Statistics

数据嗨客

基本概念

总体 | 根据一定的需求所研究的事物的全体．总体分布的数量特征为**总体参数**，也是统计推断的对象．总体中的个体称为**总体单位**.

样本 | 指的是总体的部分单位组成的子集．子集中的单位数量称为**样本容量**.

样本统计量 | 有关样本的函数，属于随机变量，而总体参数通常为常数.

通常来说，通过研究分析样本来去推断总体的特征.

统计指标 | 反映总体的数量特征的度量方式.

抽样调查 | 属于非全面调查，随机从调查对象中选择部分单位作为样本，并从样本中获取对象的总体特征.

统计分组 | 根据调查对象的特征以及研究的需求，将总体划分为不同的组．各组之间性质相异.

频数分布 | 根据统计分组，将调查总体按照某一特征（性质）归类排列．各个分组在总体中出现的次数称为**频数**.

与频数分布相关的统计指标有：

✓ 频数密度，各组频数与组距的比值；

✓ 频率，各组频数与总体单位之和的比值；

✓ 频率密度，各组频率与组距的比值.

大数定律 | 给定长度为n的独立同分布随机变量序列$\{x_1, x_2, \cdots, x_n\}$，均值为$\mu$，方差为$\sigma^2$，有 $\forall \epsilon > 0$

$$\lim_{n\to\infty} p\Big\{|\frac{1}{n}\sum_{i=1}^{n} x^i - \mu| < \epsilon\Big\} = 1$$

中心极限定理 | 给定长度为n的独立同分布随机变量序列$\{x_1, x_2, \cdots, x_n\}$．分布均值$\mathbb{E}(x) = \mu$，方差$\mathrm{var}(x) = \sigma^2$，若$n \to \infty$，

$$\frac{1}{n}\sum_{i=1}^{n} x_i \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

描述性统计

✓ 反映统计分布的集中趋势：给定$x = \{x_1, x_2, \cdots, x_n\}$，

算术平均数 | $\overline{x} = \mathbb{E}(x) = (\sum_{i=1}^{n} x_i)/n$.

加权平均数 | 给定f_i为单位x^i出现的频率，$\overline{x} = \sum_{i=1}^{n} f_i x_i$.

调和平均数 | $\overline{x} = n/\sum_{i=1}^{n}\frac{1}{x_i}$.

几何平均数 | $\overline{x} = (\prod_{i=1}^{n} x_i)^{\frac{1}{n}}$.

幂平均数 | 给定k为幂的阶数，$\overline{x} = \Big(\frac{\sum_{i=1}^{n}(x_i)^k}{n}\Big)^{\frac{1}{k}}$.

众数 | 总体分布中出现频率最高的值.

中位数 | 将总体的各单位按照特征的取值排列，处于中间位置的值，即

$$median = \begin{cases} \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{n为偶数} \\ x_{\frac{n+1}{2}} & \text{n为奇数} \end{cases}$$

✓ 反映统计分布的离中趋势：给定$\{x_1, x_2, \cdots, x_n\}$，

极差 | 使用极值反映取值的变动范围，$r = x_{max} - x_{min}$

方差 | 表示变量与平均数之间的离散程度，

$$\begin{aligned}\mathrm{var}(x) &= \mathbb{E}[(x - \overline{x})^2] \\ &= \mathbb{E}(x^2) - (\mathbb{E}(x))^2 = (\sum_{i=1}^{n}(x_i - \overline{x})^2)/n\end{aligned}$$

标准差 | 方差的平方根，$\sigma = \sqrt{(\sum_{i=1}^{n}(x_i - \overline{x})^2)/n}$.

标准差系数 | 标准差与平均数的比值，$v = \sigma/\overline{x}$.

k阶原点距 | $\mathbb{E}(x^k), k = 1, 2, \cdots$

k阶中心距 | $\mathbb{E}[(x - \mathbb{E}(x))^k], k = 1, 2, \cdots$

✓ 反映统计分布的不对称度和陡峭度：

偏度 | 表征分布对称性的统计量，为三阶标准中心距，

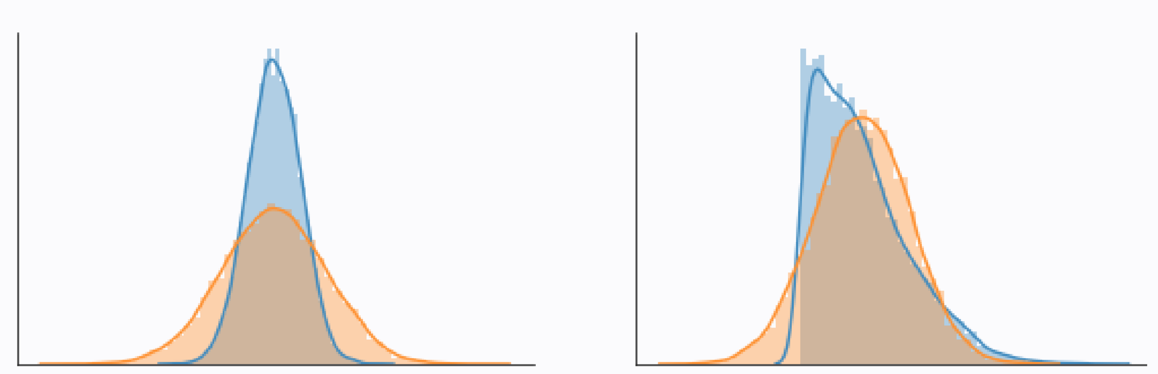
$$\mathrm{skew}(x) = \mathbb{E}\Big[(\frac{x - \mathbb{E}(x)}{\sigma})^3\Big]$$

偏度为负，密度函数左偏，长尾在左侧；偏度为正，密度函数右偏，长尾在右侧.

峰度 | 表征分布陡峭的统计量，为四阶标准中心距，

$$\mathrm{kurt}(x) = \mathbb{E}\Big[(\frac{x - \mathbb{E}(x)}{\sigma})^4\Big]$$

峰度为零，密度函数与正态分布一致；峰度为正，密度函数比正态分布陡峭.



峰度　　　　偏度

参数估计

抽样方法 | 分为重复抽样和不重复抽样．判断的标准是从总体取出的样本是否放回．与总体分布相对应的是抽样分布.

点估计 | 使用样本统计量作为相应的总体参数的估计，使用样本均值、样本方差来去估计总体的均值和方差.

评价点估计的标准：✓ 一致性 ✓ 有效性 ✓ 无偏性

无偏性 | 给定总体参数θ，估计量$\hat{\theta}$，满足$\mathbb{E}(\hat{\theta}) = \theta$. ✓ 样本均值是总体均值$\mu$的无偏估计.

✓ $(\sum_{i=1}^{n}(x_i - \overline{x})^2)/n$是总体方差$\sigma^2$的有偏估计.

✓ 由于$\sum_{i=1}^{n} x_i = n\mu, \mathrm{var}(\overline{x}) = \sigma^2/n$，$\sigma^2$的无偏估计量$\hat{\sigma^2}$为

$$\begin{aligned}\hat{\sigma^2} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= \frac{1}{n-1}\mathbb{E}\Big[\sum((x_i - \mu)^2 + (\overline{x} - \mu)^2 - 2(x_i - \mu)(\overline{x} - \mu))\Big] \\ &= \frac{1}{n-1}(\sum\mathbb{E}[(x_i - \mu)^2] + \sum\mathbb{E}[(\overline{x} - \mu)^2]) \\ &= \frac{1}{n-1}(n\sigma^2 - n\frac{\sigma^2}{n}) = \sigma^2\end{aligned}$$

给定数据集x，训练模型f，测试集预测结果为$\hat{y} = f(x)$，那么

✓ 模型MSE，$\mathrm{mse}(x) = \mathbb{E}[(y - \hat{y})^2]$

✓ 方差，$\mathrm{var}(\hat{y}) = \mathbb{E}\Big[(\hat{y} - \mathbb{E}(\hat{y}))^2\Big]$

✓ 偏差，$\mathrm{bias} = \mathbb{E}(\hat{y}) - y$

Variance-bias tradeoff 可以表示为

$$\begin{aligned}\mathrm{mse}(x) &= \mathbb{E}[(y - \hat{y})^2] = \mathbb{E}(y^2) + \mathbb{E}(\hat{y}^2) - 2y\mathbb{E}(\hat{y}) \\ &= \underline{\mathrm{var}(y) + (\mathbb{E}(y))^2} + \underline{\mathrm{var}(\hat{y}) + (\mathbb{E}(\hat{y}))^2} - 2y\mathbb{E}(\hat{y}) \\ &= \underbrace{\mathrm{var}(y)}_{variance} + \mathrm{var}(\hat{y}) + \underbrace{(y - \mathbb{E}(\hat{y}))^2}_{bias^2}\end{aligned}$$

区间估计 | 与点估计给出估计值不同，区间估计尝试估计总体参数的取值范围，并给出该取值范围成立的概率.

$$p(\Phi_1 \le x \le \Phi_2) = 1 - \alpha$$

其中，α为显著性水平，$1 - \alpha$为置信度水平.

举例，给定随机变量序列x^k，给出平均数$\overline{x}$的估计区间．假定$\overline{x} \sim \mathcal{N}(\mu, \sigma^2)$，根据显著性水平$\alpha$，通过标准正态分布表，得到临界统计量的取值$z_{\alpha/2}$，

$$p(-z_{\alpha/2} \le \frac{\overline{x} - \mu}{\sigma} \le z_{\alpha/2}) = 1 - \alpha$$
$$p(\overline{x} - \sigma z_{\alpha/2} \le \mu \le \overline{x} + \sigma z_{\alpha/2}) = 1 - \alpha$$

显著性水平α的条件下，估计区间为$[\overline{x} - \sigma z_{\alpha/2}, \overline{x} + \sigma z_{\alpha/2}]$．反之，已知$|\overline{x} - \mu| < \Delta$，那么

$$p(|\overline{x} - \mu| < \Delta) = p(|\frac{\overline{x} - \mu}{\sigma}| < \frac{\Delta}{\sigma}) = p(|z| < \frac{\Delta}{\sigma})$$

于是，临界值$z_{\alpha/2} = \Delta/\sigma$，查询标准正态分布表，可以确定显著性水平$\alpha$.

已知$\Delta = z\sigma_{\overline{x}}$，通过提高样本容量来降低平均数的误差．给定总体方差，以重复抽样的方式获取样本，确定样本容量，则有

$$\Delta = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, n = \frac{\Delta^2}{\sigma^2}z^2_{\alpha/2}$$

假设检验

假设 | 以抽样样本为依据去推断是否正确的命题．**假设检验** | 首先针对要估计的总体分布参数进行假设，然后根据抽样的样本以及小概率事件原理，来对假设正确与否作出判断.

✓ 小概率事件原理通常指小概率事件实际上在一次随机试验中不可能发生．小概率通常由显著性水平α确定，因此假设检验可以称为显著性检验.

✓ 假设检验中，需要被检验的假设称为**零假设**，记为H_0；零假设的对立假设，称为**备择假设**，记为H_1．检验假设使用的统计量称为**检验统计量**；使原假设成立的样本所在的区域，称为**接受域**，反之，则称为**否定域**.

假设检验的**一般步骤**为：

① 根据题目要求给出零假设H_0和备择假设H_1；

② 假定H_0成立，选择恰当的检验统计量；

③ 给定显著性水平α，根据检验统计量的分布、H_1以及抽样样本，计算小概率事件发生的概率；

④ 判定小概率事件是否发生：如果发生，拒绝H_0，接收H_1；反之，H_0成立.

举例，给定$x \sim \mathcal{N}(\mu, \sigma^2)$，一组样本$\{x^1, x^2, \cdots, x^n\}$，均值为$\overline{x}$，$\sigma^2$已知，基于样本进行关于$\mu$的假设检验：

✓ 零假设$H_0 : \mu = \mu_0$，备择假设$H_1 : \mu \ne \mu_0$；

✓ 检验统计量$\Phi = \frac{\overline{x} - \mu_0}{(\sigma/\sqrt{n})}$服从正态分布$\mathcal{N}(0, 1)$；

✓ 计算小概率事件的概率$p(|\Phi| > \phi_{\alpha/2}) = \alpha$；

✓ 计算$\overline{x}$和Φ，查表获取$\phi_{\alpha/2}$；

✓ 如果$|\Phi| > \phi_{\alpha/2}$，小概率事件发生，拒绝H_0，接受H_1；如果$|\Phi| < \phi_{\alpha/2}$，小概率事件没有发生，接受H_0.

单侧检验 | 上述的假设检验问题可以拆分为

$$H_0 : \mu \le \mu_0, H_1 : \mu > \mu_0 \text{左侧检验}$$
$$H_0 : \mu \ge \mu_0, H_1 : \mu < \mu_0 \text{右侧检验}.$$

两种类型的错误 | 样本数据决定假设检验的结果，由于样本抽取的随机性，会导致以下的错误

分类	接受H_0	拒绝H_0
H_0成立	正确	第一类错误
H_0不成立	第二类错误	正确

第一类错误，零假设H_0成立，但检验结果拒绝H_0，也称为**弃真错误**；第二类错误，零假设H_0不成立，检验结果接受H_0，也称为**取伪错误**.