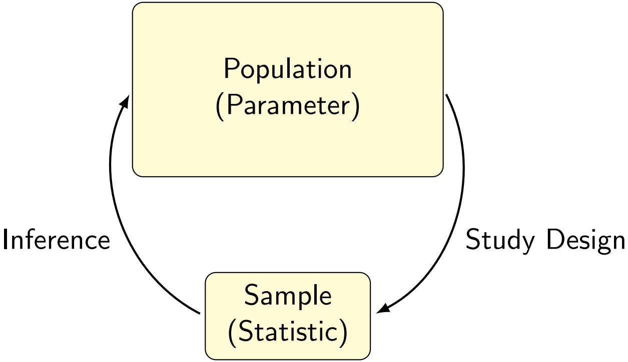
Normal Distributions

Grinnell College

March 24, 2025

Review – Inference



BIG IDEA: Parameter value is unknown → we use the statistic to estimate it

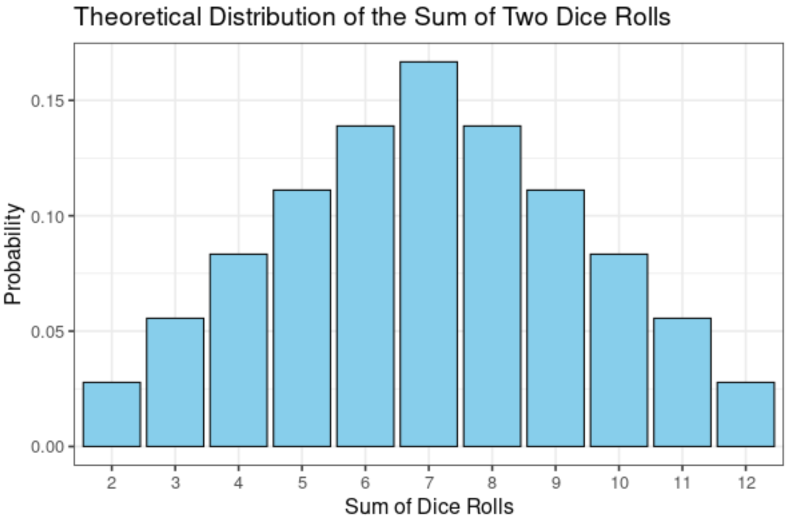
Distributions

Recall that a **distribution** tells us:

- What values
- How frequently

Most distributions are governed by **distributional parameters**: if we know these, we know everything we can about the data-generating process

Dice Sum	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$



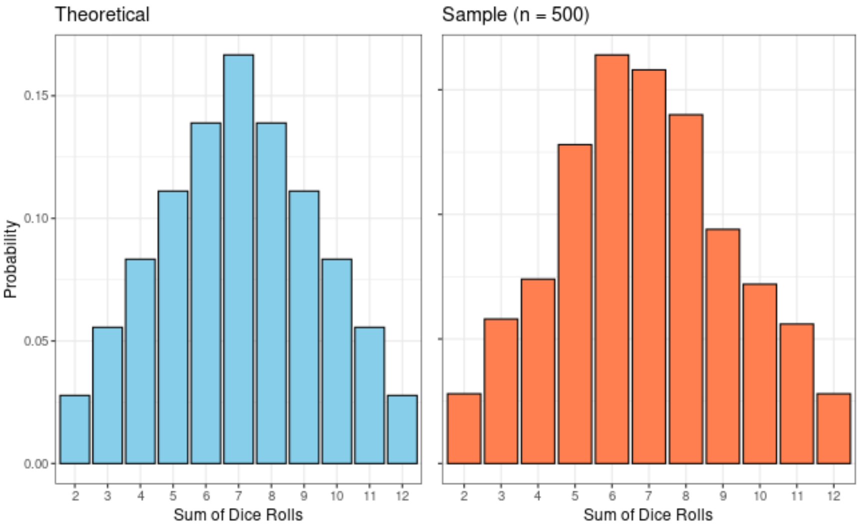
Random Samples

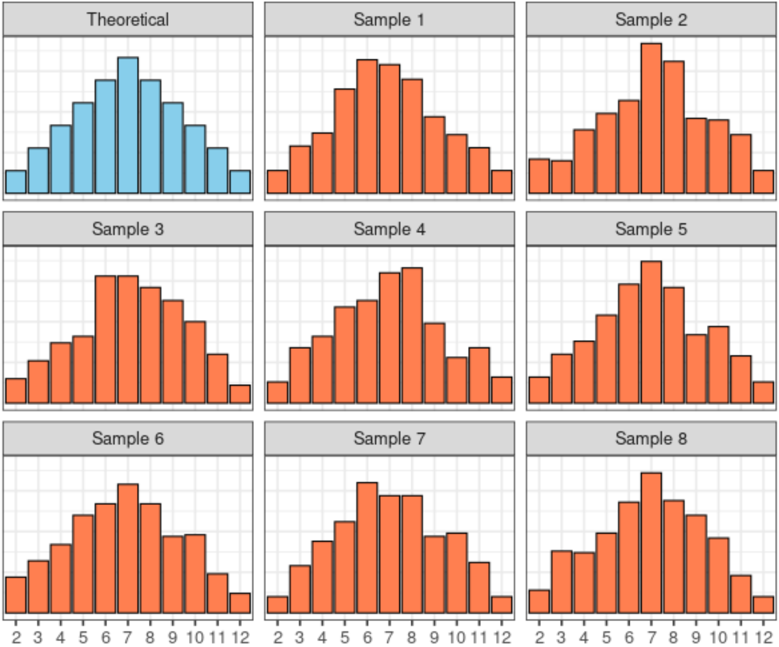
When we don't know the distributional parameters, we are instead required to take a sample from the population. If done correctly, this sample should be **representative**

The goal of any sample is to compute a statistic and perform **inference** on the unknown population parameter

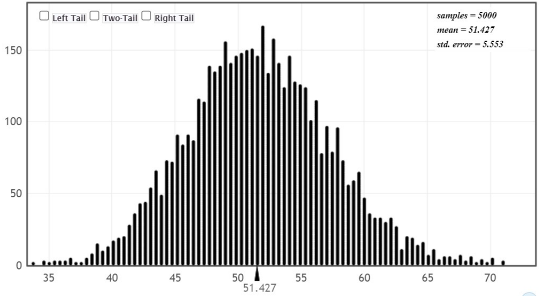
Sampling is a **random process**, and this randomness will be reflected in the values of the statistics we are able to compute

Dice Sum	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$




Theoretical
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
2 3 4 5 6 7 8 9 10 11 12
2 3 4 5 6 7 8 9 10 11 12
2 3 4 5 6 7 8 9 10 11 12

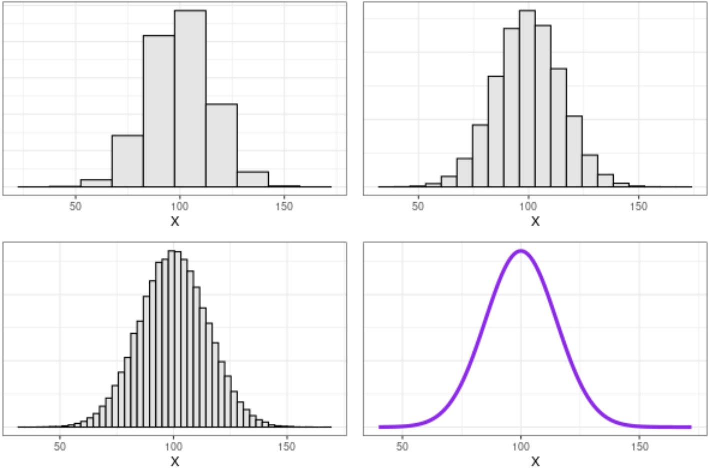
Bell-shaped Distribution



Here is another example of the shape we just saw a lot of. The distribution of a sample of Hollywood movie budgets is given above. We have previously called this unimodal and symmetric (bell-shaped)

The shape we have seen over and over in the previous slides is something we are going to see come up a lot from here on. We are going to give it a special name, and see what we can do with it: **Normal Distribution**

The Normal Distribution



Central Limit Theorem

One of the natural questions we may ask is *why* does the Normal distribution come up so often? The answer arises from something called the Central Limit Theorem (CLT), which is a mathematical rule.

CLT

The sum of many independent random variables will *approximately* follow a Normal distribution.

- We saw this just a moment ago with the sum of 2 dice
- We will see more on this in a few days

Many things that occur in nature are the result of many independent random events occurring over time to give us an outcome → Normal distribution

Normal Distribution

It turns out we only need to know two things in order to completely describe the Normal distribution

1. the mean (μ)
2. the standard deviation (σ) or variance (σ^2)

These will tell us where the center of the normal distribution is and how stretched out it should be.

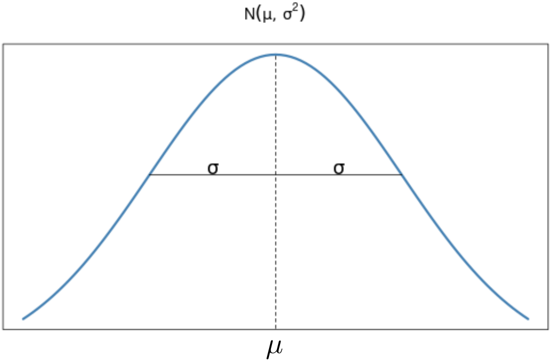
If a variable looks like a normal distribution, we will often use the following notation to say that:

- X $\sim$ N(μ, σ^2)

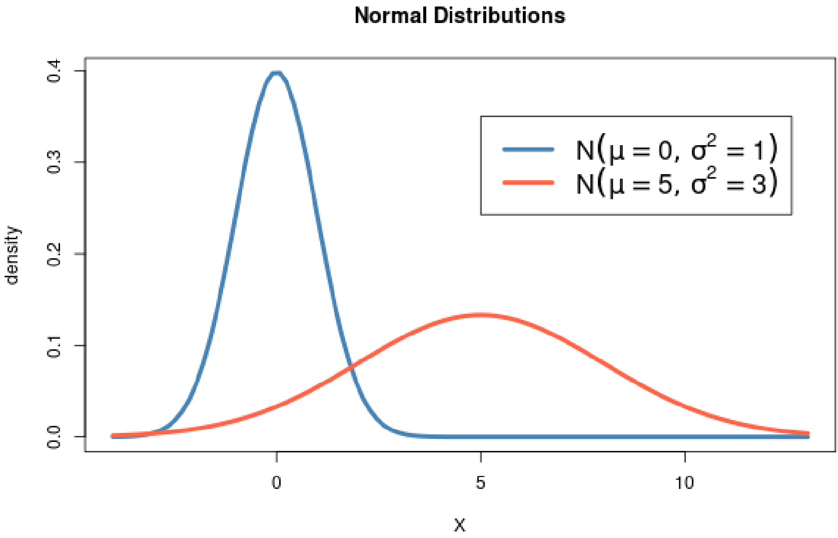
Normal Distribution

$X \sim N(\mu, \sigma^2)$

- the mean tells us where the center of the normal distribution is
- the variance tells us how spread out the distribution is



Examples


Normal Distributions
$N(\mu = 0, \sigma^2 = 1)$
$N(\mu = 5, \sigma^2 = 3)$
density
0.0
0.1
0.2
0.3
0.4
0
5
10
X

Standard Normal Distribution

When a normal distribution has mean zero and variance equal to 1, we call it a **Standard Normal Distribution** and write X $\sim$ N(0, 1).

Why? It's related to standardizing a variable like we did with Z-scores.

Suppose the variable X $\sim$ N(μ, σ^2),
then $Y = \frac{X-\mu}{\sigma} \sim$ N($\mu = 0$, $\sigma^2 = 1$)

In other words, if we standardize a normal variable (with any mean and variance) then we get back a normal variable that has $\mu = 0$ and $\sigma^2 = 1$

Probabilities

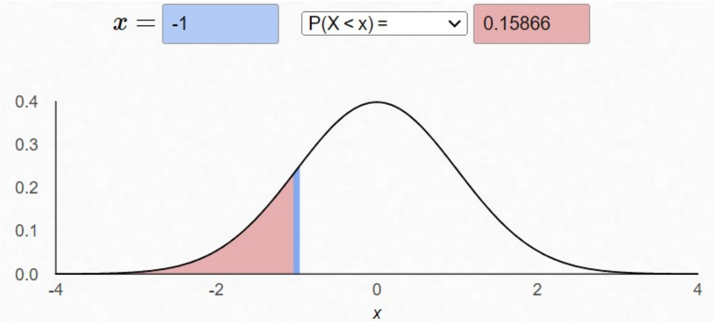
Probabilities
If our population follows a normal distribution... we can pick a case at random from our population

- probability the observation is less/greater than some value?
- probability the observation is between two values?

Note: It turns out that using a normal distribution we cannot find the probability of the case having a *specific* value, we can only use ranges of values.

Probabilities – Less than

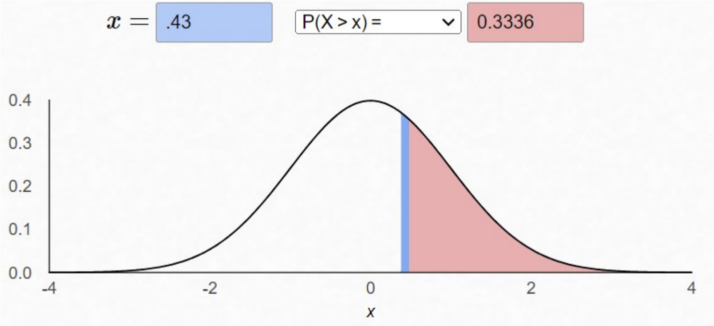
Standard Normal: $X \sim N(0, 1)$
Probability a randomly selected observation is below (less than) **-1**?



We can write this using our probability notation: $P(X < -1) = 0.15866$

Probabilities – Greater than

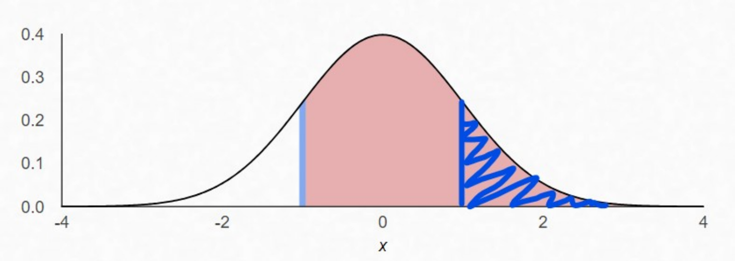
Standard Normal: X $\sim$ N(0, 1)
Probability a randomly selected observation is above (greater than) **0.43**?



$P(X > 0.43) = 0.3336$

Probabilities – Between

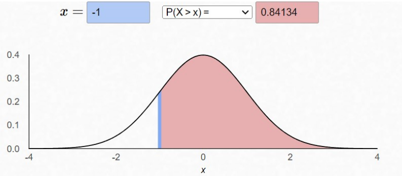
Standard Normal: X $\sim$ N(0, 1)
What about the probability that a case falls *between* -**1** and **1**?

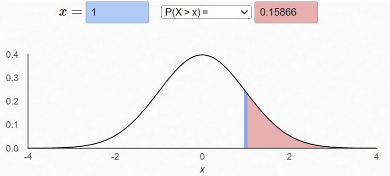


We need to do a bit more work...

Probabilities – Between

Standard Normal: $X \sim N(0, 1)$
What about the probability that a case falls *between* -**1** and **1**?





We can chop off the extra probability we don't need that is above **1**.

$$P(X \text{ is between -1 and } 1) = P(-1 < X < 1) = P(X > -1) - P(X > 1)$$
$$= 0.84134 - 0.15866 = 0.68286$$

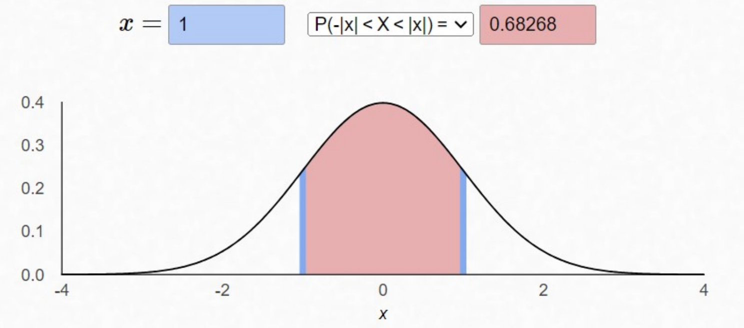
Probabilities – Between

When the values we are looking at are the same but just with different signs (like -1 and $+1$)

- We can write them in a specific way
- There is a shortcut on the app for getting the probability

Probabilities – Between

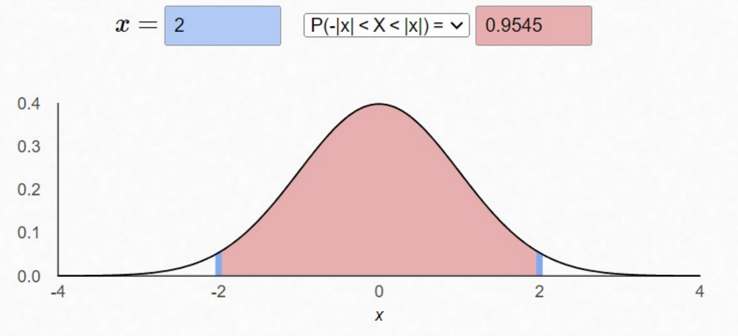
Standard Normal: X $\sim$ N(0, 1)
What about the probability that a case falls *between* **-1** and **1**?



$$P(|X| < 1) = 0.68286$$

Probabilities – Between

Standard Normal: $X \sim N(0, 1)$
What about the probability that a case falls *between* **-2** and **2**?

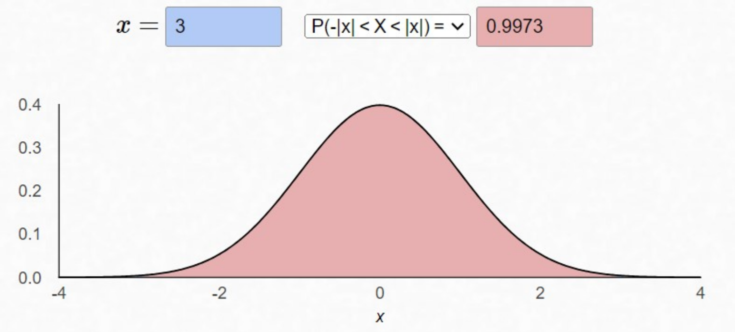


$$P(|X| < 2) = 0.9545$$

Probabilities – Between

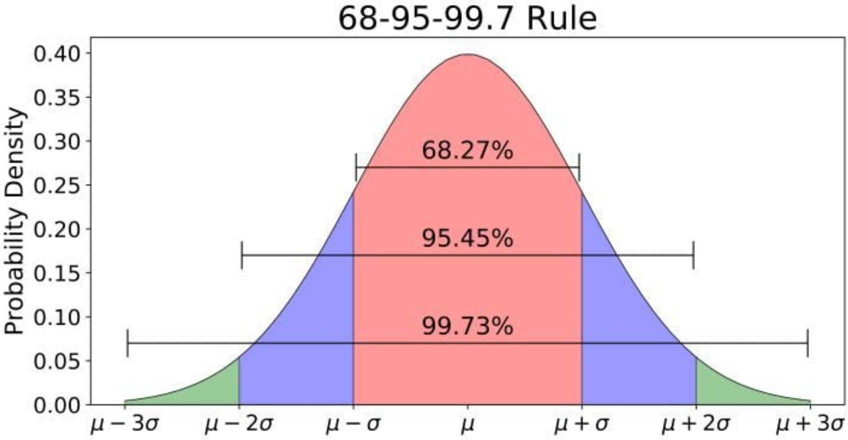
Standard Normal: X $\sim$ N(0, 1)

What about the probability that a case falls *between* -**1** and **1**?



$$P(|X| < 3) = 0.9973$$

Summary



68-95-99.7 Rule
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
68.27%
95.45%
99.73%
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

Probabilities from R

We can use the "pnorm()" function in R to get these probabilities.

- tell the function what number you are trying to find the probability more/less than
- tell the function the value of the mean
- tell the function the value of the std. dev.

Note: By default R will try to give you 'less than' probabilities (also called lower tail probabilities). To get 'greater than' probabilities, put "Lower.Tail=FALSE" into the pnorm() function.

```
> pnorm(-1, mean=0, sd=1)
[1] 0.1586553
> pnorm(-1, mean=0, sd=1, lower.tail = FALSE)
[1] 0.8413447
> pnorm(-1, mean=0, sd=1, lower.tail = FALSE)
- pnorm(1, mean=0, sd=1, lower.tail = FALSE)
[1] 0.6826895
```

Summary

We learned a bit about the Normal distribution!

- what it looks like
- how to find probabilities with it
  - less than a value
  - more than a value
  - between values

Central Limit Theorem tells us that for large samples, taking the average (or sum) of a whole bunch of random variables will (approximately) follow a Normal distribution