第7章 抽样分布

概率分布

抽样分布

7.1 概率分布

7.1.1 概率的基本概念

- **随机的三个定义：** 随机现象　随机试验　随机事件
- **概率的三个定义：** 古典概率　统计概率　主观概率

7.1.2 随机变量与概率函数

7.1.2.1随机变量

- **定义：** 一次随机试验的结果的数值性描述
- **定义域：** 样本空间；**值域：** 实数集合
- **种类：** 离散型　连续型
- **注：** 一般以 X、Y、Z 来表示
- **例：** 投掷两次硬币正面向上的次数

7.1.2.2 概率函数

- **定义：** 能够将随机变量的所有取值与相应的概率值对应起来的函数，称为此随机变量的概率函数
- **公式：** 一般用 P（x）来表示, $0 \leq P(X) \leq 1$

7.1.2.3 联系

- **随机变量函数：** 定义域样本空间---值域实数集合
- **概率函数：** 定义域实数集合---值域实数(概率值)集合
- **两者结合：** 样本空间---概率值

7.1.3 概率分布的基本概念

7.1.3.1 离散型随机变量

- **概率分布：**指某种表格、图形、公式或其他设计，并且该设计指定了该离散变量所有可能的取值及其相应概率值
- **数学期望：**描述变量取值的集中程度

$$E(X)=\sum_{i=1}^{n}x_i p_i;\quad E(X)=\sum_{i=1}^{\infty}x_i p_i$$

- **方差：**描述变量取值的分散程度

$$D(X)=E[X-E(X)]^2\quad D(X)=\sum_{i=1}^{n(\infty)}\left[x_i-E(X)\right]^2\cdot p_i$$

7.1.3.2 连续型随机变量

- **概率密度函数：** 设X为连续型随机变量，不能列出每一个x值及其相应的概率，故近似认为当变量值个数趋于无穷大而取值区间趋于0时，如果可用一个公式或运算来刻画概率分布的性质，则这个函数称为X的概率密度函数，记为*f(x)*：$f(x)\geq 0;\int_{-\infty}^{+\infty} f(x)\mathrm{d}x=1$
- **概率分布函数：** $F(x)=P(X\leq x)=\int_{-\infty}^{x} f(t)\mathrm{d}t\ \ (-\infty<x<+\infty)$
- **数学期望：** $E(X)=\int_{-\infty}^{+\infty} xf(x)\mathrm{d}x=\mu$
- **方差：** $D(X)=\int_{-\infty}^{+\infty}\left[x-E(X)\right]^2 f(x)\mathrm{d}x=\sigma^2$

7.1.4 常用的概率分布

7.1.4.1 二项分布

- **定义：** 进行 n 次贝努里试验，出现“成功”的次数的概率分布称为二项分布，记为$b(n,p)$, p为“成功”的概率, 设X为n次重复试验中某事件出现的次数，X取x的概率为

$$P\{X=x\}=C_n^x p^x q^{n-x};\ C_n^x=\frac{n!}{x!(n-x)!}\quad (x=0,1,2,\cdots,n)$$

- **期望和方差：** $E(X)=np$；$D(X)=npq$

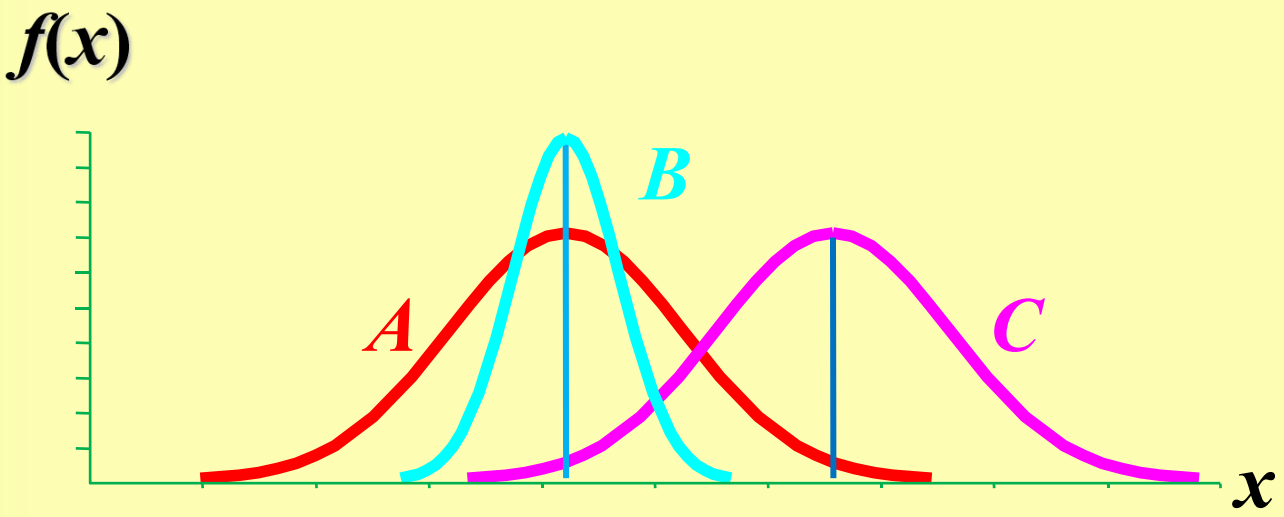
7.1.4.2 正态分布

- **概率密度函数：** $f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$ ， $-\infty<x<\infty$

$f(x)$ = 随机变量 X 的频数； μ ： 总体均值、σ^2：总体方差；
π =3.14159； e = 2.71828； x = 随机变量的取值 (-∝ $< x <$ ∝)

- **重要性：** 描述连续型随机变量的最重要的分布；经典统计推断的基础；可用于近似离散型随机变量的二项分布

- **函数性质：** $N(\mu,\sigma^2)$ 在均值μ达到最高点，关于μ对称分布；是以μ和σ为参数的分布族，μ为位置参数， σ为形状参数；f(x)>0，曲线尾端无限延伸，以横轴为渐近线



- **标准正态分布函数：** 一般的正态分布$X\sim N(\mu,\sigma^2)$，

  通过线性变换 $Z=\dfrac{X-\mu}{\sigma}$ 转化为标准正态分布$Z\sim N(0,1)$

- **标准正态分布表的使用：**

  当 $z>0$，$\phi(z)=\int_{-\infty}^{z}\varphi(x)\mathrm{d}x$

  当 $z<0$ ，$\Phi(z)=1-\Phi(-z)$

  一般的正态分布 $X\sim N(\mu,\sigma^2)$

$$P(a\le X\le b)=\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)$$

7.2 抽样分布

7.2.1 基本概念

7.2.1.1 统计推断

- **抽样：** 指从研究总体中，依据一定的原则，抽取部分总体单位组成样本，对样本数据进行测量。
- **统计推断：** 也称抽样推断，是指根据样本数据，基于概率论的相关知识，对总体的数量特征进行推断，基本方法为参数估计和假设检验。

7.2.1.2 样本统计量

部分指标，如均值、比例、方差，当用来描述样本数量特征时，就可以称为样本统计量；统计推断中的抽样方法默认为概率抽样，则**样本统计量是随机变量**

符号对应表

名称	样本	总体
容量	n	N
均值	$\bar{x}$	μ
比例	$\tilde{p}$	p
方差	s^2	σ^2
标准差	s	σ

7.2.1.3 总体参数

描述总体特征各类指标, 称为**总体参数**

7.2.1.4 抽样分布

重复概率抽样下选取样本时，由某统计量的所有可能的取值形成的频数分布，就是样本统计量的概率分布，则：

样本统计量所服从的概率分布为该统计量的抽样分布。

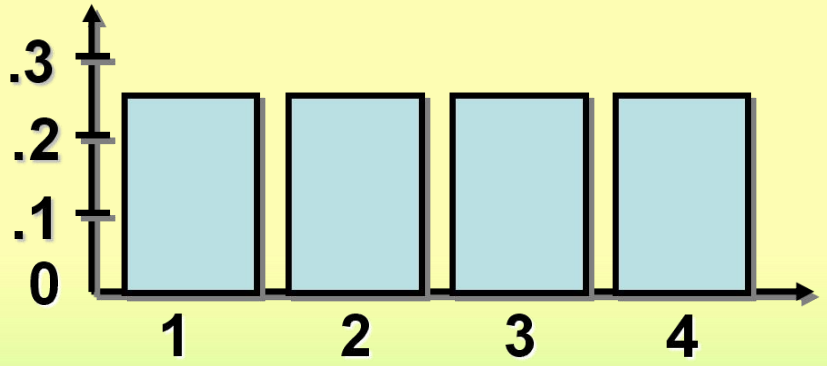
7.2.2 样本均值 $\bar{x}$ 的抽样分布

7.2.2.1 构造一个例子

- 例：设一个总体，含有4个单位，即总体容量N=4；X_1=1,X_2=2,X_3=3,X_4=4；总体均值、方差及分布如下

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = 2.5 \qquad \sigma^2 = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N} = 1.25$$

总体分布



➢ 列举所有样本：

现从总体中抽取n＝2的简单随机样本，在重复抽样条件下，共有4^2=16个样本

第一个值	第二个值			
	1	2	3	4
1	•1,1	•1,2	•1,3	•1,4
2	•2,1	•2,2	•2,3	•2,4
3	•3,1	•3,2	•3,3	•3,4
4	•4,1	•4,2	•4,3	•4,4

- 计算各样本的均值：

16个样本的均值				
第一个值	第二个值			
	1	2	3	4
1	1.0	1.5	2.0	2.5
2	1.5	2.0	2.5	3.0
3	2.0	2.5	3.0	3.5
4	2.5	3.0	3.5	4.0

- 计算样本统计量的值：

计算所有样本均值的平均值和方差(M为样本数目)

$$\mu_{\bar{x}} = \frac{\sum_{i=1}^{M} \bar{x}_i}{M} = \frac{1.0 + 1.5 + \cdots + 4.0}{16} = 2.5 = \mu$$

$$\sigma_{\bar{x}}^2 = \frac{\sum_{i=1}^{M} (\bar{x}_i - \mu_{\bar{x}})^2}{M} = \frac{(1.0-2.5)^2 + \cdots + (4.0-2.5)^2}{16} = 0.625 = \frac{\sigma^2}{n}$$

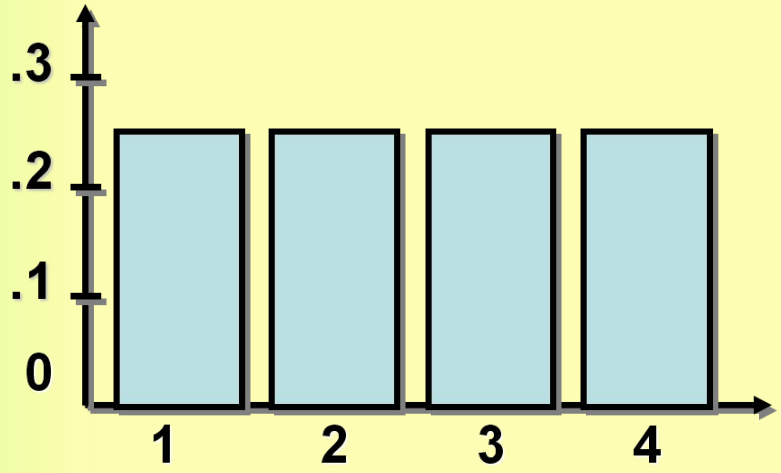
7.2.2.2 本例结论

- 数学期望：样本均值的均值等于总体均值，$\mu_{\bar{x}} = \mu$
- 方差：重复(有放回)抽样时, 样本均值的方差等于总体方差的1/n，$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$
- 方差的矫正：不重复抽样(无放回)时, 方差的矫正系数(N-n)/(N-1) 或1-n/N，$\sigma_{\bar{x}}^2 = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n} \approx (1 - \frac{n}{N}) \cdot \frac{\sigma^2}{n}$

现实中，无放回抽样的抽样比 n/N<5% 时，可忽略矫正。

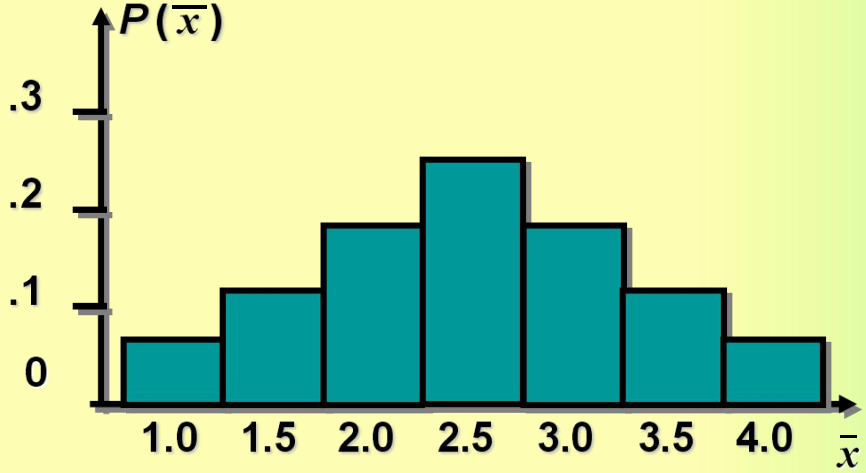
7.2.2.3 比较与结论

- 总体分布



$\mu = 2.5 \quad \sigma^2 = 1.25$

- 抽样分布



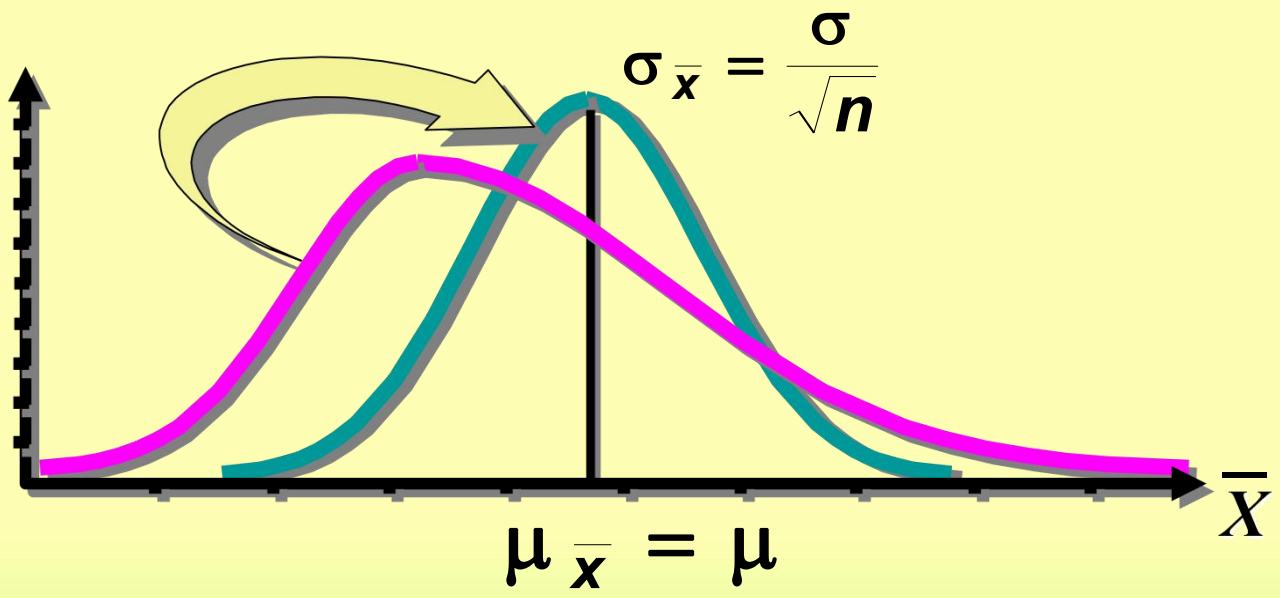
$\mu_{\bar{x}} = 2.5 \quad \sigma_{\bar{x}}^2 = 0.625$

- **结论**：当总体服从正态分布X~$N(\mu,\sigma^2)$ 时，样本统计量 $\overline{x}$ 也服从正态分布(中心极限定理)，且(忽略校正)

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

7.2.2.4 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，$\overline{x}$ 的抽样分布近似服从均值为μ、方差为σ^2/n的正态分布；一般的 $n \geq 30$



7.2.3 样本均值差 $\bar{x}_1-\bar{x}_2$ 的抽样分布

- **假设条件：** 两个总体X_1、X_2 都服从正态分布，参数为

$$X_1 \sim N(\mu_1,\sigma_1^2);\ X_2 \sim N(\mu_2,\sigma_2^2)$$

各自独立随机抽取容量为 n_1,n_2 两个样本；

或者总体不是正态分布， 则 $n_1 \geq 30$；$n_2 \geq 30$

- **抽样分布：** $\bar{x}_1 - \bar{x}_2$ 服从(近似)正态分布 (忽略校正系数)

$$E(\bar{x}_1-\bar{x}_2)=\mu_1-\mu_2 \quad \sigma_{(\bar{x}_1-\bar{x}_2)}=\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

7.2.4 样本比例 $\tilde{p}$ 的抽样分布

总体比例指总体中具有某种特征的单位在总体中所占的比例

- **假定条件：** 设p为总体比例；总体服从二项分布

  $b(np, npq)$，可以由正态分布来近似

- **抽样分布：** 样本比例 $\tilde{p} = a/n$ 近似服从正态分布

  $\tilde{p} \sim N(p, pq/n)$， 一般的 np和nq皆大于5

7.2.5 样本比例差 $\tilde{p}_1-\tilde{p}_2$ 的抽样分布

- **假设条件：** 当 n_1，$n_2 \geq 30$，$\tilde{p}_1, \tilde{p}_2$ 服从正态分布

  $N(p_1, p_1q_1/n_1)$ 和 $N(p_2, p_2q_2/n_2)$

- **抽样分布：** 样本比例差 $\tilde{p}_1-\tilde{p}_2$ 近似服从正态分布

$$N(p_1 - p_2, \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2})$$

注： $n_1p_1, n_2p_2, n_1q_1, n_2q_2 > 5$

7.2.6 样本方差s^2 的抽样分布

- **假设条件：** 设总体服从正态分布，构造样本方差 s^2 的统计量为

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

- **抽样分布：** $\chi^2(n-1)$称为自由度为$(n\text{-}1)$的卡方分布

注：$s^2 = \sum_{i=1}^{n} (x - \overline{x})^2 / (n-1)$

卡方分布：

阿贝(Abbe) 于1863年首先提出，后来由海尔墨特(Hermert)和卡·皮尔逊(K·Pearson) 分别于1875年和1900年推导得出。

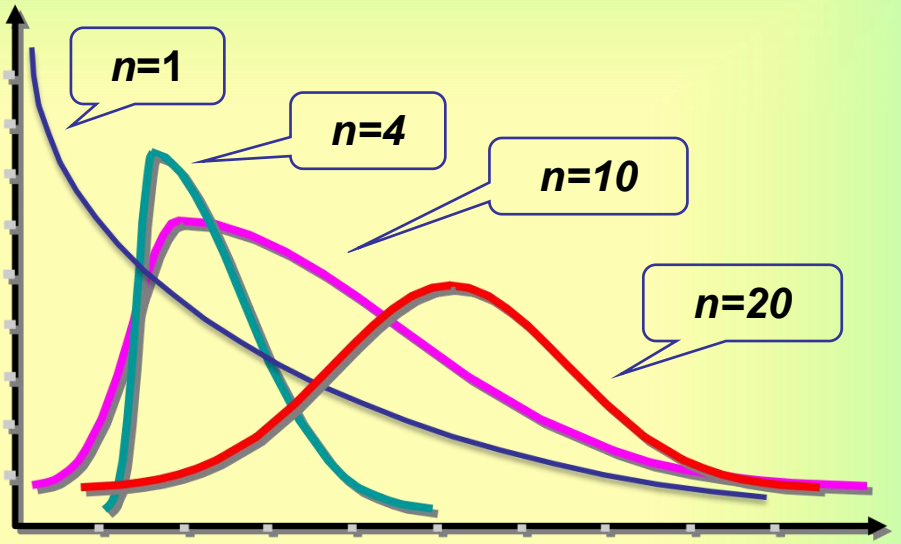
设总体 $X \sim N(\mu, \sigma^2)$ ，

则 $z = \frac{X-\mu}{\sigma} \sim N(0,1)$

令 $Y = z^2$，则 $Y \sim \chi^2(1)$；

Y_i独立同分布，$\sum_{i=1}^{n} Y \sim \chi^2(n)$

当总体$X \sim N(\mu, \sigma^2)$，从中抽取容量为n的样本，则 $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2} \sim \chi^2(n-1)$



特点：分布的变量值始终为正；分布的形状取决于自由度n的大小，由不对称的右偏分布，随着n的增大逐渐趋于对称；$E(\chi^2)=n$，$D(\chi^2)=2n$；

可加性：若$X \sim \chi^2(n_1)$，$Y \sim \chi^2(n_2)$，X和Y独立，则$X+Y \sim \chi^2(n_1+n_2)$。

7.2.7 样本方差比 s_1^2/s_2^2 的抽样分布

- **假设条件**：两个正态总体X_1 , X_2 相互独立

$$X_1 \sim N(\mu_1,\sigma_1^2);\ X_2 \sim N(\mu_2,\sigma_2^2)$$

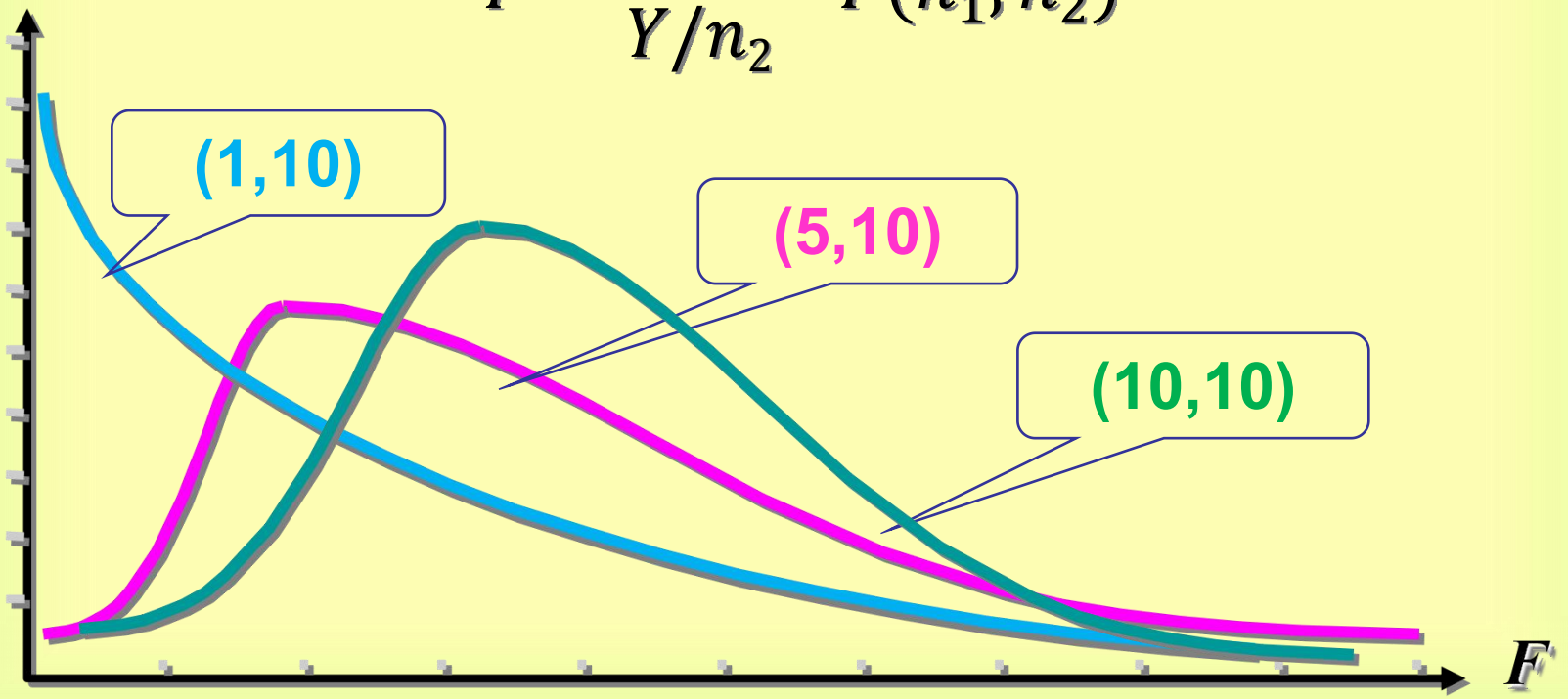
- **构造样本统计量**：$F=\dfrac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2}=\dfrac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}\sim F(n_1-1,n_2-1)$

- **抽样分布**：$F(n_1-1,n_2-1)$ 称分子自由度(第一自由度)为$(n_1\text{-}1)$、分母自由度(第二自由度)为$(n_2\text{-}1)$的F分布

F分布：

由统计学家费希尔(R.A.Fisher) 提出，以其姓氏第一个字母命名；设若$X\sim\chi^2(n_1)$，$Y\sim\chi^2(n_2)$,且X和Y相互独立，则称F为服从自由度n_1和n_2的F分布

$$F=\frac{X/n_1}{Y/n_2}\sim F(n_1,n_2)$$



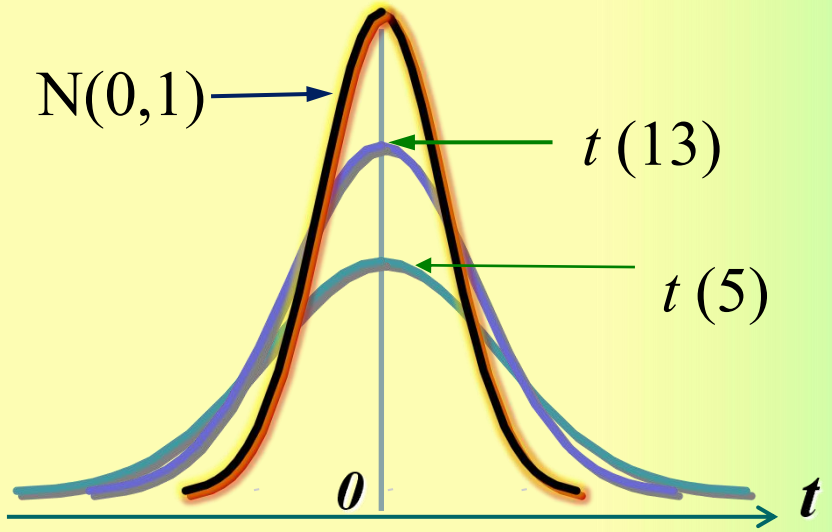
7.2.8 样本均值的t分布

➢ 构造统计量：

设X_1，X_2，…，X_n是来自 $N(\mu,\sigma^2)$的一个样本，则t 统计量服从自由度为 $(n-1)$的 t 分布

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

不同自由度的t分布与标准正态布



➢ t 分布

戈塞特(W.S.Gosset)于1908年在一篇以“Student”为笔名的论文中提出；

设X~N(0,1)，Y~$\chi^2(n)$，且X与Y独立，

则 $t=\frac{X}{\sqrt{Y/n}}\sim t(n)$

t 分布通常比标准正态分布分散，也是以纵轴为对称轴的对称分布族，随着自由度n的增大(约n=30)，逐渐趋于N(0,1)。