Independent Trials and Special Distributions

Gregory M. Shinault

Goal for this Lecture

Cover the idea of independent trials, and the associated special discrete distributions. Topics covered will include

1. the definition of independent trials and the Bernoulli distribution,
2. the definition and application of the Binomial distribution, and
3. the definition and application of the Geometric distribution.

The material of this lecture roughly corresponds to Section 2.4 of the textbook.

Independent Trials

Big Idea

Repeat a random experiment multiple times, in such a way that the experiments do not affect one another.

Suppose the independent trials we conduct only have two possible outcomes, which we call success and failure. The probability of success is p (thus the failure probability is $q = 1 - p$). We call this a Bernoulli(p) trial.

Bernoulli Distribution

Definition: We say a random variable X has the Bernoulli(p) distribution if

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p. \end{cases}$$

We denote this as $X \sim \text{Ber}(p)$. The success probability p is the only parameter of this distribution.

Equivalently, $X \sim \text{Ber}(p)$ if and only if the probability mass function is

$$p_X(k) = p^k(1-p)^{1-k}$$

for $k = 0, 1$.

Binomial Distribution

Introduction

Suppose we are conducting an early clinical drug trial. The medicine we are testing was effective in treating insomnia in 80% of lab rats. The trial is only on 8 human patients. What is the probability at least 6 patients have a positive response to the drug?

Definition

A random variable X is said to have a Binomial(n, p) distribution if it has the probability mass function

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 0, 1, 2, \ldots, n$. We denote this by $X \sim \text{Bin}(n, p)$.

The success probability p and number of trials n are the parameters of the Binomial distribution.

Connection to Bernoulli RVs

Suppose $X \sim \text{Bin}(n, p)$.

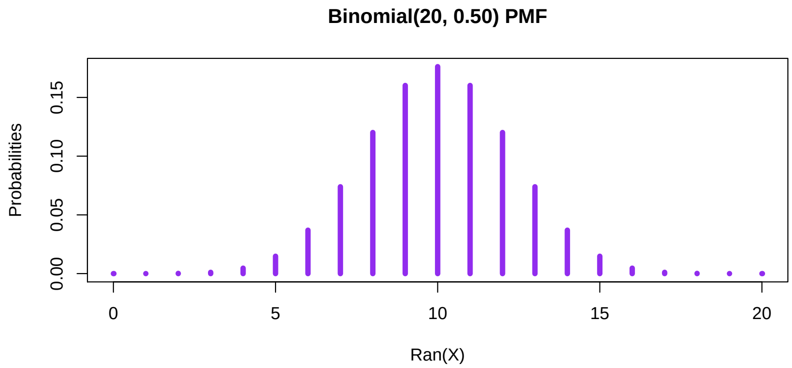
Let $X_1, X_2, \ldots, X_n$ be the results of the n trials. So $X_k = 1$ if the k-th trial was a success and $X_k = 0$ if it failed.

Then

1. $X_1, X_2, \ldots, X_n$ are all independent Bernoulli(p) RVs.
2. $X = X_1 + X_2 + \cdots + X_n$.

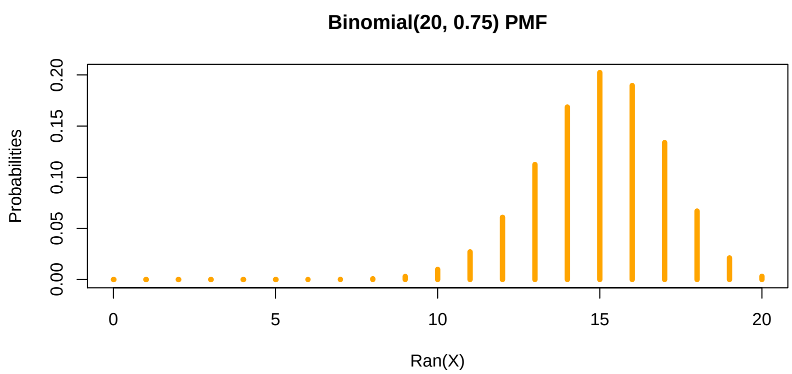
Probability Mass Function

```
PMF1 <- dbinom(0:20, size = 20, prob = 0.5)
kVals <- 0:20
plot(kVals, PMF1, type = "h", col = "purple2",
    lwd = 5, main = "Binomial(20, 0.50) PMF",
    xlab = "Ran(X)", ylab = "Probabilities")
```



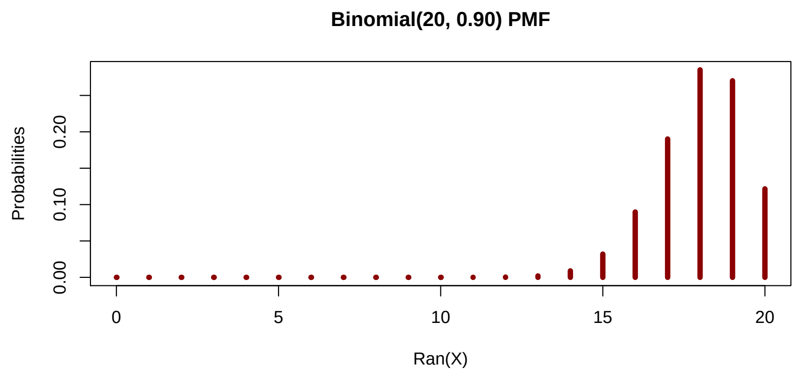
Probability Mass Function

```
PMF2 <- dbinom(0:20, size = 20, prob = 0.75)
kVals <- 0:20
plot(kVals, PMF2, type = "h", col = "orange",
    lwd = 5, main = "Binomial(20, 0.75) PMF",
    xlab = "Ran(X)", ylab = "Probabilities")
```

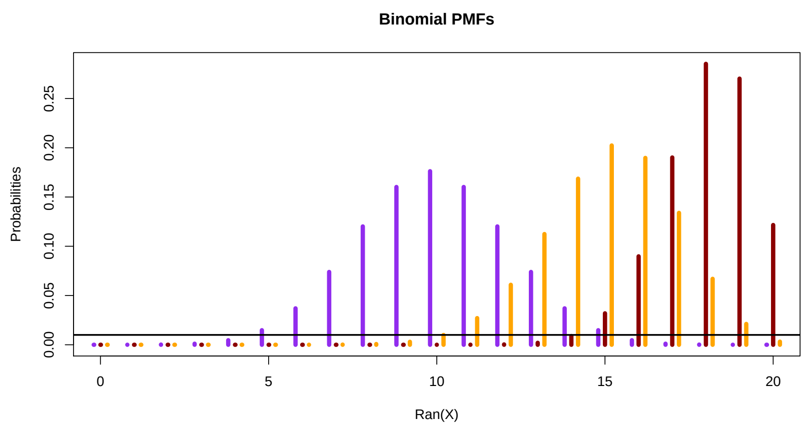


Probability Mass Function

```
PMF3 <- dbinom(0:20, size = 20, prob = 0.9)
kVals <- 0:20
plot(kVals, PMF3, type = "h", col = "darkred",
    lwd = 5, main = "Binomial(20, 0.90) PMF",
    xlab = "Ran(X)", ylab = "Probabilities")
```



Binomial Distribution



Example

You are conducting a medical trial with 80 patients of a drug that was 75% effective in treating lung cancer in mice. Let X be the number of patients that respond positively to treatment. Assuming this level of effectiveness holds for humans, what is $\mathbb{P}(X \geq 60)$?

Computation in R

The following all yield the same answer.

```
sum(dbinom(60:80, size = 80, prob = 0.75))
```

```
## [1] 0.5597063
```

```
pbinom(60 - 1, size = 80, prob = 0.75, lower.tail = FALSE)
```

```
## [1] 0.5597063
```

```
1 - pbinom(60 - 1, size = 80, prob = 0.75)
```

```
## [1] 0.5597063
```

Example

You ask 10 people if the prefer raspberry or blueberry cobbler. Raspberry is obviously way better so 70% of the entire population prefers raspberry cobbler.

- What is the probability 6 people prefer raspberry?
- Given at least 3 people prefer raspberry, what is the probability exactly 6 people prefer raspberry?
- Given the first three people you ask prefer raspberry, what is the probability that exactly 6 people prefer raspberry?

Computation in R

```
dbinom(6, size = 10, prob = 0.7)
```

```
## [1] 0.2001209
```

```
dbinom(6, size = 10, prob = 0.7)/pbinom(3 - 1,
    size = 10, prob = 0.7, lower.tail = FALSE)
```

```
## [1] 0.2004397
```

```
dbinom(3, size = 7, prob = 0.7)
```

```
## [1] 0.0972405
```

Applied Example

The Survey of Professional Forecasters is conducted by the Federal Reserve. Predictions are collected from economists and data is compiled to form a 90% confidence interval for GDP growth rate. Collect the data for the past 20 years for the confidence interval (predicted growth rate) and the actual growth rate. How well did the confidence intervals match the actual growth rate? What is the probability a set of true 90% confidence intervals would perform this way?

This analysis can be read in *The Signal and the Noise* by Nate Silver. It is an engaging book that tells stories involving the use of Bayesian statistics. This book is the most enjoyable read to accompany this course, or the website `http://fivethirtyeight.com`.

Computation in R

```
dbinom(12, size = 18, prob = 0.9)

## [1] 0.005243022

pbinom(12, size = 18, prob = 0.9)

## [1] 0.00641515
```

Hypothesis Test in R

```
binom.test(12, n = 18, p = 0.9)

##
##  Exact binomial test
##
## data:  12 and 18
## number of successes = 12, number of
## trials = 18, p-value = 0.006415
## alternative hypothesis: true probability of success is not equal to 0.9
## 95 percent confidence interval:
##  0.4099252 0.8665726
## sample estimates:
## probability of success
##              0.6666667
```

Geometric Distribution

Introduction

Conduct independent Bernoulli trials. Stop when you have the first success. What is the probability you conduct k trials?

Definition

A random variable X is said to have a Geometric(p) distribution if it has the probability mass function

$$p_X(k) = (1-p)^{k-1}p$$

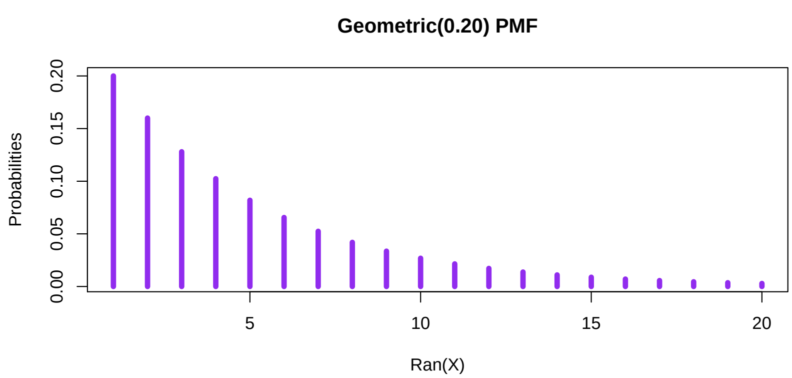
for $k = 1, 2, 3, \ldots$. We denote this by $X \sim \text{Geo}(p)$.

The success probability p is the only parameter in the Geometric distribution.

WARNING: Some sources use a slightly different convention, counting the number of failed trials.

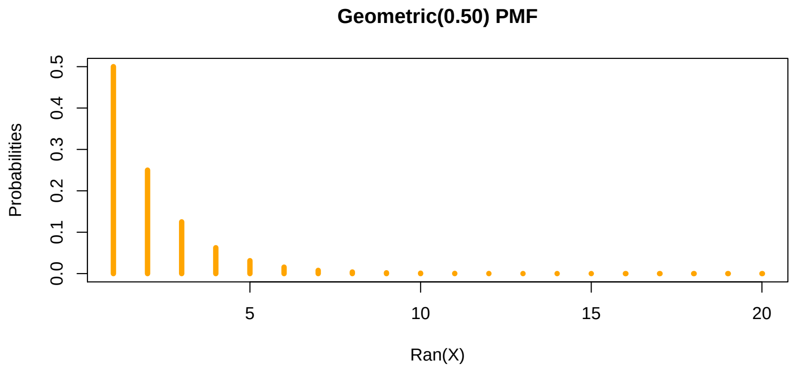
Probability Mass Function

```
GPMF1 <- dgeom(0:19, prob = 0.2)
kVals <- 1:20
plot(kVals, GPMF1, type = "h", col = "purple2",
    lwd = 5, main = "Geometric(0.20) PMF", xlab = "Ran(X)",
    ylab = "Probabilities")
```



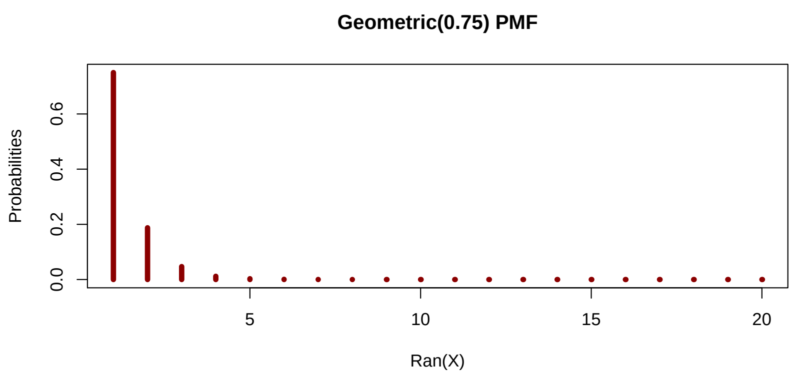
Probability Mass Function

```
GPMF2 <- dgeom(0:19, prob = 0.5)
kVals <- 1:20
plot(kVals, GPMF2, type = "h", col = "orange",
    lwd = 5, main = "Geometric(0.50) PMF", xlab = "Ran(X)",
    ylab = "Probabilities")
```

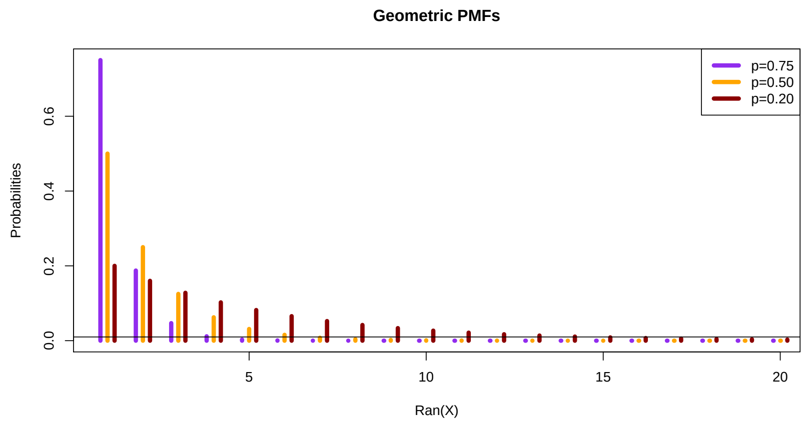


Probability Mass Function

```
GPMF3 <- dgeom(0:19, prob = 0.75)
kVals <- 1:20
plot(kVals, GPMF3, type = "h", col = "darkred",
    lwd = 5, main = "Geometric(0.75) PMF", xlab = "Ran(X)",
    ylab = "Probabilities")
```



Geometric Distribution



Example

We have a bin with 6 purple marbles and 2 orange marbles in it. We draw 4 marbles without replacement from the bin. If there are exactly 3 purple marbles in our sample we stop. If not we put the marbles back in the bin and repeat this procedure. Let X be the number of times we do not draw 3 purples. What is the probability mass function of X? What is $\mathbb{P}(X \leq 5)$?

Example

Suppose X is a Geometric(p) RV. Find an expression for $\mathbb{P}(X > k)$ in two ways:

- Use $\mathbb{P}(X > k) = \sum_{\ell=k+1}^{\infty} \mathbb{P}(X = \ell)$ and manipulate the series.
- Make an argument based upon the meaning of $\{X > k\}$ in terms of independent trials.

The Wrap Up

Summary

1. Independent trials give us ways to construct interesting random variables with special distributions (we will see at least one more).
2. The binomial disribution gives probabilities for the number of successes in a fixed number of trials.

3. The geometric distribution gives probabilities for the number of trials until the first success.

Next Step

We are done with the basics of conditional probability. Now we just have to cover a few odds and ends to finish the topic.