The normal distribution

Patrick Breheny

March 4
Histograms of infant mortality rates, heights, and cholesterol levels:

What do these histograms have in common?
Mathematicians discovered long ago that the equation

\[ y = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \]

described the histograms of many random variables.
Features of the normal curve

- The normal curve is symmetric around $x = 0$
- The normal curve is always positive
- The normal curve drops rapidly down near zero as $x$ moves away from 0
Note that the data has been standardized and that the vertical axis is now called “density”.

Data whose histogram looks like the normal curve are said to be *normally distributed* or to follow a *normal distribution*.
Probabilities from the normal curve

Probabilities are given by the area under the normal curve:
This is where the 68%/95% rule of thumb that we discussed earlier comes from:
By knowing that the total area under the normal curve is 1, we can get a rough idea of the area under a curve by looking at a plot.

However, to get exact numbers, we will need a computer.

“How much area is under this normal curve?” is an extremely common question in statistics, and programmers have developed algorithms to answer this question very quickly.

The output from these algorithms is commonly collected into tables, which is what you will have to use for exams.
Calculating the area under a normal curve, example 1

Find the area under the normal curve between 0 and 1

\[ 0.84 - 0.5 = 0.34 \]
Calculating the area under a normal curve, example 2

Find the area under the normal curve above 1

\[ 1 - 0.84 = 0.16 \]
Calculating the area under a normal curve, example 3

Find the area under the normal curve that lies outside -1 and 1

1 - (.84−.16) = .32

Alternatively, we could have used symmetry: 2(.16)=.32
A related question of interest is, “What is the \( x \)th percentile of the normal curve?”

This is the opposite of the earlier question: instead of being given a value and asked to find the area to the left of the value, now we are told the area to the left and asked to find the value.

With a table, we can perform this inverse search by finding the probability in the body of the table, then looking to the margins to find the percentile associated with it.
Calculating percentiles (cont’d)

- What is the 60th percentile of the normal curve?
- There is no “.600” in the table, but there is a “.599”, which corresponds to 0.25
- The real 60th percentile must lie between 0.25 and 0.26 (it’s actually 0.2533)
- For this class, 0.25, 0.26, or anything in between is an acceptable answer
- How about the 10th percentile?
- The 10th percentile is -1.28
Calculating values such that a certain area lies within/outside them

Find the number \( x \) such that the area outside \(-x\) and \( x\) is equal to 10%

Our answer is therefore \( \pm 1.645 \) (the 5th/95th percentile)
Reconstructing a histogram

- In week 2, we said that the mean and standard deviation provide a two-number summary of a histogram.
- We can now make this observation a little more concrete.
- Anything we could have learned from a histogram, we will now determine by approximating the real distribution of the data by the normal distribution.
- This approach is called the *normal approximation*.
The data set we will work with on these examples is the NHANES sample of the heights of 2,649 adult women.

- The mean height is 63.5 inches.
- The standard deviation of height is 2.75 inches.
Procedure: Probabilities using the normal curve

The procedure for calculating probabilities with the normal approximation is as follows:

#1 Draw a picture of the normal curve and shade in the appropriate probability

#2 Convert to standard units: letting $x$ denote a number in the original units and $z$ a number in standard units,

$$z = \frac{x - \bar{x}}{SD}$$

where $\bar{x}$ is the mean and $SD$ is the standard deviation

#3 Determine the area under the normal curve using a table or computer
Suppose we want to estimate the percent of women who are under 5 feet tall

5 feet, or 60 inches, is \( \frac{3.5}{2.75} = 1.27 \) standard deviations below the mean

Using the normal distribution, the probability of more than 1.27 standard deviations below the mean is

\[
P(x < -1.27) = 10.2\%
\]

In the actual sample, 282 out of 2,649 women were under 5 feet tall, which comes out to 10.6\%
Another example: suppose we want to estimate the percent of women who are between 5’3 and 5’6 (63 and 66 inches).

These heights are 0.18 standard deviations below the mean and 0.91 standard deviations above the mean, respectively.

Using the normal distribution, the probability of falling in this region is 39.0%.

In the actual data set, 1,029 out of 2,649 women were between 5’3 and 5’6: 38.8%.
Procedure: Percentiles using the normal curve

- We can also use the normal distribution to approximate percentiles.
- The procedure for calculating percentiles with the normal approximation is as follows:
  
  #1 Draw a picture of the normal curve and shade in the appropriate area under the curve.
  
  #2 Determine the percentiles of the normal curve corresponding to the shaded region using a table or computer.
  
  #3 Convert from standard units back to the original units:

\[ x = \bar{x} + z(SD) \]

where, again, \( x \) is in original units, \( z \) is in standard units, \( \bar{x} \) is the mean, and \( SD \) is the standard deviation.
Approximating percentiles: Example

- Suppose instead that we wished to find the 75th percentile of these women’s heights
- For the normal distribution, 0.67 is the 75th percentile
- The mean plus 0.67 standard deviations in height is 65.35 inches
- For the actual data, the 75th percentile is 65.39 inches
These examples are by no means special: the distribution of many random variables are very closely approximated by the normal distribution.

Indeed, this is why statisticians call it the “normal” distribution.

Other names for the normal distribution include the Gaussian distribution (after its inventor) and the bell curve (after its shape).

For variables with approximately normal distributions, the mean and standard deviation essentially tell us everything about the data – other summary statistics and graphics are redundant.
Other variables, however, are not approximated by the normal distribution well, and give misleading or nonsensical results when you apply the normal approximation to them.

For example, the value 0 lies 1.63 standard deviations below the mean infant mortality rate for Europe.

The normal approximation therefore predicts a probability that 5.1% of the countries in Europe will have negative infant mortality rates.
Caution (cont’d)

- As another example, the normal distribution will always predict the median to lie 0 standard deviations above the mean
- i.e., it will always predict that the median equals the mean
- As we have seen, however, the mean and median can differ greatly when distributions are skewed
- For example, according to the U.S. census bureau, the mean income in the United States is $66,570, while the median income is $48,201
Summary

- The distribution of many random variables are very closely approximated by the normal distribution
- Know how to calculate the area under the normal curve
- Know how to determine percentiles of the normal curve
- Know how to approximate probabilities for real data using the normal approximation
- Know how to approximate quantiles for real data using the normal approximation