Applying the Normal Model

Ryan Miller

Introduction

- The last presentation introduced the **Normal distribution** as a probability model for a continuous random variable

Introduction

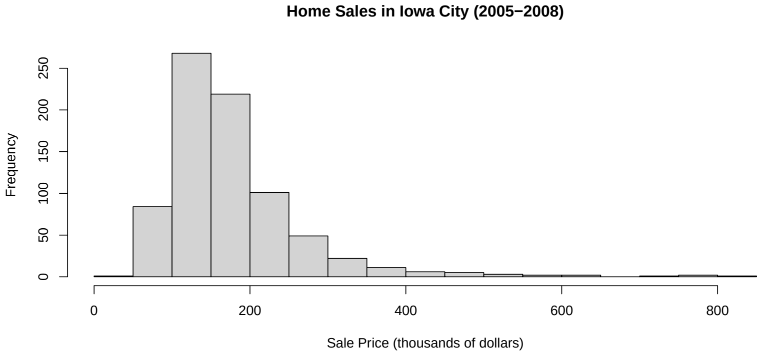
- The last presentation introduced the **Normal distribution** as a probability model for a continuous random variable
- This model is defined by two components:
  - The parameter μ, a constant that defines the *center* of the bell-curve
  - The parameter σ, a constant that defines the *spread* of the bell-curve

Introduction

- The last presentation introduced the **Normal distribution** as a probability model for a continuous random variable
- This model is defined by two components:
  - The parameter μ, a constant that defines the *center* of the bell-curve
  - The parameter σ, a constant that defines the *spread* of the bell-curve
- This presentation will go through an example that illustrates where this model is and is not appropriate

Example

- 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



Example

- Let X be a random variable denoting the sale price of a randomly selected home
  - What might you consider using as a probability model for X?

Example

- Let X be a random variable denoting the sale price of a randomly selected home
  - What might you consider using as a probability model for X?
- 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 $P(X \geq \$400k)$?

Example

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

Example

- 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

Example

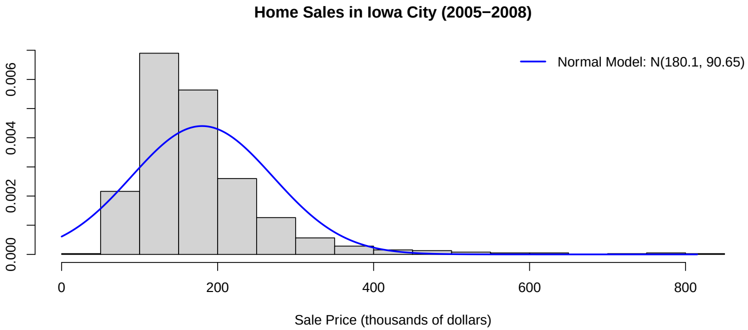
- Using StatKey, we could directly input our mean and standard deviation then calculate this right-tail probability to be 0.0076
  - We also could standardize \$400k into a Z-score of $z = 400 - 180.190.65 = 2.426$ and use the Standard Normal distribution to arrive at the same estimated probability
-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

- Primary issue with this application is that distribution of the data doesn't match the *shape* of the normal curve

Appropriateness of the Normal Model

- Primary issue with this application is that 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

- Primary issue with this application is that 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

Appropriateness of the Normal Model

- Primary issue with this application is that 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
- That said, more data (larger sample sizes) *does* impact the distributional shape of certain random variables
  - Next week, we will extend the normal model to the distribution of *sample averages*

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
  - However, 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
  - However, 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, if the random variable represents the *sample average* (rather than the data-points themselves), having more data *does* have an important impact
  - We will explore the distribution of sample averages next week