CONFIDENCE INTERVALS

- Consider the example in previous video, where the mean of the Netflix viewing hours was 6 hours, the standard deviation was 1,2 hours, and the samples of size N=100.

- We know that the best estimate for the average of the sample mean is 6 hours, but it is very likely that for a particular sample, the value will be different.

- Intuitively the question is: if I take a simple sample, in which range of values can I expect to find my sample average?

That question is not stated very well.

We will prefer to ask the following:

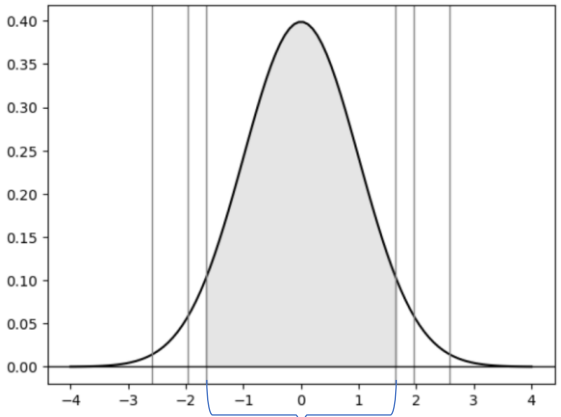
- What is a 95% confidence interval in which I will find my sample mean?

- It can be rephrased as follows: Suppose that I draw many random samples of size N, each of them giving me a different sample average. I would like to select an interval of values such that 95 % of the times, the measured sample average is within that interval.

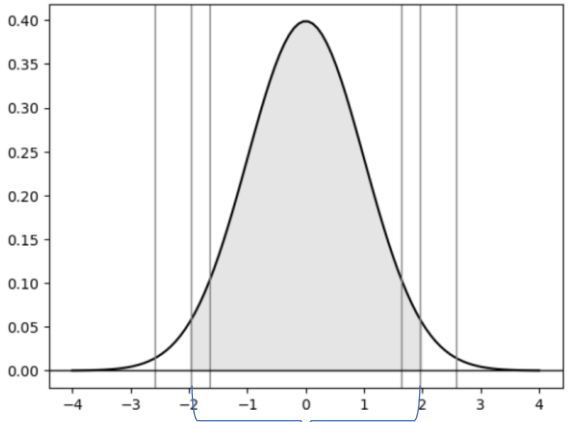
- Smaller interval => less chance to be there => less confidence (but more precision)
- Biggest interval => more chance to be there => more confidence (but less precision)

Important fact to remember: The distribution of samples means (when N is large) is approximately a normal distribution.

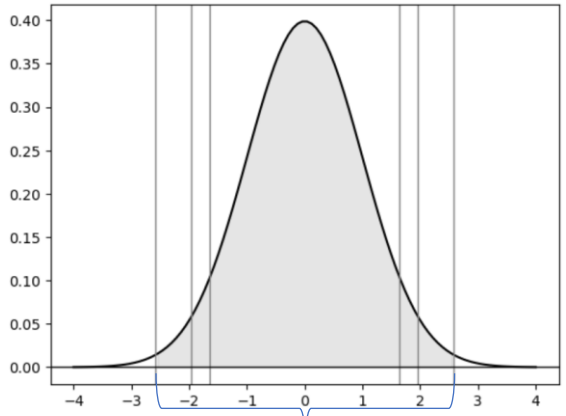
Confidence intervals for the normal distribution



90% of data between
-1.645 and 1.645 std



95% of data between
-1.96 and 1.96 std



99% of data between
-2.58 and 2.58 std

Confidence intervals for the standard normal distribution

In Python is very easy to find any interval:

```
from scipy import stats
ci_75=stats.norm.interval(0.75)
ci_80=stats.norm.interval(0.8)
ci_90=stats.norm.interval(0.9)
ci_95=stats.norm.interval(0.95)
print('75% interval: ', ci_75)
print('80% interval: ', ci_80)
print('90% interval: ', ci_90)
print('95% interval: ', ci_95)
```

```
75% interval:  (-1.1503493803760079, 1.1503493803760079)
80% interval:  (-1.2815515655446004, 1.2815515655446004)
90% interval:  (-1.6448536269514729, 1.6448536269514722)
95% interval:  (-1.959963984540054, 1.959963984540054)
```

Confidence intervals for nonstandard normal distributions

Suppose that I am sampling data from a distribution with mean=30, standard deviation=5 and the size of each sample N=50

What is a 95% confidence interval for the sample mean?

Step 1: Find the value for the 95% CI of the standard normal distribution => 1.96

Step 2: Compute the standard error $se = \frac{\sigma}{\sqrt{N}} = \frac{5}{\sqrt{50}} = 0.7$

Step 3: Create an interval centered around the mean using the previous values

$$CI = [30 - 1.96 \times 0.7\ , 30 + 1.96 \times 0.7] = [\ \mathbf{28.628}\ , \mathbf{31.372}\ ]$$

Example 1

- In the Netflix example, what is a 95% CI or the sample mean?

1. Take the population mean -> $\mu = 6$ hours
2. Compute the standard error, $\text{SE} = \frac{1,2}{\sqrt{100}} = 0,12$ hours
3. Select the value corresponding to 95% ->1,96
4. Create the interval

CI = (μ – 1,96 SE, μ + 1,96 SE) = (6 – 0,2352, 6+ 0,2352) = **(5,7648, 6,2352)**

95% if the samples will have a sample mean between 5,7648 and 6,2352

These values are so common that you should remember them:

90% confidence interval: $-1.645\sigma/\sqrt{n}$ to $+1.645\sigma/\sqrt{n}$
95% confidence interval: $-1.96\sigma/\sqrt{n}$ to $+1.96\sigma/\sqrt{n}$
99% confidence interval: $-2.58\sigma/\sqrt{n}$ to $+2.58\sigma/\sqrt{n}$

For any other CI value, you can use Python for instance.