Random Variables and Probability Models

Ryan Miller

Introduction

- Video #1
  - Introduction to random variables (discrete random variables)
- Video #2
  - Continuous random variables
- Video #3
  - When to use the Normal model

Random Variables

- We've been studying *probability* to understand the possible *outcomes* of a *random process*
  - Two important random processes are *sampling from a population*, and *assigning treatment/control groups*

Random Variables

- We've been studying *probability* to understand the possible *outcomes* of a *random process*
  - Two important random processes are *sampling from a population*, and *assigning treatment/control groups*
- Statisticians use a **random variable** to represent the *unknown numeric outcome* of a random process
  - Like any variable, you can think of a random variable, such as X, as a written placeholder for an unknown numerical value

Random Variables

- Consider the random process of flipping a fair coin
  - Because random variables must involve a numeric outcome, we can use the value "1" to represent the outcome "heads" and "0" to represent the outcome "tails"

Random Variables

- Consider the random process of flipping a fair coin
  - Because random variables must involve a numeric outcome, we can use the value "1" to represent the outcome "heads" and "0" to represent the outcome "tails"
  - We could've also mapped tails to 1 and heads to 0 without any consequence (so long as we keep track of what is what)

Random Variables

- Consider the random process of flipping a fair coin
  - Because random variables must involve a numeric outcome, we can use the value "1" to represent the outcome "heads" and "0" to represent the outcome "tails"
  - We could've also mapped tails to 1 and heads to 0 without any consequence (so long as we keep track of what is what)
- We can now define X as a random variable
  - $X = 1$ if "heads" is observed, and $X = 0$ if "tails" is observed

Probability Models

- After each touchdown in the National Football League (NFL), the scoring team gets to choose between a 1-pt and 2-pt attempt to earn additional points (on top of the 6 given for scoring a touchdown)

Probability Models

- After each touchdown in the National Football League (NFL), the scoring team gets to choose between a 1-pt and 2-pt attempt to earn additional points (on top of the 6 given for scoring a touchdown)
- We can use a random variable X to denote the number of total points the team earns from a touchdown
  - Recognize X represents a numeric outcome that is *unknowable in advance*

Probability Models

- After each touchdown in the National Football League (NFL), the scoring team gets to choose between a 1-pt and 2-pt attempt to earn additional points (on top of the 6 given for scoring a touchdown)
- We can use a random variable X to denote the number of total points the team earns from a touchdown
  - Recognize X represents a numeric outcome that is *unknowable in advance*
- Since a rule change in 2015, 9.6% of touchdowns were accompanied by zero additional points, 86.5% resulted in one additional point, and 3.9% resulted in two additional points

Probability Models

- After each touchdown in the National Football League (NFL), the scoring team gets to choose between a 1-pt and 2-pt attempt to earn additional points (on top of the 6 given for scoring a touchdown)
- We can use a random variable X to denote the number of total points the team earns from a touchdown
  - Recognize X represents a numeric outcome that is *unknowable in advance*
- Since a rule change in 2015, 9.6% of touchdowns were accompanied by zero additional points, 86.5% resulted in one additional point, and 3.9% resulted in two additional points
  - Based upon these data, we might consider following **probability model** for X:

X	6	7	8
$P(X = x)$	0.096	0.865	0.039

Probability Models

Probability models are useful because they help us understand a few key aspects of a random process:

1) **Expected Value**, or the “average” numeric outcome
2) **Variance**, or the total amount that the numeric outcomes vary from their *expected value*
3) **Standard Deviation**, or the “average” amount that numeric outcomes vary from their *expected value*

Expected Value

- The **expected value** of a random variable is denoted $E(X)$
- It describes the *expected result*, which is the sum of each possible outcome weighted by its probability

X	6	7	8
P(X = x)	0.096	0.865	0.039

- For a randomly chosen NFL touchdown, $E(X) = 6 * 0.096 + 7 * 0.865 + 8 * 0.039 = 6.94$ points

Variance

To see how much each possible outcome (6, 7, or 8 pts) varies from the expected outcome (6.94 pts) we can calculate their *squared deviations*

Points	6	7	8
Deviation	(6-6.94)^2	(7-6.94)^2	(8-6.94)^2

If we add these squared deviations, weighted by their probabilities, we get **variance**:

$$\text{Var}(X) = 0.096 * (6 - 6.94)^2 + 0.865 * (7 - 6.94)^2 + 0.039 * (8 - 6.94)^2 = 0.13$$

Variance

To see how much each possible outcome (6, 7, or 8 pts) varies from the expected outcome (6.94 pts) we can calculate their *squared deviations*

Points	6	7	8
Deviation	(6-6.94)^2	(7-6.94)^2	(8-6.94)^2

If we add these squared deviations, weighted by their probabilities, we get **variance**:

$$\text{Var}(X) = 0.096*(6-6.94)^2+0.865*(7-6.94)^2+0.039*(8-6.94)^2 = 0.13$$

Standard Deviation

Taking the square-root of the variance, we have the **standard deviation**, or the *average deviation* of outcomes from the expected value:

$$\text{SD}(X) = \sqrt{\text{Var}(X)} = \sqrt{0.13} = 0.36$$

So, we expect the average deviation (from the expected value of 6.94) of a touchdown to be 0.36 pts (not much variation)

Closing Remarks (Discrete Random Variables)

- The examples we've seen so far involve **discrete random variables**, or those where only a finite set of numeric outcomes are possible

Closing Remarks (Discrete Random Variables)

- The examples we've seen so far involve **discrete random variables**, or those where only a finite set of numeric outcomes are possible
  - For discrete random variables, we can define a **probability model** using a table
  - The information in this table can help us calculate the random variable's **expected value** and **standard deviation** better understand the underlying random process

Closing Remarks (Discrete Random Variables)

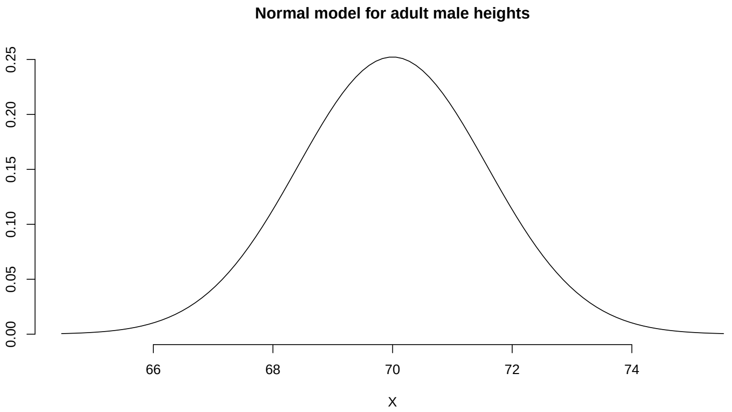
- The examples we've seen so far involve **discrete random variables**, or those where only a finite set of numeric outcomes are possible
  - For discrete random variables, we can define a **probability model** using a table
  - The information in this table can help us calculate the random variable's **expected value** and **standard deviation** better understand the underlying random process
- Next we'll look at **continuous random variables**, or those with infinitely many possible outcomes
  - As you'd expect, we'll need to introduce more sophisticated probability models to help us understand these variables

Continuous Random Variables

- Consider *randomly sampling* an adult male residing in the United States and let the random variable X denote their height (in inches)
  - Recognize that X could potentially take on infinitely many values (70.0 in, 70.01 in, 70.001 in, etc.)
- Although the probability of any individual value of X is exactly zero (technically speaking), not heights are equally likely
  - We need a continuous probability model to map the possible outcomes of X to probabilities

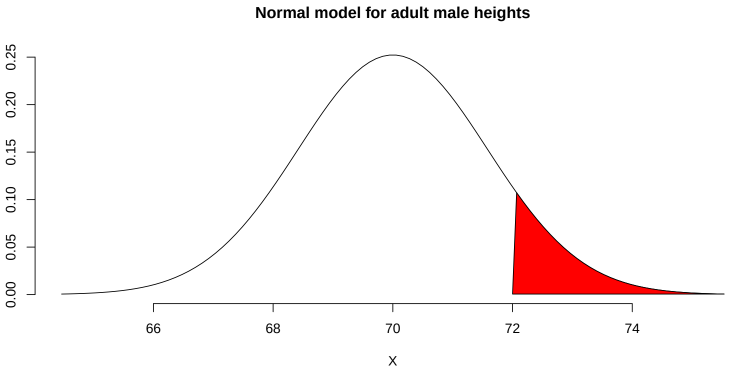
The Normal Model

- The **Normal distribution** is perhaps the most widely used probability model for continuous random variables



The Normal Model (cont)

- Under a *continuous probability model*, the probability of any single value of X is zero (as there are infinitely many possible values)
  - Thus, probabilities only make sense for intervals, for example we can represent $P(X > 72)$ using the *shaded area* shown below:



The Normal Model (cont)

- The Normal probability model is defined by the curve:

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

- The parameter μ is a constant that defines the *expected value* of the bell-curve
- The parameter σ is a constant that defines the *standard deviation* of the bell-curve (how tall or flat it appears)

The Normal Model (cont)

- The Normal probability model is defined by the curve:

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

- The parameter μ is a constant that defines the *expected value* of the bell-curve
- The parameter σ is a constant that defines the *standard deviation* of the bell-curve (how tall or flat it appears)
- There infinitely many different Normal curves, one for each combination of μ and σ
  - We will use the notation: $N(\mu, \sigma)$, for example $N(70, 2.5)$ might apply to our height example

Standardization

- Historically, statisticians wanted to avoid the possibility of infinitely many different probability models
  - This led them to **standardize** their data onto uniform, unitless scale

Standardization

- Historically, statisticians wanted to avoid the possibility of infinitely many different probability models
  - This led them to **standardize** their data onto uniform, unitless scale
- Z-scores are perhaps the most common form of standardization
  - Consider a random variable X and a Normal model defined by μ and σ
  - Under this model, the Z-score of X is calculated:

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

Z-scores

- A Z-score can be interpreted as how many *standard deviations* an *observed outcome* is above or below its *expected value*

Z-scores

- A Z-score can be interpreted as how many *standard deviations* an *observed outcome* is above or below its *expected value*
- For example, suppose X is a random variable from a $N(\mu = 70, \sigma = 2.5)$ distribution and we observe $x = 72$
  - This outcome leads to the Z-score: $z = (72 - 70)/2.5 = 0.8$
  - Therefore, a height of 72 inches is 0.8 standard deviations above what is expected (at least according to this probability model)

The Standard Normal Distribution

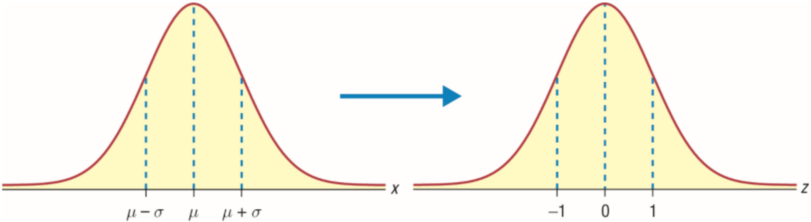
- Standardization enables us to use the **Standard Normal distribution** as a probability model in a wide variety of settings

The Standard Normal Distribution

- Standardization enables us to use the **Standard Normal distribution** as a probability model in a wide variety of settings
- For example, suppose adult male heights follow a Normal distribution centered at 70 inches with a standard deviation of 2.5 inches
  - This means, $X \sim N(70, 2.5)$
  - After standardization, $Z = \frac{X-70}{2.5} \sim N(0, 1)$

The Standard Normal Distribution

$$N(\mu,\sigma) \xrightarrow{z-transformation} N(0,1)$$



Example

Let X denote the height of a randomly chosen adult male, and assume the probability model $X \sim N(70, 2.5)$

1) Find the probability that this male's height is between 5'10 and 6'0 directly from the given Normal probability model
2) Find this same probability using Z-scores and the Standard Normal distribution

For each of these tasks, we'll utilize a new StatKey menu: StatKey Normal Curve

Example (solution)

Using Statkey:

1) On the N(70,2.5) curve, the area to the left of 70 inches (5'10) is 0.5, while the area to the left of 72 inches (6'0) is 0.788; thus, there is a 28% probability of a random adult male being between 5'10 and 6'0 under this model
2) To use the Standard Normal model, we'd the same thing, but with the preliminary step of calculating Z-scores. The Z-score for 70 inches is 0, while the Z-score for 72 increases is 0.8. On the Standard Normal curve, the area to the left of 0 is 0.5, while the area to the left of 0.8 is 0.788; again we find a 28% probability that a random adult male is between 5'10 and 6'0 under this model

Closing Remarks (Continuous Random Variables)

- Continuous random variables require a continuous probability model
  - The *Normal distribution* is a widely used probability model for these variables

Closing Remarks (Continuous Random Variables)

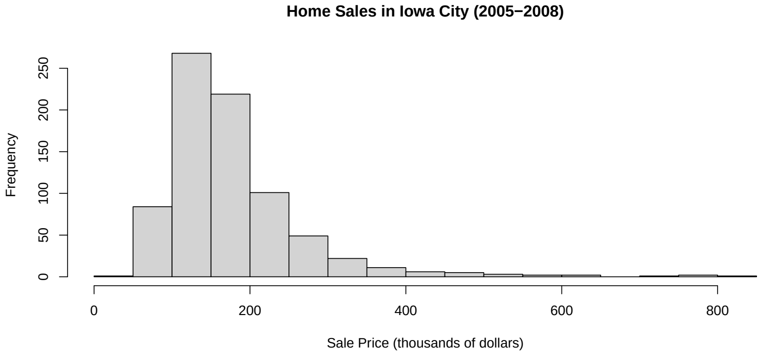
- Continuous random variables require a continuous probability model
  - The *Normal distribution* is a widely used probability model for these variables
- The Normal curve is defined by two parameters
  - The parameter μ, a constant that defines the *expected value* of the bell-curve
  - The parameter σ, a constant that defines the *standard deviation* of the bell-curve

Closing Remarks (Continuous Random Variables)

- Continuous random variables require a continuous probability model
  - The *Normal distribution* is a widely used probability model for these variables
- The Normal curve is defined by two parameters
  - The parameter μ, a constant that defines the *expected value* of the bell-curve
  - The parameter σ, a constant that defines the *standard deviation* of the bell-curve
- Standardization (ie: calculating Z-scores) allows us to work with a single Normal distribution (rather than needing to worry about infinitely many combinations of μ and σ)

How Accurate is the Normal Model?

- In this example, we'll look at the sale prices of all homes in Iowa City, IA between 2005-2008
  - The mean sale price was \$180.1k, and the standard deviation was \$90.65k



Applying the Normal Model

- Let X be a random variable denoting the sale price of a randomly selected home
- Because X is a continuous random variable, it seems reasonable to take the mean and standard deviation in our dataset and use $N(180.1, 90.65)$ as a probability model for X
  - How would you use this model to estimate $P(X \geq \$400k)$?

Applying the Normal Model

- Using StatKey, we could directly input our mean and standard deviation then calculate this right-tail probability to be 0.0076

Applying the Normal Model

- Using StatKey, we could directly input our mean and standard deviation then calculate this right-tail probability to be 0.0076
  - We also could standardize \$400k into a Z-score of $z = 400 - 180.190.65 = 2.426$ and use the Standard Normal distribution to arrive at the same estimated probability

Applying the Normal Model

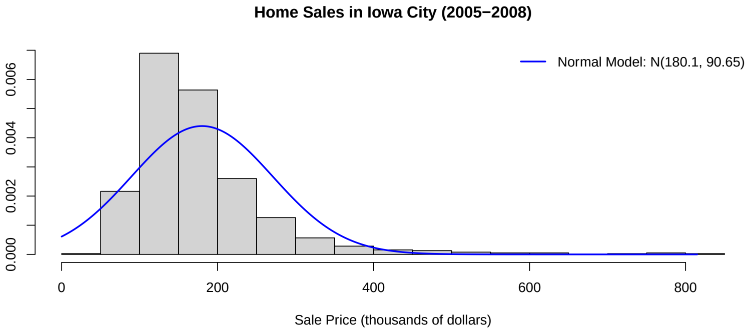
- Using StatKey, we could directly input our mean and standard deviation then calculate this right-tail probability to be 0.0076
  - We also could standardize \$400k into a Z-score of $z = 400 - 180.190.65 = 2.426$ and use the Standard Normal distribution to arrive at the same estimated probability
- However, both calculations assume the Normal model is a perfect representation of these data (or the population represented by them)
  - Is that an appropriate assumption?

Example

- The *empirical probability* of a randomly selected home selling for more than \$400k is 0.0283 (22 of 777 homes)
  - This discrepancy might not seem like much, but this is 3.7 times larger than what the Normal model suggested! (0.0076)

Example

- The *empirical probability* of a randomly selected home selling for more than $\$400k$ is 0.0283 (22 of 777 homes)
  - This discrepancy might not seem like much, but this is 3.7 times larger than what the Normal model suggested! (0.0076)



Appropriateness of the Normal Model

- In this application, the distribution of the data doesn't match the *shape* of the normal curve

Appropriateness of the Normal Model

- In this application, the distribution of the data doesn't match the *shape* of the normal curve
  - That is, even if we *center* and *scale* our normal model appropriately (ie: good choices of μ and σ), the model is incapable of representing the underlying distribution of these data

Appropriateness of the Normal Model

- In this application, the distribution of the data doesn't match the *shape* of the normal curve
  - That is, even if we *center* and *scale* our normal model appropriately (ie: good choices of μ and σ), the model is incapable of representing the underlying distribution of these data
- As an aside, notice these data contain $n = 777$ cases
  - A common misconception is that larger amounts of data tend to be normally distributed (they don't)

Appropriateness of the Normal Model

- In this application, the distribution of the data doesn't match the *shape* of the normal curve
  - That is, even if we *center* and *scale* our normal model appropriately (ie: good choices of μ and σ), the model is incapable of representing the underlying distribution of these data
- As an aside, notice these data contain $n = 777$ cases
  - A common misconception is that larger amounts of data tend to be normally distributed (they don't)
- That said, more data will improve the Normality of a special random variable, the *sample average*

Conclusion

- The Normal distribution provides a useful probability model for many, but not all, continuous random variables
  - Proper application of the Normal model requires the specification the bell-curve's center, μ, and it's spread, σ

Conclusion

- The Normal distribution provides a useful probability model for many, but not all, continuous random variables
  - Proper application of the Normal model requires the specification the bell-curve's center, μ, and it's spread, σ
  - Variables with skewed distributions cannot be appropriately modeled by the normal curve, even when using reasonable values of μ and σ

Conclusion

- The Normal distribution provides a useful probability model for many, but not all, continuous random variables
  - Proper application of the Normal model requires the specification the bell-curve's center, μ, and it's spread, σ
  - Variables with skewed distributions cannot be appropriately modeled by the normal curve, even when using reasonable values of μ and σ
- In general, having more data does not make a random variable more normally distributed
  - However, for the *sample average* (rather than the data-points themselves), having more data *does* have an important impact
  - We'll explore the *distribution of sample averages* next week