

统计知识复习

杨弘毅

创建: 2020 年 4 月 9 日

修改: 2021 年 8 月 8 日

1 Definition

Expected Value

$$\begin{aligned} E[X] &= \sum_{i=1}^n x_i p_i = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n \\ &= \int x f(x) dx \end{aligned}$$

Variance

$$\begin{aligned} \text{Var}(X) &= \text{Cov}(X, X) = \sigma_X^2 \\ &= E[(X - E[X])^2] \\ &= E[X^2 - 2X E[X] + E[X]^2] \\ &= E[X^2] - 2E[X]^2 + E[X]^2 \\ &= E[X^2] - E[X]^2 \end{aligned}$$

Covariance

$$\begin{aligned} \text{Cov}(X, Y) &= E[(X - E(X))(Y - E(Y))] \\ &= E[XY - X E[Y] - Y E[X] + E[X] E[Y]] \\ &= E[XY] - E[X] E[Y] - E[X] E[Y] + E[X] E[Y] \\ &= E[XY] - E[X] E[Y] \end{aligned}$$

Correlation Coefficient

$$\begin{aligned} \rho_{X,Y} &= \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \\ &= \frac{E[(X - E[X])(Y - E[Y])]}{\sigma_X \sigma_Y} \end{aligned}$$

2 Expected Value Properties

$$E[X + Y] = E[X] + E[Y]$$

$$E[aX] = a E[X]$$

$$E[XY] = E[X] E[Y] \quad (X, Y \text{ are independent})$$

3 Variance Properties

$$\text{Var}(X + a) = \text{Var}(X)$$

$$\text{Var}(aX) = a^2 \text{Var}(X)$$

$$\text{Var}(aX \pm bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) \pm 2ab \text{Cov}(X, Y)$$

$$\text{Var}\left(\sum_{i=1}^N X_i\right) = \sum_{i,j=1}^N \text{Cov}(X_i, X_j) = \sum_{i=1}^N \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j)$$

$$\begin{aligned} \text{Var}\left(\sum_{i=1}^N a_i X_i\right) &= \sum_{i,j=1}^N a_i a_j \text{Cov}(X_i, X_j) \\ &= \sum_{i=1}^N a_i^2 \text{Var}(X_i) + \sum_{i \neq j} a_i a_j \text{Cov}(X_i, X_j) \\ &= \sum_{i=1}^N a_i^2 \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq N} a_i a_j \text{Cov}(X_i, X_j) \end{aligned}$$

4 Covariance Properties

$$\text{Cov}(X, a) = 0$$

$$\text{Cov}(X, X) = \text{Var}(X)$$

$$\text{Cov}(X, Y) = \text{Cov}(Y, X)$$

$$\text{Cov}(aX, bY) = ab \text{Cov}(X, Y)$$

$$\text{Cov}(X + a, Y + b) = \text{Cov}(X, Y)$$

$$\text{Cov}(aX + bY, cW + dV) = ac \text{Cov}(X, W) + ad \text{Cov}(X, V) + bc \text{Cov}(Y, W) + bd \text{Cov}(Y, V)$$

5 假设检验 (Statistical hypothesis testing)

原假设 (H_0 , **null hypothesis**), 也称为零假设或虚无假设。而与原假设相反的假设称为备择假设 (H_a , **alternative hypothesis**)。假设检验的核心为反证法。在数学中, 由于不能穷举所有可能性, 因此无法通过举例的方式证明一个命题的正确性。但是可以通过举一个反例, 来证明命题的错误。在掷骰子的例子中, 在每次掷的过程相当于一次举例, 假设进行了上万次的实验, 即便实验结果均值为 3.5, 也无法证明总体的均值为 3.5, 因为无法穷举。

可以理解为原假设为希望拒绝的假设, 或反证法中希望推翻的命题。我们先构造一个小概率事件作为原假设 (H_0), 并假设其正确。如样本均值等于某值, 两个样本均值是否相等, 样本中的不同组直接是否等概率发生, 一般使用等式 (小概率) 作为原假设。如果抽样检验中小概率事件发生, 则说明原假设的正确性值得怀疑。如此时假设实验的结果 (样本) 远大于或小于理论计算结果 3.5, 即发生了小

概率事件，那么就有理由相信举出了一个反例，这时就可以否定原命题（reject the null hypothesis）。而相反，如果原假设认为均值为 3.5，在实验的过程中结果大概率不会偏离这个理论值太多，可以认为我们并没办法举出反例。由于不能直接证明原命题为真，只能说” We can not(fail to) reject the null hypothesis “，无法拒绝原命题。

在需要评估总体数据的时候，由于经常无法统计全部数据，需要从总体中抽出一部分样本进行评估。假设掷骰子一个骰子，其期望为 3.5，但假设掷骰子了 100 次，计算均值为 3.47，由于总体的理论值和样本呢的实验值可能存在偏差，误差永远存在，无法避免。那么是否可以认为么 3.47 “等于” 3.5？这时候就需要要界定一个**显著水平** (α , **significant level**)，相当于设定一个等于的阈值范围。即多小概率的事情发生，是 10% 还是 5% 的概率，使我们认为举出了一个反例，值得去怀疑原命题的正确性。当我们知道随机变量的分布时候，根据所进行的检验，我们可以根据计算出的**统计量** (**test statistic**)，由于分布已知，统计量对应了一个 **p 值** (**p-value**)，即小概率（极端）事件发生的概率，因此在图形上表示为统计量向两侧延申的线下区域。如果这个概率足够低，如小于 $\alpha = 5\%$ ，那么就有理由拒绝原假设。

用 $1 - \alpha$ 显著水平 ($1 - \alpha$)，得到值称为**置信水平** (**confidence level**) (概率大小)。置信水平越大，对应的置信区间也越大 (随机变量范围)。此时有置信水平为 $1 - \alpha$ ，假设置信区间为 (a, b) ，那么有 $P(a < \text{随机变量} < b) = 1 - \alpha$ 。对于双侧检验，有置信水平为 $1 - \alpha$ (概率大小)，两侧拒绝域分别为 $\alpha/2$ 。对于单侧检验，则有单侧拒绝域大小为 α 。

6 Chi-square distribution

假设有随机变量 X 服从标准正态分布，即有 $X \sim N(0, 1)$ ，此时有随机变量 $Q_1 = X^2$ ，则有随机变量 Q_1 服从卡方分布 (χ^2 -distribution)，由于此时只有一个随机变量，因此卡方分布自由度 (degree of freedom) 为 1，即 $Q_1 \sim \chi^2(1)$ 。如随机变量 $Q_2 = X_1^2 + X_2^2$ ，且 X_1 与 X_2 同时服从标准正态分布。则此时 Q_2 服从自由度为 2 的卡方分布，即 $Q_2 \sim \chi^2(2)$ 。

Goodness of fit

Pearson's chi-squared test

$$\chi^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i}$$

- O_i the number of observations of type i
- E_i the expected(theoretical) number of type i

7 Probability vs Likelihood

7.1 Probability

$P(\text{data} \mid \text{distribution}) = \text{area under curve}$

$P(\text{weight between 32g and 34g} \mid \text{mean} = 32 \text{ and standard deviation} = 2.5) = 0.29$

$P(\text{weight} > 34g \mid \text{mean} = 32 \text{ and standard deviation} = 2.5) = 0.21$

7.2 Likelihood

$L(\text{distribution} \mid \text{data}) = \text{value of the curve (y)}$

$L(\text{mean} = 32 \text{ and standard deviation} = 2.5 \mid \text{mouse weights 34g}) = 0.12$

$L(\text{mean} = 34 \text{ and standard deviation} = 2.5 \mid \text{mouse weights } 34\text{g}) = 0.21$

在调整了分布的 mean 之后, likelihood 最大, 在 mean=34 sigma=2.5 的正态分布中, 抽中一只 34g 的老鼠的概率最大

7.3 Maximum likelihood

测量了数只老鼠的重量, 尝试找到其分布, maximizes the likelihood 找到最大化所有观察重量 likelihood 的分布, 找到 mean 和 standard deviation