Confidence Interval: Known Sigma

```
▶ TaskLocalRNG()
```

Let us look at a normal distribution with μ = 68 and standard deviation of σ = 3: N(68, 3).

```
μ = 68
```

```
σ = 3
```

We generating 100,000 points from this distribution as shown below:

```
height_data =
▶ [71.1975, 63.8236, 71.4398, 63.3594, 65.5812, 71.8468, 66.598, 64.9816, 67.4315, 66.3028, 6
```

Point Estimates

Our goal is to estimate the population mean μ from taking a sample from the population.

If we take a sample of 10 points from this data set:

"65.28, 70.8, 67.97, 65.93, 66.79, 68.25, 67.28, 74.29, 67.28, 68.86"
and find the sample mean $\bar{x} =$ 68.27 we can see with a large enough sample size that this will be close to the population mean, $\mu =$ 68, and as the sample size increases the differnce is:

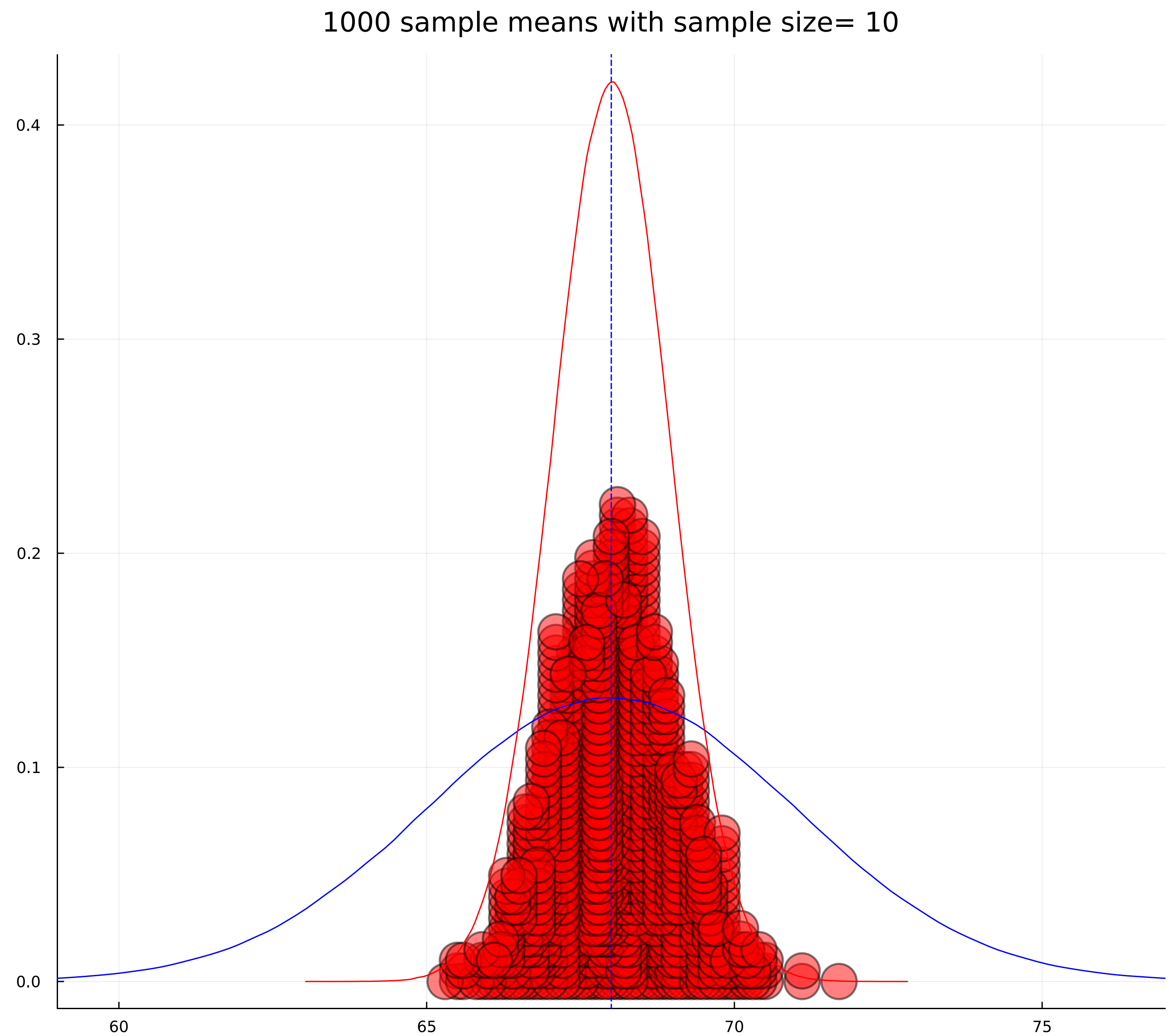
$\bar{x} - \mu =$ 68.27 - 68 $=$ 0.27

Let's take a bunch of sample means and plot them as red dots as well as the popultion mean as a dashed blue line. We can also plot the original heights distribution in blue as well as the sampling distribution in red.

Sample size 10, Max number of samples 1000

Number of sample means to plot 1000

☑ Height distribution ☑ Sampling distribution



From this we can see that the $\bar{x}$ is a good point esitmate of the population mean μ and that as the sample size increases the more we can trust this point estimate.

This leads to an important definition:
An *unbiased estimator* is a point estimate that on average equals the population parameter.

The sample mean is an unbiased estimator. We know this from the central limit theorem which states that the sampling distribution of sample means will be:

- normally distributed if the underlying distribution is normally distributed and if not then
- with a large enough sample size it will be approximately normal.

And further $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \sigma$. The key to the sample mean being an unbiased estimator is that $\mu_{\bar{X}} = \mu$ which means the sample means on average equals the population mean.

Recall that randomness being what it is, a single sample mean is not enough to conclude with a lot of confidence what the population mean is.

Suppose you were extremely unlucky and your sample of size 5 happened to grab the 5 largest values in the population:
79.13, 79.24, 80.64, 80.71, 81.01
Which has a mean $\bar{x} = 80.15$. This is clearly not a good estimate of the population mean $\mu = 68$.

Confidence Intervals

As we saw above the sample mean is sometimes closer and sometimes farther from the population mean. Instead of having a single number to estimate μ it would be nice to have a range of values that we were *reasonably sure* contained the population mean. This is our next topic.

We will come back and explain the phrase *reasonably sure* in detail later, but for now let us say that we want to be able to say that we are in some sense 80% confident that our range of values contains the mean.

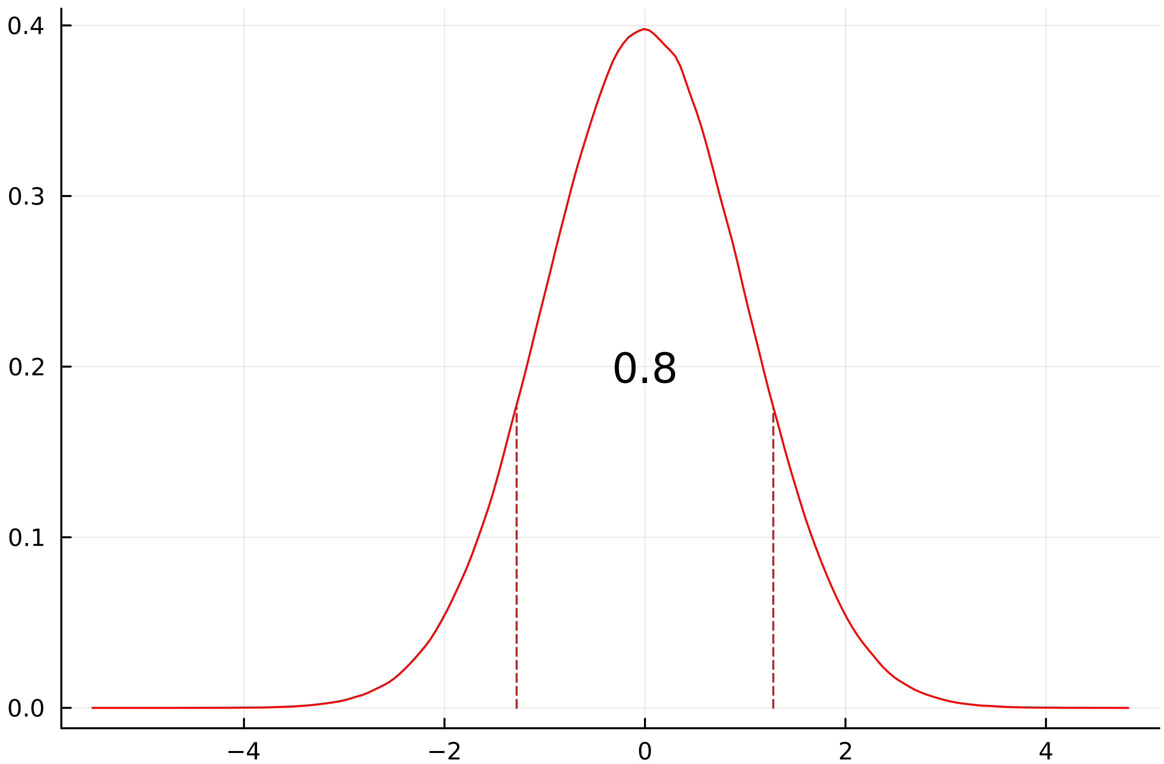
To find this range we will start by picking a center of the range and go the same distance to the left and to the right to find the endpoints. Since we are trying to estimate the population mean the most reasonable center point is the sample mean since it is the best point estimate for the population mean.

Next we need to determine how far to the left and right we need to go to be reasonably sure that the interval captures the population mean. This is sometimes called the **margin of error** (MOE). We need to find a formula for the margin of error.

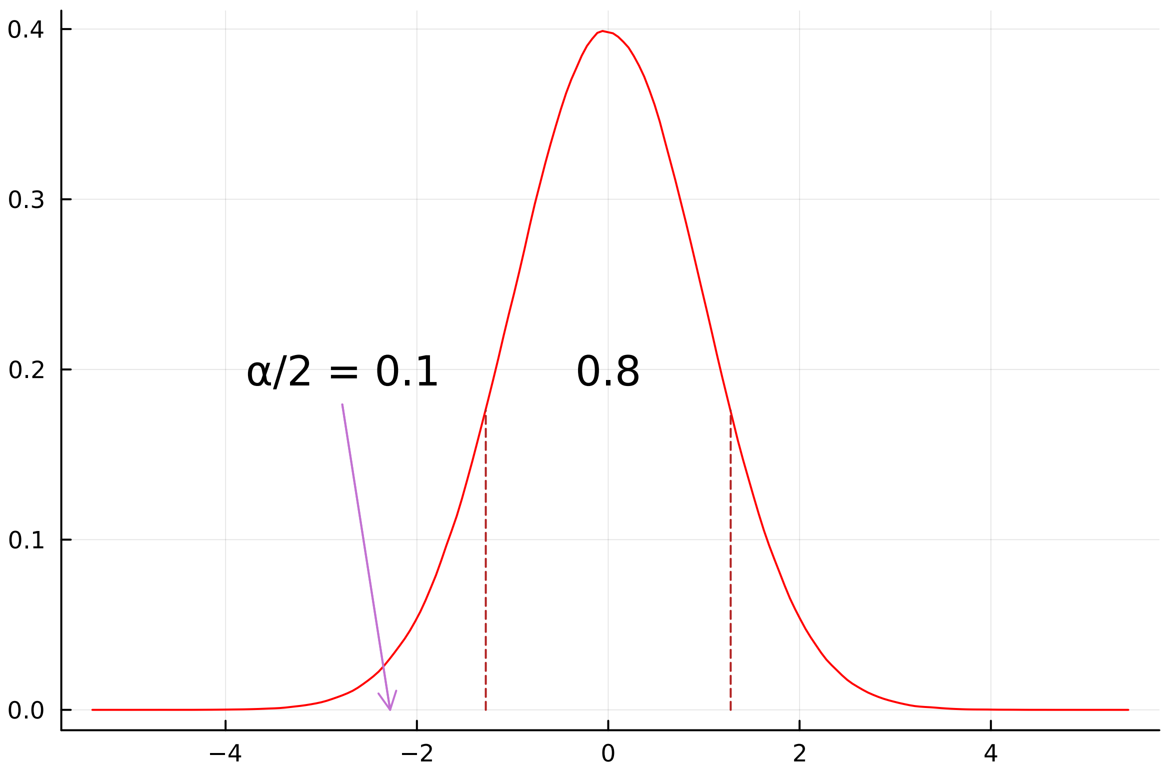
To find this formula we start by chosing a confidence level:

Confidence Level: 80%

The sampling distribution of sample means is normally distributed by the central limit theorem. What we want is shown in the following figure:



So we need to find the two z scores that will trap an area of 0.8 between them. If the area bewteen these z scores is 0.8 then the area outside of them is $\alpha = 0.2$. This is the area in both tails of the distribution. This means that the area in one tail is $\alpha/2 = 0.1$



Recall our notation that z_α is the z score such that the area to the left of it is α. So what we want to find is

$$P(-z_{0.1} < Z < z_{0.1}) = 0.8$$

This is going to get a little math intensive so feel free to jump to ❎ below.

The confidence level can be denoted by $1-\alpha$. So, in general, what we want to do is to find the z scores such that
$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$
Now let's concentrate on the expresion $-z_{\alpha/2} < Z < z_{\alpha/2}$ We know that
$z = \dfrac{\overline{x} - \mu_{\overline{X}}}{\sigma_{\overline{X}}}$
So now we have
$-z_{\alpha/2} < \dfrac{\overline{x} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} < z_{\alpha/2}$
Let us multiply both sides by $\sigma_{\overline{X}}$
$-z_{\alpha/2}\sigma_{\overline{X}} < \overline{x} - \mu_{\overline{X}} < z_{\alpha/2}\sigma_{\overline{X}}$
Multipling through by -1 flips all of the inequality signs and changes the sign of every term:
$-z_{\alpha/2}\sigma_{\overline{X}} < \mu_{\overline{X}} - \overline{x} < z_{\alpha/2}\sigma_{\overline{X}}$
Adding $\overline{x}$ to all three parts gives
$\overline{x} - z_{\alpha/2}\sigma_{\overline{X}} < \mu_{\overline{X}} < \overline{x} + z_{\alpha/2}\sigma_{\overline{X}}$
And finally replacing $\mu_{\overline{X}}$ with μ and $\sigma_{\overline{X}}$ with $\dfrac{\sigma}{n}$ we get:
$\overline{x} - z_{\alpha/2}\dfrac{\sigma}{n} < \mu < \overline{x} + z_{\alpha/2}\dfrac{\sigma}{n}$

❎ Welcome back.
So the formula for calculating a confidence interval is given by:

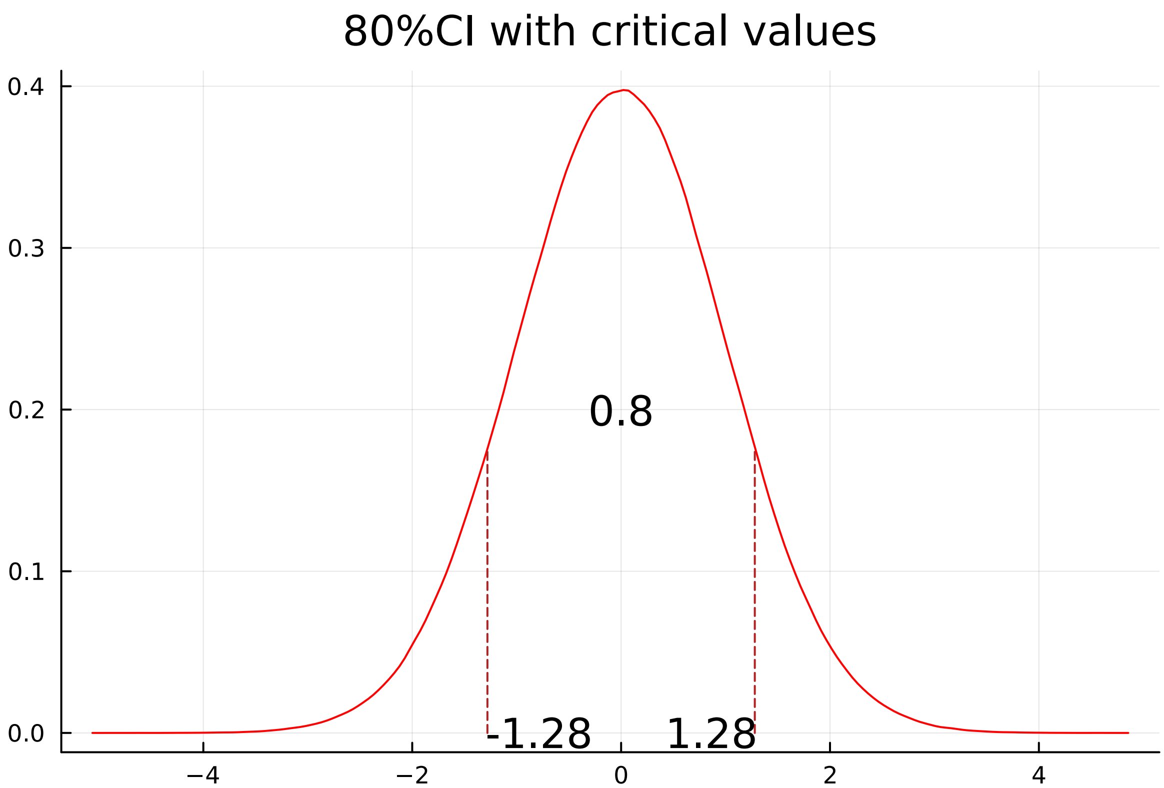
$$\overline{x} - z_{\alpha/2}\frac{\sigma}{n} < \mu < \overline{x} + z_{\alpha/2}\frac{\sigma}{n}$$

The margin of error is given by:

$$\mathrm{MOE} = z_{\alpha/2}\frac{\sigma}{n}$$

Confidence Level: 80%

$$\alpha = 0.2, \quad \alpha/2 = 0.1, \quad z_{\alpha/2} = 1.28$$



So if we know the confidence level, 80%, we will know that $\alpha/2 = 0.1$ and then we can find $z_{\alpha/2}$. We call $z_{\alpha/2}$ the critical value.

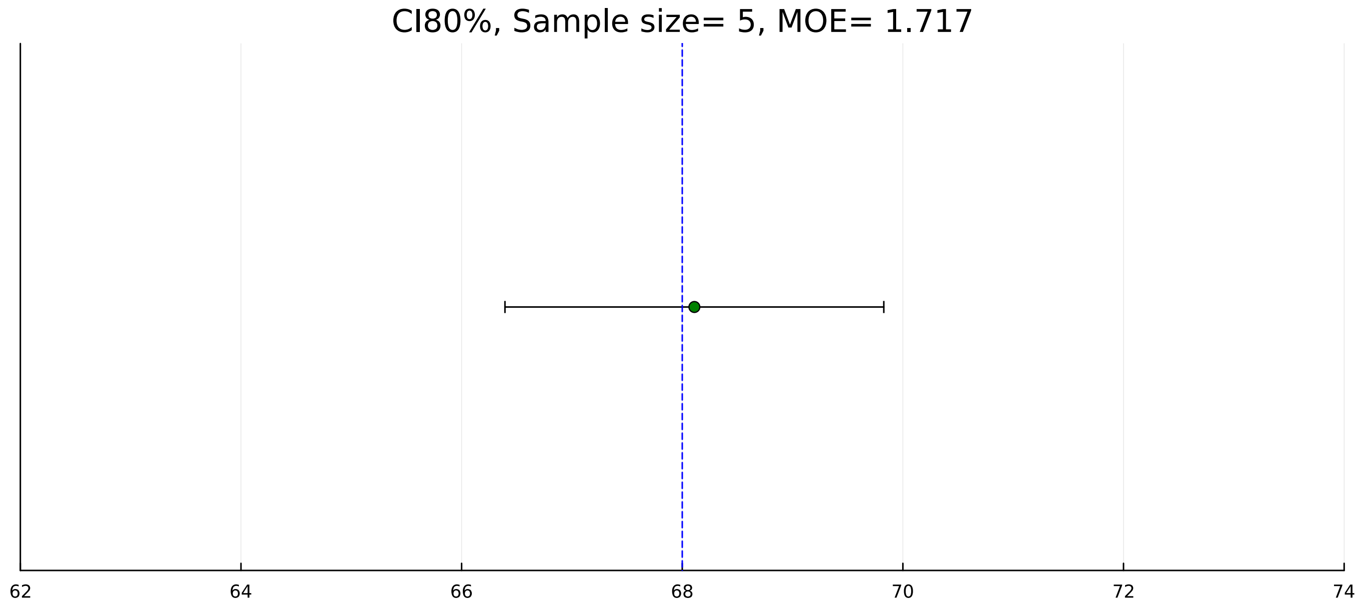
Let's look at what changing the sample size does to a confidence interval:

Sample size: 5, Confidence Level: 80%

Gather another sample

Sample data: 68.06, 65.77, 67.11, 67.03, 72.56

$$\bar{x} = 68.11, \quad \text{MOE} = z_{\alpha/2}\frac{\sigma}{\sqrt{5}} = 1.717$$



Asking what is the proabaility that a confidence interval contains the population mean is a meaningless question. A confidence interval either contains the population mean μ or it does not.

So either $\mu = 68$ is between 66.39 and 69.83 or it is not.

From the above information we can see that it is true that the population mean $\mu = 68$ is in the confidence interval

(66.39 , 69.83)

Each time we collect a new sample and recompute the sample mean and therefore the confidence interval the population mean is either in the confidence interval or it is not.

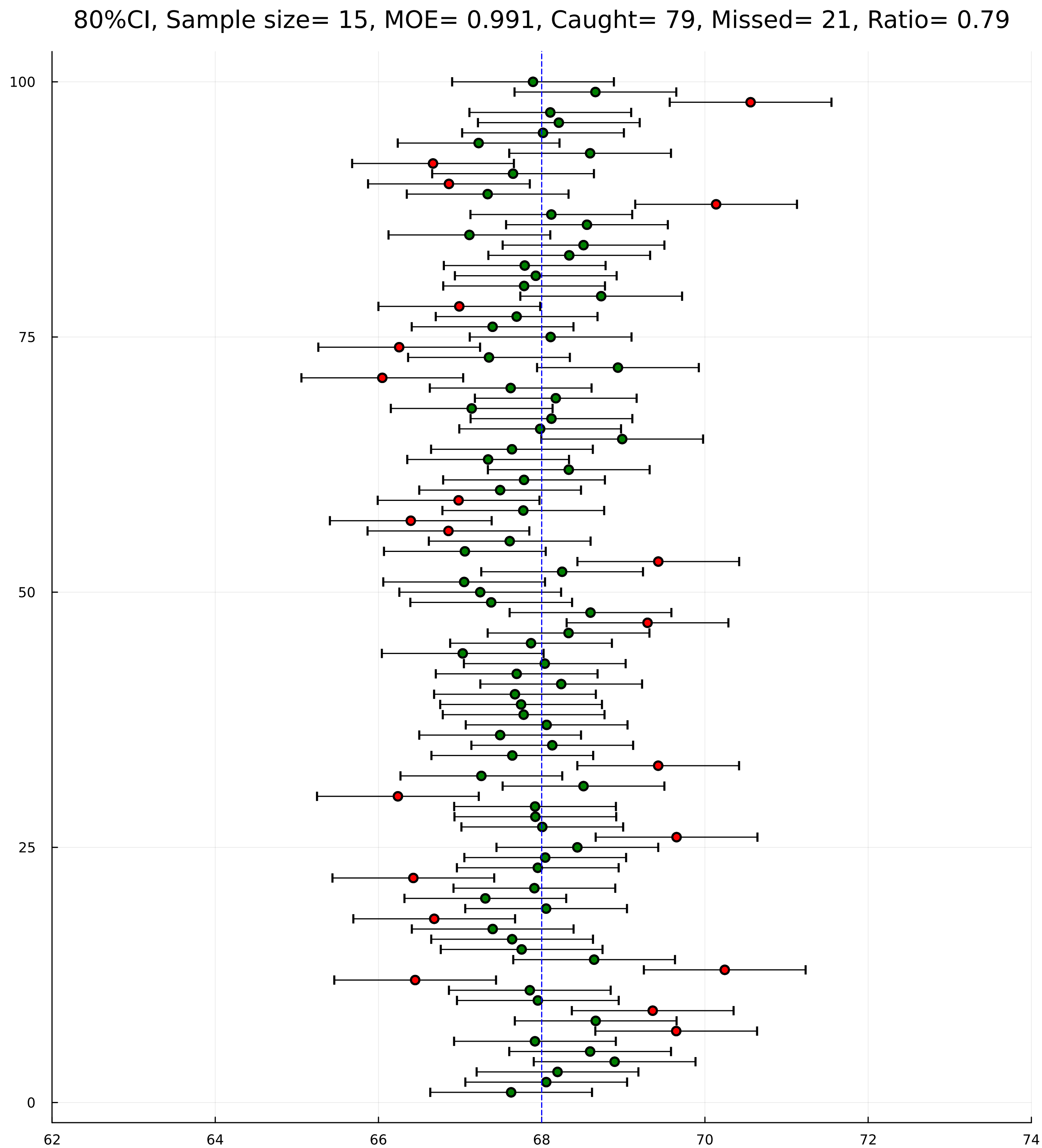
So what does it mean to have a 80% confidence interval?

Instead of just collecting a sample, finding the sample mean, computing the confidence interval, and ploting that one confidence interval, we will repeat this process 100 times and plot all of the confidence intervals on the same graph as shown below with the known population parameter...

Confidence Level: 80%

$$\alpha = 0.2, \quad \textbf{Critical value: } z_{\alpha/2} = 1.28$$

Number of CIs 100, Sample size: 15



Of course the whole point of a confidence interval is that you don't know what the population mean is and you are trying to estimate it by giving a range of values that are likely to contain it.

NOTE: To avoid round off errors keep at least 6 decimals places in intermediate calculations.

Minimum Sample Size

Suppose that we know the population standard deviation (magic, sorcery, devilery!), what confidence level we want, and the desired margin of error. The question becomes what sample size would we need?

The formula is

$$n = \left\lceil \left(\frac{z_{\alpha/2} \cdot \sigma}{\mathrm{MOE}} \right)^2 \right\rceil$$

Confidence Level: 80%, $\sigma =$ 50.0, $\mathrm{MOE} =$ 50.0

$$n = \left\lceil \left(\frac{1.28 \cdot 50.0}{50.0} \right)^2 \right\rceil = 2.0, \quad n \approx 1.6384$$

Hiding the code to get sample means here 👀

Here be the hidden code to set colors and calculate caught ratio and possibly a 🐉