随机性和概率

- **相对频率**: 事件可以反复被观察并被确认是否发生。如果实验所含观察次数较少，同一事件每次实验得到的相对频率会不等。但随着观察次数的增加，频率趋于相等，这时就能用频率来表示概率
- **主观概率**: 基于个人信念、经验、直觉或主观判断对某一事件发生的可能性给出的数值度量，而不是基于长期频率或等可能模型。反映的是一种主观可能性。
- **伯努利大数定律**: $\forall \varepsilon, \eta \in (0,1)$, $\mathbb{P}\left(\left|\frac{\mu_n}{n} - p\right| < \varepsilon\right) > 1 - \eta$
- **泊松分布**: 适用于单位时间或单位面积内某事件发生次数的概率分布。
- 在 n 重伯努利试验中, 设 μ_n 为 n 次试验中事件 A 发生的次数, 则当 n 充分大时,

$$\mathbb{P}\left(a \le \frac{\mu_n - np}{\sqrt{np(1-p)}} \le b\right) \approx \Phi(b) - \Phi(a)$$

- **t-分布**: 当总体服从正态分布, 且总体标准差未知时, 可用样本标准差代替总体标准差。设 $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, 则称

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

服从自由度为 $n-1$ 的 t 分布, 记为 $T \sim t(n-1)$. 自由度越大, t 分布越接近标准正态分布。

- **χ^2-分布**: 从正态总体样本 $\mathcal{N}(0,1)$ 中抽取 n 个样本 $X_1, \ldots, X_n$, 称随机变量 $\chi^2 = X_1^2 + \cdots + X_n^2$ 服从自由度为 n 的 χ^2-分布。
  - 用于检验观察到的数据与预期数据之间是否有显著差异。
  - 帮助我们判断实际发生的事件是否偶然, 或者是否有某种规律。
  - 自由度越大，越趋于对称，形状越接近正态分布曲线的形状。
- **F-分布**: 设 $X \sim \chi^2(n_1)$, $Y \sim \chi^2(n_2)$, 且 X 与 Y 相互独立, 则随机变量 $F = \frac{X/n_1}{Y/n_2}$ 服从自由度为 (n_1, n_2) 的 F 分布。在方差分析、回归分析的显著性检验中都有着重要的地位。
- **使用概率来核对假设**: 设原假设为 H_0, 备择假设为 H_1. 通过样本数据计算出检验统计量, 并根据其概率分布计算出 p 值。如果 p 值小于预先设定的显著性水平 α, 则拒绝原假设 H_0, 否则不拒绝 H_0.
- **期望和方差**
  - **二项分布**: $X \sim \mathrm{Bin}(n,p)$, $\mathbb{E}[X] = np$, $\mathrm{Var}(X) = np(1-p)$.
  - **泊松分布**: $X \sim \mathrm{Poisson}(\lambda)$, $\mathbb{E}[X] = \lambda$, $\mathrm{Var}(X) = \lambda$.
  - **正态分布**: $X \sim \mathcal{N}(\mu, \sigma^2)$, $\mathbb{E}[X] = \mu$, $\mathrm{Var}(X) = \sigma^2$.
  - **χ^2-分布**: $X \sim \chi^2(n)$, $\mathbb{E}[X] = n$, $\mathrm{Var}(X) = 2n$.
  - **t-分布**: $X \sim t(n)$, $\mathbb{E}[X] = 0$ (当 $n > 1$), $\mathrm{Var}(X) = \frac{n}{n-2}$ (当 $n > 2$). 不满足条件则期望方差不存在。
  - **F-分布**: $X \sim F(n_1, n_2)$, $\mathbb{E}[X] = \frac{n_2}{n_2 - 2}$ (当 $n_2 > 2$), $\mathrm{Var}(X) = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$ (当 $n_2 > 4$). 不满足条件则期望方差不存在。
- **辛钦大数定律**: 随机变量 X 的数学期望为 μ, $x_1, \ldots, x_n$ 是 X 的独立同分布样本, 则对于任意的 $\varepsilon > 0$, 有

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\right| < \varepsilon\right) = 1$$

数据的解读

- **点估计**: 用样本均值估计总体均值, 用样本方差估计总体方差。
  - 无偏性: 如果有无穷多个样本，样本统计量的期望等于总体参数的真值,则称这种样本统计量为该总体参数的无偏估计。
  - 有效性: 重复抽样所得的估计值不应有太大的波动 (若干次抽样结果，要求有小的标准差)。
  - 对 n 重伯努利试验, 样本百分比是对总体百分比有效的无偏估计, 样本均值是对总体均值的有效无偏估计, 样本方差是对总体方差的无偏估计。
- **区间估计**: 设 Π 为某个总体参数, X 为该参数的样本估计量, x 是 X 的一个样本值。若对于给定的**置信水平** α (通常取 0.95 或 0.99), 有

$$\mathbb{P}(|X - \Pi| < \varepsilon) = \alpha$$

则称区间 $(x - \varepsilon, x + \varepsilon)$ 为参数 Π 的置信区间。

- **置信度的含义**
  - 总体参数值：固定、未知
  - 大多数情况下，人们只抽取一个样本，没人能够知道这个样本产生的置信区间是否包含参数的真实值。
  - 抽取 m 组样本, 这些样本可以构造出总体参数的 m 个置信区间, 其中大约有 αm 个置信区间包含总体参数的真实值。
- **置信区间的长度**: 与置信水平成正比, 与样本容量成反比。
- **总体比例的置信区间**: 考虑 $\mathbb{P}(|p - p'| < \varepsilon) = \alpha$, 则有

$$\Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{p'(1-p')}}\right) = \frac{1+\alpha}{2}$$

置信区间为 $(p' - \varepsilon, p' + \varepsilon)$.

- **总体均值的置信区间**: 已知总体方差时,

$$\begin{aligned}\mathbb{P}(|\bar{X} - \mu| < \varepsilon) &= \mathbb{P}\left(-\frac{\varepsilon}{\sigma/\sqrt{n}} \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le \frac{\varepsilon}{\sigma/\sqrt{n}}\right) \\ &= 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) - 1 = \alpha\end{aligned}$$

置信区间为 $(\bar{X} - \varepsilon, \bar{X} + \varepsilon)$.
总体方差未知时, 使用 t-分布代替正态分布。

$$\begin{aligned}\mathbb{P}(|\bar{X} - \mu| < \varepsilon) &= \mathbb{P}\left(-\frac{\varepsilon}{S/\sqrt{n}} \le \frac{\bar{X} - \mu}{S/\sqrt{n}} \le \frac{\varepsilon}{S/\sqrt{n}}\right) \\ &= 2t_{n-1}\left(\frac{\varepsilon\sqrt{n}}{S}\right) - 1 = \alpha\end{aligned}$$

- **总体方差的置信区间**: 从正态总体 $\mathcal{N}(\mu, \sigma^2)$ (μ, σ 均未知) 中抽取样本 $X_1, \ldots, X_n$, 则有 $\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$. 计算置信区间的方法为:

$$\mathbb{P}(\chi^2 < a) = \mathbb{P}(\chi^2 > b) = \frac{1-\alpha}{2} \implies a, b$$

$$\mathbb{P}\left(\frac{(n-1)S^2}{b} < \sigma^2 < \frac{(n-1)S^2}{a}\right) = \alpha \implies \text{置信区间}$$

- **两个参数的差异的置信区间**: 可以用来研究两个群体是否有所不同或一个群体是否随时间发生了变化。
- **总体比例差异的置信区间**: 样本 X 和 Y 分别有 n_1 和 n_2 个观测值, 样本比例分别为 p_1 和 p_2, 则
  - 当 n_1 足够大时, $\frac{p_1-\Pi_1}{\sqrt{p_1(1-p_1)/n_1}}$ 近似服从标准正态分布。
  - 当 n_2 足够大时, $\frac{p_2-\Pi_2}{\sqrt{p_2(1-p_2)/n_2}}$ 近似服从标准正态分布。
  - 当 n_1, n_2 足够大时, $\frac{(p_1-p_2)-(\Pi_1-\Pi_2)}{\sqrt{p_1(1-p_1)/n_1+p_2(1-p_2)/n_2}}$ 近似服从标准正态分布。
  - $\Phi(\varepsilon) = \frac{1+\alpha}{2}$, 置信区间为 $(p_1 - p_2 - \varepsilon\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}, p_1 - p_2 + \varepsilon\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}})$
- **均值差异(方差不等)**: $A = S_1^2/n_1, B = S_2^2/n_2$, $df = \frac{(A+B)^2}{\frac{A^2}{n_1-1}+\frac{B^2}{n_2-1}}$, $t = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{A+B}} \sim t(df)$. $t(\varepsilon, df) = \frac{1+\alpha}{2}$, 置信区间为 $(\bar{X} - \bar{Y} - \varepsilon\sqrt{A+B}, \bar{X} - \bar{Y} + \varepsilon\sqrt{A+B})$.
- **均值差异(方差相等)**: $t(\varepsilon, n_1 + n_2 - 2) = \frac{\alpha+1}{2}$, $(\bar{X} - \bar{Y} - \varepsilon S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{X}-\bar{Y}+\varepsilon S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}})$, $S = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$.
- **假设检验**: 关注总体参数是否等于某个特殊值。
  - **零假设**: 通常表示“无效”或“无差异”的情况, 记为 H_0, 一般是"参数=值"。
  - **备择假设**: 与零假设相对立, 记为 H_1, 一般是"参数≠值""参数>值"或"参数<值"。
  - **检验方法**: 在零假设成立的前提下, 计算检验统计量的概率分布, 并根据样本数据计算出检验统计量的实际值, 进而计算出 p 值。如果 p 值小于预先设定的显著性水平 $1-\alpha$, 则拒绝零假设 H_0, 否则不拒绝 H_0.
  - **双边检验**: 适用于备择假设为“参数≠值”的情况, 显著水平一般选为 0.05.
  - **单边检验**: 适用于备择假设为“参数>值”或“参数<值”的情况, 显著水平一般选为 0.025.

变量间的关系

- 是否有关系? 关系强度? 样本 $\Longrightarrow$ 总体? 因果?
- 自变量: 解释变量 $\rightarrow$ 因变量: 响应变量
- **因果关系**: 1. 时间顺序: 自变量 $\rightarrow$ 因变量; 2. 因变量随自变量变化而变化; 3. 排除其他可能影响因变量的因素。
- **χ^2-分析**:

$$\chi^2 = \frac{n|ad-bc|}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

  数据更多时, 计算

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{\left(n_{i,j} - \frac{n_{i,\cdot}\times n_{\cdot,j}}{n}\right)^2}{\frac{n_{i,\cdot}\times n_{\cdot,j}}{n}}$$

  满足 $\chi^2 \sim \chi^2((m-1)\times(k-1))$.
- **Cramer's V**: $V = \sqrt{\frac{\chi^2}{n\times\min(r-1,c-1)}}$, r: 行数, c: 列数。$V \in [0,1]$, 越接近 1 表示变量间关系越强。
- **相关分析**: $r = \mathrm{Cov}(X,Y)/\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}$.
- **回归分析**: $y = kx + b$, $k = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}$, $b = \bar{Y} - k\bar{X}$.
- **总平方和、残差平方和、回归平方和**: $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$—自变量、残差变量 $\rightarrow$ 因变量, $RSS = \sum_{i=1}^{n}(y_i - kx_i - b)^2$—残差变量 $\rightarrow$ 因变量, $RegrSS = RSS - TSS = \sum_{i=1}^{n}(b + kx_i - \bar{y})^2$—自变量的效应.
- **总体回归系数**: $s = \frac{RSS}{n-2}$, $T = \frac{k-K}{s/\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \sim t_{n-2}$. 零假设: $K = 0$.

3. 紧凑表格示例

标准正态分布

α	$(1+\alpha)/2$	z
0.90	0.95	1.645
0.95	0.975	1.960
0.99	0.995	2.576

$t-$ 分布

df	**0.90**	**0.95**	**0.99**
9	1.833	2.262	3.250
19	1.729	2.093	2.861
29	1.699	2.045	2.756
∞	1.645	1.960	2.576

标准正态分布 CDF 对照

z	$\Phi(z)$
0.0	0.5000
1.0	0.8413
1.5	0.9332
2.0	0.9772