Part 5

抽样与期望

郎大为 J.D. Power

Outlines

抽样与期望

- 期望
    - 连续与离散变量
    - 均值
    - 方差
- 抽样
    - 随机抽样
- 大数定律与中心极限定理

期望

期望

两个人对赌，赌资100法郎，规定谁赢够3局胜利，拿走所有赌资

- 第一个人赢了两局
- 第二个人赢了一局
- 这时因故暂停了赌博
- 假设两个人水平相当，赌资应该如何分配？

期望

- **期望** 或者称 **均值** 是随机变量的分布的中心
- 对离散随机变量X 有分布列(PMF) $p(x)$, 期望的定义为

$$E[X] = \sum_x xp(x).$$

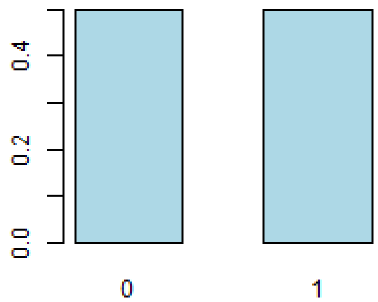
- $E[X]$ 代表了这个离散分布的加权平均 $\{x, p(x)\}$

例子

- 掷硬币,假设得到的是随机变量 X, 0代表反面, 1代表反面
- X的期望是

$$E[X] = .5 \times 0 + .5 \times 1 = .5$$

- 二者的概率一样,所以等价于平均值 $.5$



例子

- 如果扔一枚骰子,随机变量 X 代表了扔出的点数
- X的期望是?

$$E[X] = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5$$

连续变量的期望

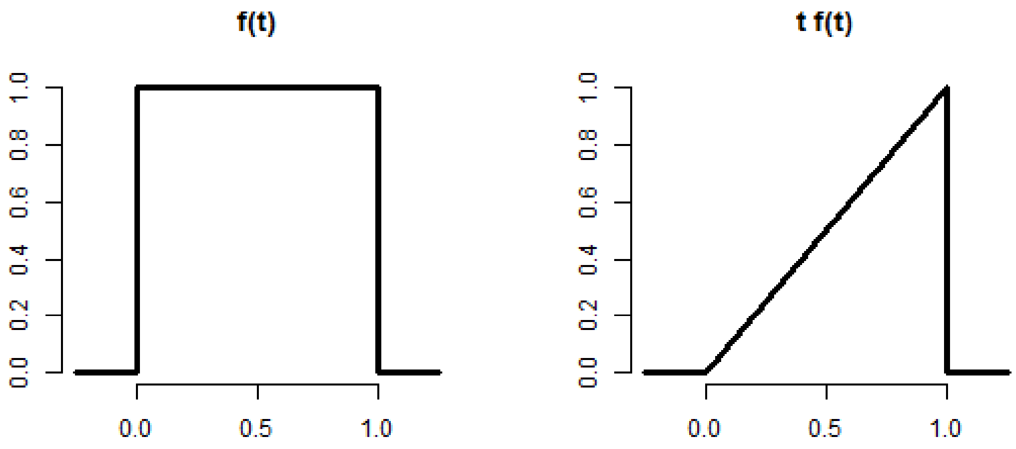
- 对于连续随机变量 X, 概率密度函数为 f, 期望的定义为

$$E[X] = \text{the area under the function} \quad tf(t)$$

- 将期望的定义推广到了连续的场合
- 分布的中心

例子

- 假设一个随机变量的概率密度函数为 $f(x) = 1$, x的取值为0到1
- 如果随机变量X 服从这个分布,它的期望是?



期望的性质

- 如果 a 和 b 不是随机变量,而 X, Y 是两个随机变量
  - $E(a) = a$
  - $E[aX + b] = aE[X] + b$
  - $E[X + Y] = E[X] + E[Y]$

例子

- X 是扔一枚硬币的结果, Y是从0到1生成的一个随机数,二者和的期望是?

$$E[X+Y]=E[X]+E[Y]=.5+.5=1$$

- 一枚骰子掷两次,均值的期望是?
- 令 X_1 和 X_2 是两次的结果

$$E[(X_1+X_2)/2]=\frac{1}{2}(E[X_1]+E[X_2])=\frac{1}{2}(3.5+3.5)=3.5$$

例子

1. 令 $X_i \ i = 1, \ldots, n$是从一个总体中生成的抽样,相互独立,均值为 μ
2. 计算样本X_i均值的期望

$$
\begin{aligned}
E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] &= \frac{1}{n}E\left[\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} E[X_i] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu = \mu.
\end{aligned}
$$

Remark

- 样本均值是对总体均值的一个**估计**
- 因此, **样本均值**的期望就是总体的均值 μ
  - 也就是样本均值想预测的结果
- 当期望与想预测的参数一致的时候
  - 这个估计就是一个无偏估计

方差

- 方差是随机变量*离散程度*的度量
- 如果X 是一个随机变量,均值为μ, X的方差定义为

$$Var(X) = E[(X-\mu)^2]$$

随机变量到均值的距离的平方

- 方差较大的随机变量离散的程度更高

方差的性质

- 方便计算的形式

$$Var(X) = E[X^2] - E[X]^2$$

- 如果 a 为常数 $Var(aX) = a^2 Var(X)$
- 方差的平方根为 **标准差**
- 标准差的单位与 X 相同

Example

- 掷一枚骰子的方差

Example

- 掷一枚骰子的方差
  - $E[X] = 3.5$
  - $E[X^2] = 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + 3^2 \times \frac{1}{6} + 4^2 \times \frac{1}{6} + 5^2 \times \frac{1}{6} + 6^2 \times \frac{1}{6} = 15.17$
- $Var(X) = E[X^2] - E[X]^2 \approx 2.92$

Example

- 伯努利分布的方差

Example

- 伯努利分布的方差
  - $E[X] = 0 \times (1-p) + 1 \times p = p$
  - $E[X^2] = E[X] = p$
- $Var(X) = E[X^2] - E[X]^2 = p - p^2 = p(1-p)$

Interpreting variances

- Chebyshev's inequality is useful for interpreting variances
- This inequality states that

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

- For example, the probability that a random variable lies beyond k standard deviations from its mean is less than $1/k^2$

$$\begin{aligned} 2\sigma &\to 25\% \\ 3\sigma &\to 11\% \\ 4\sigma &\to 6\% \end{aligned}$$

- Note this is only a bound; the actual probability might be quite a bit smaller

Example

- IQs are often said to be distributed with a mean of 100 and a sd of 15
- What is the probability of a randomly drawn person having an IQ higher than 160 or below 40?
- Thus we want to know the probability of a person being more than 4 standard deviations from the mean
- Thus Chebyshev's inequality suggests that this will be no larger than 6\%
- IQs distributions are often cited as being bell shaped, in which case this bound is very conservative
- The probability of a random draw from a bell curve being 4 standard deviations from the mean is on the order of 10^{-5} (one thousandth of one percent)

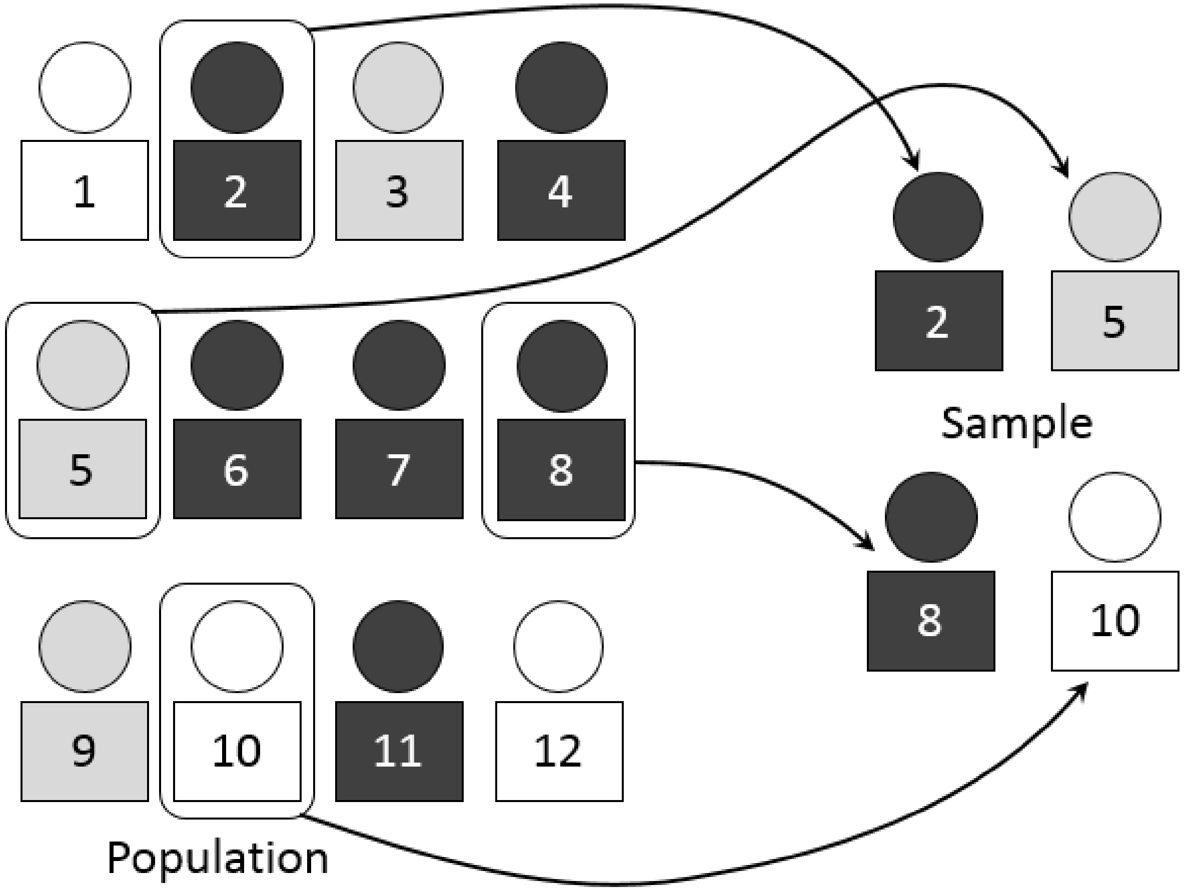
Example

- A former buzz phrase in industrial quality control is Motorola's "Six Sigma" whereby businesses are suggested to control extreme events or rare defective parts
- Chebyshev's inequality states that the probability of a "Six Sigma" event is less than $1/6^2 \approx 3\%$
- If a bell curve is assumed, the probability of a "six sigma" event is on the order of 10^{-9} (one ten millionth of a percent)

抽样

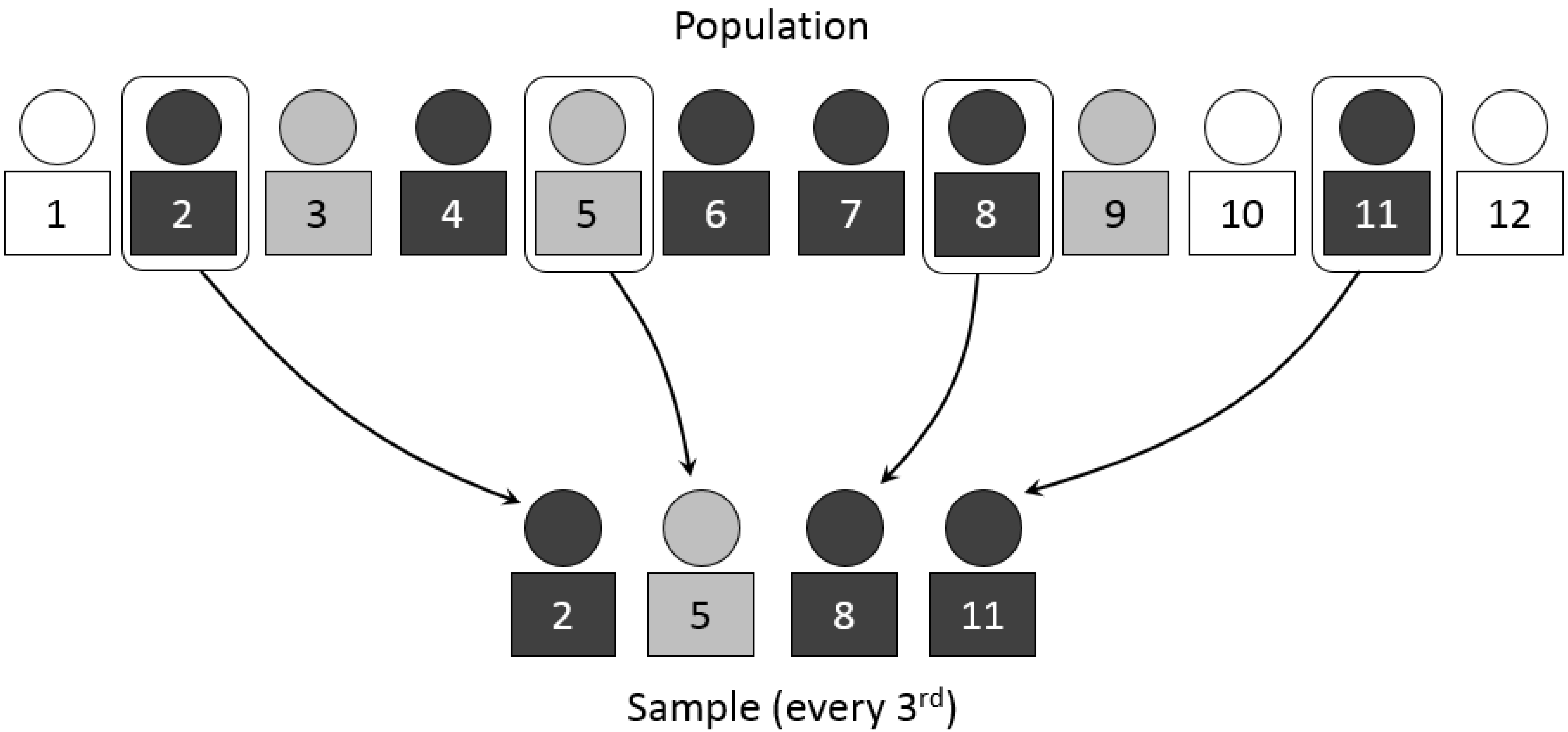
抽样

简单随机抽样



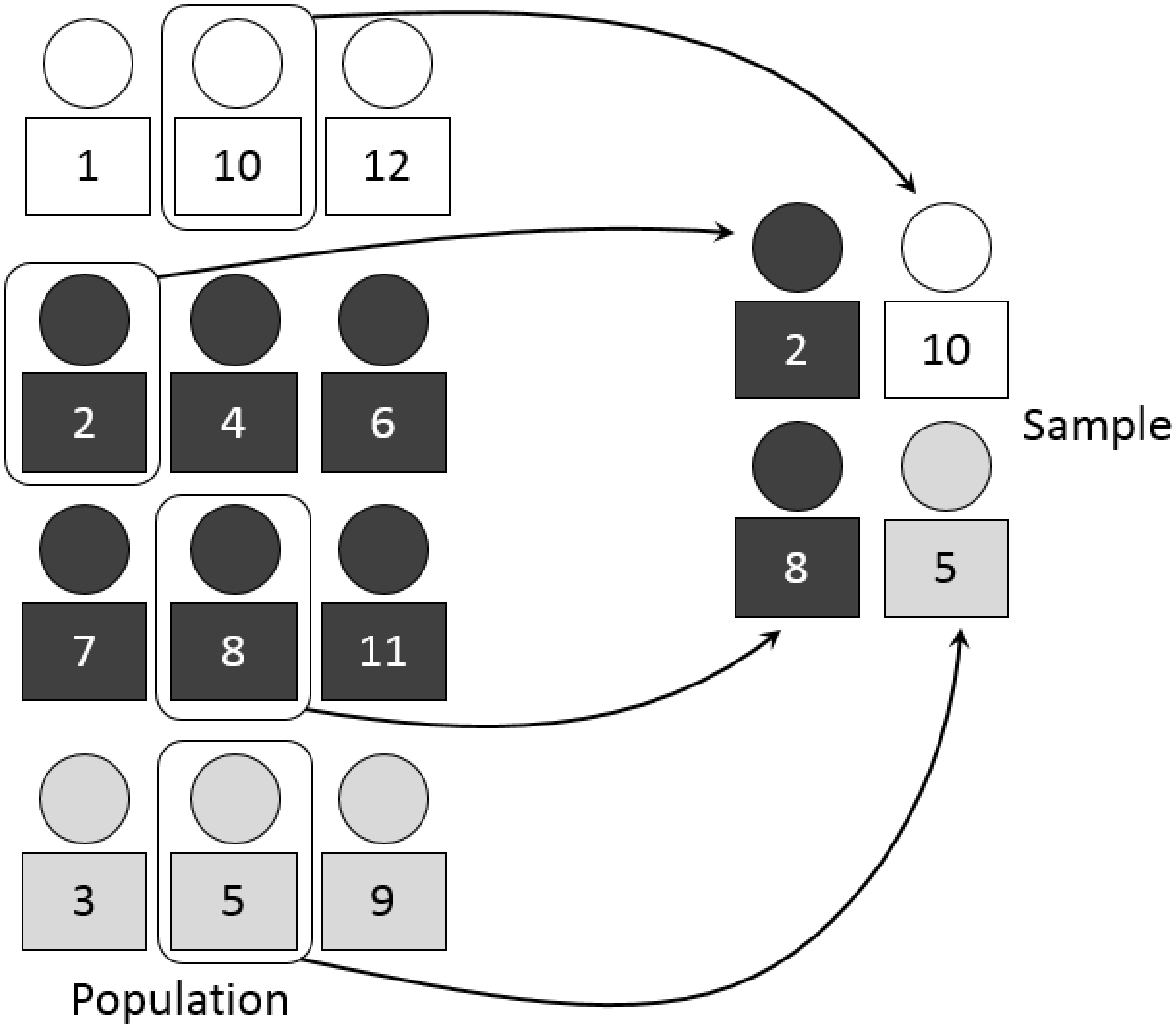
抽样

系统抽样



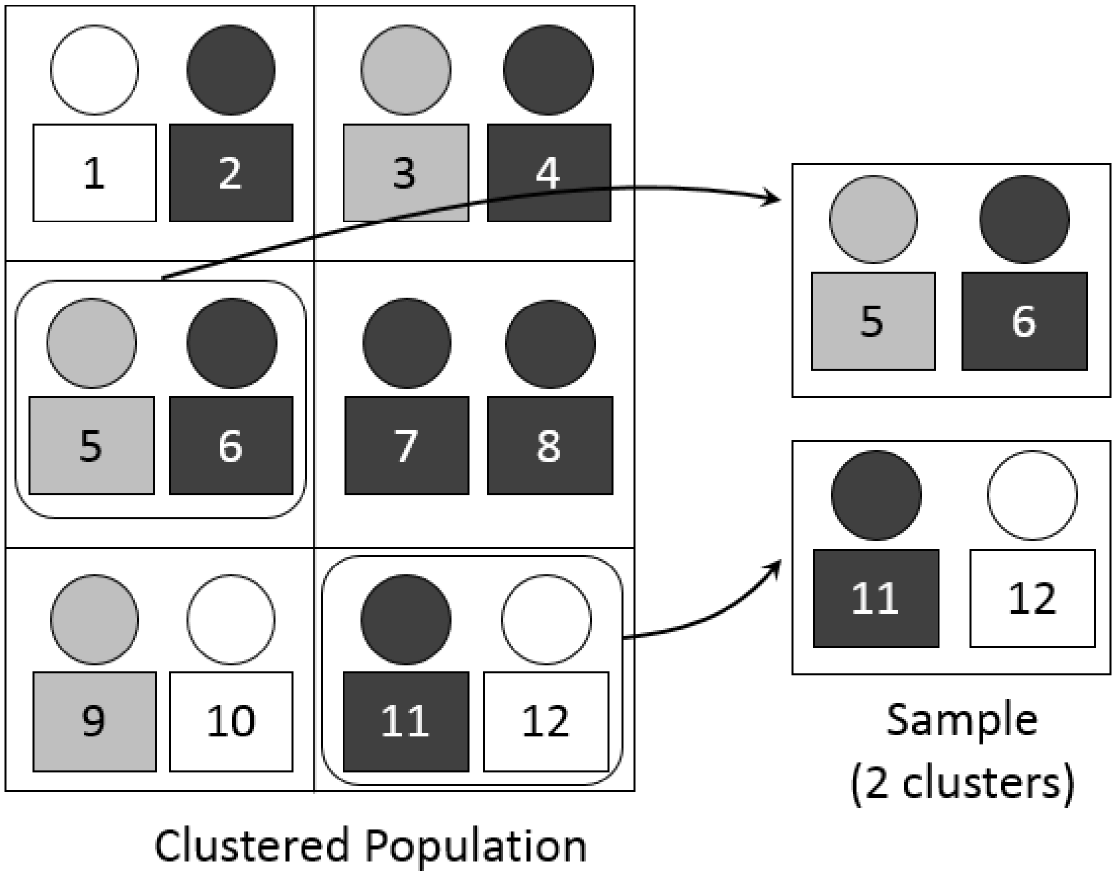
抽样

分层抽样



抽样

整群抽样


1
2
3
4
5
6
7
8
9
10
11
12
5
6
11
12
Sample
(2 clusters)
Clustered Population

渐进性

Asymptotics

- Asymptotics 指抽样数量趋近与无穷大时所表现出的性质
- (Asymptopia is my name for the land of asymptotics, where everything works out well and there's no messes. The land of infinite data is nice that way.)
- Asymptotics are incredibly useful for simple statistical inference and approximations
- Asymptotics generally give no assurances about finite sample performance
- Asymptotics form the basis for frequency interpretation of probabilities (the long run proportion of times an event occurs)
- To understand asymptotics, we need a very basic understanding of limits.

数值极限

Numerical limits

- 设想以下的一个序列
  - $a_1 = .9$,
  - $a_2 = .99$,
  - $a_3 = .999$, ...
- 这个序列收敛到 1
- 极限的定义：对任何一个确定的距离，我们能找到一个点，序列在这个点后到极限的距离都比给定的距离小

随机变量的极限

Limits of random variables

- 随机变量的问题要稍微难一点
- 令$\bar{X}_n$为前n个iid样本的均值
  - 比如$\bar{X}_n$是n次投硬币结果的均值
- $\bar{X}_n$ converges in probability to a limit if for any fixed distance the probability of $\bar{X}_n$ being closer (further away) than that distance from the limit converges to one (zero)

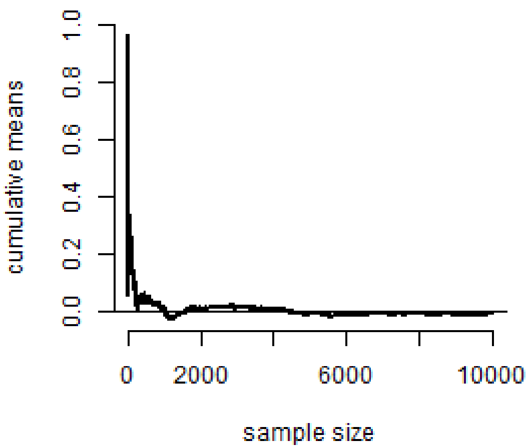
大数定律

The Law of Large Numbers

- 建立一个收敛到某个极限的随机序列比较难
- 大数定律**Law of Large Numbers**有个有效的结论
    - 如果$X_1, \ldots X_n$ 是一个总体的iid抽样，这个总体的均值是μ ，方差是 σ^2
    - $\bar{X}_n$ 依概率收敛到 μ

Law of large numbers in action

```
n <- 10000; means <- cumsum(rnorm(n)) / (1  : n)
plot(1 : n, means, type = "l", lwd = 2,
     frame = FALSE, ylab = "cumulative means", xlab = "sample size")
abline(h = 0)
```



中心极限定理

The Central Limit Theorem

- 中心极限定理是统计学中最重要的定理之一
- 中心极限定理描述了iid随机变量均值**在样本增加的时候**，趋近于一个标准正态分布
- 令 $X_1, \ldots, X_n$ 是从一个均值为 μ ，方差为 σ^2 产生的iid随机变量
- 令 $\bar{X}_n$ 是样本均值
- 在当样本量n变大时， $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$趋近于一个标准正态分布
- 具体的形式

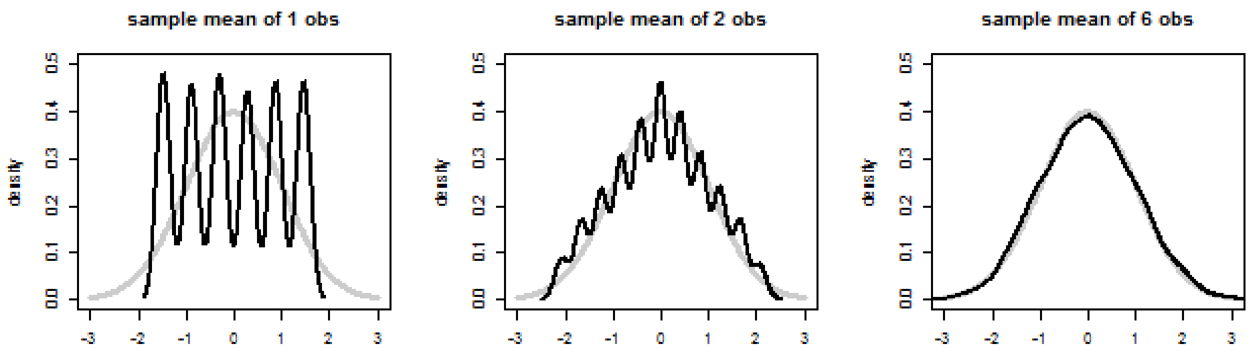
$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\text{Estimate} - \text{Mean of estimate}}{\text{Std. Err. of estimate}}.$$

Example

- 掷一颗均匀的骰子
- 令 X_i 为这个骰子第 i 次的结果
- 均值为 $\mu = E[X_i] = 3.5$
- $Var(X_i) = 2.92$
- 均值的方差 $\sqrt{2.92/n} = 1.71/\sqrt{n}$
- 均值的标准误差

$$\frac{\bar{X}_n - 3.5}{1.71/\sqrt{n}}$$

模拟n次骰子的均值

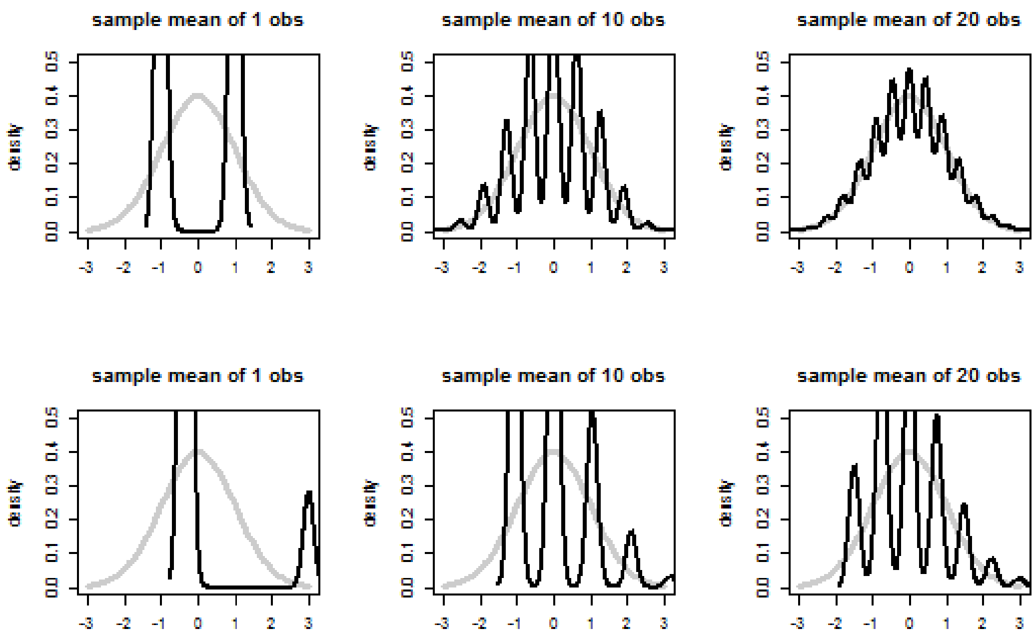


sample mean of 1 obs
sample mean of 2 obs
sample mean of 6 obs
density
0.0
0.1
0.2
0.3
0.4
0.5
-3
-2
-1
0
1
2
3

硬币中心极限定理2

- 令 X_i 取值为 0 或者 1 ，为第 i^{th} 次硬币的结果。
- 样本的均值, $\hat{p}$, 投硬币结果的均值
- 样本的均值和方差为$E[X_i] = p$和 $Var(X_i) = p(1-p)$
- 均值的标准误差为 $\sqrt{p(1-p)/n}$
- $$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

例子


sample mean of 1 obs
sample mean of 10 obs
sample mean of 20 obs
sample mean of 1 obs
sample mean of 10 obs
sample mean of 20 obs
density
0.0
0.1
0.2
0.3
0.4
0.5
-3
-2
-1
0
1
2
3

CLT in practice

- 实际操作中，CLT经常被用作一个近似：

$$P\left(\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}\leq z\right)\approx\Phi(z).$$

- 1.96 与 $.975^{th}$ 分位数
- 例子：

$$\begin{aligned}.95 &\approx P\left(-1.96\leq\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}\leq 1.96\right)\\ &= P\left(\bar{X}_n+1.96\sigma/\sqrt{n}\geq\mu\geq\bar{X}_n-1.96\sigma/\sqrt{n}\right),\end{aligned}$$

置信区间

Confidence Interval

- 依据CLT, 一个随机区间

$$\bar{X}_n \pm z_{1-\alpha/2}\sigma/\sqrt{n}$$

  包含 μ 的概率为 $100(1-\alpha)$%, 其中 $z_{1-\alpha/2}$ 是 $1-\alpha/2$ 标准正态分布的分位数

- 这个区间被称为μ 的 $100(1-\alpha)$% **置信区间**
- 可以用s 代替 σ

抽样性质

- 在一个伯努利实验中，每个X_i 的取值是 0 或者 1， 取1的概率为 p 则方差为 $\sigma^2 = p(1-p)$
- 区间的形式

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

- 注意 $p(1-p) \leq 1/4$， $0 \leq p \leq 1$
- 令 $\alpha = .05$ ，有 $z_{1-\alpha/2} = 1.96 \approx 2$ 则

$$2\sqrt{\frac{p(1-p)}{n}} \leq 2\sqrt{\frac{1}{4n}} = \frac{1}{\sqrt{n}}$$

- 所以 $\hat{p} \pm \frac{1}{\sqrt{n}}$ 是对 p 的CI估计