生物统计学（VI）：统计推断

西安交通大学数学与统计学院

May, 2018

一个总体频率 p 的区间估计与点估计

在置信度 $P = 1 - \alpha$ 下，若 σ_p 已知，

- 对一个总体频率 p 的区间估计为

$$(\hat{p} - u_\alpha \sigma_p, \hat{p} + u_\alpha \sigma_p)$$

- 其置信区间下限 L_1 与上限 L_2 分别为

$$L_1 = \hat{p} - u_\alpha \sigma_p, L_2 = \hat{p} + u_\alpha \sigma_p$$

- 总体频率 p 的点估计 L 为

$$L = \hat{p} \pm u_\alpha \sigma_p$$

一个总体频率 p 的区间估计与点估计

在置信度 $P = 1 - \alpha$ 下，若样本容量较小或者 np, nq 小于 30，需做连续性矫正，其矫正公式为

$$\left(L_1 = \hat{p} - u_\alpha \sigma_p - \frac{0.5}{n}, L_2 = \hat{p} + u_\alpha \sigma_p + \frac{0.5}{n} \right)$$

总体频率 p 的点估计 L 为

$$L = \hat{p} \pm u_\alpha \sigma_p \pm \frac{0.5}{n}$$

- 若 σ_p 未知，用 $s_{\hat{p}}$ 来估计
- 当 $n < 30$ 时，u_α 值用 t_α 值 $(df = n - 1)$ 来代替。

一个总体频率 p 的区间估计与点估计

例 4.20

调查 100 株玉米，得到受玉米螟危害的植株为 20 株。试进行置信度为 95% 的玉米螟危害率的区间估计与点估计。

- 本题 $n = 100, x = 20, \hat{p} = 0.2, np = 20 < 30$，因此需进行连续性矫正；
- 因 σ_p 未知，故用 $s_{\hat{p}}$ 来估计，计算 $s_{\hat{p}}$ 如下：

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.20 \times (1-0.20)}{100}} = 0.04$$

- 当 $\alpha = 0.05$ 时，$u_{0.05} = 1.96$，于是，置信度为 95% 的玉米螟危害率的区间估计为

$$\begin{aligned} L_1 &= \hat{p} - u_\alpha s_{\hat{p}} - \frac{0.5}{n} = 0.2 - 1.96 \times 0.04 - \frac{0.5}{100} = 0.117 \\ L_2 &= \hat{p} + u_\alpha s_{\hat{p}} + \frac{0.5}{n} = 0.2 + 1.96 \times 0.04 + \frac{0.5}{100} = 0.283 \end{aligned}$$

一个总体频率 p 的区间估计与点估计

例 4.20

玉米螟危害率的点估计为

$$L = \hat{p} \pm u_\alpha \sigma_p \pm \frac{0.5}{n} = 0.2 \pm 1.96 \times 0.04 \pm \frac{0.5}{100} = 0.2 \pm 0.083$$

因此，玉米螟危害率为 $0.117 \sim 0.283$，这个估计的置信度为 95%.

两总体频率差数 $p_1 - p_2$ 的区间估计与点估计

- 若总体频率未知，常进行两个样本频率的比较，用 $s_{\hat{p}_1-\hat{p}_2}$ 估计 $\sigma_{p_1-p_2}$.
- 在进行两个总体频率差数 $p_1 - p_2$ 的区间估计和点估计时，一般应明确两个频率有显著差异才有意义。
- 在置信度 $P = 1 - \alpha$ 下，两个总体频率差数 $p_1 - p_2$ 的区间估计为

$$(\hat{p}_1 - \hat{p}_2) - u_\alpha s_{\hat{p}_1-\hat{p}_2}, (\hat{p}_1 - \hat{p}_2) + u_\alpha s_{\hat{p}_1-\hat{p}_2})$$

- 其置信区间上下限为

$$\left[L_1 = (\hat{p}_1 - \hat{p}_2) - u_\alpha s_{\hat{p}_1-\hat{p}_2}, L_2 = (\hat{p}_1 - \hat{p}_2) + u_\alpha s_{\hat{p}_1-\hat{p}_2}\right]$$

- 两个总体频率差数的点估计 L 为

$$L = (\hat{p}_1 - \hat{p}_2) \pm u_\alpha s_{\hat{p}_1-\hat{p}_2}$$

两总体频率差数 $p_1 - p_2$ 的区间估计与点估计

- 当 $5 < np$ 或 $np < 30$ 时，需进行连续性矫正，则两总体频率差数 $p_1 - p_2$ 的区间估计为

$$\left((\hat{p}_1 - \hat{p}_2) - u_\alpha s_{\hat{p}_1 - \hat{p}_2} - \frac{0.5}{n_1} - \frac{0.5}{n_2}, \right.$$
$$\left. (\hat{p}_1 - \hat{p}_2) + u_\alpha s_{\hat{p}_1 - \hat{p}_2} + \frac{0.5}{n_1} + \frac{0.5}{n_2} \right)$$

- 其置信区间的上下限为

$$L_1 = (\hat{p}_1 - \hat{p}_2) - u_\alpha s_{\hat{p}_1 - \hat{p}_2} - \frac{0.5}{n_1} - \frac{0.5}{n_2}$$
$$L_2 = (\hat{p}_1 - \hat{p}_2) + u_\alpha s_{\hat{p}_1 - \hat{p}_2} + \frac{0.5}{n_1} + \frac{0.5}{n_2}$$

- 两个总体频率差数的点估计 L 为

$$L = (\hat{p}_1 - \hat{p}_2) \pm u_\alpha s_{\hat{p}_1 - \hat{p}_2} \pm \left(\frac{0.5}{n_1} + \frac{0.5}{n_2} \right)$$

两总体频率差数 $p_1 - p_2$ 的区间估计与点估计

例 4.21

研究地势对小麦锈病发病率的影响。调查低洼地麦田 378 株，其中锈病株 342 株；调查高坡地麦田 396 株，其中锈病株 313。试进行置信度为 99% 的两块麦田锈病发病率差数的区间估计与点估计。

- 可计算得出 $\hat{p}_1 = 0.905, \hat{p}_2 = 0.790, s_{\hat{p}_1 - \hat{p}_2} = 0.026$
- 由于 $n_1, n_2 > 30$, 且 np, nq 均大于 30，故当 $P = 1 - \alpha = 0.99$ 时，$u_{0.01} = 2.58$，可以用 $s^2_{\hat{p}_1 - \hat{p}_2}$ 来估计 $\sigma^2_{p_1 - p_2}$, 无需进行连续性矫正。

两总体频率差数 $p_1 - p_2$ 的区间估计与点估计

- 所以置信度为 99% 的两块麦田锈病发病率差数的区间估计

$$\begin{aligned} L_1 &= (\hat{p}_1 - \hat{p}_2) - u_\alpha s_{\hat{p}_1 - \hat{p}_2} \\ &= (0.905 - 0.790) - 2.58 \times 0.026 = 0.048 \\ L_2 &= (\hat{p}_1 - \hat{p}_2) + u_\alpha s_{\hat{p}_1 - \hat{p}_2} \\ &= (0.905 - 0.790) + 2.58 \times 0.026 = 0.182 \end{aligned}$$

- 点估计为

$$(\hat{p}_1 - \hat{p}_2) \pm u_\alpha s_{\hat{p}_1 - \hat{p}_2} = 0.115 \pm 0.067$$

- 因此，低洼地麦田锈病率比高坡地高 $0.048 \sim 0.182$，这个估计的置信度为 99%

样本方差的同质性检验

- 我们已经讨论了样本平均数、频率的假设检验
- 平均数、频率表示的是计量数据资料的中心位置，但其代表性的好坏，与样本资料中各个观测值的变异程度密切相关，而方差是表示变异度的一个重要统计数。
- 对样本平均数、频率的假设都是以方差的同质性为前提的，否则假设检验的结论将是不正确的。

方差同质性

方差的同质性，又称方差齐性（homogeneity of variance），就是指各个总体的方差是相同的。方差同质性检验（homogeneity test）就是要从各样本的方差来推断其总体方差是否相同。

样本方差的同质性检验

一个样本方差的同质性检验

- 我们知道，从标准正态总体中抽取 k 个独立 u^2 之和为 χ^2，即

$$\chi^2 = \sum \left(\frac{x-\mu}{\sigma} \right)^2 = \frac{1}{\sigma^2} \sum (x-\mu)^2$$

- 当用样本平均数 $\bar{x}$ 估计总体平均数 μ 时，则有

$$\chi^2 = \frac{1}{\sigma^2} \sum (x-\bar{x})^2$$

样本方差

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$$

上式可变为

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

该式中，分子表示样本的离散程度，分母表示总体方差，χ^2 服从自由度为 $n-1$ 的 χ^2 分布。

样本方差的同质性检验

若提出无效假设 $H_0 : \sigma^2 = \sigma_0^2$，及备择假设 $H_A : \sigma^2 \neq \sigma_0^2$，那么我们可以计算 χ^2-分布的临界值 χ_α^2，

- 若 $\chi^2 < \chi_\alpha^2$，则接受 H_0：认为样本所属总体方差与已知总体方差相同
- 反之，则接受 H_A：认为样本所属总体方差与已知总体方差不同

样本方差的同质性检验

例 4.22

已知某农田收到重金属的污染，经抽样测定其铅浓度为 4.2,4.5,3.6,4.7,4.0,3.8,3.7,4.2($\mu g/g$)，方差为 $0.150(\mu g/g)^2$。经试验收到污染的农田铅浓度的方差是否与正常农田铅浓度的方差 $0.065(\mu g/g)^2$ 相同。

- 此题中正常农田铅浓度方差 $\sigma_0^2 = 0.065(\mu g/g)^2$，为一个样本方差与已知总体方差的同质性检验
- 假设 $H_0 : \sigma^2 = \sigma_0^2 = 0.065$, 即受到污染的农田铅浓度的方差与正常农田铅浓度的方差相同；$H_A : \sigma^2 \neq \sigma_0^2$
- 确定显著性水平 $\alpha = 0.05$
- 检验计算 $\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(8-1)\times 0.150}{0.065} = 16.154$
- 推断：查表，可知当 $df = n - 1 = 7$ 时，$\chi^2_{0.05} = 14.07$。我们的计算 $\chi^2 > \chi^2_{0.05}$, 故否定 H_0，接受 H_A, 即样本方差与总体方差是不同质.

样本方差的同质性检验

两个样本方差的同质性检验

- 假设两个样本的样本容量分别为 n_1 和 n_2，方差分别为 s_1^2 与 s_2^2（一般把数值较大的样本方差叫 s_1^2），总体方差分别为 σ_1^2 和 σ_2^2，当检验 σ_1^2 与 σ_2^2 是否同质时，可用 F-检验
- 回忆一下，当两样本所属总体均服从正态分布，且两样本的抽样是随机的和独立的，其 F 值等于两样本方差 s_1^2 与 s_2^2 的比值，即

$$F = \frac{s_1^2}{s_2^2}$$

  并服从 $df_1 = n_1 - 1, df_2 = n_2 - 1$ 的 F-分布。
- 当 $F < F_\alpha$，接受 $H_0 : \sigma_1^2 = \sigma_2^2$, 即认为两样本的方差是同质
- 当 $F > F_\alpha$，接受 $H_A : \sigma_1^2 \neq \sigma_2^2$, 即认为两样本的方差不是同质

样本方差的同质性检验

例 4.23

两个小麦品种千粒重 (g) 的调查结果如下所示：

品种甲：50，47，42，43，39，51，43，38，44，37

品种乙：36，38，37，38，36，39，37，35，33，37

检验这两个小麦品种千粒重的方差是否同质。

样本方差的同质性检验

本例中，$s_1^2 = 22.933, s_2^2 = 2.933, n_1 = n_2 = 10$，为两个样本方差的同质性检验

- 假设 $H_0 : \sigma_1^2 = \sigma_2^2$，即两小麦品种的千粒重方差相同；
  $H_A : \sigma_1^2 \neq \sigma_2^2$
- 确定显著性水平 $\alpha = 0.01$
- 检验计算

$$F = \frac{s_1^2}{s_2^2} = \frac{22.933}{2.933} = 7.819$$

- 推断：查表，
  $df_1 = 10 - 1 = 9, df_2 = 9, F_{0.01} = 5.35, F > F_{0.01}$，故否定 H_0，接受 H_A, 即认为两个小麦品种千粒重的方差不同质。

多个样本方差的同质性检验

对 3 个或以上样本方差进行同质性检验，一般采用巴特勒检验法(Bartlett test)，

- 假设 $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots \sigma_k^2$，即 k 个样本的方差是同质的；$H_A : \sigma_1^2, \cdots, \sigma_k^2$ 不相等。
- 对 k 个独立样本方差 $s_1^2, s_2^2, \cdots, s_k^2$，求其合并方差 s_p^2, 矫正数 C 和 χ^2 如下

$$
\begin{aligned}
s_p^2 &= \frac{\sum_{i=1}^{k} s_i^2 (n_i - 1)}{\sum_i n_i - 1} \\
C &= 1 + \frac{1}{3(k-1)} \left[\sum_i \frac{1}{n_i - 1} - \frac{1}{\sum_i (n_i - 1)} \right] \\
\chi^2 &= \frac{2.3026}{C} \left[\log_{10}(s_p^2) \sum_i (n_i - 1) - \sum_i (n_i - 1) \log_{10}(s_i^2) \right]
\end{aligned}
$$

多个样本方差的同质性检验

上式 χ^2 服从 $df = k-1$ 的 χ^2-分布，其中 $2.3026 = \ln 10$。对确定的显著水平 α，如果 $\chi^2 < \chi^2_\alpha$，则接受 H_0，认为 $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$；如果 $\chi^2 > \chi^2_\alpha$，则应否定 H_0。

例 4.24

在某种树的抗压强度试验中，以相同加压速率测试 4 种不同比重的该树种的抗压强度，所取样本容量分别为 4，3，3，4，计算样本的方差分别为 144.25，66.34，39.00，158.33。试检验这 4 个样本所来自的总体方差是否相同

多个样本方差的同质性检验

- 此例为多个样本方差的同质性检验
- 假设 $H_0 : \sigma_1^2 = \cdots = \sigma_4^2$, 即 4 个方差是同质的；$H_A$:4 个方差不同质
- 确定显著性水平 $\alpha = 0.05$
- 检验计算

$$\begin{aligned} s_p^2 &= 111.842 \\ C &= 1.174 \\ \chi^2 &= 1.146 \end{aligned}$$

- 推断：查表，当 $df = 4 - 1 = 3$ 时，$\chi^2_{0.05} = 7.81$，我们得到的 $\chi^2 < \chi^2_{0.05}$，故接受 H_0，即认为这 4 个样本的方差是同质的。

Questions?