

5.3 The Normal Distribution

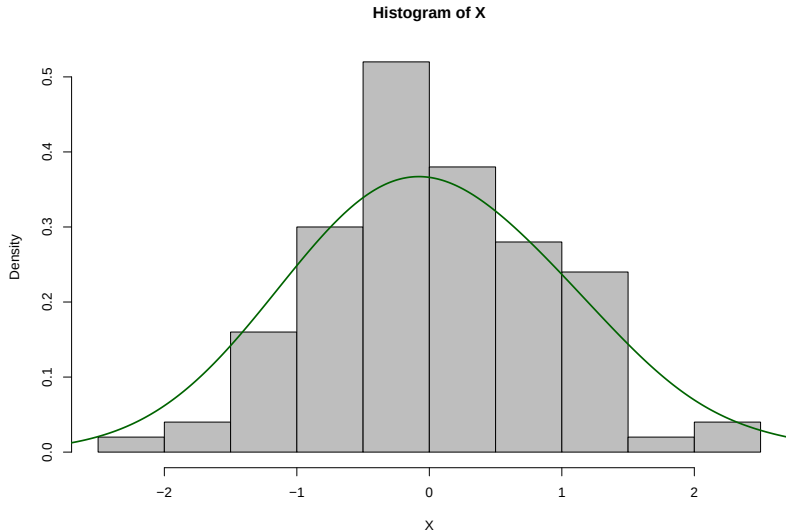
Dr. Lauren Perry

Goals

1. Use z scores to compare observations on different scales.
2. Calculate probabilities for a normal distribution using area under the curve.
3. Calculate normal distribution percentiles.

Density Curves

We represent the shape of a continuous variable using a **density curve**. This is like a histogram, but with a smooth curve:



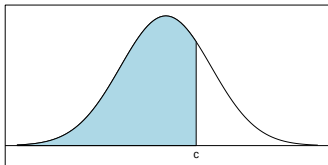
Density Curves

Properties:

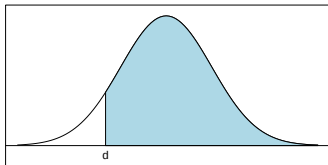
1. The curve is always above the horizontal axis (because probabilities are always nonnegative).
2. The total area under the curve equals 1.

Density Curves

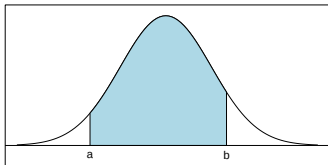
The proportion of all possible observations that lie within a specified range equals the corresponding area under the density curve.



$$P(X < c)$$



$$P(X > d)$$



$$P(a < X < b)$$

Normal Curves

- ▶ A **normal curve** is a special type of density curve that has a “bell-shaped” distribution.
- ▶ A variable is **normally distributed** or has a **normal distribution** if its distribution has the shape of a normal curve.

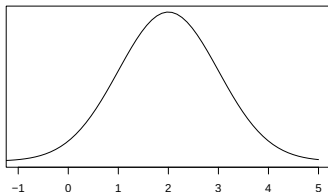
Why “normal”? Because it appears so often in practice!

Normal Distributions

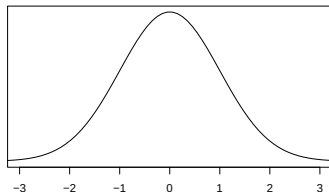
Normal distributions. . .

- ▶ are fully determined by parameters mean μ and standard deviation σ .
- ▶ are symmetric and centered at μ .
- ▶ have spreads that depend on σ .

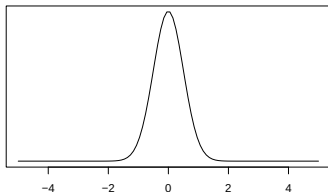
Normal Distributions



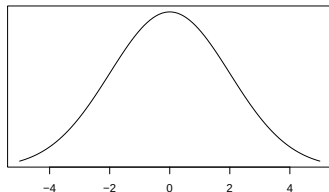
Normal($\mu = 2$, $\sigma = 1$)



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



Normal($\mu = 0$, $\sigma = 0.5$)



Normal($\mu = 0$, $\sigma = 2$)

Normal Distributions

To check whether a variable is (approximately) normally distributed,

1. Check the histogram to see if it is symmetric and bell-shaped.
2. Estimate the parameters: μ using $\bar{x}$ and σ using s .

Z-Scores

- ▶ We standardize a variable using

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

- ▶ This is also called a **z-score**.
- ▶ Standardizing using this formula will *always* result in a variable with mean 0 and standard deviation 1.
- ▶ These are useful for comparing values which are originally on very different scales.

Note about Z-Scores

A z-score tells us how many standard deviations an observation is from the mean.

- ▶ A positive z-score $z > 0$ is *above* the mean.
- ▶ A negative z-score $z < 0$ is *below* the mean.

Example: $z = -0.23$ is 0.23 standard deviations below the mean.

Checkpoint

- ▶ ACT scores have mean 20.8 and standard deviation 5.8.
- ▶ SAT scores have mean 1500 and standard deviation 300.
- ▶ If Jose scored a 27 on his ACT and Alice scored a 1850 on her SAT, who would you say got the better score?

Empirical Rule for Variables

For any (approximately) normally distributed variable,

1. Approximately 68% of all possible observations lie within one standard deviation of the mean: $\mu \pm \sigma$.
2. Approximately 95% of all possible observations lie within two standard deviations of the mean: $\mu \pm 2\sigma$.
3. Approximately 99.7% of all possible observations lie within three standard deviations of the mean: $\mu \pm 3\sigma$.

The Standard Normal Distribution

- ▶ A **standard normal distribution** is a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.
- ▶ If X is approximately normal, then the standardized variable Z will have a standard normal distribution.

Note: when we z-score a variable, we preserve the area under the curve properties!

► If X is Normal(μ, σ), then

$$P(X < c) = P\left(Z < \frac{c - \mu}{\sigma}\right) = P(Z < z).$$

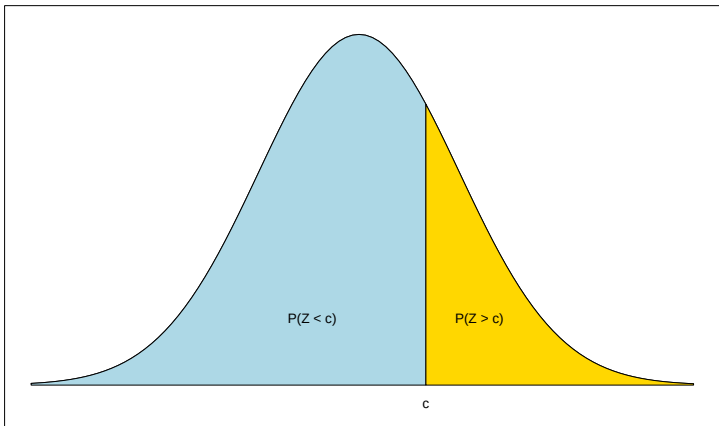
Area Under the Standard Normal Curve

Properties:

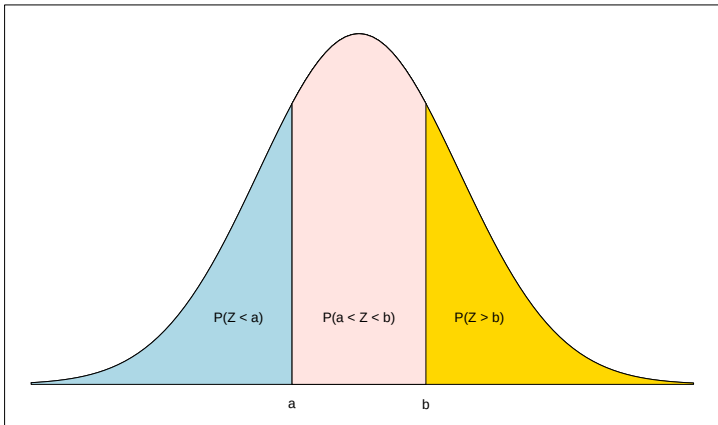
1. Total area under the curve is 1.
2. The curve extends infinitely in both directions, never touching the horizontal axis.
3. Symmetric about 0.
4. Almost all of the area under the curve is between -3 and 3.

We work with **cumulative probabilities** or probabilities of the form $P(Z < z)$.

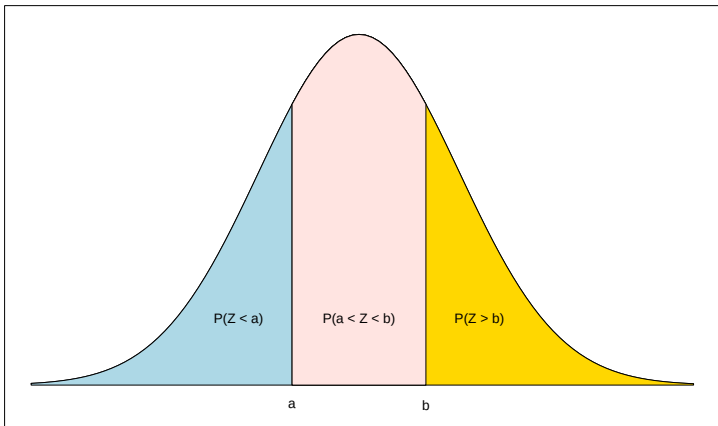
We will use the fact that the total area under the curve is 1 to find probabilities like $P(Z > c)$:



We can also use this concept to find $P(a < Z < b)$.



Total area = 1



Total area = 1

Notice that

$$1 = P(Z < a) + P(a < Z < b) + P(Z > b)$$

From

$$1 = P(Z < a) + P(a < Z < b) + P(Z > b)$$

we can write

$$P(a < Z < b) = 1 - P(Z > b) - P(Z < a)$$

Since we just found that

$$P(Z > b) = 1 - P(Z < b)$$

we can replace $1 - P(Z > b)$ with $P(Z < b)$, and get

$$P(a < Z < b) = P(Z < b) - P(Z < a).$$

Key Cumulative Probability Concepts

- ▶ $P(Z > c) = 1 - P(Z < c)$
- ▶ $P(a < Z < b) = P(Z < b) - P(Z < a)$
- ▶ The normal distribution is symmetric, so
 $P(X < \mu) = P(X > \mu) = 0.5$.

Now that we can get all of our probabilities written as *cumulative* probabilities, we're ready to use software to find the area under the curve!

Finding Area Under the Curve: Applets

Rossman and Chance Normal Probability Calculator

Example

ACT scores are well-approximated by a normal distribution with mean 20.8 and standard deviation 5.8.

1. Find the probability that a randomly selected test taker scores above a 23.
2. Find the probability that a randomly selected test taker scores between an 18 and a 22.

Determining Normal Distribution Probabilities

1. Sketch the normal curve for the variable.
2. Shade the region of interest and mark its delimiting x -value(s).
3. Find the z -score(s) for the value(s).
4. Use an applet (or the `pnorm` command in R) to find the associated area.

Example

Find the proportion of SAT-takers who score between 1150 and 1300. Assume that SAT scores are approximately normally distributed with mean $\mu = 1100$ and standard deviation $\sigma = 200$.

Rossman and Chance Normal Probability Calculator

Percentiles

We can also find the *observation* associated with a percentage/proportion.

Recall: The w th **percentile** p_w is the observation that is higher than $w\%$ of all observations

$$P(X < p_w) = w$$

Finding a Percentile

1. Sketch the normal curve for the variable.
2. Shade the region of interest and label the area.
3. Use the applet (or R - see below) to determine the z-score for the area.
4. Find the x-value using z , μ , and σ .

Note that if $z = \frac{x-\mu}{\sigma}$, then $x = \mu + z\sigma$.

Example

SAT scores are approximately Normal($\mu = 1100$, $\sigma = 200$). Find the 90th percentile for SAT scores.

Rossman and Chance Normal Probability Calculator