生物统计学

第五章 参数估计

云南大学 生命科学学院

會澤百家 至公天下

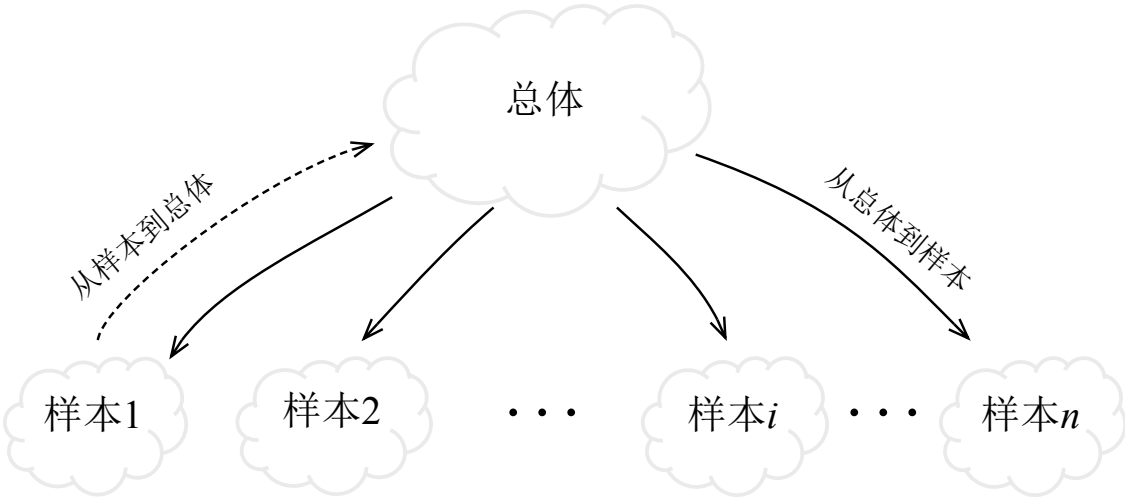


图 **4.1** 统计学的研究方向

从样本到总体的方向，通过抽样分布，由一个样本或一系列样本来推断总体的特征，称为统计推断(statistical inference)，

包括参数估计和假设检验。

参数估计 (parameter estimation)，是指由样本结果对总体参数在一定概率水平上做出的估计。

- 点估计(point estimation)
- 区间估计(interval estimation)

1. 点估计
2. 估计量的优良性
3. 区间估计

点估计是依据样本估计总体分布中未知参数或未知参数的函数的统计推断方法。

假设从某总体中抽取的独立随机样本 $x_1, x_2, \cdots, x_n$，要依据这些样本观测值对总体参数 $\theta_1, \theta_2, \cdots, \theta_k$ 中的未知项做出估计。

为了估计 θ_1，可构造统计量

$$\hat{\theta}_1 = f(x_1, x_2, \cdots, x_n) \tag{5.1}$$

作为 θ_1 的估计量。

点估计问题的关键在于为估计方法确定一个原理，

依据该原理构造出来的点估计方法，

要比其它方法存在统计性质上的优越性。

定义 (5.1)

假设 $x_1, x_2, \cdots, x_n$ 是从个体数为 N 的总体中抽出的一组随机样本，总体的概率密度函数可表示为 $f(x|\theta_1, \cdots, \theta_k)$，其中 $\theta_1, \cdots, \theta_k$ 是待估计的总体参数。则 k 阶总体原点矩为

$$A_k = \frac{\sum_{i=1}^{N} x_i^k}{N}, \quad k = 1, 2, \cdots \tag{5.2}$$

相应地，k 阶样本原点矩为

$$a_k = \frac{\sum_{i=1}^{n} x_i^k}{n}, \quad k = 1, 2, \cdots \tag{5.3}$$

定义 (5.2)

结合总体 1 阶原点矩，有 k 阶总体中心矩

$$M_k = \frac{\sum_{i=1}^{N}(x_i - A_1)^k}{N}, \quad k = 1, 2, \cdots \tag{5.4}$$

相应地，结合样本 1 阶原点矩，有 k 阶样本中心矩

$$m_k = \frac{\sum_{i=1}^{n}(x_i - a_1)^k}{n}, \quad k = 1, 2, \cdots \tag{5.5}$$

辛钦大数定理：当 n 足够大时，随机变量的算术平均数将接近于总体平均数，即样本平均数稳定于总体平均数。

因此，样本 1 阶原点矩 a_1 可以作为总体 1 阶原点矩 A_1 的点估计。

$$E(a_k) = E(\frac{\sum_{i=1}^{n} x_i^k}{n}) = \frac{E(\sum_{i=1}^{n} x_i^k)}{n} = \frac{\sum_{i=1}^{n} E(x_i^k)}{n} \tag{5.6}$$

辛钦大数定理：当 n 足够大时，随机变量的算术平均数将接近于总体平均数，即样本平均数稳定于总体平均数。

因此，样本 1 阶原点矩 a_1 可以作为总体 1 阶原点矩 A_1 的点估计。

$$E(a_k) = E(\frac{\sum_{i=1}^{n} x_i^k}{n}) = \frac{E(\sum_{i=1}^{n} x_i^k)}{n} = \frac{\sum_{i=1}^{n} E(x_i^k)}{n} \tag{5.6}$$

大数定律为总体平均数的点估计作了理论支撑。

样本 2 阶中心距 $m_2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$，按照方差的期望形式定义，有

$$m_2 = E\left([x - E(x)]^2 \right) \tag{5.8}$$

样本 2 阶中心距 $m_2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$，按照方差的期望形式定义，有

$$m_2 = E\left([x - E(x)]^2 \right) \tag{5.8}$$

$$\vdots$$

$$E(m_2) = \frac{n-1}{n} M_2 \tag{.}$$

$$
\begin{aligned}
M_2 &= E(m_2) \times \frac{n}{n-1} \\
&= E(\frac{n}{n-1} \times \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}) \\
&= E(\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}) \\
&= E(s^2)
\end{aligned}
\tag{5.16}
$$

假设一个试验产生了 3 个观测值：x_1, x_2, x_3。

计算样本平均数和样本方差得：

$$\begin{aligned} \overline{x} &= \frac{x_1 + x_2 + x_3}{3} \\ s^2 &= \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2}{3 - 1} \end{aligned} \tag{5.17}$$

假设一个试验产生了 3 个观测值：x_1, x_2, x_3。

计算样本平均数和样本方差得：

$$\begin{aligned} \overline{x} &= \frac{x_1 + x_2 + x_3}{3} \\ s^2 &= \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2}{3 - 1} \end{aligned} \tag{5.17}$$

3 个观测值中必有其一受到样本平均数的限制，而不能自由变化。能够自由变化观测值就有 $3 - 1 = 2$ 个，此即自由度的含义。

5.2 估计量的优良性

- 无偏性
- 相合性
- 有效性

- 总体平均数的区间估计 (总体方差已知)
- 总体平均数的区间估计 (总体方差未知)
- 总体比率的区间估计
- 总体方差的区间估计

假设从某一个已知参数平均数为 μ、方差为 σ^2 的总体中随机抽取一组样本，样本平均数服从正态分布 $N(\mu, \frac{\sigma^2}{n})$，标准化后得统计量

$$z = \frac{\overline{x} - \mu}{\sigma_{\overline{x}}} = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \tag{5.24}$$

服从标准正态分布。

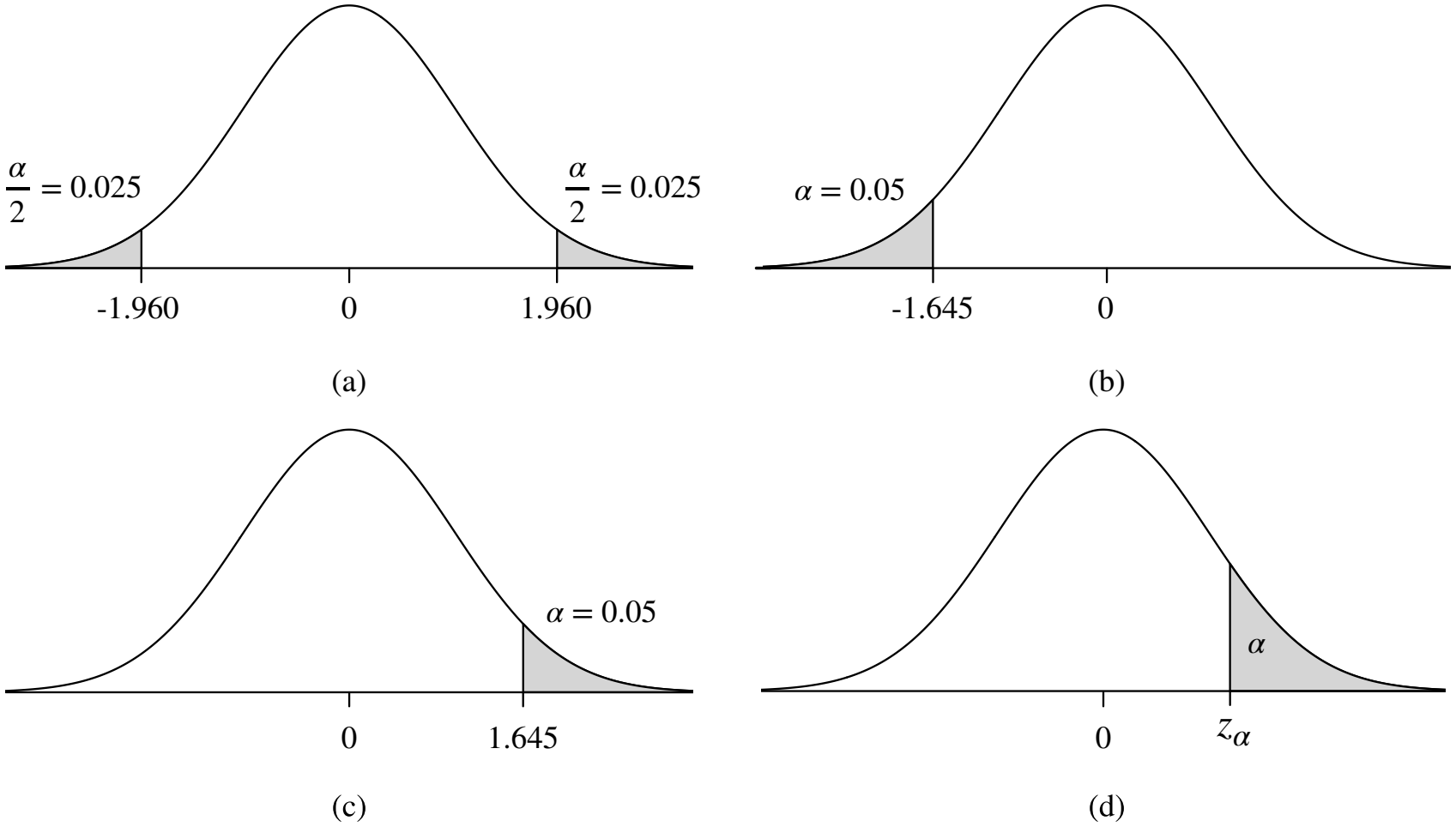
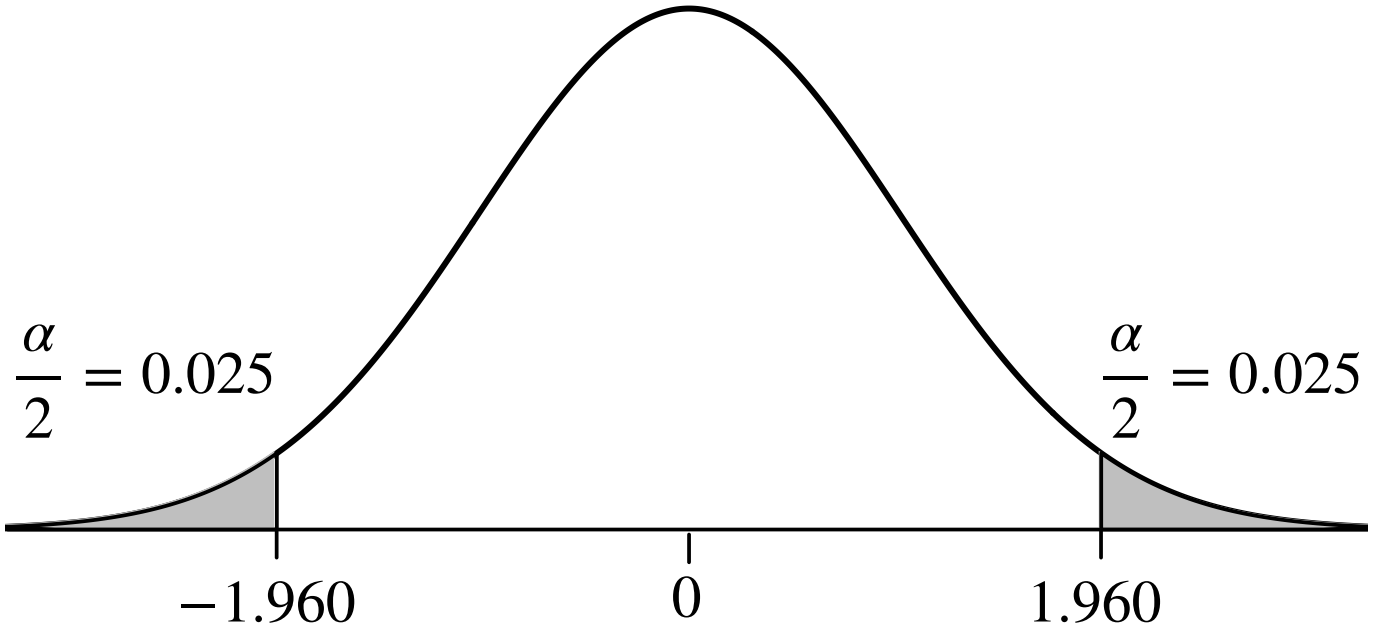
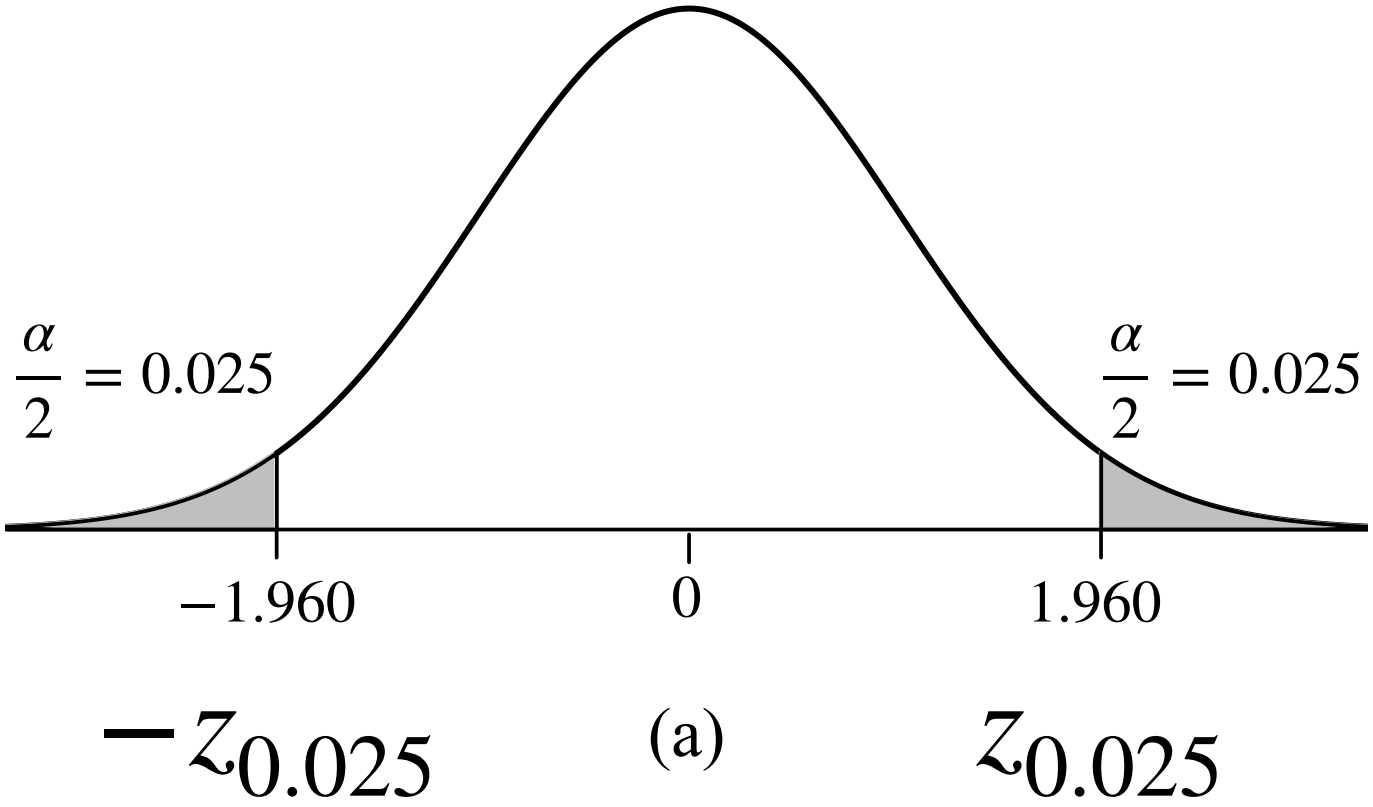


图 **3.15** 标准正态分布的上侧与下侧分位数



(a)


$\frac{\alpha}{2} = 0.025$
$\frac{\alpha}{2} = 0.025$
−1.960
0
1.960
$-z_{0.025}$
(a)
$z_{0.025}$

$$\begin{aligned} P(z \leq z_{0.025}) - P(z \leq -z_{0.025}) &= 0.95 \\ P(-z_{0.025} \leq z \leq z_{0.025}) &= 0.95 \end{aligned} \tag{5.25}$$

$$
\begin{aligned}
P(-z_{0.025} \leq \frac{\overline{x}-\mu}{\sigma/\sqrt{n}} \leq z_{0.025}) &= 0.95 \\
P(-z_{0.025}\frac{\sigma}{\sqrt{n}} \leq \overline{x}-\mu \leq z_{0.025}\frac{\sigma}{\sqrt{n}}) &= 0.95 \\
P(\overline{x}-z_{0.025}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x}+z_{0.025}\frac{\sigma}{\sqrt{n}}) &= 0.95
\end{aligned}
\tag{5.26}
$$

定义 (5.7)

当从平均数为 μ、方差为 σ^2 的总体中抽取容量为 n 的样本时，由样本平均数构成的区间

$$\left[\overline{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \quad \overline{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] \tag{1}$$

称为总体平均数 μ 的 $1-\alpha$ 置信区间 (confidence interval)，$1-\alpha$ 称为置信度。

例 (5.1)

测得某批小麦 25 个随机样本的平均蛋白质含量 $\overline{x} = 14.5\%$，已知总体标准差 $\sigma = 2.5\%$。试对该批小麦蛋白质含量进行置信度为 0.95 的区间估计。

例 (5.1)

测得某批小麦 25 个随机样本的平均蛋白质含量 $\bar{x} = 14.5\%$，已知总体标准差 $\sigma = 2.5\%$。试对该批小麦蛋白质含量进行置信度为 0.95 的区间估计。

```
> z.0.025 <- qnorm(p = 0.025, lower.tail = FALSE); z.0.025
[1] 1.959964
> 14.5 - z.0.025 * (2.5 / sqrt(25))
[1] 13.52002
> 14.5 + z.0.025 * (2.5 / sqrt(25))
[1] 15.47998
```

在总体方差未知时，用样本标准差代替总体标准差，
标准化统计量由 z 变为

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}} \tag{5.29}$$

服从自由度为 $n-1$ 的 t 分布。

定义 (5.8)

当从平均数为 μ、方差未知的总体中抽取容量为 n 的样本时，由样本平均数构成的区间

$$\left[\overline{x} - t_{\frac{\alpha}{2},\mathrm{df}}\frac{s}{\sqrt{n}},\quad \overline{x} + t_{\frac{\alpha}{2},\mathrm{df}}\frac{s}{\sqrt{n}}\right] \tag{5.30}$$

称为总体平均数 μ 的 $1-\alpha$ 置信区间。其中 s 为样本标准差，t 分布的自由度 df 为 $n-1$。

例 (5.2)

随机抽测 5 年生杂交杨树 16 株，算得平均树高 9.27 米，样本标准差 1.4 米。试对树高进行置信度为 0.95 的区间估计。

例 (5.2)

随机抽测 5 年生杂交杨树 16 株，算得平均树高 9.27 米，样本标准差 1.4 米。试对树高进行置信度为 0.95 的区间估计。

```
> t.0.025 <- qt(p = 0.025, df = 15, lower.tail = F); t.0.025
[1] 2.13145
> 9.27 - t.0.025 * (1.4 / sqrt(16))
[1] 8.523993
> 9.27 + t.0.025 * (1.4 / sqrt(16))
[1] 10.01601
```

对总体比率作点估计，可以进行 n 次试验并记录事件发生的次数 m，当 n 很大时，事件发生的样本比率 $\hat{p} = \frac{m}{n}$ 就可作为总体比率 p 的点估计。

中心极限定理表明标准化统计量

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \tag{5.31}$$

当 n 很大时近似服从标准正态分布。所以有

$$P(-z_{\frac{\alpha}{2}} \leq \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \leq z_{\frac{\alpha}{2}}) \approx 1 - \alpha \tag{5.32}$$

中心极限定理表明标准化统计量

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \tag{5.31}$$

当 n 很大时近似服从标准正态分布。所以有

$$P(-z_{\frac{\alpha}{2}} \leq \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \leq z_{\frac{\alpha}{2}}) \approx 1 - \alpha \tag{5.32}$$

$$P(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) \approx 1 - \alpha \tag{5.33}$$

定义 (5.9)

n 次试验中事件发生 m 次，当 n 很大时，由样本比率 $\hat{p} = \frac{m}{n}$ 构成的区间

$$\left[\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \tag{5.34}$$

称为总体比率 p 的 $1-\alpha$ 置信区间。其中 $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ 为样本比率的标准误 $s_{\hat{p}}$。

对样本容量 n 的要求是对二项分布进行正态近似的关键条件。

- 如果 np 或 $n(1-p)$ 小于 5，不能使用正态近似。
- np 和 $n(1-p)$ 均大于 30 时，近似效果最佳。
- np 或 $n(1-p)$ 小于 30 时，近似效果需要连续性矫正因子矫正，同时抽样分布也需要根据样本量 n 的情况，选择标准正态分布 $(n \geq 30)$ 或 t 分布 $(n < 30)$。

$$\left[\hat{p} - z_{\frac{\alpha}{2}} s_{\hat{p}} - \frac{0.5}{n}, \hat{p} + z_{\frac{\alpha}{2}} s_{\hat{p}} + \frac{0.5}{n}\right] \tag{5.35}$$

$$\left[\hat{p} - t_{\frac{\alpha}{2},n-1} s_{\hat{p}} - \frac{0.5}{n}, \hat{p} + t_{\frac{\alpha}{2},n-1} s_{\hat{p}} + \frac{0.5}{n}\right] \tag{5.36}$$

例 (5.3)

假设从母亲患有乳腺癌的 50～54 岁妇女群体中随机选择 10000 人，筛查发现有 400 人患有乳腺癌。试对该妇女群体的患乳腺癌的患病率进行置信度为 0.95 的区间估计。

例 (5.3)

假设从母亲患有乳腺癌的 50～54 岁妇女群体中随机选择 10000 人，筛查发现有 400 人患有乳腺癌。试对该妇女群体的患乳腺癌的患病率进行置信度为 0.95 的区间估计。

```
> z.0.025 <- qnorm(p = 0.025, lower.tail = FALSE); z.0.025
[1] 1.959964
> 0.04 - z.0.025 * sqrt(0.04 * (1 - 0.04) / 10000)
[1] 0.03615927
> 0.04 + z.0.025 * sqrt(0.04 * (1 - 0.04) / 10000)
[1] 0.04384073
```

例 (5.4)

针对一块玉米田随机调查 100 株玉米，发现受玉米螟虫害的植株共 21 株。试对发病率进行置信度为 0.95 的区间估计。

例 (5.4)

针对一块玉米田随机调查 100 株玉米，发现受玉米螟虫害的植株共 21 株。试对发病率进行置信度为 0.95 的区间估计。

```
> z.0.025 <- qnorm(p = 0.025, lower.tail = FALSE); z.0.025
[1] 1.959964
> 0.21 - z.0.025 * sqrt(0.21 * (1-0.21)/100) - 0.5/100
[1] 0.1251691
> 0.21 + z.0.025 * sqrt(0.21 * (1-0.21)/100) + 0.5/100
[1] 0.2948309
```

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \tag{5.37}$$

服从自由度 $\mathrm{df} = n-1$ 的 χ^2 分布。类似地，有

$$P(\chi^2_{1-\frac{\alpha}{2},\mathrm{df}} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{\frac{\alpha}{2},\mathrm{df}}) = 1-\alpha \tag{5.38}$$

其中 $\chi^2_{\frac{\alpha}{2},\mathrm{df}}$ 和 $\chi^2_{1-\frac{\alpha}{2},\mathrm{df}}$ 分别是自由度为 $n-1$ 的 χ^2 分布的上侧 $\frac{\alpha}{2}$ 分位数和上侧 $1-\frac{\alpha}{2}$ 分位数。

定义 (5.40)

当从方差为 σ^2 的总体中抽取容量为 n 的样本时，由样本方差构成的区间

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

称为总体方差 σ^2 的 $1-\alpha$ 置信区间。其中 χ^2 分布的自由度 df 为 $n-1$。

例 (5.5)

已知某水稻田受到重金属污染，抽样测定其镉含量 (单位：$\mu g/g$) 分别为 3.6、4.2、4.7、4.5、4.2、4.0、3.8 和 3.7。试对该农田水稻镉含量的总体方差做置信度为 0.95 的区间估计。

例 (5.5)

已知某水稻田受到重金属污染，抽样测定其镉含量 (单位：$\mu g/g$) 分别为 3.6、4.2、4.7、4.5、4.2、4.0、3.8 和 3.7。试对该农田水稻镉含量的总体方差做置信度为 0.95 的区间估计。

```
> cadmium <- c(3.6, 4.2, 4.7, 4.5, 4.2, 4.0, 3.8, 3.7)
> cadmium.var <- var(cadmium)
> chisq.0.025 <- qchisq(p = 0.025, df = 7, lower.tail = F)
> chisq.0.975 <- qchisq(p = 0.975, df = 7, lower.tail = F)
> (8 - 1) * cadmium.var / chisq.0.025
[1] 0.06549463
> (8 - 1) * cadmium.var / chisq.0.975
[1] 0.6206102
```


SELF
LEARNING

解决区间估计问题的关键是标准化统计量(standardized statistic, s.s.),
一些数理统计的资料还称其为枢轴量(pivotal quantity, pivot)。

解决区间估计问题的关键是标准化统计量(standardized statistic, s.s.)，
一些数理统计的资料还称其为枢轴量(pivotal quantity, pivot)。

s.s. 是一个关于可观测的样本信息和不可观测的总体参数的函数，
且该函数具有不依赖任何未知参数的概率分布。

解决区间估计问题的关键是标准化统计量(standardized statistic, s.s.)，
一些数理统计的资料还称其为枢轴量(pivotal quantity, pivot)。

s.s. 是一个关于可观测的样本信息和不可观测的总体参数的函数，
且该函数具有不依赖任何未知参数的概率分布。

$$\text{s.s.} = f(x, \theta) \sim \textbf{sampling distribution}$$

其中 x 表示可观察的样本信息，θ 表示不可观察的未知总体参数。

频率学派思想下的置信区间，本身是一个随机区间。

$$[\overline{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$$

频率学派思想下的置信区间，本身是一个随机区间。

$$[\overline{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$$

- “置信区间有 95% 的可能性包含总体参数”
- “总体参数有 95% 的可能性落在置信区间内”

本章小结

1. 点估计
   - 矩法
   - 最大似然法
2. 估计量的优良性
3. 区间估计
   - 单总体的区间估计
   - 双总体的区间估计
   - 关于置信区间的理解