**The Basic Statistical Problem**

We have a “large” collection of objects (population) and a feature of that population that we would like to know the value of.

**Example**: The population is the set of all parts coming off a production line. What percent of the parts produced are defective?

p = proportion of defective parts produced

The population is too large to test all of the parts produced. So, what can we do?

We take a random sample of 50 parts produced and test them.

= proportion of defective parts in the sample

We use as an estimate for the value of p. Is a good estimate for p? Maybe yes and maybe no. How likely is it that we get a that is a good estimate for p? That is a question for statistical/mathematical analysis. It requires an investigation into randomness and probability.

**Probability**

**Random Process/Experiment**

**Sample Space**

**Event**

**Probability**

**Random Variable**

**Discrete RV**

**Continuous RV**

**Probability for a Discrete Random Variable**

**Calculating Probabilities**

**Empirically**

**Using simulation (done by R)**

**Example1** Toss a fair coin 20 times. Let X = number of heads produced

**rflip(n)** is a an R procedure that simulates n tosses of a fair coin.

> rflip(20)

Flipping 20 coins [ Prob(Heads) = 0.5 ] ...

T T T T H H H H H H H H T H T T H T H T

Number of Heads: 11 [Proportion Heads: 0.55]

**do(k)** is an R procedure that causes a procedure to be done k times.

* do(5)\*rflip(20)

n heads tails prop

1 20 11 9 0.55

2 20 10 10 0.50

3 20 8 12 0.40

4 20 13 7 0.65

5 20 12 8 0.60

**We simulate 20 tosses of a fair coin 10000 times**

> tosses20<-do(10000)\*rflip(20)

> head(tosses20)

n heads tails prop

1 20 8 12 0.40

2 20 13 7 0.65

3 20 10 10 0.50

4 20 8 12 0.40

5 20 12 8 0.60

6 20 12 8 0.60

**We have R count the number of times each value from 0 to 20 is produced by our 10000 simulations**

> tally(~heads,data=tosses20)

1 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

1 9 45 150 379 737 1162 1632 1762 1562 1278 759 341 133 39 10

20

1

> tally(~heads,data=tosses20,format="prop")

1 3 4 5 6 7 8 9 10 11 12

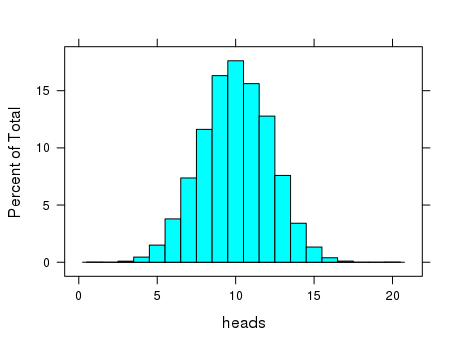
0.0001 0.0009 0.0045 0.0150 0.0379 0.0737 0.1162 0.1632 0.1762 0.1562 0.1278

13 14 15 16 17 20

0.0759 0.0341 0.0133 0.0039 0.0010 0.0001

**We have R plot a histogram for the number of heads**

> histogram(~heads,data=tosses20,type="percent",width =1)

****

**The estimated probability distribution for X = number of heads in 20 tosses of a fair coin**

x 1 3 4 5 6 7 8 9 10 11 12

P(X = x) 0.0001 0.0009 0.0045 0.0150 0.0379 0.0737 0.1162 0.1632 0.1762 0.1562 0.1278

13 14 15 16 17 20

0.0759 0.0341 0.0133 0.0039 0.0010 0.0001

**Example 2** Toss a biased coin 10 times and let X = number of heads. Assume the biased coin has a probability of .25 of producing a head**.**

**We simulate 10 tosses of the biased coin.**

> rflip(10,prob=.25)

Flipping 10 coins [ Prob(Heads) = 0.25 ] ...

T T T H T T H H T T

Number of Heads: 3 [Proportion Heads: 0.3]

**We simulate 10000 performances, find the estimated probability distribution for X, and the histogram.**

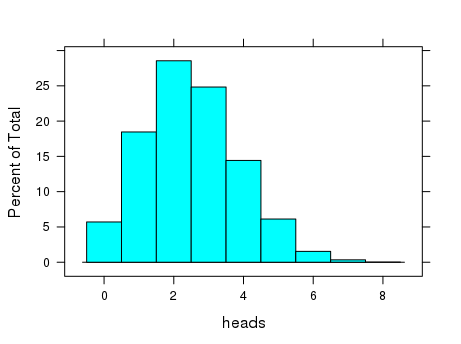
> tosses10<-do(10000)\*rflip(10,prob=.25)

> tally(~heads,data=tosses10, format ="prop")

0 1 2 3 4 5 6 7 8

0.0571 0.1846 0.2855 0.2483 0.1443 0.0612 0.0154 0.0034 0.0002

* histogram(~heads,data=tosses10,type="percent", width = 1)



**What is (approximately) the probability of getting at least 5 heads?**

**Example 3** Toss of a pair of fair dice

X = sum of the results

**1:n produces the integers 1,2,…,n**

> 1:10

[1] 1 2 3 4 5 6 7 8 9 10

**We can generate the result of tossing a fair die 10 times**

> resample(1:6,10)

[1] 5 1 6 5 3 2 2 2 3 1

**We now simulate 10000 tosses of a fair die twice.**

> die1<-resample(1:6,10000)

> die2<-resample(1:6,10000)

> head(die1)

[1] 5 1 3 3 1 5

**Next, we create the sum of the results of die1 and die2.**

> s<-die1+die2

> head(s)

[1] 7 2 5 8 3 8

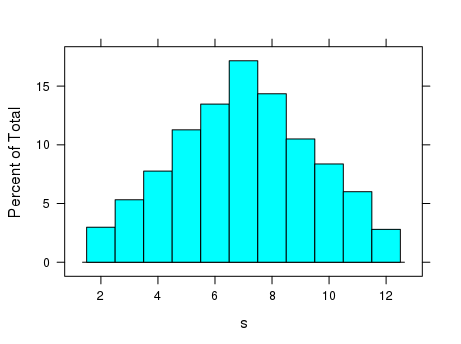
**We create the estimated probability distribution for X and the histogram.**

> tally(~s,format="prop")

2 3 4 5 6 7 8 9 10 11 12

0.0298 0.0532 0.0776 0.1128 0.1347 0.1716 0.1435 0.1050 0.0837 0.0601 0.0280

* histogram(~s,type="percent",width = 1)

****

**Estimate the probability that the sum of dice is more than 9.**

**Example 4** A pair of fair dice are tossed. Use 10000 simulations to estimate the probability that the dice turn up with the same value.

> die1<-resample(1:6,10000)

> die2<-resample(1:6,10000)

> head(die1)

[1] 5 1 3 3 1 5

> head(die2)

[1] 2 1 2 5 2 3

> head(die1==die2)

[1] FALSE TRUE FALSE FALSE FALSE FALSE

> tally(die1==die2,format = "prop")

TRUE FALSE

0.1647 0.8353

**Example 5** John has a free throw shooting average of 85%. He has been fouled in the act of shooting a 3-point shot as time runs out. John has 3 free throws coming. His team is behind by 2 points. Estimate the probability that his team does not lose in regulation time. Use 10000 simulations.

**Exercises 3**

1. a. A fair coin is tossed 8 times. Use 10,000 simulations to estimate the probability distribution for the random variable X = number of heads.

b. Using your result in (a), estimate the probability that at least 5 heads are produced.

c. Repeat (a) and (b) for a biased coin where the probability of a head is .66.

2. Three fair dice are tossed. Let X be the sum of the values produced.

a. Use 10,000 simulations to estimate the probability distribution of X.

b. Use your result in (a) to estimate the probability that the sum is less than 10.

3. Carol has a batting average in softball of .440. If she has 5 at bats in a game, what is the probability that she gets exactly 2 hits? What is the probability that she gets at least 2 hits? Simulate her performance in 10000 games to estimate these probabilities.

4. An ordinary deck of playing cards contains 52 cards with 4 suits of 13 cards each. Three cards are dealt from a well-shuffled deck. What is the probability that all three are spades? Estimate this probability by simulating this activity 10000 times. You can model the 52 cards as the integers 1-52, where the values 1-13 are the spades. The **resample()** procedure in R allows the same value to be produced more than once. The **sample()** procedure produces values that are different, so you will have to use sample() to simulate the dealing of the three cards.

> ex<-sample(1:52,3)

> ex

[1] 21 48 20

To check whether a card is a spade (value <14)

> ex<14

[1] FALSE FALSE FALSE

To count the number of cards that are spades (this converts the truth values to integers 1 = TRUE and 0 = FALSE and adds the values)

|  |
| --- |
| > sum(ex<14)  [1] 0  Use **do()** to repeat this 10000 times and store the results.   * sim<-do(10000)\*sum(sample(1:52,3)<14)   Use **head(sim)** to see what this data frame looks like.  > head(sim)  Finally, use **tally** to find the appropriate proportions.  > tally(~sim,format="prop")  sim  0 1 2 3  0.4127 0.4306 0.1412 0.0155 |
| **Your Exercise for (4):**  Use 10000 simulations to estimate the probability that in a five card hand, at least four of the cards  are red (either a heart or a diamond). Include the R-commands you used and the outputs. |