

## 可汗笔记

---

### 均值与方差公式

---

$$\text{(总体均值)} \quad \mu = \frac{\sum_{i=1}^N x_i}{N} = \frac{x_1 + x_2 + \dots + x_N}{N}$$

$$\text{(样本均值)} \quad \bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

$$\text{(总体方差)} \quad \sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} = \frac{\sum_{i=1}^N x_i^2}{N} - \mu^2$$

$$\text{(样本方差)} \quad s^2 = s_{n-1}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

**n-1 :** 你所取的数学有可能不包含实际总体均值，此时样本方差低估了总体方差，但完全有可能总体均值在样本之外

Eg. 2 2 3 3

均值 =  $(2+2+3+3)/4 = 2.5$

$$\text{总体方差} = \frac{(2-2.5)^2 + (2-2.5)^2 + (3-2.5)^2 + (3-2.5)^2}{4}$$

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

$$= \frac{\sum_{i=1}^N x_i^2 - 2\mu \sum_{i=1}^N x_i + \mu^2 \sum_{i=1}^N 1}{N}$$

$$= \frac{\sum_{i=1}^N x_i^2}{N} - \frac{2\mu \sum_{i=1}^N x_i}{N} + \frac{\mu^2 N}{N}$$

$$= \frac{\sum_{i=1}^N x_i^2}{N} - 2\mu^2 + \mu^2$$

$$= \frac{\sum_{i=1}^N x_i^2}{N} - \mu^2$$

## 随机变量

---

离散随机变量：值是有限的

连续随机变量：值有无限多个

$n-1$ ：你所取的数学有可能不包含实际总体均值，此时样本方差低估了总体方差，但完全有可能总体均值在样本之外

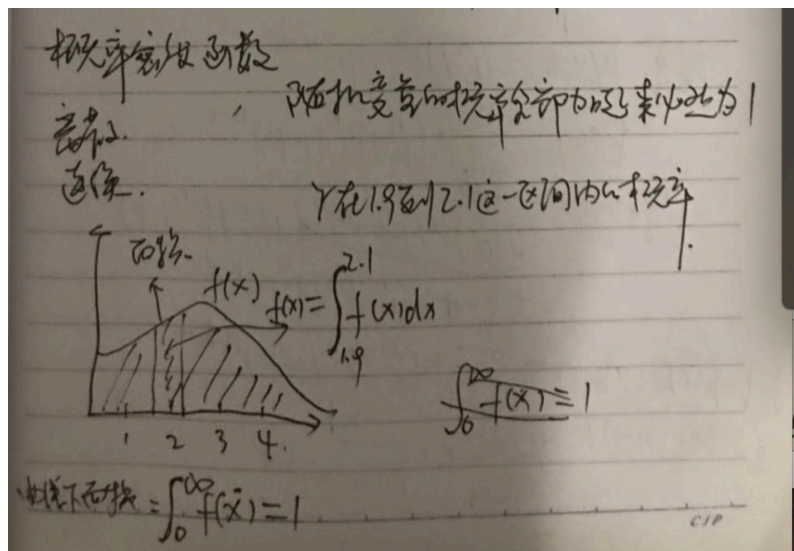
## 概率密度函数

---

用来计算连续随机变量的概率

概率密度函数： $\int_0^{\infty} f(x) = 1$

概率密度函数下方面积必然等于 1



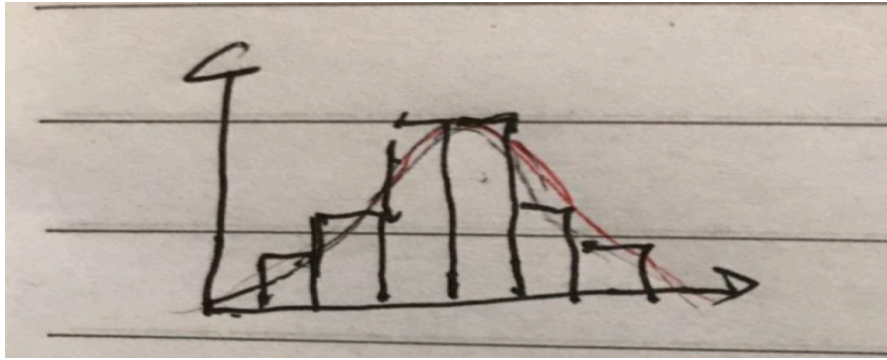
## 二项及正态分布

---

离散的情况将得到二项分布

连续的情况将得到正态分布

二项分布图如下：



二项式概率： $p(x=k) = \binom{n}{k} p^k (1-p)^{n-k}$

二项分布期望值： $E(x) = \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} = np$ ,  $n$ : 试验次数,  $p$ : 每次成功的概率

期望值  $E(x) = n \cdot p$ , 随机变量的期望值其实就是总体均值, 只是总体是无穷尽的, 无法全部求和取平均值

## 正态分布 (高斯分布)

二项分布试验足够多时, 会很接近正态分布

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

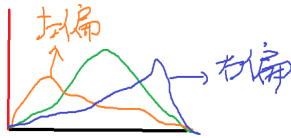
正态分布/高斯函数.

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \rightarrow z\text{分数}$$

标准误差 (抽样标准差)  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$  或  $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$

$$z\text{分数} = \frac{x-\mu}{\sigma}$$





Eg: the 2007 AP Statistics examination scores were not normally distributed, with

$\mu = 2.8, \sigma = 1.34$ , What is the approximate z-score that corresponds to an exam score of 5 (the scores range from 1-5)

$$\text{解: } \frac{5 - 2.8}{1.34} = 1.64$$

- 1、尾部向正，为正(右)偏态分布，尾部向负，为负(左)偏态分布
- 2、如果峰度为正，尾部会较肥，同时峰值为较尖，负峰态则尾部较小，中间更平滑
- 3、Z 分数：表示离均值有多少个标准差。‘Z 分数既可以用在正态分布，也可以用在非正态分布
- 4、经验法则：68-95-99.7，即均值左右一个标准差的概率是 68%，两个标准差是 95%，3 个标准差内是 99.7%
- 5、标准正态分布:  $\mu = 0, \sigma = 1$
- 6、随着样本容量的增大，会发生两件事，一是更接近正态分布，二是标准差更小

EG: The average male drinks 2L of water when active outdoors (with a standard deviation of 0.7L). You are planning a full day nature trip for 50 men and will bring 110 L of water. What is the probability that you will run out?

解:  $P(\text{average water use per man is } > 2.2\text{L/m})$

$$\mu_{\bar{x}} = \mu = 2\text{L}$$

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.7}{\sqrt{50}} = 0.099$$

$$\text{Z 分数} = \frac{2.2 - 2}{0.099} = 2.02$$

$$P(\bar{x} \text{ will be more water 2.02 std. deviation above the mean}) = 1 - 0.9783 = 2.17\%$$

查阅 Z 表格 2.02 对应值为 0.9783，它是小于该 Z 分数处的面积

## 中心极限定律

---

样本的频率非常接近正态分布，随着  $n$  值越大，接近的越好

## 大数定律

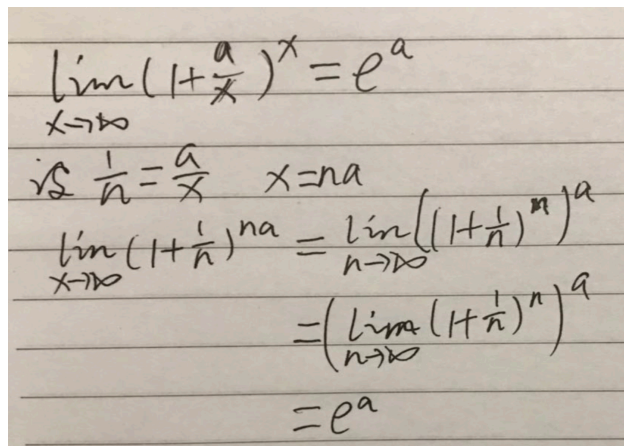
---

当  $n$  趋于  $\infty$  时，样本均值趋于总体样本，即期望值趋于真正的值

## 泊松过程

---

### 公式推导 1



Handwritten derivation of the limit definition of  $e^a$ :

$$\lim_{x \rightarrow \infty} \left(1 + \frac{a}{x}\right)^x = e^a$$

Let  $\frac{1}{n} = \frac{a}{x}$   $x = na$

$$\lim_{x \rightarrow \infty} \left(1 + \frac{1}{n}\right)^{na} = \lim_{n \rightarrow \infty} \left(\left(1 + \frac{1}{n}\right)^n\right)^a$$
$$= \left(\lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n\right)^a$$
$$= e^a$$

### 公式推导 2

$$E(x) = \lambda = np \Rightarrow p = \frac{\lambda}{n}$$

$$P(X=k) = \lim_{n \rightarrow \infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

$$= \lim_{n \rightarrow \infty} \frac{n!}{(n-k)!k!} \cdot \frac{\lambda^k}{n^k} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}$$

$$= \frac{n \cdot (n-1) \cdot (n-2) \cdots (n-k+1)}{n^k} \cdot \frac{\lambda^k}{k!} \cdot \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}$$

$$\lim_{x \rightarrow a} f(x)g(x) = \lim_{x \rightarrow a} f(x) \cdot \lim_{x \rightarrow a} g(x)$$

$$= \lim_{n \rightarrow \infty} \frac{n^k + \dots}{n^k} \cdot \frac{\lambda^k}{k!} \cdot \lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}$$

$$= 1 \cdot \frac{\lambda^k}{k!} e^{-\lambda} \cdot 1$$

$$P(X=k) = \lim_{n \rightarrow \infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

$$= \frac{\lambda^k}{k!} e^{-\lambda}$$

λ = 平均每小时到达的车辆数，即每小时正好有 λ 辆车到达的概率

$$\frac{9^2}{2!} = \frac{81}{2} \cdot e^{-9}$$

$$\bar{x}_n = \frac{x_1 + x_2 + \dots + x_n}{n}$$

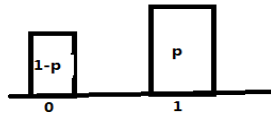
$$\bar{x}_n \rightarrow E(X) \quad n \rightarrow \infty$$

$$\bar{x}_n \rightarrow \mu$$



## 伯努利分布均值和方差公式实例

---



$$\mu = (1-p) \cdot 0 + p \cdot 1 = p$$

$$\sigma^2 = (1-p)(0-p)^2 + p(1-p)^2$$

$$= (1-p)p^2 + p(1-2p+p^2)$$

$$= p^2 - p^3 + p - 2p^2 + p^2$$

$$= p - p^2$$

## 误差范围

---

eg. In a local teaching district a technology grant is available to teachers in order to install a cluster of four computers in their classrooms. From the 6250 teachers in the district, 250 were randomly selected and asked if they felt that computers were an essential teaching tool for their classroom. Of those selected, 142 teachers felt that computers were an essential teaching tool.

1. Calculate a 99% confidence interval for the proportion of teachers who felt that computers are an essential teaching tool
2. How could the survey be changed to narrow the confidence interval but to maintain the 99% confidence interval?

250 sampled: 142 good=1    108 not good=0

样本均值： $\bar{x} = \frac{1 \cdot 142 + 0 \cdot 108}{250} = 0.568$

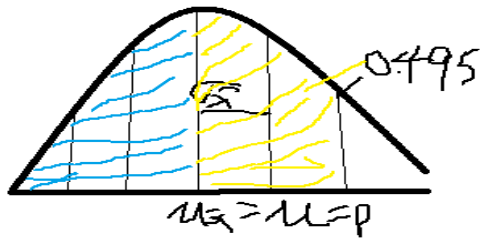
样本方差： $S^2 = \frac{142(1-0.568)^2 + 108(0-0.568)^2}{250-1} = 0.246$

样本标准差： $S = 0.50$

Confident that: 抽样分布标准差近似于样本标准差除以根号下样本容量：

$$\sigma_x \approx \frac{0.50}{\sqrt{250}} = 0.031$$

99%的一半是 0.495



0.5+0.495=0.995 即上面绿色和黄色部分和，为所求  
查询 0.995 对应的 Z 分数为 2.58

99% chance that a random  $\bar{x}$  is within 2.58  $\sigma_{\bar{x}}$  of  $p$ (样本均值落在抽样分布均值左右 2.58 个标准差范围内)

“confident” 99% chance that  $\bar{x}$  (0.568) is within  $2.58 \times 0.031 = 0.08$  of the population

也就是：“confident” 99% chance that  $p$  is within 0.08 of the 0.568

上限:  $0.568 + 0.08 = 0.648$

下限:  $0.568 - 0.08 = 0.488$

答案 1：The true percentage of teachers that like the computers is between 48.8% and 64.8%

实际上有 48.8%到 64.8%的老师认为计算机是必备的

答案 2：增大样本容量,标准差越小，而置信区间是加减一定倍数的标准差，范围自然也会减小

Eg: 7 patients' s blood pressures have been measured after having been given a new drug for 3 months. They had blood pressure increases of 1.5, 2.9, 0.9, 3.9, 3.2, 2.1 and 1.9. Construct a 95% confidence interval for the true expected blood pressure increase for all patients in a population

$$\bar{x} = 2.34$$

$$S = 1.04$$

使用样本标准差  $s$  来估计总体标准差  $s \approx \sigma = 1.04$  (这里的标准差  $s$  not good,  $n$  is small, in general, this is considered a bad estimate if  $n$  is less than 30.  $n$  小于 30 通常被认为是糟糕的估计)

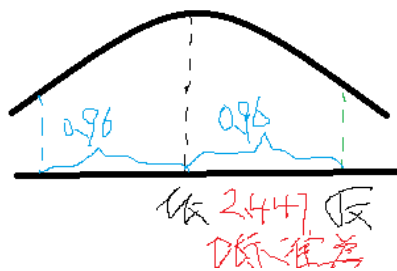
$t$  分布一般为小样本容量时置信区间的更好估计所设计的，它为正态分布差不多，不过，尾部较肥

查询  $t$  表格，自由度为  $6 = n - 1$  在 95% 对应的数据为 2.447, 它对应两侧 2.447 个标准差，即距离均值 2.447 个标准差

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} = \frac{1.04}{\sqrt{7}} = 0.39$$

要求抽样分布上包含 95% 面积的这个区域，我们需要用  $0.39 \times 2.447 = 0.96$





有 95% 几率 2.34 在抽样分布实际均值周围 0.96 范围内，即 95% 几率均值在样本均值 2.34 周围 0.96 范围内

置信区间下限是  $2.34 - 0.96 = 1.38$

上限  $2.34 + 0.96 = 3.3$

所以置信区间是 1.38 到 3.3

Eg. A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug, subjecting each to neurological stimulus, and recording its response time. The neurologist knows that the mean response time of rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats' response times is 1.05 seconds with a sample standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time?

$H_0$ : Drug has no effect  $\mu = 1.2s$

$H_1$ : Drug has an effect  $\mu \neq 1.2s$  when the drug is given

$$\sigma_x = \frac{\sigma}{\sqrt{100}} \approx \frac{s}{\sqrt{100}} = \frac{0.5}{\sqrt{100}} = 0.05$$

标准差估计值  $\bar{\sigma}_x = 0.05$

想想，得到 1.05 秒的概率是多少？或者说 1.05 秒离抽样分布均值有多少个标准差远，以及均值周围这么多标准差远之内的概率是多少，首先求这离均值有多少个标准差远，这其实就是求一个 z 分数，z 统计量离均值有多远呢

$$Z = \frac{\text{均值} - 1.05}{\text{标准差估计值}} = \frac{1.2 - 1.05}{0.05} = 3$$

3 个标准差内的置信区间为 99.7%，那么  $1 - 99.7\% = 0.3\%$ ，下面蓝色部分即为所求， $0.3\% = 0.003$  概率非常小， $< 0.05$ ，所以拒绝零假设，所以药物是有效果的

上面的为双侧检验，即正负两边

Eg. 单侧检验，只检验一边

$H_0$ : Drug has no effect,  $\mu = 1.2s$

$H_1$ : Drug lowers response,  $\mu < 1.2s$



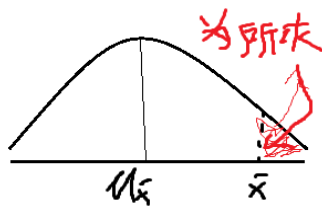
### t 统计与 z 统计

需要求出该值离均值有多少个标准差远，做法是用样本均值减去实际均值，然后除以抽样分布的标准差

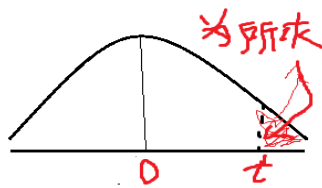
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \text{ (总体标准差除以样本容量的平方根)}$$

$$Z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{\sigma}{\sqrt{n}}} \text{ (对离均值有多少个标准差远的最好度量)}$$

$$Z \approx \frac{\bar{x} - \mu_{\bar{x}}}{\frac{S}{\sqrt{n}}} \text{ (n>30 时, 将服从正态分布)}$$



$$t \approx \frac{\bar{x} - \mu_{\bar{x}}}{\frac{S}{\sqrt{n}}} \text{ (n<30 时, 将为 t 分布)}$$



### 小样本假设检验

The mean emission of all engines of a new design needs to be below 20 ppm if the design is to meet new emission requirements. Ten engines are manufactured for testing purposes, and the emission level of each is determined. The emission data is:

15.6, 16.2, 22.5, 20.5, 16.4, 19.4, 16.6, 17.9, 12.7, 13.9

Does the data supply sufficient evidence to conclude that this type of engine meets the new standard? Assume we are willing to risk a Type 1 error with probability=0.01

H0:  $\mu = 20\text{ppm}$

H1:  $\mu < 20\text{ppm}$

Reject H0 (样本均值得到 17.17 的概率若小于 1% 就拒绝零假设)

$$t = \frac{17.17 - 20}{\frac{2.98}{\sqrt{10}}} = -3$$



Eg. We want to test the hypothesis that more than 30% of U.S. households have internet access (with a significance level of 5%). We collect a sample of 150 households and find that 57 have access.

### 随机变量之差的方差

设  $X, Y$

$$E(X) = \mu_x$$

$$E(Y) = \mu_y$$

$$\text{Var}(X) = E((X - \mu_x)^2) = \sigma_x^2 \quad \text{Var}(Y) = E((Y - \mu_y)^2) = \sigma_y^2$$

$$Z = X + Y$$

$$A = X - Y$$

$$E(Z) = E(X + Y) = E(X) + E(Y)$$

$$E(A) = E(X - Y) = E(X) - E(Y)$$

$$\mu_z = \mu_x + \mu_y$$

$$\mu_A = \mu_x - \mu_y$$

$$\text{Var}(Z) = \text{Var}(X) + \text{Var}(Y) \quad \sigma_A^2 = \sigma_{X-Y}^2 = \sigma_{X+(-Y)}^2 = \sigma_x^2 + \sigma_{-Y}^2$$

$$\sigma_{-Y}^2 = \text{Var}(-Y) = E((-Y - E(-Y))^2)$$

$$= E((-1)^2 (Y - E(Y))^2)$$

$$= E((Y - E(Y))^2)$$

$$= \sigma_y^2$$

$$\sigma_z^2 = \sigma_{X+Y}^2 = \sigma_x^2 + \sigma_y^2 \quad \sigma_A^2 = \sigma_{X-Y}^2 = \sigma_x^2 + \sigma_y^2$$

Eg. We're trying to test whether a new, low-fat diet actually helps obese people lose weight. 100 randomly assigned obese people are assigned to group 1 and put on the low fat diet. Another 100 randomly assigned obese people are assigned to group 2 and put on a diet of approximately the same amount of food, but not as low in fat. After 4 months, the mean weight loss was 9.31 lbs. for group 1 ( $s=4.67$ ) and 7.40 lbs ( $s=4.04$ ) for group 2.

### 线性回归

$$y = mx + b$$

m: 斜率 b: 截距

$$m = \frac{\overline{xy - xy}}{x^2 - (\bar{x})^2}$$

$$b = \bar{y} - m\bar{x}$$

eg. (-2, -3), (-1, -1), (1, 2), (4, 3)

$$\bar{x} = \frac{-2 + (-1) + 1 + 4}{4} = \frac{1}{2}$$

$$\bar{y} = \frac{-3 + (-1) + 2 + 3}{4} = \frac{1}{4}$$

$$\overline{xy} = \frac{6 + 1 + 2 + 12}{4} = \frac{21}{4}$$

$$\overline{x^2} = \frac{4 + 1 + 1 + 16}{4} = \frac{11}{2}$$

$$m = \frac{\overline{xy - xy}}{x^2 - (\bar{x})^2} = \frac{\frac{21}{4} - \frac{1}{2} * \frac{1}{4}}{\frac{11}{2} - (\frac{1}{2})^2} = \frac{41}{42}$$

$$b = \bar{y} - m\bar{x} = \frac{1}{4} - \frac{41}{42} * \frac{1}{2} = -\frac{5}{21}$$

$$y = \frac{41}{42}x - \frac{5}{21}$$

协方差  $Cov(x, y) = \overline{xy} - \bar{y}\bar{x}$

$$M = \frac{\overline{xy - xy}}{x^2 - (\bar{x})^2} = \frac{\overline{xy - xy}}{x^*x - \bar{x}^*x} = \frac{Cov(x, y)}{Cov(x)} = \frac{Cov(x, y)}{Var(x)}$$

## 方差分析

---

1 2 3

3 5 5

2 3 6

1 4 7

n: 每组个数, m: 共有几组

$$\bar{x}_1 = \frac{3+2+1}{3} = 2 \quad \bar{x}_2 = \frac{5+3+4}{3} = 4 \quad \bar{x}_3 = \frac{5+6+7}{3} = 6$$

$$\bar{x} = \frac{3+2+1+5+3+4+5+6+7}{9} = 4$$

$$\begin{aligned} \text{组内均值(总均值)SST} &= (3-4)^2 + (2-4)^2 + (1-4)^2 + (5-4)^2 + (3-4)^2 + \\ &\quad (4-4)^2 + (5-4)^2 + (6-4)^2 + (7-4)^2 \\ &= 30 \end{aligned}$$

自由度:  $mn-1$

$$\begin{aligned} \text{SSW (组内平方和)} &= (3-2)^2 + (2-2)^2 + (1-2)^2 + (5-4)^2 + (3-4)^2 + \\ &\quad (4-4)^2 + (5-6)^2 + (6-6)^2 + (7-6)^2 \end{aligned}$$

$$=6$$

自由度：m(n-1)

SSB (组间平方和) :组均值相对总均值的波动。

$$\begin{aligned} \text{SSB} &= (2-4)^2 + (2-4)^2 + (2-4)^2 + (4-4)^2 + (4-4)^2 + (4-4)^2 + (6-4)^2 + (6-4)^2 + (6-4)^2 \\ &= 24 \end{aligned}$$

自由度：m-1

$$\text{SST} = \text{SSW} + \text{SSB}$$

$$\text{总自由度 } mn-1 = m(n-1) + (m-1)$$

$$F\text{-statistic} = \frac{\frac{\text{SSB}}{m-1}}{\frac{\text{SSW}}{m(n-1)}}$$

## 演绎推理

归纳推理是寻找规律或趋势，然后推广、

演绎推理是从一些数据或事实出发，演绎得到其它正确的事实

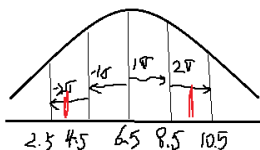
## 切比雪夫定理

At least  $(1 - \frac{1}{k^2})\%$  of all observations lie within k deviations of the mean

Eg. what proportion of people spend between 3.3 and 9.7 minutes at the station.

$$\mu = \frac{3.3 + 9.7}{2} = 6.5$$

$$\sigma = \sqrt{\frac{(3.3-6.5)^2 + (9.7-6.5)^2}{2}} = 2$$



$$\text{离均值有多少个标准差: } \frac{9.7-6.5}{2} = 1.6 \quad \frac{3.3-6.5}{2} = -1.6$$

$$(1 - \frac{1}{k^2})\% = (1 - \frac{1}{1.6^2})\% = 61\%$$