n_2-1)}\right)$$



单侧置信区间

双侧置信区间

- 对于未知参数θ, 我们给出两个统计量 $\underline{\theta} = \underline{\theta}(X_1, X_2, \dots, X_n)$ 和 $\overline{\theta} = \overline{\theta}(X_1, X_2, \dots, X_n)$, 得到θ的双侧置信区间 $(\underline{\theta}, \overline{\theta})$

$$P\{\underline{\theta}(X_1, X_2, \dots, X_n) < \theta < \overline{\theta}(X_1, X_2, \dots, X_n)\} \geq 1 - \alpha$$

单侧置信区间

- 在某些实际问题中，我们只关心“上限”或者“下限”

  电池/灯泡的平均寿命;有害物质的平均含量

单侧置信下限

$$P\{\theta > \underline{\theta}(X_1, X_2, \dots, X_n)\} \geq 1-\alpha \longrightarrow (\boxed{\underline{\theta}}, +\infty)$$

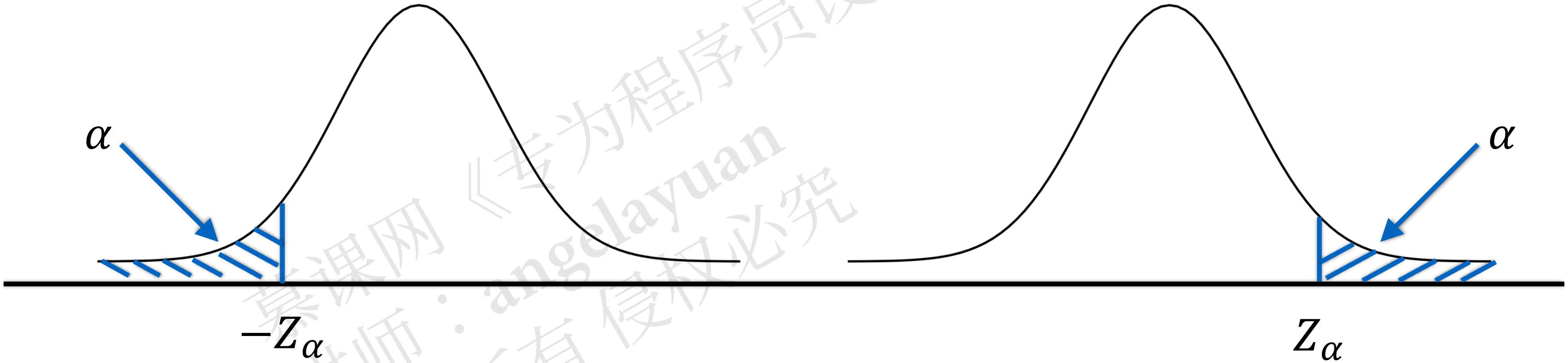
$$P\{\theta < \overline{\theta}(X_1, X_2, \dots, X_n)\} \geq 1-\alpha \longrightarrow (-\infty, \boxed{\overline{\theta}})$$

置信水平为1-α的

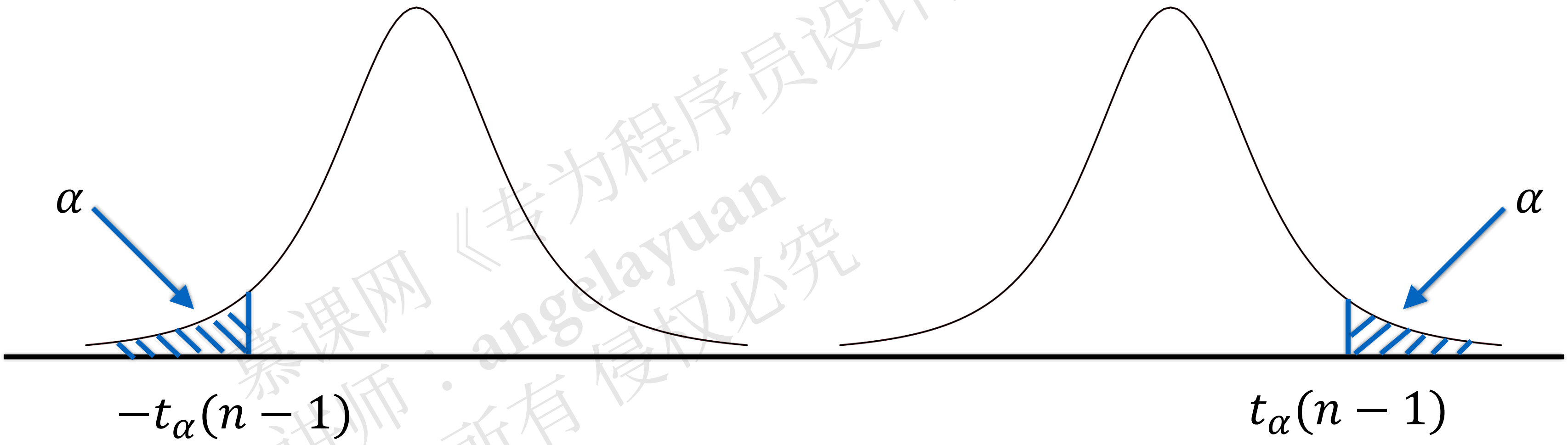
单侧置信区间

单侧置信上限

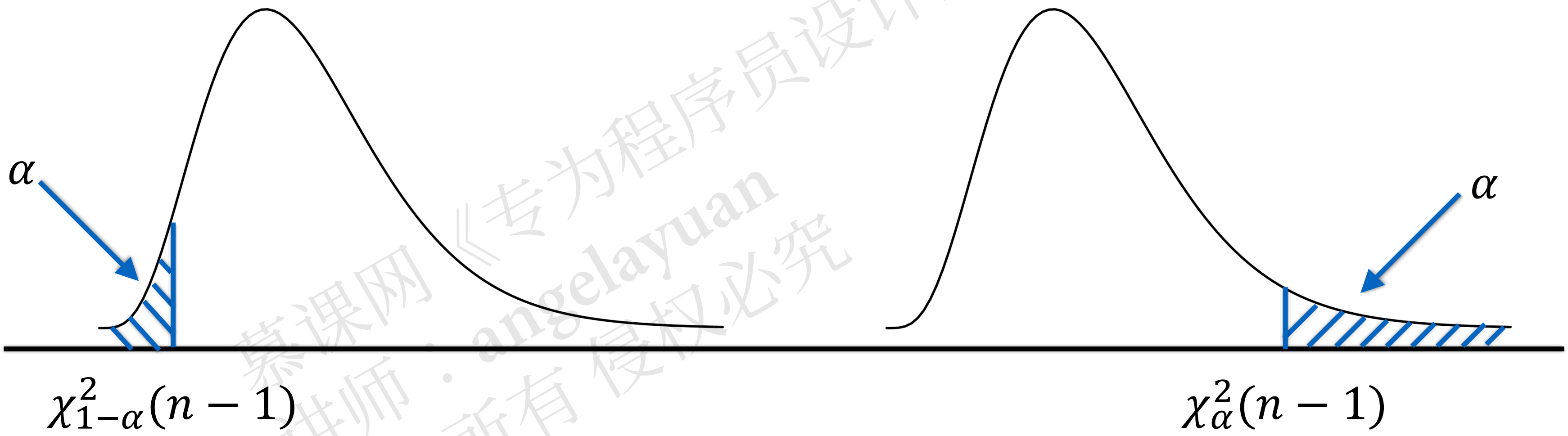
单侧置信区间 - 标准正态分布



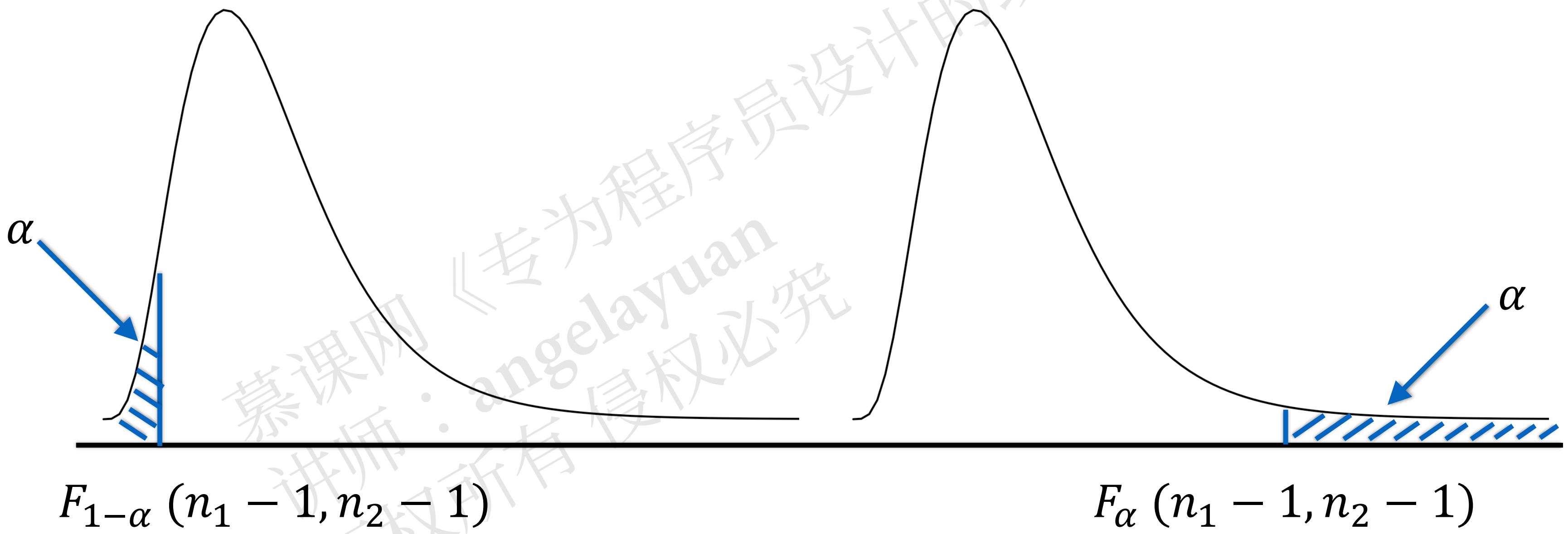
单侧置信区间 - t分布



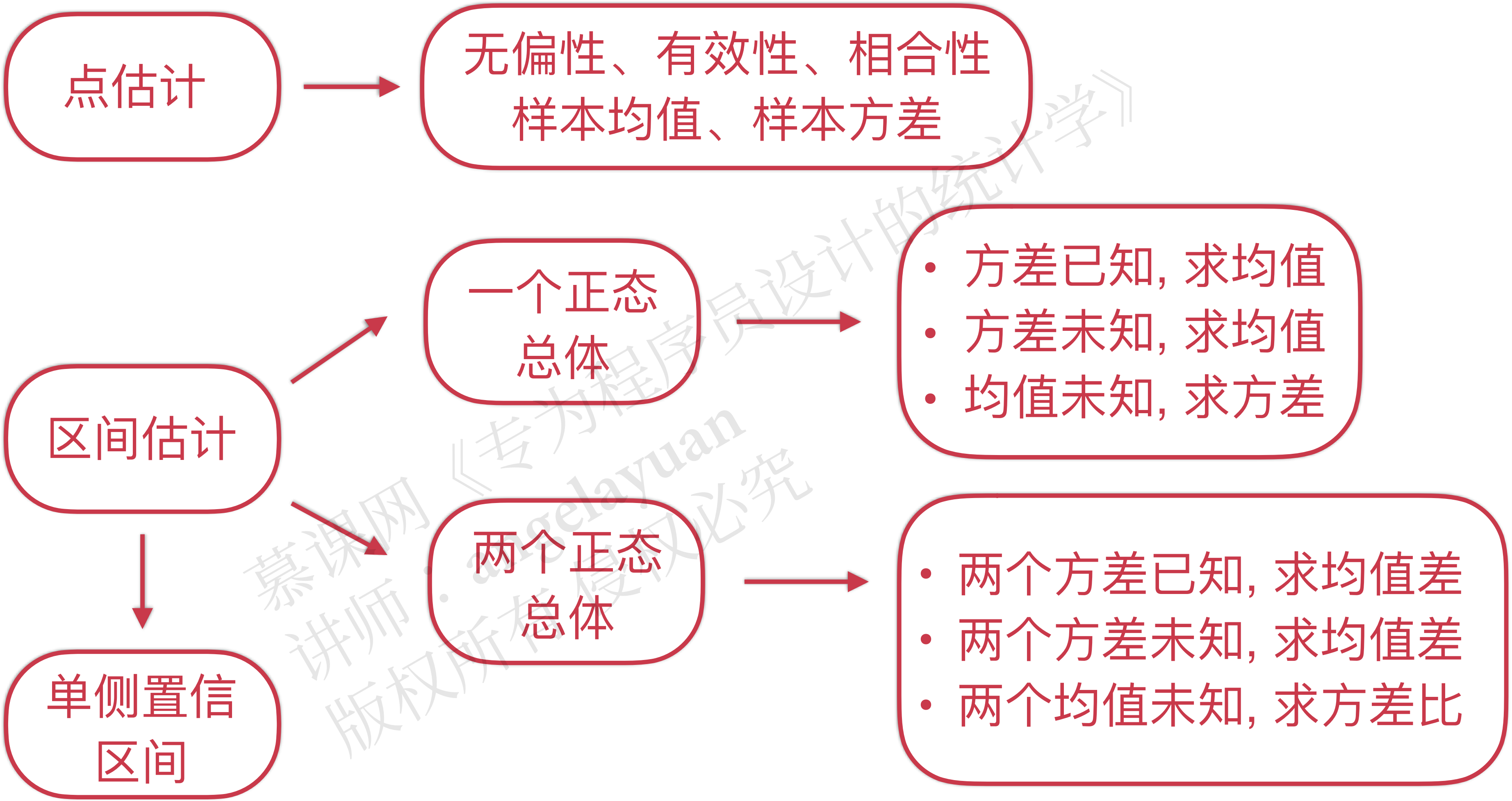
单侧置信区间 - 卡方分布



单侧置信区间 - F分布



本章小结


点估计
无偏性、有效性、相合性
样本均值、样本方差
区间估计
一个正态总体
• 方差已知，求均值
• 方差未知，求均值
• 均值未知，求方差
两个正态总体
• 两个方差已知，求均值差
• 两个方差未知，求均值差
• 两个均值未知，求方差比
单侧置信区间

非正态总体 或 统计量的抽样分布未知

还可以使用非参数的方法来寻找置信区间，留到非参数章再讲