Elementary Statistics: Math 080

Jordan Hanson

July 23, 2020

Whittier College Department of Physics and Astronomy

Summary

Summary

1. Topics from Chapter 4: 4.1 - 4.4
   - Discrete random variables
   - Expectation values and standard deviations
   - The binomial distribution
   - The geometric distribution
2. Topics from Chapter 6: 6.1 - 6.4
   - 2.1 The normal and standard normal distributions
   - 2.2 Using normal distributions

Discrete Random Variables

Discrete Random Variables

A discrete random variable is a property of data that can be counted with integers.

Examples:

- Times a baby eats per day
- Number of students in a class
- The number of wins a team has in a season
- *The number of calories we ate yesterday* - We may think of this as discrete if we have to round to the nearest calorie

Discrete Random Variables

Bin	n	$P(x)$	$x * P(x)$	$(x-\mu)^2 P(x)$
0-10k	13			
10-20k	15			
20-30k	20			
30-40k	11			
40-50k	9			
50-60k	9			
60-70k	6			
70-80k	7			
80-90k	5			
90-100k	3			
100k+	2			
Totals	100			

Table 1: Wage data for 100 Los Angeles County workers.

Discrete Random Variables

$P(X)$ is a **Probability distribution function** of a discrete random variable. PDFs are tools for answering questions like:

1. What is the probability that a random individual in LA County earns yearly wages in the top 5 categories of Tab. 1?
2. What is the probability that a random individual in LA County earns yearly wages in the bottom 5 categories of Tab. 1?
3. What is the *expectation value* of Tab. 1?
4. What is the *standard deviation* of Tab. 1?

Discrete Random Variables

Bin	n	$P(x)$	$x * P(x)$	$(x-\mu)^2 P(x)$
0	45			
1	190			
2	410			
3	220			
4	80			
5	55			
Totals	1000			

Table 2: Number of cars owned by 1,000 California citizens.

Discrete Random Variables

Consider Tab. 2 above.

1. What is the probability that a random Californian has 2 or fewer cars, according to Tab. 2?
2. Suppose a random Californian owns 4 cars. How many standard deviations above the mean is this, according to Tab. 2?

Discrete Random Variables

Bin	n	$P(x)$	$x * P(x)$	$(x-\mu)^2 P(x)$
200-300k	110			
300-400k	130			
400-500k	140			
500-750k	270			
750-1000k	100			
1000k+	250			
Totals	1000			

Table 3: Values of 1,000 residential properties in Los Angeles County.

Discrete Random Variables

Consider Tab. 3 and Tab. 1 above.

1. Consider the wage distribution of Tab. 1, and consider the home value distribution of Tab. 3. What is the average home value divided by the average yearly wage? What statistical fact does this reveal?
2. Typically, residents of California devote 30-40 percent of their budget to housing. Take 35 percent as a good estimate, and apply it to the prior calculation. How many years must someone work for the average wage to purchase an average home?
3. For more data and interesting figures, see `https://datausa.io/profile/geo/los-angeles-county-ca`

Discrete Random Variables

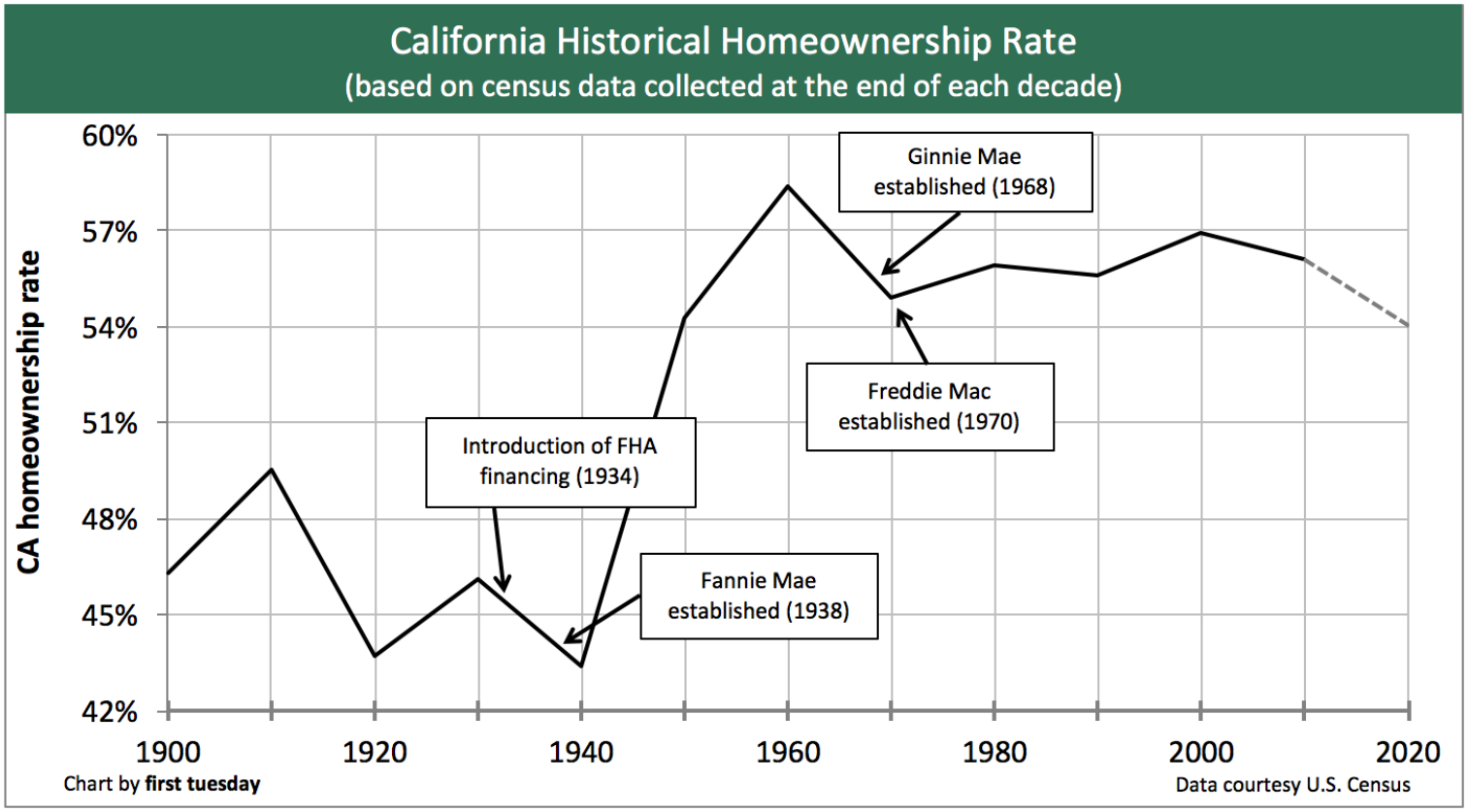

Figure 1: The effect of the FHA on home ownership in California across several decades.

The Binomial Distribution

The Binomial Distribution

Suppose we suspect a discrete random variable data set is binomially-distributed. The mean is $\mu = 2.1$ and the standard deviation is $\sigma = 1.0$. Suppose this data had 10 trials. Independently, someone tells us that the probability of a trial being successful in a similar experiment was 0.4. Is this data binomially-distributed?

- A: Yes, the mean follows $\mu = Np$ within one standard deviation.
- B: No, the mean does not follow $\mu = Np$ within one standard deviation.
- C: Yes, the mean implies a probability of success of 0.4.
- D: No, the standard deviation is too large to make the determination.

The Binomial Distribution

Suppose we are working with a biological experiment to predict the behavior of a small aquatic creature when its environment is inside a large magnetic field. The creature can either choose to go left (in the direction of the compass) or right (opposite to the compass). We conduct 10 trials on the same individual, and then repeat on 10 different individuals. How would you determine if the creatures are following the magnetic field?

- A: Count the number of times in all runs, all trials, that the individuals go left. If it is more than half the time, $p > 50\%$.
- B: Count the number of times per run that the individuals go left. Calculate the expectation value of those frequencies, and derive p.

Probability Distributions Can Be Continuous

Continuous Random Variables

Suppose instead of *counting trials*, we measure a number. Identify below: discrete random variable or continuous random variable?

1. Measuring the heights of a sample of people
2. Measuring the number of home runs a baseball player earns per season
3. Measuring the number of hands a poker player wins per game won
4. Measuring the wind speed on the top of a mountain

Continuous Random Variables

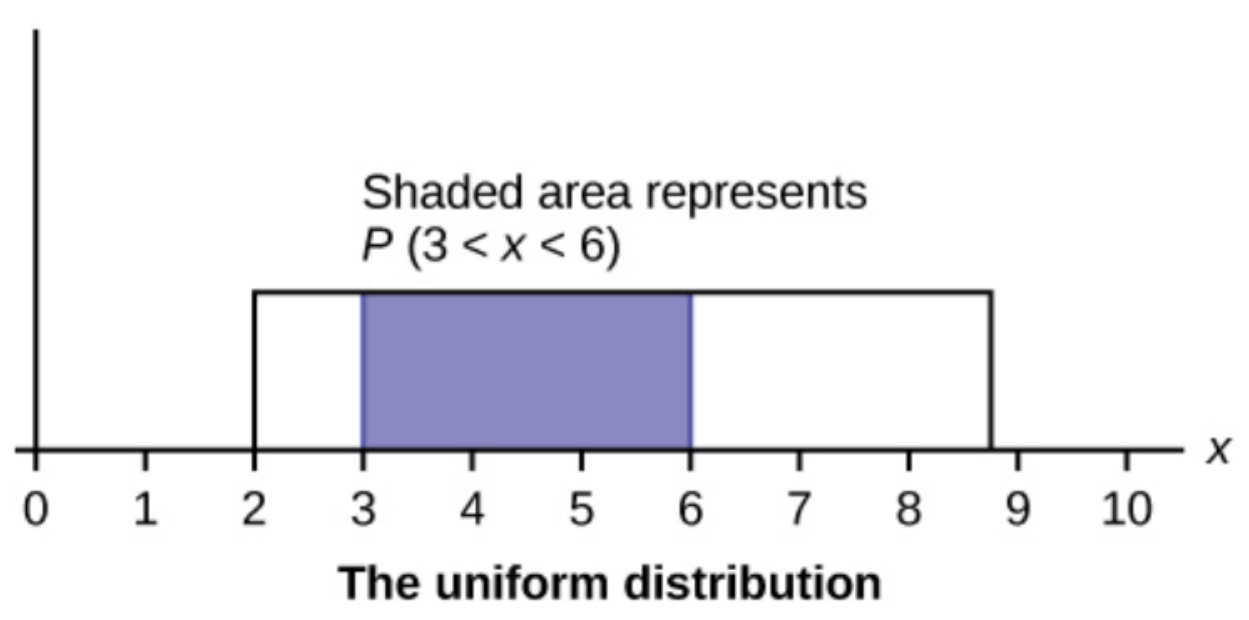

Figure 2: A uniform PDF of a *continuous random variable.*

- How do we *normalize* the frequencies?
- How do we calculate probabilites?
- What are the expectation value and standard deviation?

Continuous Random Variables

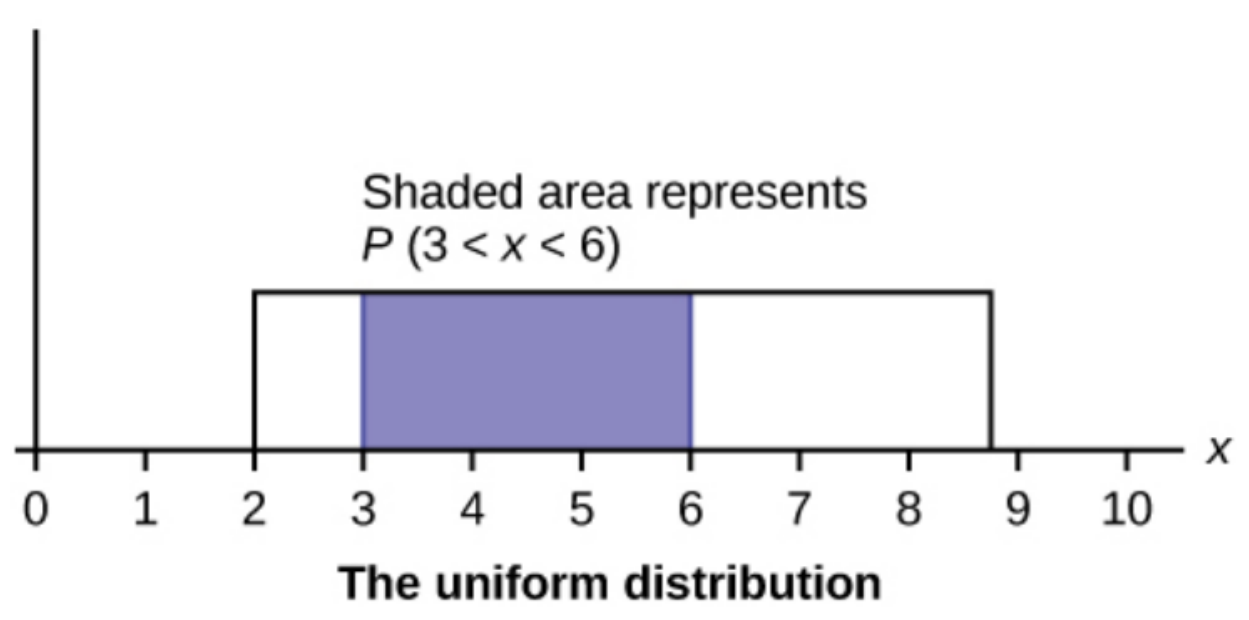

Figure 3: A uniform PDF of a *continuous random variable.*

- Normalization: $p(x) = 1/(b-a)$
- Probability that $x_1 < x < x_2$? $(x_2 - x_1)/(x_2 + x_1)$
- $E[x] = \mu = (b+a)/2$, $\sqrt{Var[x]} = \sigma = \sqrt{(b-a)^2/12}$

Continuous Random Variables

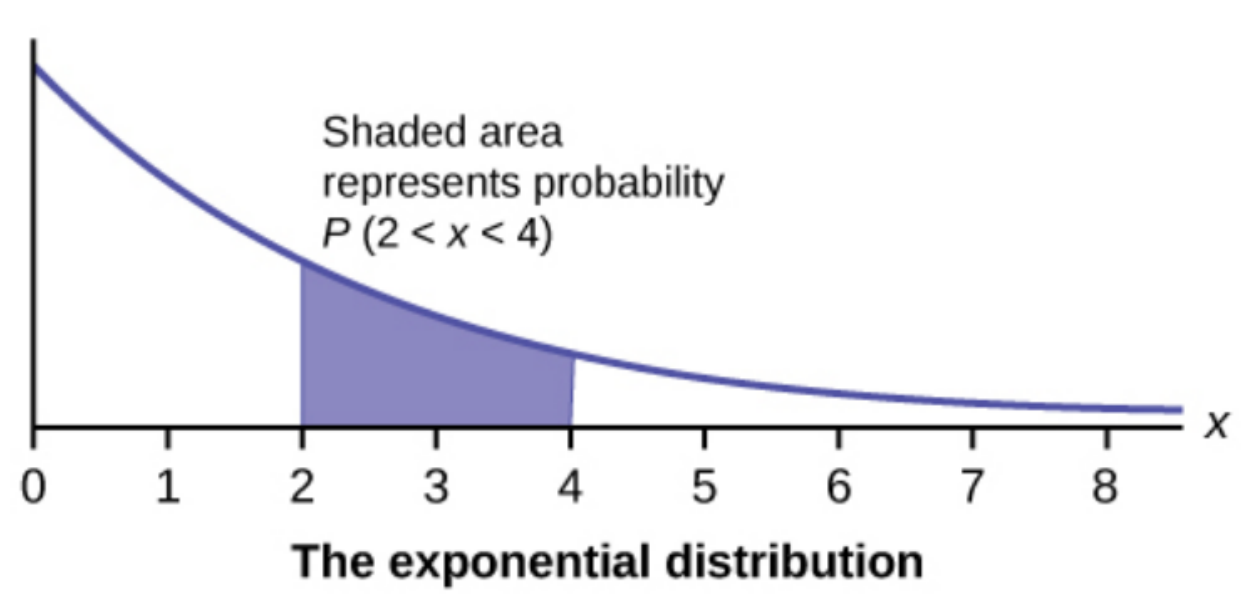

Figure 4: An exponential PDF of a *continuous random variable.*

Continuous Random Variables

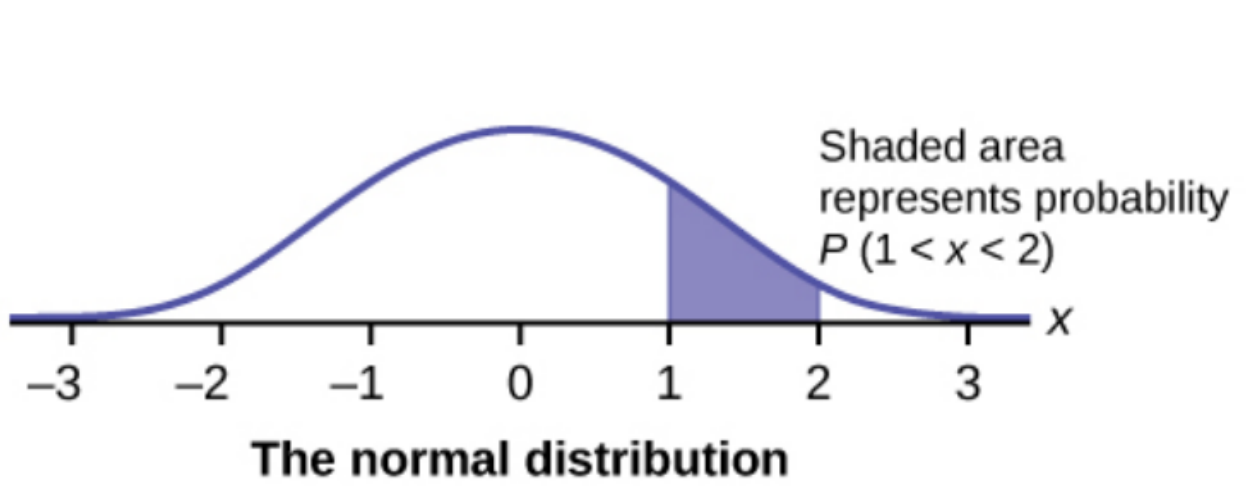

Figure 5: A normal distribution PDF of a *continuous random variable.*

Continuous Random Variables

Suppose we record the volume of milk a baby drinks per feeding when between the ages of 0-3 months. The volumes are **uniformly** distributed between 3.0 and 4.0 ounces per feeding. We record 100 measurements.

- What is the normalization? Or, graph the PDF.
- What is the mean volume?
- What is the standard deviation?
- What is the probability of finding a measurement in the set between 3.2 and 3.5 ounces per feeding?

Continuous Random Variables

Suppose we record the volume of milk a baby drinks per feeding when between the ages of 0-3 months. The volumes are **uniformly** distributed between 3.0 and 4.0 ounces per feeding. We record 100 measurements.

- What are the quartiles of the data?
- What is the 90th percentile of the data?

Interactive Questions

Interactive Questions

Suppose we are looking at a distribution (histogram/PDF) of a baseball team's batting average (probability a player gets a hit), and it is uniformly distributed. The lowest value is 0.200 and the highest is 0.400. What is the mean of the distribution?

- A: 0.200
- B: 0.300
- C: 0.400
- D: 0.350

Interactive Questions

Suppose we are looking at a distribution (histogram/PDF) of a baseball team's batting average (probability a player gets a hit), and it is uniformly distributed. The lowest value is 0.200 and the highest is 0.400. What is the median?

- A: 0.200
- B: 0.300
- C: 0.400
- D: 0.350

Interactive Questions

Suppose we are looking at a distribution (histogram/PDF) of a baseball team's batting average (probability a player gets a hit), and it is uniformly distributed. The lowest value is 0.200 and the highest is 0.400. What is the probability of being between 0.300 and 0.350?

- A: 10 percent
- B: 15 percent
- C: 25 percent
- D: 50 percent