DATA 252 / DATA 551: Homework 2 Solution

1. Suppose a random variable X can take values $\{1, 2, 3\}$ according to the probabilities $\{0.2, 0.3, 0.5\}$.

   a. Simulate 10000 observations from the distribution of X. Calculate the mean of the 10000 simulated draws.

   b. Simulate another 10000 observations from the distribution of X, with the condition that you are only allowed to make random draws *uniformly* from $(0, 1)$ (that is, if you use R, you cannot use the `sample` function, but can only use `runif`).

SOLUTION: a. The mean should be very close to $0.2 + 0.3 \times 2 + 0.5 \times 3 = 2.3$.

```
mysim=sample(c(1,2,3),size=10000,prob=c(0.2,0.3,0.5),replace=T)
mean(mysim)
```

```
## [1] 2.2973
```

   b. Suppose U is a uniform random draw from $(0, 1)$. To draw a value x from the distribution of X, we need to
      - Set $x = 1$ if $U < 0.2$;
      - Set $x = 2$ if $0.2 \leq U < 0.5$;
      - Set $x = 3$ if $0.5 \leq U$.

```
sim.unif=runif(10000)
sim.x=sim.unif
sim.x[sim.unif<0.2]=1
sim.x[sim.unif<0.5 & sim.unif>=0.2]=2
sim.x[sim.unif>=0.5]=3
mean(sim.x)  #verify: the mean should be close to 2.3
```

```
## [1] 2.3028
```

2. There is a 2% click rate for an online advertisement. The number of people that will click on the advertisement among the next 100 people can be modeled by a binomial distribution.

   a. What are the parameters of the binomial distribution?

   b. Discuss the binomial model assumptions in the context of the problem.

   c. Use your computer to calculate, among the next 100 people: the probability that exactly two people click the advertisement; the probability that no one clicks the advertisement; the probability that 5 or more people click the advertisement.

      SOLUTION: We model it as a binomial distributuion with parameters $n = 100$ and $p = 0.02$. There are two assumptions. (1) We assume that each person clicks the

advertisement *independently* of each other; that is, one person's action does not affect the action of the next person. This assumption is reasonable if the 100 people are randomly sampled. We want to avoid situations like sampling a group of friends because, for instance, if one person clicks the advertisement, her friend might be more likely to also click it. (2) We also assume that each person has the *same probability* of clicking the advertisement. This assumption may be invalid because some populations (like certain age groups, certain regions, one gender, etc.) can have different probabilities of clicking the advertisement. It is therefore important to carefully define the population when conducting this type of research.

```
dbinom(2,size=100,prob=0.02)  #P(exactly two)
```

```
## [1] 0.2734139
```

```
dbinom(0,size=100,prob=0.02)  #P(no one)
```

```
## [1] 0.1326196
```

```
1-sum(dbinom(c(0:4),size=100,prob=0.02))  #P(five or more)
```

```
## [1] 0.05083045
```

3. The number of times you get a 6 if you roll a fair, six-sided die for five times can be modeled by a binomial distribution.

   a. What are the parameters of the binomial distribution?

   b. Discuss the binomial model assumptions in the context of the problem. If the die were biased, can we still use a binomial distribution?

   c. What is the probability that you get more than two 6's in five rolls?

   d. Design a simulation to verify your answer in c.

SOLUTION: We model it as a binomial distributuion with parameters $n = 5$ and $p = 1/6$. We assume (1) each roll is independent of each other and (2) the probability of rolling a 6 is the same for each roll. We can still use a binomial model if the die were biased, as long as the probability of rolling a 6 stays the same.

```
1-sum(dbinom(c(0,1,2),size=5, prob=1/6))  #P(more than two 6's)
```

```
## [1] 0.03549383
```

For verifying the above probability, sample algorithm and R code given below.

- Step 1: Randomly draw five numbers, each number can be $\{1, 2, \ldots, 6\}$ with equal probability; record the number of 6's in the five numbers as N^k;
- Step 2: Repeat for $k = 1, \ldots$ to the desired number of simulated runs.

To verify the probability of rolling more than two 6's, calculate the proportion in $\{N^k\}_{k=1,2,...}$ that are more than two.

```
nsim=10000
result.numberofsix=rep(NA,nsim) #store the number of 6's in each simulation
for(k in 1:nsim){
  onerun=sample(x=c(1:6),size=5,replace=T) #simulate five rolls
  result.numberofsix[k]=sum(onerun==6) #number of 6's in the five numbers
}
mean(result.numberofsix>2) #proportion of simulated games with more than two 6's
```

```
## [1] 0.0325
```

4. In a clinical trial of 20 patients, the probability that an individual patient will respond to a certain drug is $p = 0.25$. You want to use a binomial model to study the possible outcomes of the clinical trial.

   a. Simulate 1000 random draws from a binomial distribution with $n = 20$ and $p = 0.25$. What do the simulated numbers represent in this context?

   b. In real life, the response rate p is usually unknown and of interest to researchers. One way is to assign a random distribution on p (for example, using a model called a beta-binomial distribution). Run a simulation as follows: first, randomly sample a value $\tilde{p}$ uniformly from $(0, 1)$; second, randomly sample a value of X from a binomial distribution with $n = 20$ and $p = \tilde{p}$. Repeat for a large number of times. Use your simulation result to estimate the chance that fewer than 5 patients will respond to the drug in a clinical trial of 20 patients.

SOLUTION: a. Each simulated number corresponds to a simulated clinical trial of 20 patients, where the number represents how many patients repond to the certain drug.

```
mysim=rbinom(1000,size=20,p=0.25)
head(mysim)  #print out the first few simulated numbers
```

```
## [1] 5 4 8 2 4 5
```

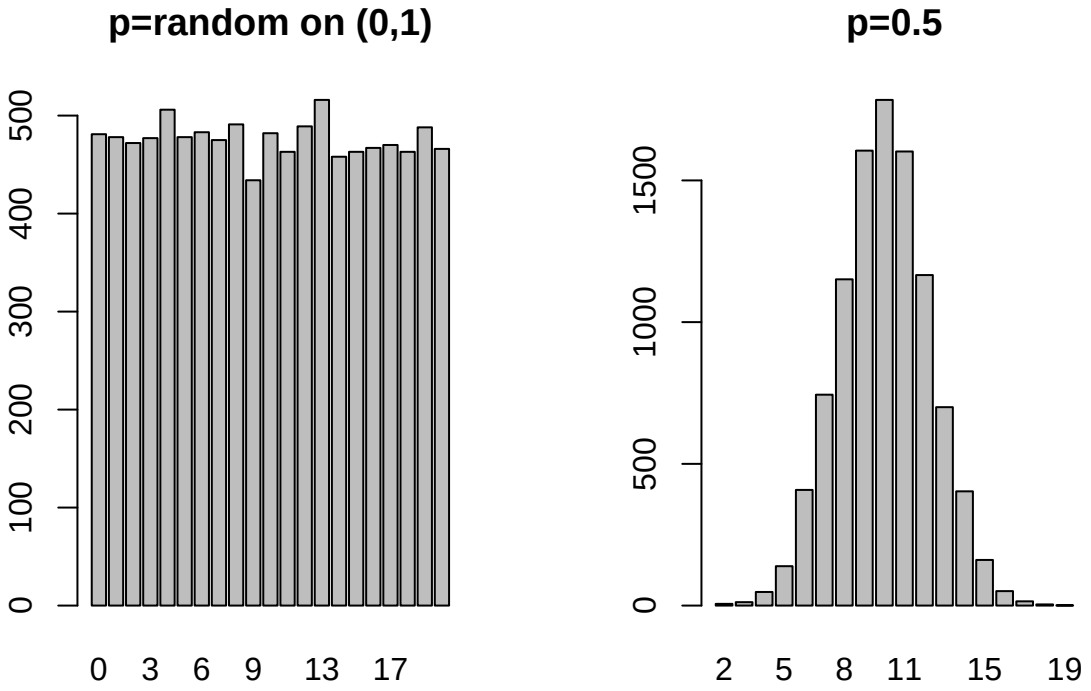
   b. Your answer should be close to $5/21 = 0.238$. Using the simulated result, it is also easy to see that under this model, the number of patients who respond to the drug is uniformly distributed on $\{0, 1, \ldots, 20\}$. In comparison, under a binomial model with $p = 0.5$, the number of patients who respond to the drug is more likely to be around 10. The two models have the same *expected* number of patients who respond to the drug (which is 10).

```
nsim=10000

#store the number of patients that respond to the drug in each simulated trial
result.patient=rep(NA,nsim)
```

```r
for(i in 1:nsim){
  ptilde=runif(1)
  result.patient[i]=rbinom(1,size=20,prob=ptilde)
}

mean(result.patient<5)
```

```
## [1] 0.2414
```

```r
#compare distribution with a binomial model
par(mfrow=c(1,2))
plot(as.factor(result.patient),main='p=random on (0,1)')
plot(as.factor(rbinom(nsim,size=20,prob=0.5)), main='p=0.5')
```



5. Read *Introduction to Probability*, page 96 to page 101 (stop at "Hypothesis Testing") for some additional explanations on the binomial distribution.

Solution NA.