

Statistics 406 Problem Set 2

Due in lab, Tuesday September 26

1. A Bernoulli trial is a random variable that can take on only two values, which are traditionally denoted 0 and 1. The distribution is characterized by a single parameter p , called the "success probability," which is the probability of observing 1 (the probability of observing 0 is $1 - p$).
 - (a) Derive a formula for the expected value and variance of a Bernoulli trial with success probability p . Since a Bernoulli trial is a discrete random variable, you will need to use the discrete analogues of the expected value and variance formulas given in the notes. The expected value for a discrete random variable X is

$$EX = \sum_x x \cdot P(X = x),$$

where the sum runs over the sample space. The variance is

$$\text{var}(X) = \sum_x (x - EX)^2 P(X = x).$$

Solution:

$$\begin{aligned} EX &= 0 \cdot P(X = 0) + 1 \cdot P(X = 1) \\ &= 0 \cdot (1 - p) + 1 \cdot p \\ &= p \end{aligned}$$

$$\begin{aligned} \text{var } X &= (0 - p)^2 \cdot P(X = 0) + (1 - p)^2 \cdot P(X = 1) \\ &= p^2(1 - p) + (1 - p)^2 p = p(1 - p) \end{aligned}$$

- (b) Write a simulation program in R in which the variance of the sample mean of n Bernoulli trials is estimated. Your program should consider all combinations of sample sizes $n = 10, 20, 40, 80$ and success probabilities $p = 0.1, 0.2, 0.5, 0.7, 0.9$. To simulate a vector of n Bernoulli trials with success probability p in R use

```
BT <- 1*(runif(n) < p)
```

Your final result should be a table in which the value calculated in the simulation appears next to the theoretical value based on the result from part (a).

Solution:

```

V <- NULL

for (p in c(0.1,0.2,0.5,0.7,0.9))
{
  for (n in c(10,20,40,80))
  {
    X <- array(1*(runif(n*1000) < p), c(n,1000))
    M <- apply(X, 2, mean)
    V <- rbind(V, c(p, n, var(M), p*(1-p)/n))
  }
}

```

Here is the result:

	[,1]	[,2]	[,3]	[,4]
[1,]	0.1	10	0.008313554	0.009000
[2,]	0.1	20	0.004577375	0.004500
[3,]	0.1	40	0.002140868	0.002250
[4,]	0.1	80	0.001153108	0.001125
[5,]	0.2	10	0.015763353	0.016000
[6,]	0.2	20	0.007606717	0.008000
[7,]	0.2	40	0.003797297	0.004000
[8,]	0.2	80	0.002122472	0.002000
[9,]	0.5	10	0.024443804	0.025000
[10,]	0.5	20	0.012006904	0.012500
[11,]	0.5	40	0.006335755	0.006250
[12,]	0.5	80	0.002897738	0.003125
[13,]	0.7	10	0.022692442	0.021000
[14,]	0.7	20	0.010656254	0.010500
[15,]	0.7	40	0.005228128	0.005250
[16,]	0.7	80	0.002697397	0.002625
[17,]	0.9	10	0.009386146	0.009000
[18,]	0.9	20	0.004385385	0.004500
[19,]	0.9	40	0.002354078	0.002250
[20,]	0.9	80	0.001130681	0.001125

- Suppose we observe 25 independent values from a certain distribution, and then inadvertently list each value twice. This produces a list of 50 values, of which only 25 are distinct. We then pass this list onto another person, who thinks that all 50 values are independent measures. Will the precision of \bar{X} from this list be determined by a sample size of 25, or by a sample size of 50? Use simulation to find out (you may assume that the 25 'real' measurements are independent and follow a standard normal distribution).

Solution:

```
X <- array(rnorm(50*1000), c(50,1000))
X[26:50,] <- X[1:25,]
M <- apply(X, 2, mean)
V <- var(M)
```

I get a variance of $0.042 \approx 1/25$. Thus the sampling behavior of \bar{X} in this case is similar to what would occur with a sample size of 25 independent values. If 50 independent values were observed, then $\text{var } \bar{X} = 0.02$.