Probability Distributions

Jerome Dumortier

07 October 2024

Lecture Overview

Random variables

- Probability distributions
- Expected value (mean) and variance

Discrete distributions

- Bernoulli
- Binomial
- Poisson

Continuous distributions

- Uniform
- Normal
- t/Student

Random Variables

Overview

A random variable is a variable whose value depends on chance

- Number of heads from flipping a coin 20 times
- Number after rolling a die
- Number of passengers showing up to a flight

Discrete random variables

- A random variable X is discrete if it can assume only a finite or countable infinite number of distinct values

Continuous random variables

- Can take an infinite number of values

Discrete versus continuous random variables

Discrete random variables

- Number of students in a class
- Number of children in a family
- Number of calls to a 911 dispatcher within a 24 hour period

Continuous random variables

- Temperature in a week from today
- Value of the S&P 500
- Average height of IUPUI students

It is sometimes easier to assume continuity even if the variable seems discrete, e.g., home values in Indianapolis.

Examples of Random Variables

Simulate rolling a die 600 times and the random variable being the number of sixes.

```
dierolls = sample(1:6,size=600,replace=TRUE)
table(dierolls)
```

```
## dierolls
##   1   2   3   4   5   6
##  94 112 107 100  88  99
```

Random Variables and Probability Distribution

A probability distribution is a combination of outcomes of a random variable and associated probabilities. For example, let the random variable X be the number of heads from flipping a coin seven times:

X	0	1	2	3	4	5	6	7
$Pr(X)$	0.01	0.05	0.16	0.27	0.27	0.16	0.05	0.01

The sum of all the probabilities associated with the mutually exclusive outcomes is equal to 1.

Expected Value and Variance

Definition

Think of the expected value as a weighted average. If X is a discrete random variable then the expected value of X, i.e., $E(X)$, is written as

$$E(X) = \sum_i x_i \cdot Pr(X = x_i)$$

If X is a continuous random variable, then calculus is needed to calculate the expected value and those details are in the lecture notes. The variance can be calculated as follows:

$$Var(X) = E(X - E(X))^2 = E(X^2) - E(X)^2$$

Both equations give you the variance. Sometimes one of the equations is more convenient to use. Note that $E(X^2) \neq E(X)^2$.

Example Setup

Suppose you are working for a car dealership. For the last year, you calculated the number of cars sold per day and came up with the following probability distribution:

X	0	1	2	3	4	5
$Pr(X)$	0.10	0.15	0.15	0.30	0.25	0.05

Example Calculations

x_i	$Pr(x_i)$	$x_i \cdot Pr(x_i)$	$x_i - \mu$	$(x_i - \mu)^2$	$Pr(x_i) \cdot (x_i - \mu)^2$
0	0.10	0.00	-2.60	6.76	0.68
1	0.15	0.15	-1.60	2.56	0.38
2	0.15	0.30	-0.60	0.36	0.05
3	0.30	0.90	0.40	0.16	0.05
4	0.25	1.00	1.40	1.96	0.49
5	0.05	0.25	2.40	5.76	0.29
Sum		2.60			1.94

Hence $Var(X) = 1.94$ and $\sigma = \sqrt{1.94} = 1.393$.

Discrete Probability Distributions

Bernoulli Distribution

Characteristics of the Bernoulli distribution:

- Simplest discrete probability distribution
- Two outcomes: "Success" and "Failure"
- One parameter: p

Probability mass function:

$$Pr(X = 1) = p$$

And thus we also have $Pr(X = 0) = 1 - p$.

Geometric Distribution: Definition

The geometric distribution models the number of trials until the first success in a series of Bernoulli trials.

- How many trials are needed to get the first success?
- Each trial has two outcomes: "Success" (with probability p) or "Failure" (with probability $1-p$).
- The trials are independent and thus, probability of success remains constant across trials.

Mathematical equation:

$$P(X=k)=(1-p)^{k-1}\cdot p$$

where X is the number of trials until the first success, p is the probability of success on each trial, and k is the trial number where the first success occurs.

Geometric Distribution: Example

Assume $p = 0.2$, the we have the following:

$$P(X = 3) = (1 - 0.2)^{3-1} \cdot 0.2 = (0.8)^2 \cdot 0.2 = 0.128$$

There is a 12.8% chance the first success occurs on the third trial. Expected value (i.e., average number of trials needed to get the first success) and variance:

$$E(X) = \frac{1}{p}$$

$$Var(X) = \frac{1 - p}{p^2}$$

Geometric Distribution: R

If $p = 0.2$, then we have the following:

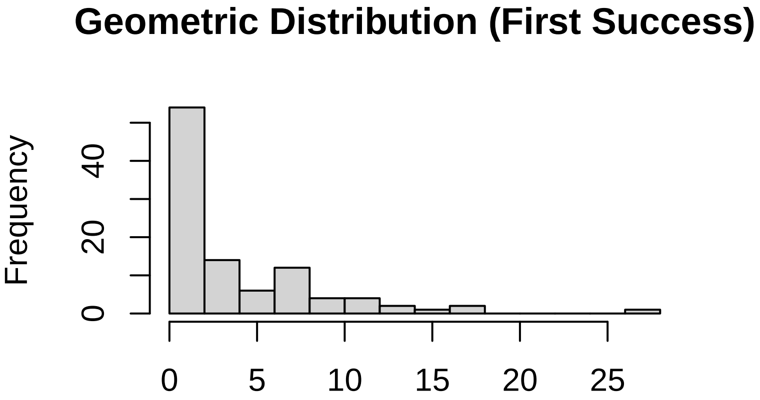
$$E(X) = \frac{1}{0.2} = 5$$

On average, 5 trials for the first success.

```
geometrictrials = rgeom(100,prob=0.2)
```

Geometric Distribution: Histogram

```
hist(geometrictrials,main="Geometric Distribution (First Success)",
     xlab="Number of Attempts",breaks=10)
```



Binomial Distribution: Definition

Characteristics of the Binomial distribution:

- Closely related to the Bernoulli Distribution
- "Repeated" Bernoulli outcomes
- Two parameters: n and p
- k number of success

Probability mass function:

$$Pr(X = k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}$$

The mean is $\mu = n \cdot p$.

Binomial: Use

When is the Binomial Distribution appropriate? A situation must meet the following conditions for a random variable X to have a binomial distribution:

- You have a fixed number of trials involving a random process; let n be the number of trials.
- You can classify the outcome of each trial into one of two groups: success or failure.
- The probability of success is the same for each trial. Let p be the probability of success, which means $1 - p$ is the probability of failure.
- The trials are independent, meaning the outcome of one trial does not influence the outcome of any other trial.

Binomial: Example I

Suppose you didn't study for a multiple choice exam. There are 10 questions with five possible answers each. Only one answer per question is correct. What is the probability that you get 6 correct answers?

$$Pr(X = k) = \frac{10!}{6! \cdot (10-6)!} \cdot 0.2^6 \cdot (1-0.2)^{10-6}$$

Or simply in R:

```
dbinom(6,10,0.2)
```

```
## [1] 0.005505024
```

Binomial in R: Probability Density Function

The probability density function (PDF) for the binomial distribution in R is written as `dbinom(x,n,p)`. Consider the following probabilities:

- Probability of 9 heads ($x = 9$) from 16 coin flips ($n = 16$)
- Probability of 0 to 16 heads from 16 coin flips

```
dbinom(9,16,0.5)
dbinom(0:16,16,0.5)
```

Binomial in R: Cumulative density function

The cumulative density function (CDF) for the binomial distribution in R is written as `pbinom(x,n,p)`. Consider the following probabilities:

- Probability of getting up to three heads from flipping a coin ten times
- Cumulative probabilities for getting 0 through 10 heads

```
pbinom(3,10,0.5)
pbinom(0:10,10,0.5)
```

Binomial: Example II

Suppose that 85% of Hoosiers are wearing a seat belt. You are a police officer and pulling over 20 cars. What is the probability that at least (!) 15 people are wearing a seat belt?

```
1-pbinom(14,20,0.85)
```

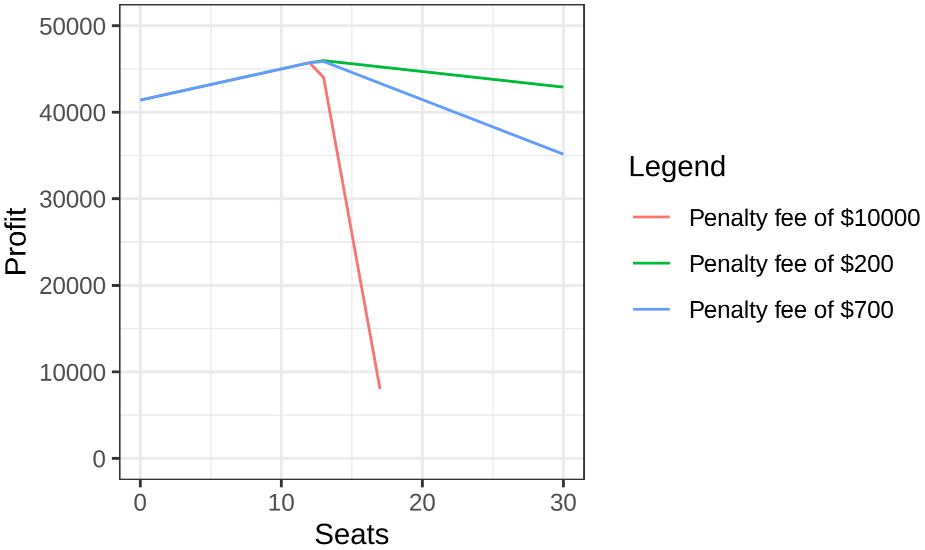
```
## [1] 0.932692
```

While using the binomial distribution, be very careful on how to interpret the results. The probability of at least 15 people wearing a seatbelt means that you are interested in the cumulative probability of 15, 16, 17, 18, 19, and 20 people wearing a seat belt. That probability is 0.933.

Binomial: Overbooking Flights I

The binomial distribution can be used to analyze the issue of overbooking. Assume that an airline as a plane with a seating capacity of 115. The ticket price for each traveler is \$400. The airline can overbook the flight, i.e., selling more than 115 tickets, but has to pay \$700 in case a person has a valid ticket but needs to be re-booked to another flight. There is a probability of 10% that a booked passenger does not show up. The results for overbooking between 0 and 30 seats are shown on the next slide.

Binomial: Overbooking Flights II



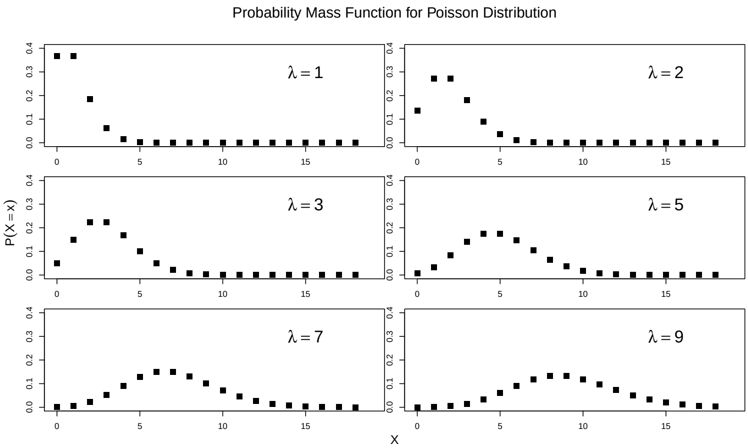
Poisson Distribution

By construction, the Poisson distribution (named after Simeon Denis Poisson, 1781-1840) is used for count data, i.e., $0, 1, 2, \ldots$. The probability mass function for the Poisson distribution is given by:

$$P(X = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$$

An example of the Poisson distribution for different parameter values is shown on the next slide.

Poisson Distribution Example



Poisson Distribution: PDF and CDF

The PDF and CDF of the Poisson Distribution in R are written as `dpois(x,lambda)` and `ppois(x,lambda)`, respectively. Consider the following probabilities:

- Probability of exactly four ($x = 4$) customers coming to your store when the average is six ($\text{lambda} = 6$)
- Probability of four or less ($x = 4$) customers coming to your store when the average is six ($\text{lambda} = 6$):

```
dpois(4,6)
```

```
## [1] 0.1338526
```

```
ppois(4,6)
```

```
## [1] 0.2850565
```

Continuous Probability Distributions

Properties:

- Probability of a particular event is zero!
- The area under the probability curve is 1.

Examples

- Uniform distribution
- Bell curve a.k.a. Normal distribution a.k.a. Gaussian Distribution
- Student's t-distribution

Uniform Distribution

The uniform distribution has two parameters, i.e., a and b. If $a <$b, a random variable X is said to have a uniform probability distribution on the interval (a, b) if and only if the density function of X is

$$f(x) = \frac{1}{b - a}$$

Examples:

- $a = 10$ and $b = 40$ then $Pr(25 < x < 30) = 1/6$
- Arrival of your online delivery during your lunch break

Normal Distribution: Introduction

The random variable X is said to be normally distributed with mean μ and variance σ^2 (abbreviated by x $\sim$ N[μ, σ^2] if the density function of x is given by

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The normal probability density function is bell-shaped and symmetric. The curve is derived from the binomial distribution:

- Galton Board

Standardizing a normal distribution to make it N(0,1) by calculating z, i.e.,

$$z = \frac{X-\mu}{\sigma}$$

z represents the distance from the mean expressed in units of the standard deviation.

Normal Distribution: Example

Suppose that we have a random variable with $\mu = 75$ and $\sigma = 10$. If we are interested in the probability $Pr(60 < x < 70)$ then we have to proceed in three steps:

1. Calculate the probability that $Pr(x < 60)$
2. Calculate the probability that $Pr(x < 70)$
3. Take the difference between the two probabilities

This can be achieved in one step with R:

```
pnorm(70,75,10)-pnorm(60,75,10)
```

```
## [1] 0.2417303
```

Student Distribution orn t-Distribution: Characteristics

The t-distribution is very similar to the Standard Normal:

- The t-distribution is continuous, symmetric, and bell-shaped.
- The shape (flatness/steepness) depends on the degrees of freedom.
- For very large degrees of freedom (i.e., ∞), the t-distribution is identical to the Standard Normal.

The important aspect of the t-distribution are the tails which are weighted heavier.

Student Distribution orn t-Distribution: Graphical Representation