生物统计学

第四章 抽样试验与抽样分布

云南大学 生命科学学院

會澤百家 至公天下

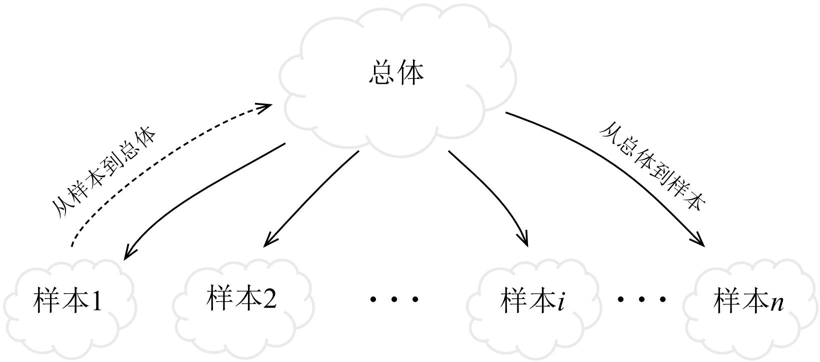

图 **4.1** 统计学的研究方向

1. 抽样试验
2. 单一总体样本统计量的分布
3. 两个总体样本统计量的分布
4. 抽样分布的分类

抽样必须符合随机原则，

即保证总体中的每一个个体在一次抽样中都有相同的概率被选中。

抽样必须符合随机原则，
即保证总体中的每一个个体在一次抽样中都有相同的概率被选中。

对于无限总体和个体数量极大的有限总体，抽样试验 (sampling experiment) 抽取部分个体后不会影响后续抽出样本被抽中的概率，可以保障抽样的随机性。

抽样必须符合随机原则，
即保证总体中的每一个个体在一次抽样中都有相同的概率被选中。

对于无限总体和个体数量极大的有限总体，抽样试验 (sampling experiment) 抽取部分个体后不会影响后续抽出样本被抽中的概率，可以保障抽样的随机性。

对于个体数量较少的有限总体，需进行放回式的重置抽样 (sampling with replacement)。

```
> sample_norm <- rnorm(n = 5000, mean = 0, sd = 1)
> hist(sample_norm)
```

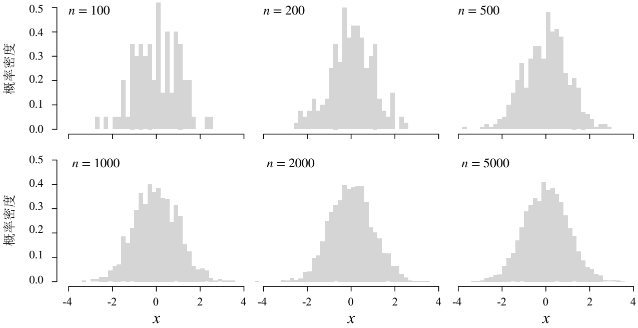

图 **4.2** 抽样自标准正态分布的样本分布 (n 为样本容量)

假设对一个由 10 个数字 $(0, 1, 2, \cdots, 9)$ 构成的总体 $(\mu = 4.5, \sigma^2 = 8.25, \sigma \approx 2.872)$，进行重置抽样，每次抽取 5 个数字。

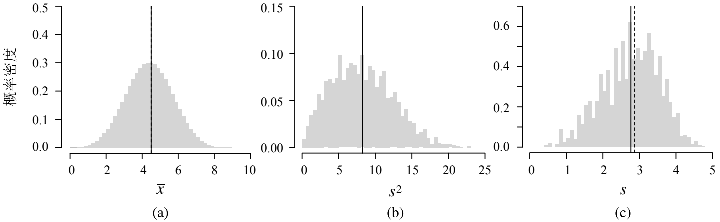

图 **4.3** 统计量的样本分布

统计学上，统计量的概率分布称为抽样分布 (sampling distribution)，包括样本平均数的分布、样本方差的分布等。

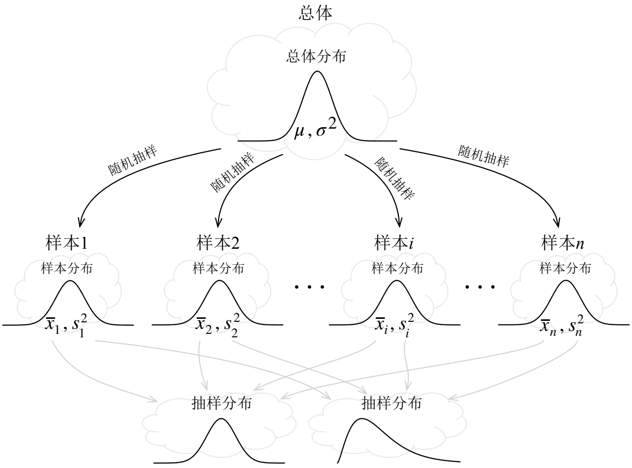

图 **4.4** 总体分布、样本分布和抽样分布的关系

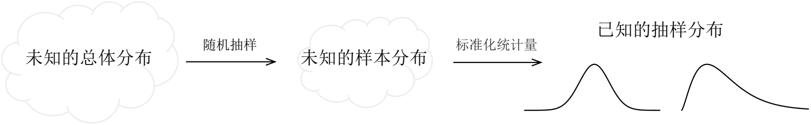

图 **4.5** 数据统计分析的基本逻辑

4.1 抽样试验

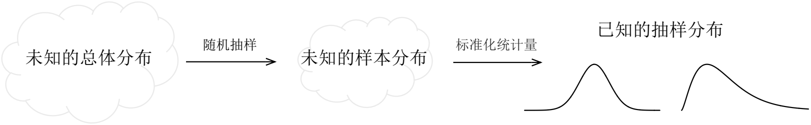

图 **4.5** 数据统计分析的基本逻辑

抽样分布可以解决以下两个问题：

- 总体 参数未知 时，通过样本统计量对总体参数做出具有概率意义的估计（参数估计）；
- 总体 参数已知（或假设已知） 时，对样本统计量的概率行为做出判断（假设检验）。

从某一已知平均数为 μ、方差为 σ^2 的总体 (不限于正态总体) 中随机抽取样本，根据中心极限定理，样本平均数服从 $N(\mu, \frac{\sigma^2}{n})$。

$$z = \frac{\overline{x} - \mu}{\sigma_{\overline{x}}} = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \tag{4.1}$$

服从标准正态分布，即 $z \sim N(0, 1)$。其中 $\sigma_{\overline{x}}$ 称为平均数的标准误 (standard error of the mean)，或总体标准误。

$$
\begin{aligned}
\operatorname{Var}(\overline{x}) &= \operatorname{Var}\left(\frac{x_1+x_2+\cdots+x_n}{n}\right) \\
&= \operatorname{Var}\left(\frac{1}{n}x_1+\frac{1}{n}x_2+\cdots+\frac{1}{n}x_n\right) \\
&= \frac{1}{n^2}\operatorname{Var}\left(x_1+x_2+\cdots+x_n\right) \\
&= \frac{1}{n^2}\left(\operatorname{Var}(x_1)+\operatorname{Var}(x_2)+\cdots+\operatorname{Var}(x_n)\right) \\
&= \frac{1}{n^2}(\sigma^2+\sigma^2+\cdots+\sigma^2) \\
&= \frac{\sigma^2}{n}
\end{aligned}
\tag{4.2-4.5}
$$

从某一已知平均数为 μ、方差未知的总体（不限于正态总体）中随机抽取样本，样本平均数标准化后，得标准化统计量

$$t = \frac{\overline{x} - \mu}{s_{\overline{x}}} = \frac{\overline{x} - \mu}{s/\sqrt{n}} \tag{4.7}$$

服从自由度 $\mathrm{df} = n - 1$ 的 t 分布，即 $t \sim t(n-1)$。

自由度为 $n-1$ 的 t 分布有概率密度函数:

$$f(x)=\frac{\mathit{\Gamma}(\frac{n}{2})}{\sqrt{\pi(n-1)}\mathit{\Gamma}(\frac{n-1}{2})}\left(1+\frac{x^2}{n-1}\right)^{-\frac{n}{2}} \tag{4.8}$$

其中 $\mathit{\Gamma}(x)=\int_0^{+\infty}t^{x-1}\mathrm{e}^{-t}\mathrm{d}t \quad (x>0)$ 称为 Gamma 函数，是阶乘函数在实数域（包括复数域）上的扩展。

t 分布在自由度 $n-1>1$ 时有数学期望 0，在自由度 $n-1>2$ 时有方差 $\frac{n}{n-2}$。

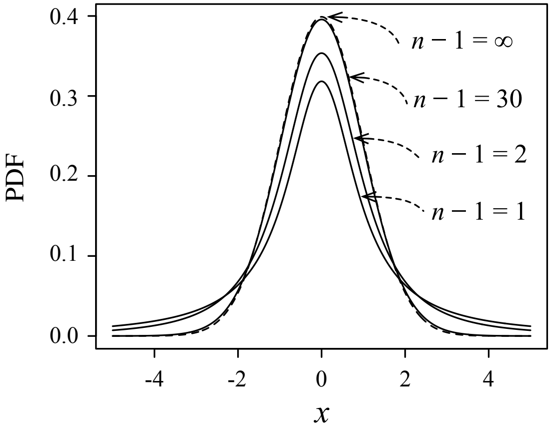

图 **4.6** 不同自由度的 t 分布

t 分布有以下性质：

1. 密度曲线关于平均数 $\mu_t = 0$ 的峰值点左右对称，两侧递降。
2. 密度曲线的形态受自由度 $\mathrm{df} = n - 1$ 的制约，每个自由度对应一条密度曲线。
3. 与标准正态分布密度曲线相比，t 分布密度曲线峰值点低于标准正态分布，而双侧尾部高于标准正态分布。当自由度 $\mathrm{df} \geq 30$ 时，t 分布曲线接近标准正态分布曲线，当 $\mathrm{df} \to \infty$ 时，两种密度曲线完全重合。
4. 概率的归一性决定了 t 分布密度曲线之下的（与 x 轴所夹的）面积为 1。

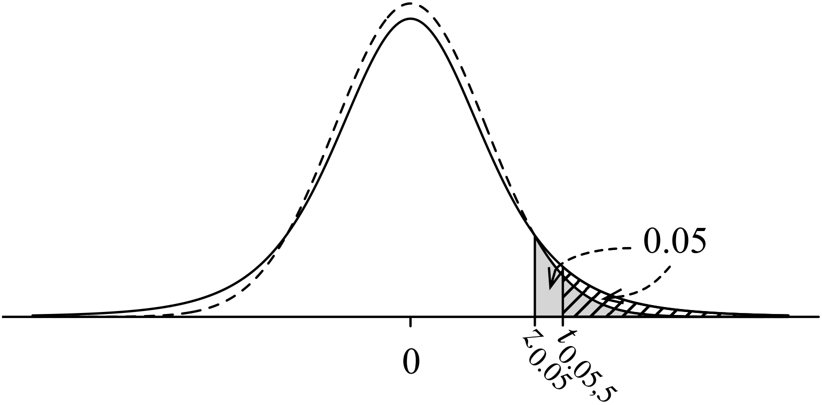

图 **4.7** t 分布与标准正态分布上侧分位数的比较

William S. Gosset (1876-1937)

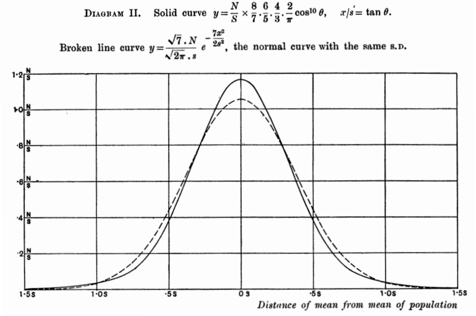

图 **4.8** 1908 年文章中 Gosset 绘制的正态分布（实线）和 t 分布（虚线，自由度为 9）的密度曲线

服从两点分布 $B(1,p)$ 的随机变量用来描述一次试验中事件 A 发生的情况，其中事件发生的概率 p，也就是总体比率(population proportion)。

伯努利试验重复 n 次，记录事件 A 发生的次数 m，$\frac{m}{n}$ 即为样本比率(sample proportion)，记作 $\hat{p}$。

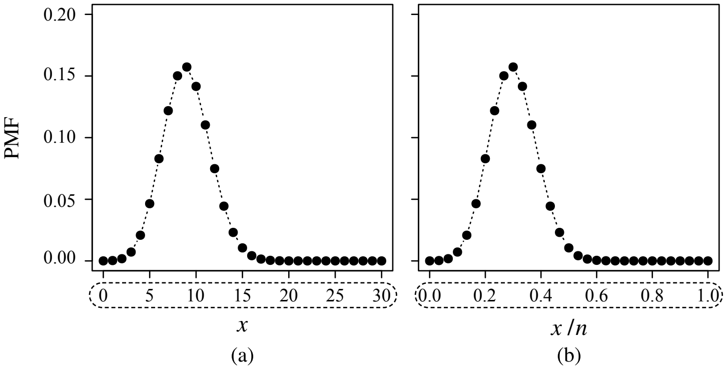

图 **4.9** 二项分布 (a) 与样本比率 (b) 的分布

据De Moivre-Laplace 中心极限定理，n 重伯努利试验中服从二项分布的事件发生 m 次，当 n 很大时，有 $\frac{m-np}{\sqrt{np(1-p)}}$ 以标准正态分布为极限。

因 $\hat{p}=\dfrac{m}{n}$，所以

$$\frac{n\hat{p}-np}{\sqrt{np(1-p)}}=\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \tag{4.9}$$

近似服从标准正态分布。

从标准正态分布 $N(0,1)$ 中抽取 n 个独立的样本 $z_1, z_2, \cdots, z_n$，取平方后求和得统计量

$$\chi^2 = \sum_{i=1}^{n} z_i^2 \tag{4.10}$$

服从自由度为 $\mathrm{df} = n$ 的 χ^2 分布，即 $\chi^2 \sim \chi^2(n)$。

自由度为 n 的 χ^2 分布有概率密度函数:

$$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} \mathrm{e}^{-\frac{x}{2}}, \quad x \geq 0 \tag{4.11}$$

服从自由度为 n 的 χ^2 分布的随机变量 X 有数学期望 n 和方差 $2n$。

$$\chi^2 = \sum_{i=1}^{n} z_i^2 = \sum_{i=1}^{n} (\frac{x_i - \mu}{\sigma})^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{4.12}$$

当总体平均数 μ 未知时，用样本平均数代替，进而有

$$\chi^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{4.13}$$

又因 $s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$，上式可变为

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \tag{4.14}$$

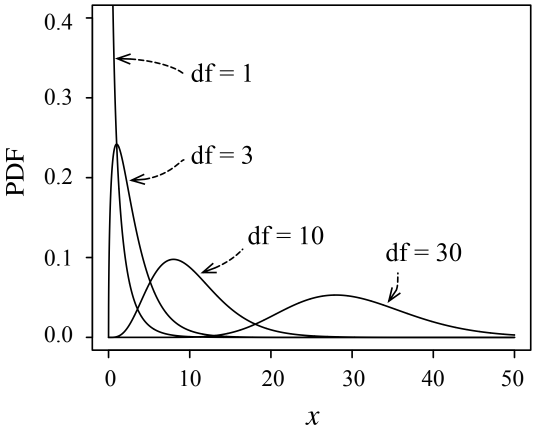

图 **4.10** 不同自由度的 χ^2 分布

χ^2 分布有以下性质：

1. χ^2 分布在区间 $(0, +\infty)$ 内其密度曲线呈右偏斜的形态；
2. χ^2 分布偏斜度随自由度的降低而增大，当自由度增大时会趋于左右对称；
3. χ^2 分布无论自由度有多大，密度曲线下的面积为 1（概率的归一性）；
4. χ^2 分布具有可加性，如 $X \sim \chi^2(n_1)$，$Y \sim \chi^2(n_2)$，且 X 和 Y 两随机变量相互独立，那么 $X + Y \sim \chi^2(n_1 + n_2)$。

从平均数分别为 μ_1 和 μ_2、方差分别为 σ_1^2 和 σ_2^2 的两个总体中，独立抽取容量为 n_1 和 n_2 的两组样本，则两组样本的平均数之差服从正态分布，标准化后得统计量

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\overline{x}_1 - \overline{x}_2}} \tag{4.17}$$

服从标准正态分布，即 $z \sim N(0,1)$。其中 $\sigma_{\overline{x}_1-\overline{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$。

$$
\begin{aligned}
\mathrm{Var}(\overline{x}_1-\overline{x}_2) &= \mathrm{Var}(\overline{x}_1)+\mathrm{Var}(-\overline{x}_2)=\mathrm{Var}(\overline{x}_1)+(-1)^2\mathrm{Var}(\overline{x}_2)\\
&= \mathrm{Var}(\overline{x}_1)+\mathrm{Var}(\overline{x}_2)\\
&= \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}
\end{aligned}
\tag{4.19}
$$

如两总体方差未知，则有标准化统计量

$$t=\frac{(\overline{x}_1-\overline{x}_2)-(\mu_1-\mu_2)}{s_{\overline{x}_1-\overline{x}_2}}=\frac{(\overline{x}_1-\overline{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \tag{4.23}$$

如果总体方差同质 (homogeneity)，
则两组样本的样本方差可进行加权平均，得样本合并方差 (pooled variance)：

$$s_{\mathrm{p}}^2 = \frac{n_1 - 1}{n_1 + n_2 - 2} s_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} s_2^2 \tag{4.21}$$

样本合并方差可作为未知总体方差的估计，计算样本标准误

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_{\mathrm{p}}^2}{n_1} + \frac{s_{\mathrm{p}}^2}{n_2}} \tag{4.22}$$

$$t=\frac{(\overline{x}_1-\overline{x}_2)-(\mu_1-\mu_2)}{s_{\overline{x}_1-\overline{x}_2}}=\frac{(\overline{x}_1-\overline{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_{\mathrm{p}}^2}{n_1}+\frac{s_{\mathrm{p}}^2}{n_2}}} \tag{4.23}$$

服从自由度为 $\mathrm{df}=(n_1-1)+(n_2-1)=n_1+n_2-2$ 的 t 分布，即 $t\sim t(n_1+n_2-2)$。

如果总体方差不同质，

则有标准化统计量

$$t' = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{s_{\overline{x}_1 - \overline{x}_2}} = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{4.24}$$

近似服从自由度为 df' 的 t 分布。

$$\mathrm{df}' = \frac{1}{\frac{R^2}{n_1 - 1} + \frac{(1-R)^2}{n_2 - 1}}, \quad \left(R = \frac{s_1^2/n_1}{s_1^2/n_1 + s_2^2/n_2}\right) \tag{4.25}$$

设两个随机变量分别服从自由度为 n_1-1 的 $\chi^2(n_1-1)$ 分布和 n_2-1 的 $\chi^2(n_2-1)$ 分布，有统计量

$$F=\frac{\frac{\chi^2(n_1-1)}{n_1-1}}{\frac{\chi^2(n_2-1)}{n_2-1}} \tag{4.32}$$

服从双自由度 n_1-1 和 n_2-1 的 F 分布。

因为 $\chi^2(n_1-1)=\frac{(n_1-1)s_1^2}{\sigma_1^2}$，$\chi^2(n_2-1)=\frac{(n_2-1)s_2^2}{\sigma_2^2}$，所以

$$F=\frac{\frac{(n_1-1)s_1^2}{\sigma_1^2}\times\frac{1}{n_1-1}}{\frac{(n_2-1)s_2^2}{\sigma_2^2}\times\frac{1}{n_2-1}}=\frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}}=\frac{s_1^2}{s_2^2}\times\frac{\sigma_2^2}{\sigma_1^2} \tag{4.33}$$

自由度 n_1 和 n_2 的 F 分布有概率密度函数:

$$f(x) = n_1^{\frac{n_1}{2}} n_2^{\frac{n_2}{2}} x^{\frac{n_1}{2}-1} \frac{\Gamma(\frac{n_2+n_1}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})} (n_1 x + n_2)^{-\frac{n_2+n_1}{2}}, \quad x > 0 \tag{4.34}$$

在 $n_2 > 0$ 时有数学期望 $\frac{n_2}{n_2-2}$，在 $n_2 > 4$ 时有方差 $\frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$。

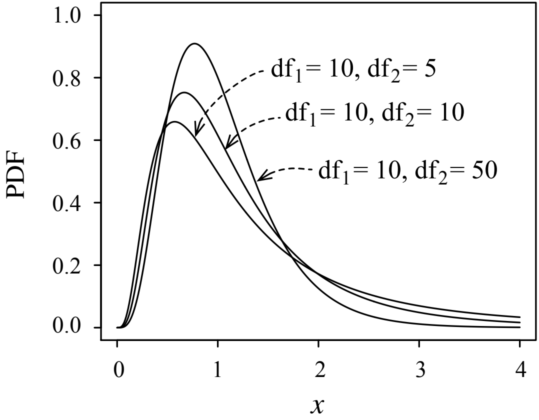

图 **4.12** 不同自由度的 F 分布

F 分布具有以下性质：

1. F 分布在区间 $(0,+\infty)$ 内其密度曲线呈右偏斜的形态；
2. 设 $X \sim F(n_1,n_2)$，则 $\dfrac{1}{X} \sim F(n_2,n_1)$；
3. 设 $X \sim t(n)$，则 $X^2 \sim F(1,n)$。

- 精确抽样分布

  当总体分布已知时，如果对任意样本容量的样本 $x_1, x_2, x_3, \cdots, x_n$，都能推导出统计量 $T(x_1, x_2, x_3, \cdots, x_n)$ 的抽样分布的数学表达式，这样的抽样分布就称为精确抽样分布。

- 渐进抽样分布

  当样本容量 n 无限大时，统计量 $T(x_1, x_2, \cdots, x_n)$ 的极限分布，称为渐进抽样分布。

- 近似抽样分布

  当精确分布和渐进分布都难以得到，或它们难以应用时，还可设法获得统计量 $T(x_1, x_2, \cdots, x_n)$ 的近似抽样分布。

本章小结

1. 抽样试验

2. 单一总体样本统计量的分布

   样本平均数的分布

   样本比率的分布

   样本方差的分布

3. 两个总体样本统计量的分布

   平均数之差的分布

   样本比率之差的分布

   样本方差之比的分布

4. 抽样分布的分类