# The Normal Probability Model

Ryan Miller

# Outline

1. Approximating the binomial distribution
2. The Normal Curve
3. Applications of the Normal Model

# Introduction

- Previously, we worked with an example involving the sampling of $n = 2$ socks from a large population where 30% of socks were black
  - We used $X$ to denote the number of black socks in our sample, and we wrote out a probability model for $X$

# Introduction

- Previously, we worked with an example involving the sampling of $n = 2$ socks from a large population where 30% of socks were black
  - We used $X$ to denote the number of black socks in our sample, and we wrote out a probability model for $X$
- We also saw that $X$ could represented by a mathematical function (the *binomial distribution function*):

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

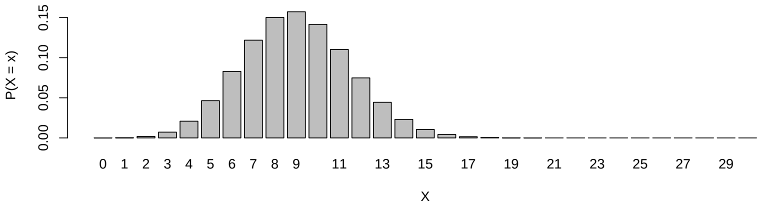
This function is useful, because it can be easily applied to larger samples. . .

# Introduction (cont.)

Consider a sample of $n = 30$ socks:

| $x$ | 0 | 1 | 2 | ... | 30 |
|---|---|---|---|---|---|
| $P(X = x)$ | $\binom{30}{0}(.3)^0(.7)^{30}$ | $\binom{30}{1}(.3)^1(.7)^{29}$ | $\binom{30}{2}(.3)^2(.7)^{28}$ | ... | $\binom{30}{30}(.3)^{30}(.7)^0$ |

Visually, we can graph these probabilities:



# The Normal Curve

The **Normal distribution** is perhaps the most widely used probability model:

# Continuous Random Variables

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

# Continuous Random Variables

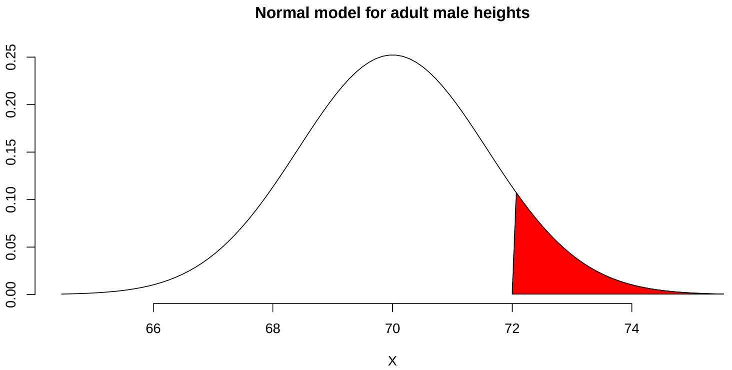
- The Normal probability model is defined by the curve:

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

- The parameter $\mu$ is a constant that defines the *expected value* of the bell-curve
- The parameter $\sigma$ 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 $\mu$ and $\sigma$
  - We will use the notation: $N(\mu, \sigma)$, for example $N(70, 2.5)$ (adult male heights)

# Normal Probability Calculations

- 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:



# Normal Curves in StatKey

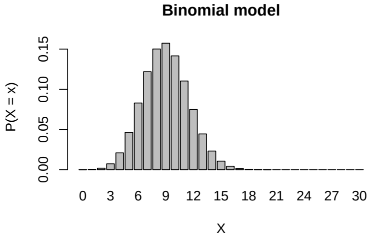
- To work with the Normal Curve, we'll utilize a new StatKey menu: StatKey Normal Curve
- As practice, verify that $P(X > 72) = 0.212$ for a $N(70, 2.5)$ distribution

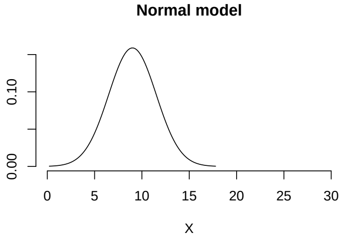
# Normal Approximation of the Binomial

- Previously, we briefly saw how the *expected value* of a binomial random variable was $E(X) = n * p$

# Normal Approximation of the Binomial

- Previously, we briefly saw how the *expected value* of a binomial random variable was $E(X) = n * p$
- Similarly, the *standard deviation* of a binomial random variable can be calculated using a mathematical formula, $SD(X) = \sqrt{n * p * (1 - p)}$
  - Thus, sampling $n = 30$ socks from a large population with 30% black socks can be approximated by a $N(9, 2.51)$ curve:



# Practice

40-weeks is considered a full-term pregnancy, but babies born prematurely often survive. For example, babies born at 24-weeks are estimated to have 60% survival rate.

1) Consider a hospital system that delivers $n = 50$ babies aged 24-weeks every year. Let $X$ denote the number of these babies who survive. What are the *expected value* and *standard deviation* of $X$?
2) Use StatKey to display a Normal Model of this scenario. Then, use this model to estimate the probability that fewer than half of these babies survive (ie: 24 or fewer survivors)

# Practice (solution)

1) $E(X) = 50 * 0.6 = 30$, $SD(X) = \sqrt{50 * 0.6 * 0.4} = 3.464$
2) Using a $N(30, 3.464)$ model, $P(X \leq 24) = 0.042$

# Standardization

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

# Standardization

- Historically, statisticians wanted to avoid the possibility of infinitely many different probability models
  - This led them to **standardize** their data onto single, unit-free scale
- $Z$-scores are perhaps the most common form of standardization
  - Consider a random variable $X$ and a Normal model defined by $\mu$ and $\sigma$
  - 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

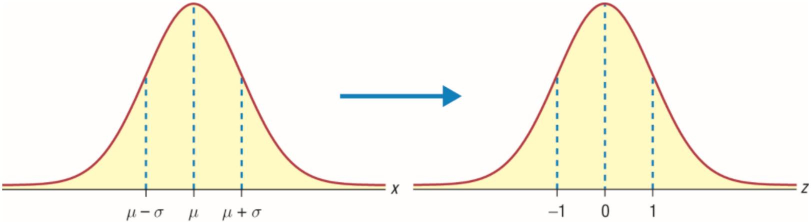
- Standardization enables us to use the **Standard Normal distribution** as a probability model in a wide variety of settings

# 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

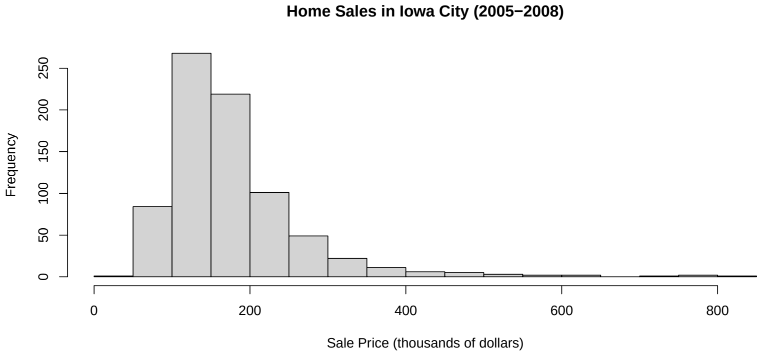
# Example (solution)

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

# 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 \$400$k$ 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

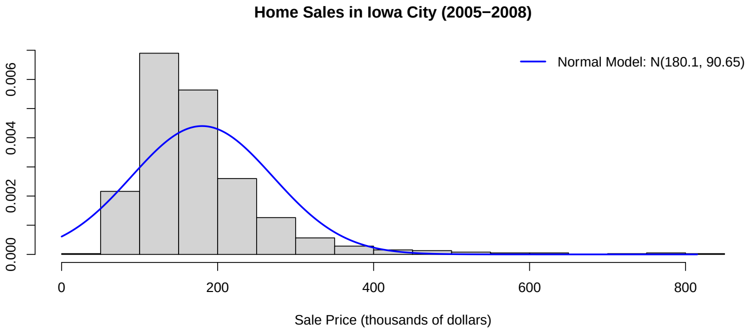
- Using StatKey, we could directly input our mean and standard deviation then calculate this right-tail probability to be 0.0076
  - We also could standardize \$400$k$ into a Z-score of $z = 400 - 180.190.65 = 2.426$ and use the Standard Normal distribution to arrive at the same estimated probability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

# Example

- The *empirical probability* of a randomly selected home selling for more than \$400$k$ 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 \$400$k$ 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 $\mu$ and $\sigma$), 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 $\mu$ and $\sigma$), 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 $\mu$ and $\sigma$), 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, $\mu$, and it's spread, $\sigma$

# 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, $\mu$, and it's spread, $\sigma$
  - Variables with skewed distributions cannot be appropriately modeled by the normal curve, even when using reasonable values of $\mu$ and $\sigma$

# 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, $\mu$, and it's spread, $\sigma$
  - Variables with skewed distributions cannot be appropriately modeled by the normal curve, even when using reasonable values of $\mu$ and $\sigma$
- 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