Probabilistic Modeling and Statistical Computing Fall 2015

September 2, 2015

G. W. Leibniz on Computing, 1685

It is unworthy of excellent men to lose hours like slaves in the labour of calculation which could safely be relegated to anyone else if machines were used.

D. Bernoulli on Probabilistic Models, 1713

We define the art of conjecture, or stochastic art, as the art of evaluating as exactly as possible the probabilities of things, so that in our judgments and actions we can always base ourselves on what has been found to be the best, the most appropriate, the most certain, the best advised; this is the only object of the wisdom of the philosopher and the prudence of the statesman.

A. De Moivre on Statistics, 1718

And thus in all cases it will be found, that although Chance produces irregularities, still the Odds will be infinitely great, that in the process of Time, those irregularities will bear no proportion to the recurrence of that Order which natural results from Original Design.

Contents of the Course

- Probabilistic modeling: about four weeks
- Basic statistical methods: about four weeks
- Some advanced topics: about four weeks
- Using R (simulation, statistical computerization, probability calculator)

Connections to Other Courses

- All further **Probability** and **Statistics** courses
- **Optimization** provides methods
- **Statistical Learning Theory** uses tools from this course and goes beyond them
- **Simulation** extends some methods much further
- **Regression Analysis** extends some methods much further

How to Succeed in This Course

- Do weekly homework sets (30 %)
- Come to all classes and participate (10 %)
- Midterm test 60 minutes with in-class and take-home components (20 %)
- Comprehensive final exam with in-class and take-home components(40 %)
- *Work hard, work with others, stay in touch*
- *Do practice exercises, keep your R skills sharp, visit the Blackboard site*

A Game of Skill

- Try to toss the ball into the bucket until you have succeeded five times.
- Record your progress.
- See you in 10 minutes!

Let's Hear How You Did

- Who succeeded five times?
- How did it happen? Which details did you record?
- Which details are possibly important?

What can we predict?

- Success in your next toss?
- At least one success in your next 10 tosses?
- Time until your next success?
- How can you improve your prediction?

A Probabilistic Model

... *that is good enough for simulation*

What does each assumption mean?
How could it be violated?

- Each toss can either succeed or not. Success is considered a random event.
- Successive tosses are independent. You cannot learn about the success of the next toss from the past - or from the future.
- Success happen with a fixed probability p.

A first round of data reduction = forgetting inessential things

Implement this in R

Write a function that outputs a 1 (success, probability p) or a 0 (fail, probability $1 - p$).

Use random function `runif()`. It produces a uniformly distributed random $u \in (0, 1)$.

```
mytoss = function(p){
    u <- runif(1)
    x <- as.numeric(u < p)
    return(x)
}
```

A more complicated simulation

A function that simulates tosses until the first success and returns the number of attempts.

```
myattempts = function(p){
    counter <- 1
    while (mytoss(p) == 0){
       counter <- counter + 1
    }
    return(counter)
}
```

Assessing the simulation tools

Does this work?

Performance?

Why is the `myattempts` simulation problematic?

How about your personal p?

Shauna recorded successes in tosses # 3, # 15, # 16, # 20, # 31. What can we say about her success rate p?

Do you think that $p > .5$?

What range of p's is compatible with the data?

Can we get some answers with a simulation?

Simulation

Do one or several simulations for a range of p.

Check for which p simulated results are similar to the observed results.

How should we record the results of the simulation?

How should we assess "similarity"?

Simulation

Do one or several simulations for a range of p.

Check for which p simulated results are similar to the observed results.

How should we record the results of the simulation?

How should we assess "similarity"?

Simulation

Do one or several simulations for a range of p.

Check for which p simulated results are similar to the observed results.

How should we record the results of the simulation?

How should we assess "similarity"?

A bit of theory.

Probability model for a single attempt:

The first success occurs in toss j with probability

$$\boxed{p_j = p(1-p)^{j-1}}$$

for $j = 1, 2, 3, \ldots$.

Why?

What is the name of this distribution?

A bit of theory.

Probability model for a single attempt:

The first success occurs in toss j with probability

$$\boxed{p_j = p(1-p)^{j-1}}$$

for $j = 1, 2, 3, \ldots$.

Why?

What is the name of this distribution?

Expected value

Let X be the random variable "number of toss in which the first success occurs".

$$\mathcal{E}(X) = \sum_{j=1}^{\infty} j \cdot P(X = j) = \sum_{j=1}^{\infty} jp(1-p)^{j-1} = ???$$

Ask Mathematica!

$$\mathcal{E}(X) = \frac{1}{p} \quad \text{or} \quad p = \frac{1}{\mathcal{E}(X)}$$

Expected value

Let X be the random variable "number of toss in which the first success occurs".

$$\mathcal{E}(X) = \sum_{j=1}^{\infty} j \cdot P(X = j) = \sum_{j=1}^{\infty} jp(1-p)^{j-1} = ???$$

Ask Mathematica!

$$\boxed{\mathcal{E}(X) = \frac{1}{p} \quad \text{or} \quad p = \frac{1}{\mathcal{E}(X)}}$$

Expected value II

Fix k. Let Y be the number of tosses until k successes. Then

$$Y = X_1 + X_2 + \ldots X_k$$

where the X_i are independent realizations of X (count until for success). **Why? So what is Y?**

Therefore

$$\mathcal{E}(Y) = \mathcal{E}(X_1) + \mathcal{E}(X_2) + \cdots + \mathcal{E}(X_k) = \frac{k}{p}$$

Does this make sense? Say it in words!

Expected value II

Fix k. Let Y be the number of tosses until k successes. Then

$$Y = X_1 + X_2 + \ldots X_k$$

where the X_i are independent realizations of X (count until for success). **Why? So what is Y?**

Therefore

$$\boxed{\mathcal{E}(Y) = \mathcal{E}(X_1) + \mathcal{E}(X_2) + \cdots + \mathcal{E}(X_k) = \frac{k}{p}}$$

Does this make sense? Say it in words!

Practice Questions

- You roll two dice until you get two 6's. How many rolls on average until you succeed?
- About 0.4% of the phones that you are selling are defective. How many sales until you have had 10 complaints?
- About 1.3% of the visitors to your website click through on the ad at the top. How many visitors until 1,000 click-throughs?

Plug-In Estimator

We know that $\boxed{\mathcal{E}(Y) = \frac{k}{p} \quad or \quad p = \frac{k}{\mathcal{E}(Y)}}$

We can observe Y from the data.

Shauna recorded successes in tosses # 3, # 15, # 16, # 20, # 31.

Thus $Y = 31$ was observed. Our best guess for $\mathcal{E}(Y)$ is $y = 31$. Use this to estimate p.

$$\boxed{\hat{p} = \frac{k}{y} = \frac{5}{31} \approx .1613}$$

Plug-In Estimator

We know that $\boxed{\mathcal{E}(Y) = \frac{k}{p} \quad or \quad p = \frac{k}{\mathcal{E}(Y)}}$

We can observe Y from the data.

Shauna recorded successes in tosses # 3, # 15, # 16, # 20, # 31.

Thus $Y = 31$ was observed. Our best guess for $\mathcal{E}(Y)$ is $y = 31$. Use this to estimate p.

$$\boxed{\hat{p} = \frac{k}{y} = \frac{5}{31} \approx .1613}$$

Data reduction II

Suggests that we only need to record the total number of tosses until k successes. We do not need to record the intermediate results.

Try to reach k successes, and repeat this n times. Observe $y_1, y_2, \ldots y_n$ tosses in these n rounds.

Estimate $\mathcal{E}(Y) \approx \bar{y} = \frac{y_1+y_2+\cdots+y_n}{n}$.

Plug in estimator

$$\boxed{\hat{p} = \frac{k}{\bar{y}}}$$

Data reduction II

Suggests that we only need to record the total number of tosses until k successes. We do not need to record the intermediate results.

Try to reach k successes, and repeat this n times. Observe $y_1, y_2, \ldots y_n$ tosses in these n rounds.

Estimate $\mathcal{E}(Y) \approx \bar{y} = \frac{y_1+y_2+\cdots+y_n}{n}$.

Plug in estimator

$$\hat{p} = \frac{k}{\bar{y}}$$

Data reduction II

Suggests that we only need to record the total number of tosses until k successes. We do not need to record the intermediate results.

Try to reach k successes, and repeat this n times. Observe $y_1, y_2, \ldots y_n$ tosses in these n rounds.

Estimate $\mathcal{E}(Y) \approx \bar{y} = \frac{y_1+y_2+\cdots+y_n}{n}$.

Plug in estimator

$$\boxed{\hat{p} = \frac{k}{\bar{y}}}$$

Practice Questions

- Rick got to 5 successes four times. Here are his results:
  $y_1 = 6, y_2 = 16, y_3 = 8, y_4 = 8$. What is $\hat{p}$?
- Rick therefore got to 1 success 20 times. Here are his results:
  $1, 2, 1, 1, 1, 3, 2, 6, 1, 4, 3, 1, 1, 1, 2, 2, 2, 1, 1, 2$.
  What is $\hat{p}$?
- Rick therefore got to 20 successes once. It took him 38 tosses. What is $\hat{p}$?
- **Explain and generalize.**

Assess this estimator

Bias? *Is there a systematic error?*

Use a simulation

Pick a p. Simulate many Y's. Compute $\hat{p} = \frac{k}{Y}$ in each case. Check if the estimates average to the original p.

Spread? *What is the range of this estimate?*

Use a simulation

Pick a p. Simulate many Y's. Compute $\hat{p} = \frac{k}{Y}$ in each case. Make a histogram or box plot.

Theoretical Bias

The quantity $\hat{p}$ is a random variable. Then

$$\mathcal{E}(\hat{p}) - p$$

is the **bias**, i.e. "average" deviation of the estimate from the true value.

For $k = 1$ and a single round, we use $\hat{p} = \frac{1}{Y}$, thus

$$\mathcal{E}(\hat{p}) = \mathcal{E}\left(\frac{1}{Y}\right) = \sum_{j=1}^{\infty} \frac{p(1-p)^{j-1}}{j} = \frac{-p \log p}{1-p}$$

Theoretical Bias II

This can be done for any k. For example, if $k = 5$

$$\mathcal{E}(\hat{p}) = \frac{10p^5 \log(p)}{(p-1)^5} - \frac{25p^3 - 23p^2 + 13p - 3}{12(p-1)^4}$$

This is definitely not p.

So there is a nonzero bias!

Bias Plots

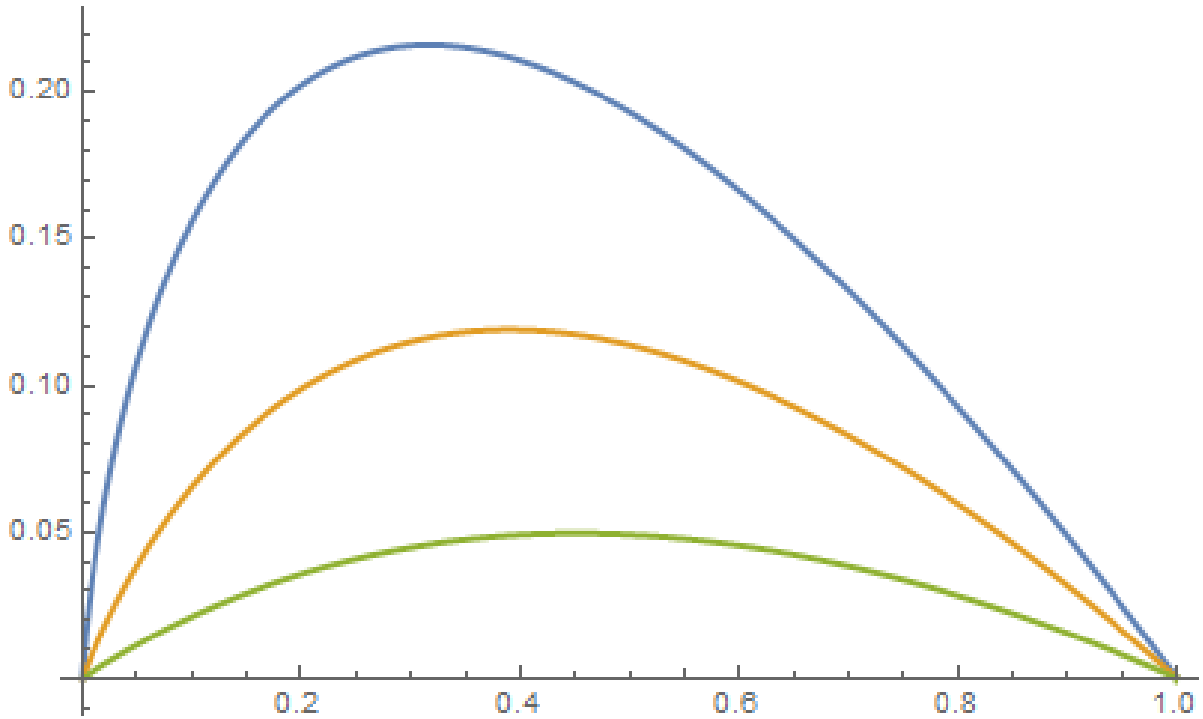

Figure : Bias $\mathcal{E}(\hat{p}) - p$ for plug-in estimator, for $k = 1, 2, 5$.

Questions?

Practice Questions

Remember Shauna? It took her 31 tosses to get to 5 successes.

Vivian argues as follows:

Shauna had 5 successes in 31 trials. I learned in my statistics course that the best estimate for p is $\hat{p} = \frac{5}{31}$ and that this estimate is **unbiased**. And Wikipedia confirms this.

So is there a bias or not? Reconcile our results with what Vivian says.