The Normal Distribution

August 8, 2019

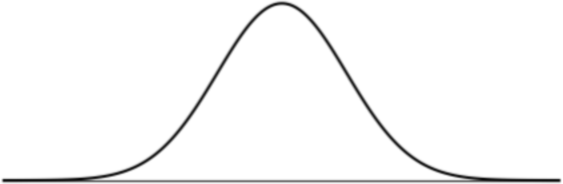
Distributions of Random Variables

- We've spent the past week talking about random variables.
- We've also talked about probability distributions.

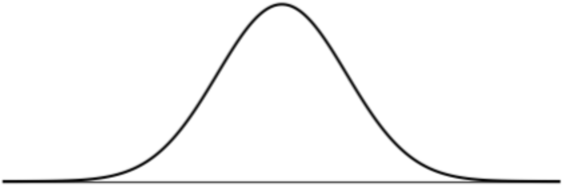
In Chapter 4, we are going to put these two concepts together to think about some common distributions that we use to model random variables.

The Normal Distribution

We start our discussion with the **normal distribution**. This is one of the most common distributions you will see in practice.



The Normal Distribution



Normal distributions are always...

- Symmetric.
- Unimodal.
- "Bell curves".

Variables such as SAT scores closely follow the normal distribution.

The Normal Distribution

The normal distribution has most measurements falling somewhere near the middle - or average - and values get less and less likely as we move further into the tails.

Variables such as SAT scores closely follow the normal distribution.

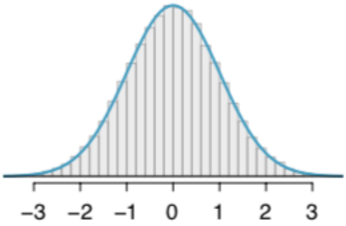
Normal Distributions

- Many variables are nearly normal, but none are exactly normal.
- While not perfect for any single problem, the normal distribution is very *useful* for a variety of problems.
- We will use it in data exploration and to solve important problems in statistics.

The Normal Distribution Model

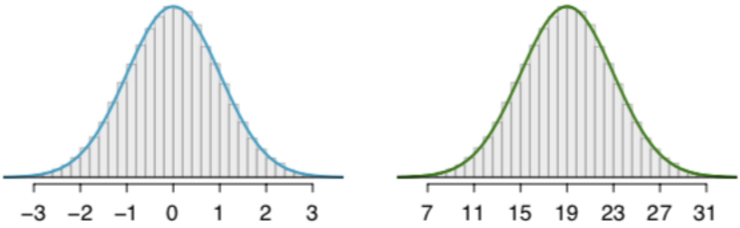
- The symmetric, unimodal, bell-shaped curve of the normal distribution can vary based on:
  - Mean
  - Standard deviation
- These adjustable details are called **model parameters**.

Parameters: Normal Distribution



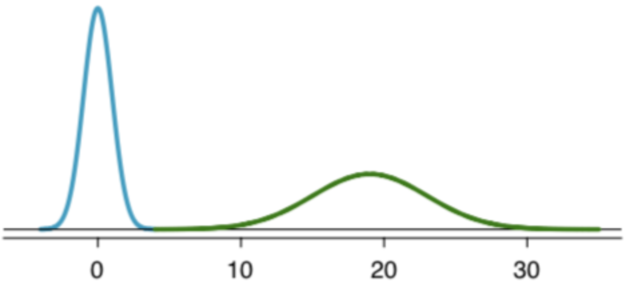
- Changing the mean shifts the curve to the left or right.
- Changing the standard deviation stretches or constricts the curve.
    - (This can make the peak appear narrower or flatter.)

Parameters: Normal Distribution



- The distribution on the left has $\mu = 0$ and $\sigma = 1$.
- The distribution on the right has $\mu = 19$ and $\sigma = 4$
- These look exactly the same because the scale of the axis has been adjusted.

Parameters: Normal Distribution



- These are the same two distributions, now on the same axis.
- Now we can see that the shift of the mean from 0 to 19 moves the distribution to the right.
- The change in standard deviation from 1 to 4 flattens the distribution.

Normal Distribution Notation

For a normal distribution with mean μ and standard deviation σ, we write

$$N(\mu, \sigma)$$

For a variable X with a normal distribution, we may write

$$X \sim N(\mu, \sigma).$$

where "$\sim$" denotes "is distributed".

Normal Distribution Notation

For a normal distribution with mean 19 and standard deviation 4, we write

$$N(\mu = 19, \sigma = 4)$$

- The mean and standard deviation describe a normal distribution fully and exactly.
- This is what we mean by a distribution's **parameters**.

Standard Normal Distribution

The **standard normal distribution** is a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

$$N(\mu = 0, \sigma = 1)$$

Standardizing with Z-Scores

We often want to put data onto a standardized scale, which can make comparisons more reasonable.

Example: SAT and ACT

The distribution of SAT and ACT scores are both nearly normal. The table shows the mean and standard deviation for total scores on each.

	SAT	ACT
Mean	1100	21
SD	200	6

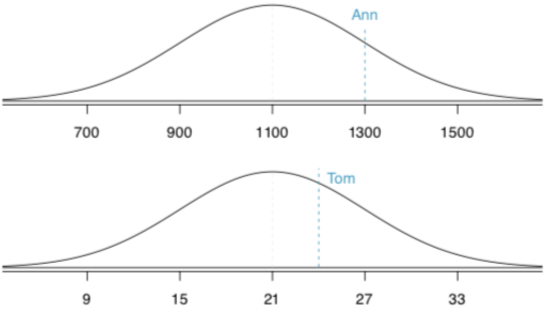
Suppose Ann scored 1300 on her SAT and Tom scored 24 on his ACT. Who performed better?

Example: SAT and ACT

- We can use the standard deviation to help us figure out who performed better.
- Ann's SAT score is 1 standard deviation above average.
  - $1100 + 200 = 1300$
- Tom's ACT score is 0.5 standard deviations above average.
  - $21 + 0.5 \times 6 = 24$
- If you remember taking either test and being told your percentile, that's the same idea!

Example: SAT and ACT

We can also plot the normal distributions with scaled axes:



Now we can see that Ann tends to do better with respect to everyone else than Tom does, so her score is better.

Standardizing with Z-Scores

Our example got at a standardization technique called a Z-score.

- This method is commonly employed with normal distributions, but could also be used more generally.
- The **Z-score** of an observation is defined as the number of standard deviations it falls above or below the mean.
  - If the observation is one standard deviation above the mean, its Z-score is 1.
  - If it is 1.5 standard deviations below the mean, then its Z-score is -1.5.

Standardizing with Z-Scores

We compute the Z-score for an observation x that follows a distribution with mean μ and standard deviation σ using

$$z = \frac{x - \mu}{\sigma}$$

Example: Standardizing with Z-Scores

The SATs had a mean score of $\mu_{SAT} = 1100$ and a standard deviation of $\sigma_{SAT} = 200$. For Ann's SAT score of 1300, the Z-score is

$$z_{Ann} = \frac{x_{Ann} - \mu_{SAT}}{\sigma_{SAT}} = \frac{1300 - 1100}{200} = 1$$

Example: Standardizing with Z-Scores

The ACTs has mean $\mu = 21$ and standard deviation $\sigma = 6$. Use Tom's ACT score, 24, to find his Z-score.

Z-Scores

- Observations above the mean always have positive Z-scores.
- Observations below the mean always have negative Z-scores.
- If an observation is equal to the mean, the Z-score is always 0.

Example

Let X represent a random variable from $N(\mu = 3, \sigma = 2)$

$$X \sim N(\mu = 3, \sigma = 2)$$

and suppose we observe $x = 5.19$.

1. Find the Z-score of x.
2. Use the Z-score to determine how many standard deviations above or below the mean x falls.

Example

We know from the problem statement that $\mu = 3$, $\sigma = 2$, and our observed value is $x = 5.19$. So

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{5.19 - 3}{2} \\ &= 1.095. \end{aligned}$$

Example

Using our definition of a Z-score, $z = 1.095$ means that the observations x is 1.095 standard deviations *above* the mean.

We know that x is above the mean because the Z-score is positive.

Example: Brushtail Possums

Head lengths of brushtail possums follow a normal distribution with mean 92.6 mm and standard deviation 3.6 mm.

Compute the Z-scores for possums with head lengths of 95.4 mm and 85.8 mm.

Example: Brushtail Possums

Let Y be the head lengths of brushtail possums. We say that $Y \sim N(\mu = 92.6, \sigma = 3.6)$.

For a head length of 95.4 mm, the Z-score will be

$$\begin{aligned} z &= \frac{y - \mu}{\sigma} \\ &= \frac{95.4 - 92.6}{3.6} \\ &= 0.78. \end{aligned}$$

Example: Brushtail Possums

Let Y be the head lengths of brushtail possums. We say that $Y \sim N(\mu = 92.6, \sigma = 3.6)$.

For a head length of 85.8 mm, the Z-score will be

$$\begin{aligned} z &= \frac{y - \mu}{\sigma} \\ &= \frac{85.8 - 92.6}{3.6} \\ &= -1.89. \end{aligned}$$

Example: Brushtail Possums

- The possum with a head length of 95.4 mm is 0.78 standard deviations *above* the mean ($z = 0.78$).
- The possum with a head length of 85.8 mm is 1.89 standard deviations *below* the mean ($z = -1.89$).

Z-Scores and Unusual Observations

- We can use Z-scores to identify potentially unusual observations.
- An observation x_1 is *more unusual* than another observation x_2 is further from the mean.
- If z_1 and z_2 are the corresponding Z-scores, x_1 is more unusual than x_2 if
$$|z_1| > |z_2|$$
- This technique is especially useful for symmetric distributions.

Example: Brushtail Possums

We decided that

- The possum with a head length of 95.4 mm is 0.78 standard deviations *above* the mean ($z = 0.78$).
- The possum with a head length of 85.8 mm is 1.89 standard deviations *below* the mean ($z = -1.89$).

Since

$$|-1.89| > |0.78|,$$

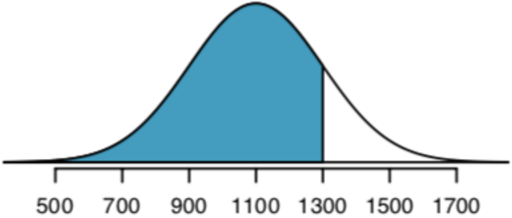
we say the possum with the head length of 85.8 mm is more unusual than the other possum.

Finding Tail Areas

- Yesterday, we talked about using the area under a curve to think about proportions.
- Determining the area under the tail of a distribution is very useful in statistics!
- For example, your SAT percentile is the fraction of people who scored lower than you.

Finding Tail Areas

We can visualize a tail area as the curve and shading shown.



- This is the distribution for SAT scores with Ann's score as the cutoff point, at $x = 1300$.
- The area to the left of x is the percentile.

Finding Tail Areas

There are several techniques for finding tail areas:

1. Integrate.
2. Use a graphing calculator.
3. Use a probability table.
4. Use a statistical software.

Finding Tail Areas: Integration

The function that creates our normal distribution curve is

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

Don't write this down. We won't use it. In fact, it's impossible to integrate this by hand!

Finding Tail Areas: Graphing Calculator

You are not required to have a graphing calculator, so you won't be required to use one for tail probabilities.

However, you can find a video of how to use a graphing calculator to calculate tail probabilities at

www.openintro.org/videos

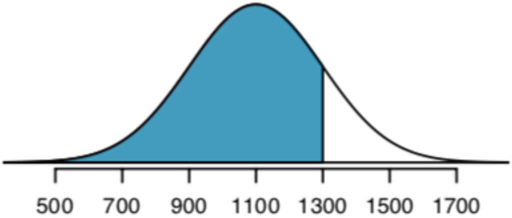
Finding Tail Areas: Probability Tables

Probability tables are often used in classrooms but these days they are rarely used in practice.

Appendix C.1 in your textbook contains such a table and a guide for how to use it.

Finding Tail Areas: Software

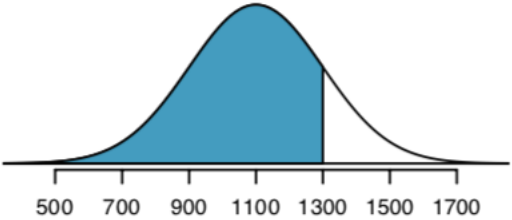
Since we can't integrate by hand, we can have a computer integrate for us!



In R, we could find the area shown using the following command, which takes in the Z-score and returns the lower tail area:

```
> pnorm(1)
[1] 0.8413447
```

Finding Tail Areas: Software



We can specify the cutoff explicitly if we also note the mean and standard deviation:

```
> pnorm(1300, mean = 1100, sd = 200))
[1] 0.8413447
```

Finding Tail Areas

- For quizzes and exams, you will be provided with information from `R`.
- I will do the work in `R`, but you will need to use a Z-score to pick the correct tail probability from a list.

For example

Z-score	Lower Tail Area
1	0.8413
1.5	0.9332

Finding Tail Areas

- We will solve all normal distribution problems by first calculating Z-scores.
- We do this because it will help us when we move on to Chapter 5.
- Therefore all tail area information will be provided in terms of Z-scores (as in the previous slide).

Example: Normal Probability

Cumulative SAT scores are well-approximated by a normal model, $N(\mu = 1100, \sigma = 200)$.

Shannon is a randomly selected SAT taker, and nothing is known about her SAT aptitude. What is the probability Shannon scores at least 1190 on her SATs?

Normal Probability

This brings up a crucial point:

- The area under a distribution curve is 1.
- This corresponds to the probabilities in a discrete probability distribution summing to 1!

So when we want to know the probability Shannon scores at least 1190 on her SATs, we are interested in $P(X < 1190)$.

Example: Normal Probability

SATs well approximated by $N(\mu = 1100, \sigma = 200)$

- First, we want to draw and label a picture of the normal distribution.
- These do not need to be exact to be useful.
  - We will see this in a moment when I try to draw on the board.
- We are interested in the chance she scores above 1190, so we shade the upper tail.

Example: Normal Probability

To find the area of the shaded section

- First calculate the Z-score

$$Z = \frac{x - \mu}{\sigma} = \frac{1190 - 1100}{200} = 0.45$$

- Then find the lower tail probability (using a statistical software or other method).
  - The area left of $Z = 0.45$ is 0.6736.

Example: Normal Probability

To find the area above $Z = 0.45$, $\mathrm{P}(Z > 0.45)$ we can use the complement,

$$P(Z > 0.45) = 1 - P(Z < 0.45),$$

Example: Normal Probability

This is one minus the area of the lower tail:

$$1 - 0.6737 = 0.3264$$

So the probability Shannon scores at least 1190 is 32.64%.

Finding Areas to the Right

- Software programs usually return the area to the left (left tail) when given a Z-score.
- To get the area to the right
  1. Find the area to the left.
  2. Subtract this area from one.

Recommendation

Draw a picture first; find the Z-score second.

- Draw and label the normal curve and shade the area of interest.
- This helps to
  1. Provide a general estimate of the probability.
  2. Set up your problem correctly.
- *Then* you can identify the appropriate Z-score and probabilities.

Example

Edward earned a 1030 on his SAT. What is his percentile?

Example

Edward earned a 1030 on his SAT. What is his percentile? Recall that his percentile is the percent of people who score lower than Edward.

First, we want to draw a picture. Recall that cumulative SAT scores are well-approximated by a normal model, $N(\mu = 1100, \sigma = 200)$

Example

Identifying the mean $\mu = 1100$, the standard deviation $\sigma = 200$, and the cutoff for the tail area $x = 1030$ makes it easy to compute the Z-score:

$$Z = \frac{x - \mu}{\sigma} = \frac{1030 - 1100}{200} = -0.35$$

Using `R`, we get a (left) tail area of 0.3632.

So Edward is at the 36th percentile.

Example

Use the results of the previous example to compute the proportion of SAT takers who did better than Edward.

Example

Use the results of the previous example to compute the proportion of SAT takers who did better than Edward.

Let's revise our picture.

Example

We know that 36.32% of test-takers do worse than Edward. So

$$\begin{aligned} P(\text{better than Edward}) &= 1 - P(\text{not better than Edward}) \\ &= 1 - 0.3632 \\ &= 0.6368 \end{aligned}$$

Percentiles

- So far, we've talked about finding a percentile based on an observation.
- Now we want to think about finding the observation corresponding to a particular percentile.
- For example, suppose you want to get into a graduate school whose incoming students usually score above the 80th percentile on the GRE.
    - We might be interested in estimating what score corresponds to the 80th percentile.

Example: Percentiles

Based on a sample of 100 men, the heights of male adults in the US is nearly normal with mean 70.0" and standard deviation 3.3".

Erik's height is at the 40th percentile. How tall is he?

Example: Percentiles

Heights are approximately normal $N(\mu = 70, \sigma = 3.3)$. Erik is at the 40th percentile.

First, we want to draw our picture.

Example: Percentiles

Heights are approximately normal $N(\mu = 70, \sigma = 3.3)$. Erik is at the 40th percentile.

- Before, we knew the Z-score and used it to find the area.
- Now, we know the area and must find the Z-score.

Using `R`, we obtain the corresponding Z-score of $z = -0.25$.

Example: Percentiles

Heights are approximately normal $N(\mu = 70, \sigma = 3.3)$. Erik is at the 40th percentile.

Now we have the corresponding Z-score of $z = -0.25$ and can use the Z-score formula to find Erik's height:

$$-0.25 = z_{Erik} = \frac{x_{Erik} - \mu}{\sigma} = \frac{x_{Erik} - 70}{3.3}$$

Example: Percentiles

With a little algebra, we can solve for x_{Erik}:

$$x_{Erik} = -0.25 \times 3.3 + 70 = 69.175$$

So Erik is about 5'9.

Example: Percentiles

What is the adult male height at the 82nd percentile?

As always, we begin by drawing our picture.

Example: Percentiles

What is the adult male height at the 82nd percentile?

We need to find the Z-score at the 82nd percentile

- This will be a positive value and can be found using software as $z = 0.92$.

Example: Percentiles

What is the adult male height at the 82nd percentile?

Finally, the height x is found using the Z-score formula with the known mean $\mu = 70$, standard deviation $\sigma = 3.3$, and Z-score $z = 0.92$:

$$0.92 = z = \frac{x - \mu}{\sigma} = \frac{x - 70}{3.3}$$

and so $x = 0.92 \times 3.3 + 70 = 73.04$

Example: Percentiles

What is the adult male height at the 50th percentile?

As always, we begin by drawing our picture.

The 50th Percentile

- When we talked about measures of center, we noted that the 50th percentile is the median.
- Because the normal distribution is *symmetric*, the mean and median will be equal.
- This means that for the normal distribution the 50th percentile will always be μ.

Example

Adult male heights follow $N(70.0, 3.3)$.

1. What is the probability that a randomly selected male adult is at least 6'2 (74 inches)?
2. What is the probability that a male adult is shorter than 5'9" (69 inches)?

Let's start by drawing a picture for each.

Example

Adult male heights follow $N(70.0, 3.3)$. What is the probability that a randomly selected male adult is at least 74 inches?

First, we calculate the Z-score:

$$z_{74} = \frac{74 - 70}{3.3} = 1.21$$

Using software, the left tail area is 0.8869, but we want the probability that he is *at least* 74 inches:

$$1 - 0.8869 = 0.1131$$

Example

Adult male heights follow $N(70.0, 3.3)$. What is the probability that a male adult is shorter than 69 inches?

First, we calculate the Z-score:

$$z_{74} = \frac{69 - 70}{3.3} = -0.30$$

Using software, the left tail area is 0.3821. We want the probability that he is shorter than 69 inches, so this is the value we want.

Interval Probabilities

What is the probability that a random adult male is *between* 69 and 74 inches?

First, let's draw a picture. We will compare this picture to the two from the previous example.

Interval Probabilities

What is the probability that a random adult male is *between* 69 and 74 inches?

The total area under the curve is 1. We've already calculated

$$P(\text{height} > 74)$$

and

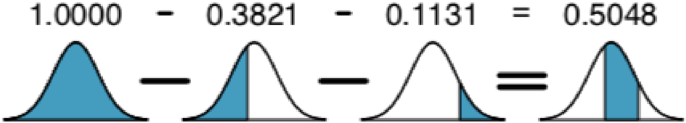
$$P(\text{height} < 69).$$

We want to calculate

$$P(69 < \text{height} < 74).$$

Interval Probabilities

We can use our drawings to visualize what we want to calculate:



So the probability of being between 69 and 74 inches tall is about 50.5%.

Example

SAT scores follow $N(1100, 200)$. What percent of SAT takers get between 1100 and 1400?

We'll start with a picture.

Example

We want the area between the two tails, so we are going to calculate the tail areas and then subtract them from one.

We'll start with $P(\text{score} < 1100)$. SAT scores follow $N(1100, 200)$.

- Notice that this is the mean.
- We know that for the normal distribution, the mean and median are the same.
- So we know that this is the 50th percentile.

So, $P(\text{score} < 1100) = 0.5$

Example

We want the area between the two tails, so we are going to calculate the tail areas and then subtract them from one.

Now we'll examine $P(\text{score} > 1400)$. SAT scores follow $N(1100, 200)$. The Z-score is

$$z = \frac{1400 - 1100}{200} = 1.5$$

Using `R`, the corresponding percentile is 0.9332, but we want the upper tail:

$$1 - 0.9332 = 0.0668$$

Example

Finally, we will subtract both of these tail probabilities from one to get the area between the two percentiles:

$$1 - 0.5 - 0.0668 = 0.4332$$

So 43.32% of SAT takers get scores between 1100 and 1400.

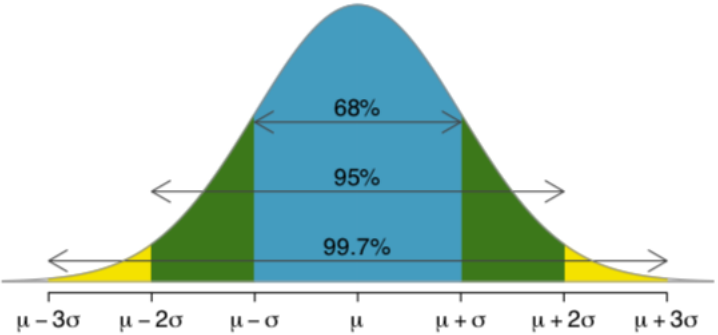
The 68-95-99.7 Rule

The 68-95-99.7 Rule is a good general rule for thinking about the normal distribution.

- 68% of the observations will fall within 1 standard deviation of the mean
- 95% of the observations will fall within 2 standard deviations of the mean
- 99.7% of the observations will fall within 3 standard deviations of the mean

This can be useful when trying to make a quick Z-score estimate without access to software.

The 68-95-99.7 Rule



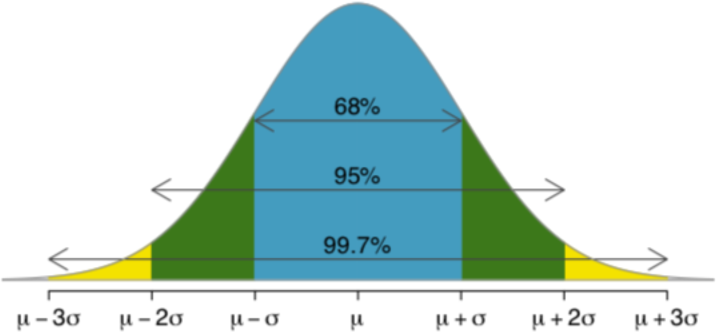
Outliers

We can also use Z-score and the 68-95-99.7 Rule to look for outliers.

- We expect 95% of the observations to fall within 2 standard deviations, so observations outside of this are *unusual*.
- We expect 99.7% of the observations to fall within 3 standard deviations, so observations outside of this are very unusual or *outliers*.

We can certainly have observations outside of 3 or 4 standard deviations from the mean, but the probability of being further than 4 standard deviations from the mean is about 1-in-15,000.

The 68-95-99.7 Rule



We will confirm these probabilities in Lab 5.