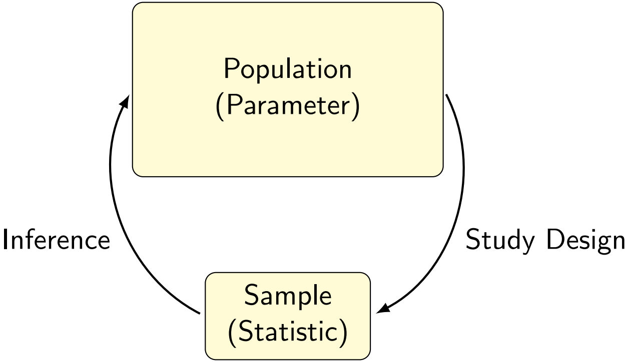
Normal Distributions

Grinnell College

October 14, 2024

Review – Inference

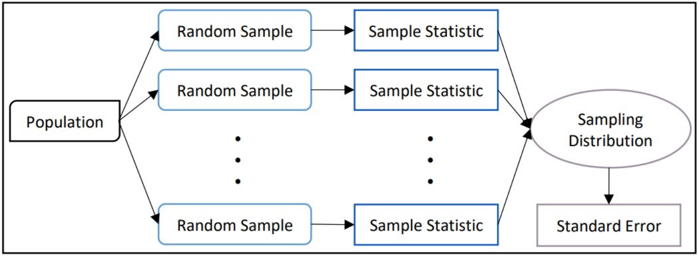


BIG IDEA: Parameter value is unknown $\rightarrow$ we use the statistic to estimate it

Review – Sampling Distribution

If we had the ability to make many different samples we could plot the statistics from each.

- This gives us an idea of the variability of the statistics



The **Standard Error** is the std. dev. of the sampling distribution

- measures variability of statistics

Sampling Distribution

To make the sampling distribution, we had to take a whole lot of different samples.

- Are there any issues with this?
- Would you actually want to go and take 5,000 different samples?

What now?

Ok, so we can't just go and take a whole bunch of random samples...

This means we can't get the standard error!

- so we can't actually quantify how far the statistic away is? Wasn't that the whole point?!

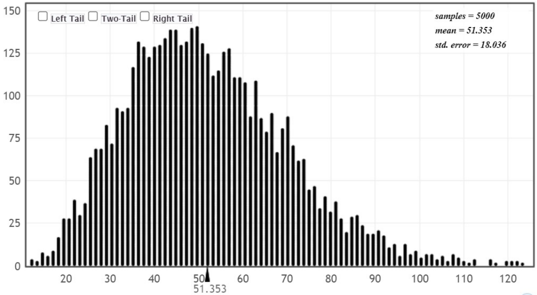
What the heck do we do now?

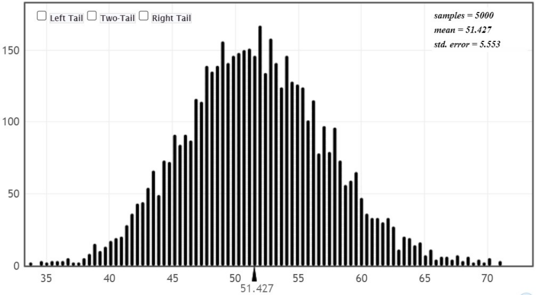
Sampling Distribution Shape

All hope is not lost. Think back to the shape of the sampling distribution.

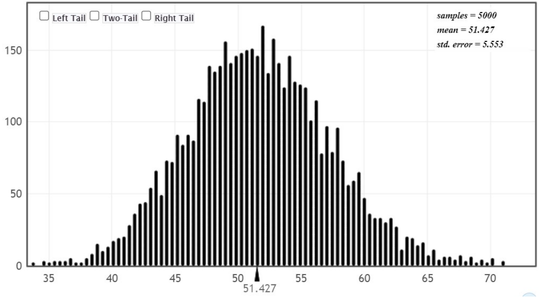
Big question: What happened to the shape of the sampling distribution as the sample size increased?

Movie Budgets Example





Bell-shaped Distribution

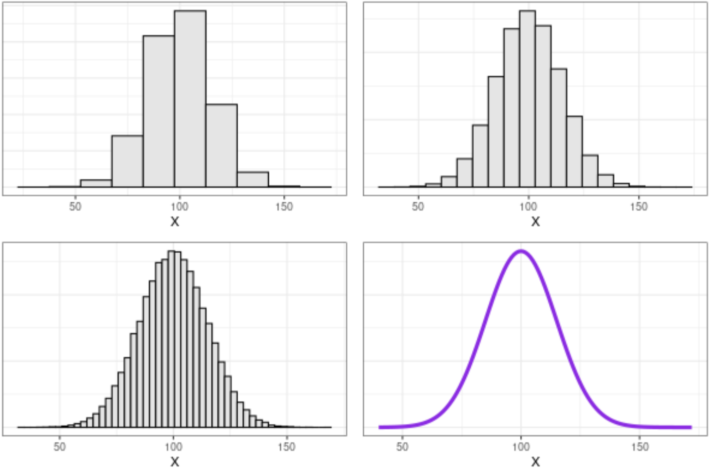


The bell-shaped distribution we see in the sampling distribution for Movie Budgets is something that happens a lot.

It turns out there is a reason for that, which we will cover shortly.

For now, we are going to give it a special name, and see what we can do with it.

The 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) <u>or</u> variance (σ^2)

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

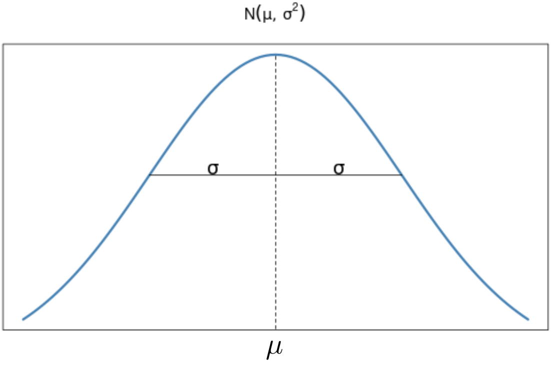
If a variable looks like a normal distribution, we will often use the following notation to say that:

- X $\sim$ N(μ, σ^2)

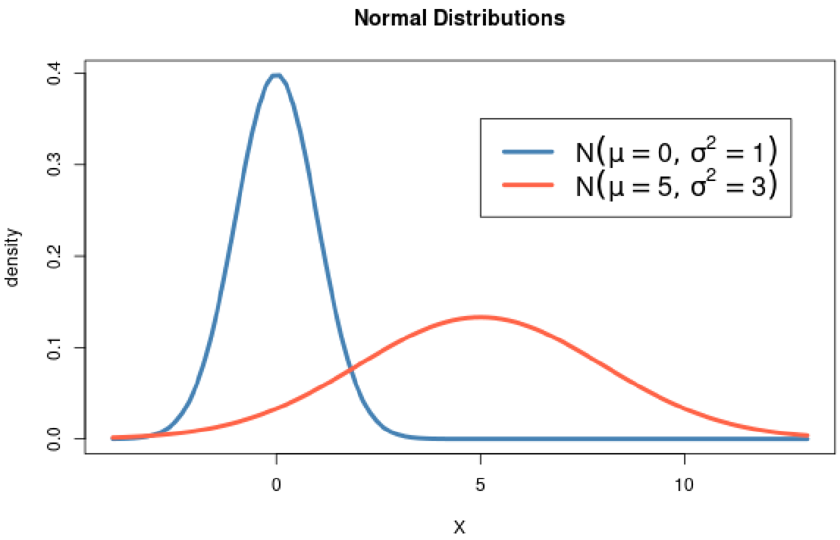
Normal Distribution

X $\sim$ N(μ, σ^2)

- the <u>mean</u> tells us where the center of the normal distribution is
- the <u>variance</u> 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 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

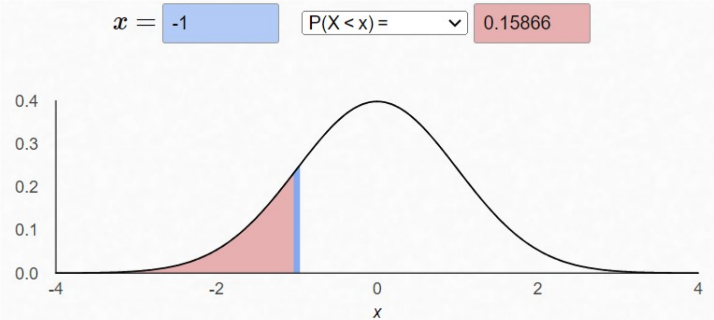
If our population follows a normal distribution... we can pick a case at random from our population

- 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

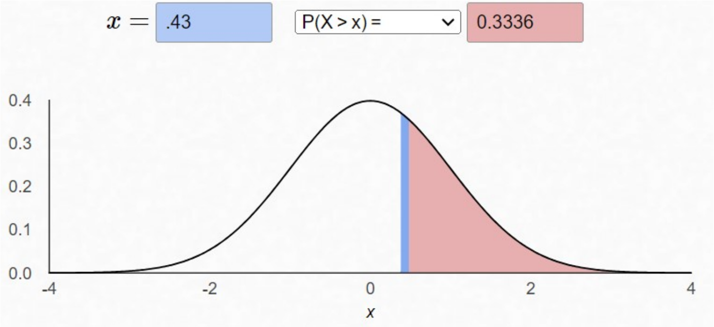
Standard Normal: $X \sim N(0, 1)$
Probability a randomly selected observation is below (less than) **-1**?



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

Probabilities – Greater than

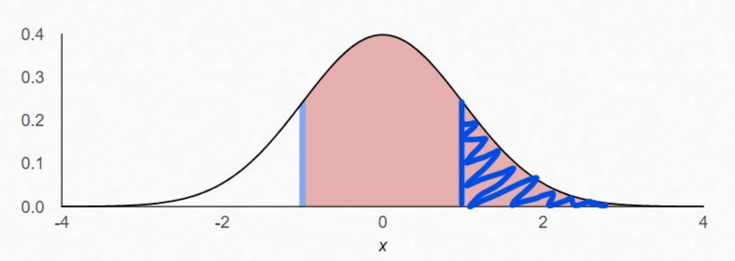
Standard Normal: X $\sim$ N(0, 1)
Probability a randomly selected observation is above (greater than) **0.43**?



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**?

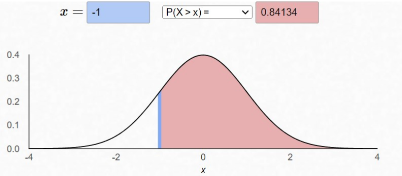


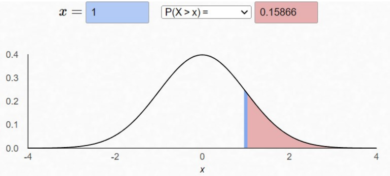
We need to do a bit more work...

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**.

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

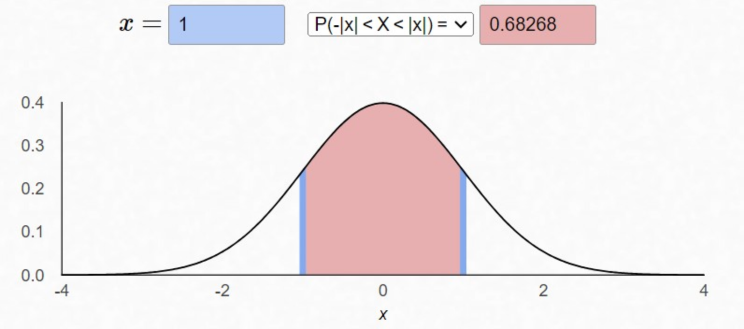
Probabilities – Between

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

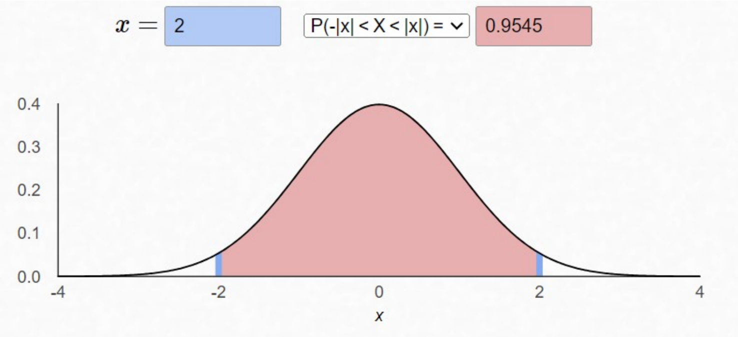
Standard Normal: $X \sim N(0, 1)$
What about the probability that a case falls *between* **-1** and **1**?



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

Probabilities – Between

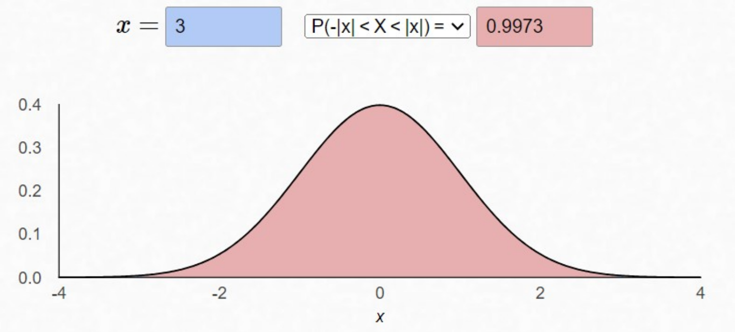
Standard Normal: X $\sim$ N(0, 1)
What about the probability that a case falls *between* **-2** and **2**?



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

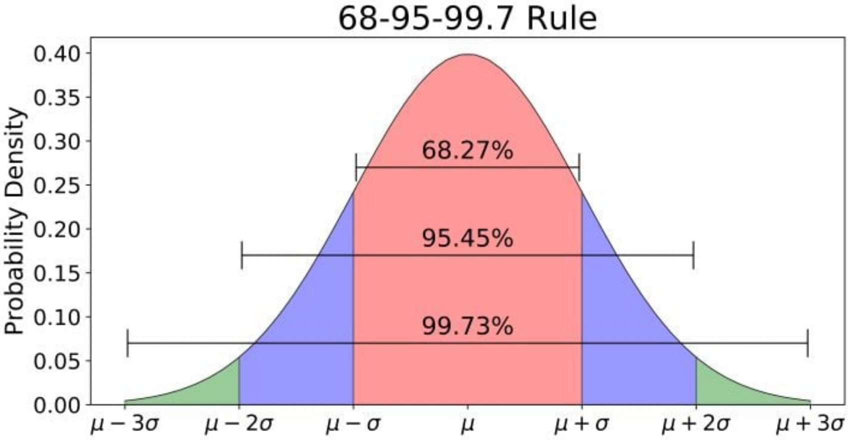
Probabilities – Between

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
```

Central Limit Theorem

The **Central Limit Theorem (CLT)** is (possibly) the most important result in all of statistics. It states:

1. If variable X has mean μ and std.dev. σ, and
2. If the number of observations in the sample (n) is large
3. then the sampling distribution for $\overline{X}$ (sample mean) is Normal with mean μ and standard error $\sigma/\sqrt{n}$.

$$\overline{X} \sim \mathrm{N}(\mu, \, \sigma^2/n)$$

Central Limit Theorem

Important bits:

- CLT doesn't require the pop. distribution look Normal
- What is considered large?
  - A recommendation for being "sufficiently large" when working with means is often to have at least 30 cases in your sample
  - If the data are approximately normal or symmetric, a smaller sample size (10 to 20) may be sufficient
  - If the data are skewed and/or have extreme outliers, the sample size may need to be higher than 30; possible more than 45. If the skew and outliers are very extreme, the sample size may need to be higher than around 200

Summary

We learned a bit about the Normal distribution!

- what it looks like
- how to find probabilities with it
- how it relates to the sampling distribution (CLT)

Central Limit Theorem tells us that for large samples $\overline{X} \sim \mathrm{N}(\mu, \sigma^2/n)$

We don't need to take 5,000 samples to get the **Standard Error** any more! We have a formula:

- $\mathrm{SE} = \sigma/\sqrt{n}$