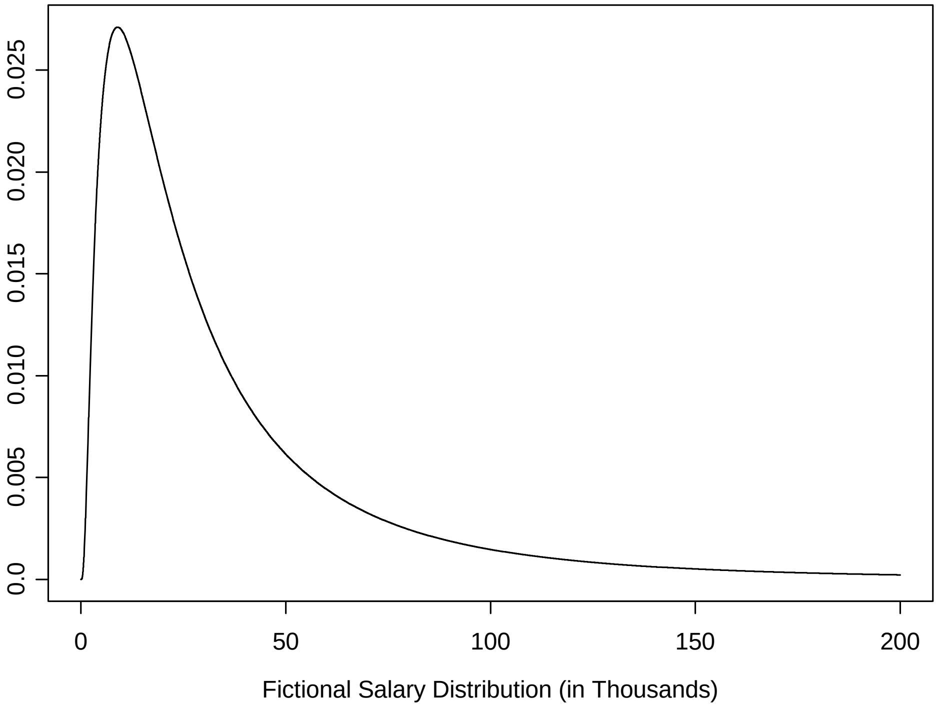
Example:

Suppose we have this fictional distribution (population) of salaries in the US:



This is a "lognormal" distribution and it has a mean of 40K and variance of 2749.25 (sd=52.43).

Its percentiles are

%	5	10	25	50	75	90	95
Salary(K)	4.7	6.7	12.4	24.3	47.6	87.4	125.7

The population mean, 40K, is at the 70th percentile.

Non-normality and the normal approximation

Now suppose I want to study this population and estimate the population mean. To do this I take a sample of 10 people:

Here are my sample data (already sorted; in K):

```
7.0; 14.3; 15.7; 19.2; 22.2; 26.9; 42.9; 52.6;
56.1; 159.3
```

And here is the stem and leaf output:

```
N = 10; Median = 24.55; Quartiles = 15.7, 52.6
Mean = 41.62; sd = 44.59

(Decimal point is 1 place to the right of the colon)
   0 : 7
   1 : 469
   2 : 27
   3 :
   4 : 3
   5 : 36

High: 159.3
```

And to estimate the mean, I'll use a confidence interval.

Because I don’t know the true population variance, possible 95% Confidence intervals for the population mean are:

$$\overline{X} \pm t^{9}_{0.025}\sqrt{\frac{S^2}{n}} \quad \text{or} \quad \overline{X} \pm Z_{0.025}\sqrt{\frac{S^2}{n}}$$

(9.72, 73.52) or (13.98, 69.23)

(Here the t-deviate is 2.26 and the z-deviate is 1.96)

In this case, neither interval will provide exactly 95% coverage probability, but both can be considered *approximate* confidence intervals.

Why? Because:

(1) the *t-interval* depends on the assumption that the underlying data are normally distributed (or, at the least, that they come from a symmetric distribution), which is clearly not the case here.

(2) the *z-interval* is will only have (approximately) the right coverage probability in large samples.

Why is the z-interval only approximate? Because the *z-interval* assumes that the distribution of the sample means will be symmetric *and* normally distributed.

Technically this only happens when the data themselves come from a normal distribution. But the Central limit theorem tells us that in large samples the distribution of the sample means will be very close to normal if the sample size is large enough.

The catch is that the definition of “large enough” depends on the context. When the underlying distribution is highly skewed, a very large sample size of 100 or 1000 will be needed. But if the underlying distribution is fairly symmetric, then we might only need 20 or 30 observations.

Aside (Pet peeve):

Confidence intervals are written as (9.7, 73.52) because this has a precise mathematical meaning.

Technically, even [9.7, 73.52] means something different, but this is acceptable.

What is not acceptable is “(a-b)”, which is “a minus b” (no matter how long the dash is). (People who be talking better English might like this well, but we be known who is right.)

But often the best thing to do is to use words
“The 95% CI was a to b.” or “(a to b).”

Ok, enough already; this is not English class. Show me the data!!

Back to our example:

The only way to tell if our approximation is a good one is to compare the distribution of the sample mean from samples of size 10 with its normal distribution approximation.

It is basically intractable to get a nice mathematical formula that represents the distribution of the sample means (of any size, in this case 10) from the lognormal distribution.

But it turns out that I can get a very good guess by simply doing this experiment over and over again. I just need to remember to save the sample mean each time. Then I can look at the distribution of those sample means.

Here is how I - actually my computer - does it:

1) Draw a sample of size 10 from the population.
2) Calculate the sample mean and save it.
3) Repeat steps #1 and #2 until I have collected a ridiculously large number of sample means, say 10,000.
4) Thank my computer for working so hard.

Now the collection of 10,000 sample means represents the *empirical distribution* of the sample mean from salary samples of size 10. Now I can examine this distribution to find out how large or small the sample means tend to be when they come from a sample of size 10.

(1) It turns out empirical distributions of this size are very accurate depictions of the true state of nature.

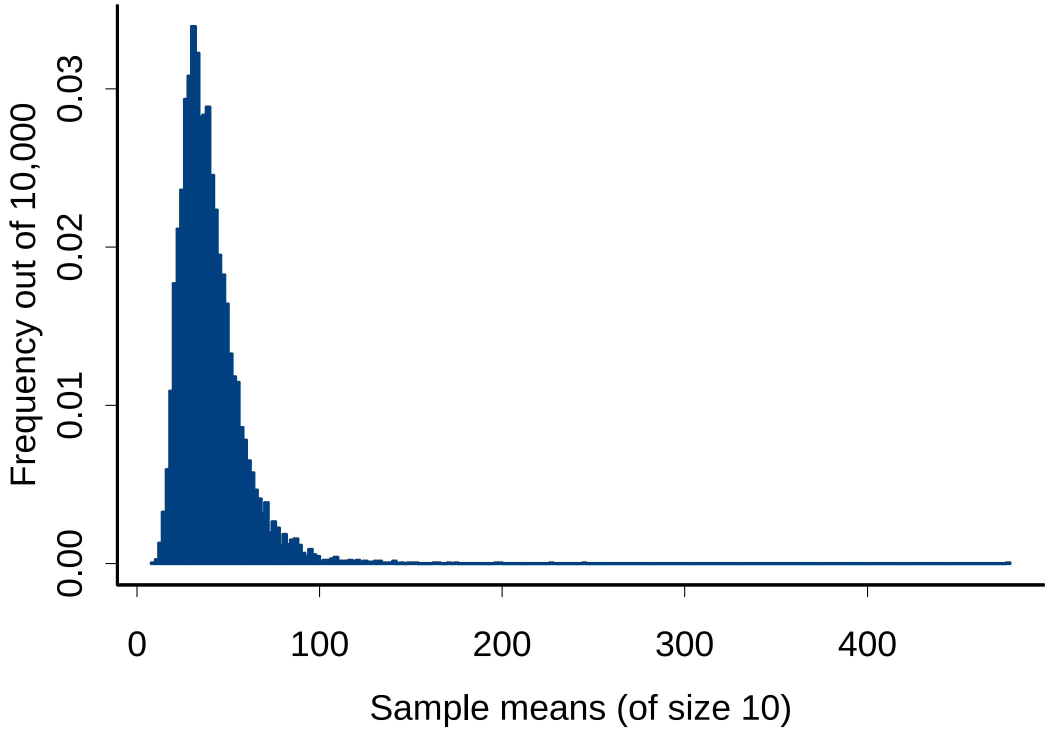
(2) We can do this for any experiment to learn about the distribution of their sample means or sample median or any parameter.

Side note (yes another one):
The ability to do this kind of brute force analysis has led to branch of statistics called “bootstrapping” techniques. Basically, they re-sample the data over and over again and then examine the empirical distributions of the sample mean or whatever.

This idea has actually been around for a long time, but the recent explosion in computing power has made it practical to carryout. (Graduate students are particularly thankful for these recent computing advances!)

Show me the data!!

Ok, here is the empirical distribution of 10,000 sample means (each from a sample of 10 salaries):



This does not look too bad, as it appears fairly symmetric.

The empirical percentiles are:

%	2.5	5	10	50	90	95	97.5
Salary(K)	18.4	20.5	23.0	36.8	60.0	70.5	81.6

This means that 10% of the sample means are below 23K, 95% of sample means are below 70.5K etc.

The average of the sample means was 40.04 and the standard deviation of the sample means (standard error!) was 17.37.

So what?

What we are interested in is how well the normal approximation does in this case. We get the normal approximation from the Central limit theorem, which tells us that

if X_1, X_2, …, X_n are i.i.d. from some distribution then

$\overline{X}_n \overset{\text{approx}}{\sim} N\left(E(X), \frac{Var(X)}{n}\right)$ in large samples.

and if we don't know the (true) variance, we can use the sample variance as an estimate and this approximation still holds. (WOW!)

In our example this means that

$$\overline{\mathrm{X}}_n \overset{\text{approx}}{\sim} \mathrm{N}(40, 17.37)$$

(Notice that I replaced the standard error with the estimated standard error from our empirical distribution of the sample means.)

So this means I can calculate the theoretical percentiles under this normal approximation. For example, suppose I want the 90th percentile.

$$\begin{aligned} P(\overline{\mathrm{X}}_n < ?) &= P\left(\frac{\overline{\mathrm{X}}_n - \mu}{\sigma/\sqrt{n}} < \frac{? - \mu}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z < \frac{? - 40}{17.37}\right) \\ &= .90 \end{aligned}$$

And because P(Z<1.28) = 0.9 we have that

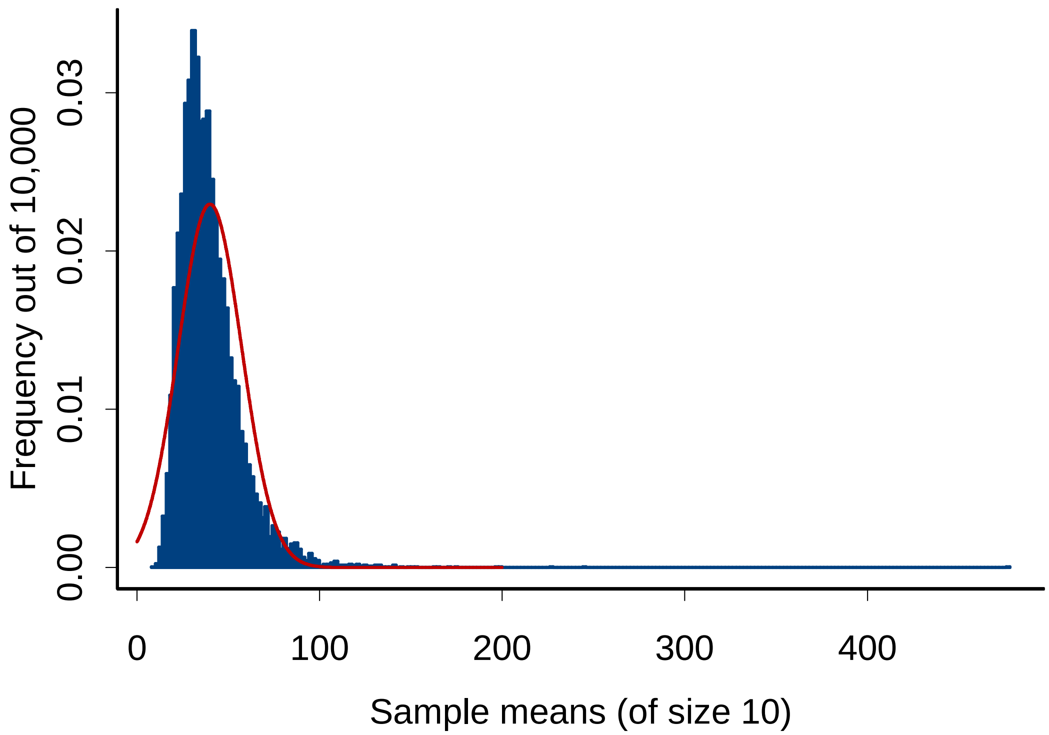
$$\begin{aligned} 1.28 &= \frac{? - 40}{17.37} \\ ? &= 40 + 1.28(17.37) \\ &= 62.23 \end{aligned}$$

Meaning that 90% of samples means (of size 10) should be less than 62.23; that is, if the normal approximation is accurate (10 is “large enough”).

So I did this calculation for all the percentiles of interest and this is what I get:

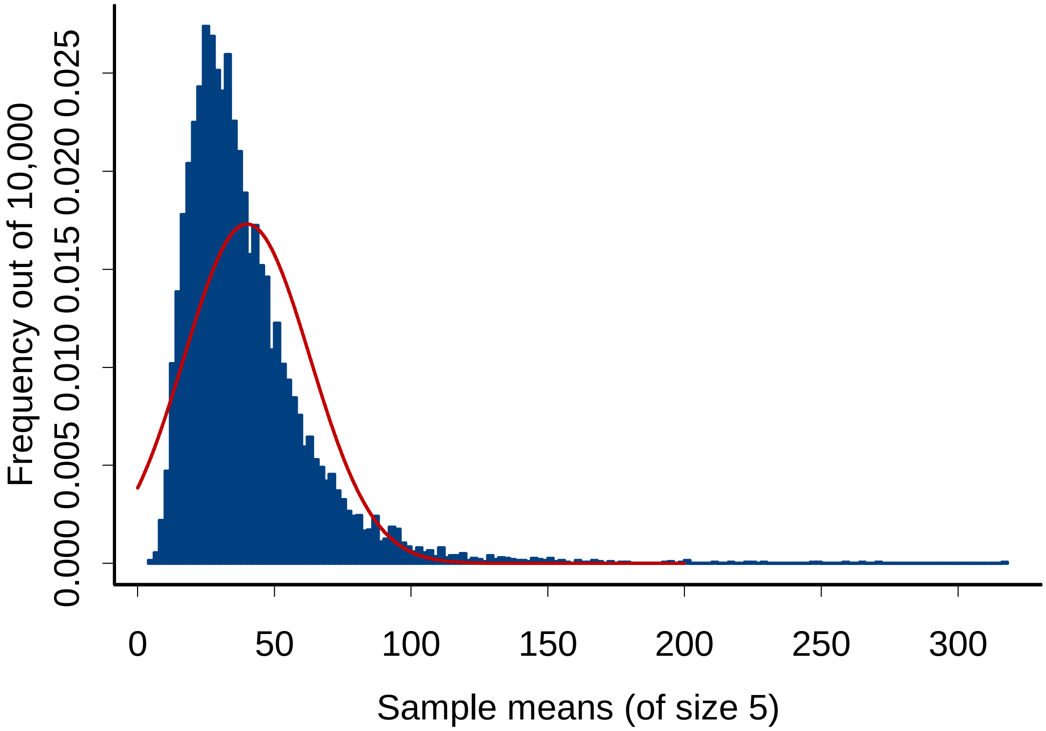
10 samples	Percentile (%)						
Salary(K)	**2.5**	**5**	**10**	**50**	**90**	**95**	**97.5**
Empirical	18.4	20.5	23.0	36.8	60.0	70.5	81.6
Normal	5.95	11.4	17.7	40.0	62.3	68.6	74.0

And here is a picture of what is going on:



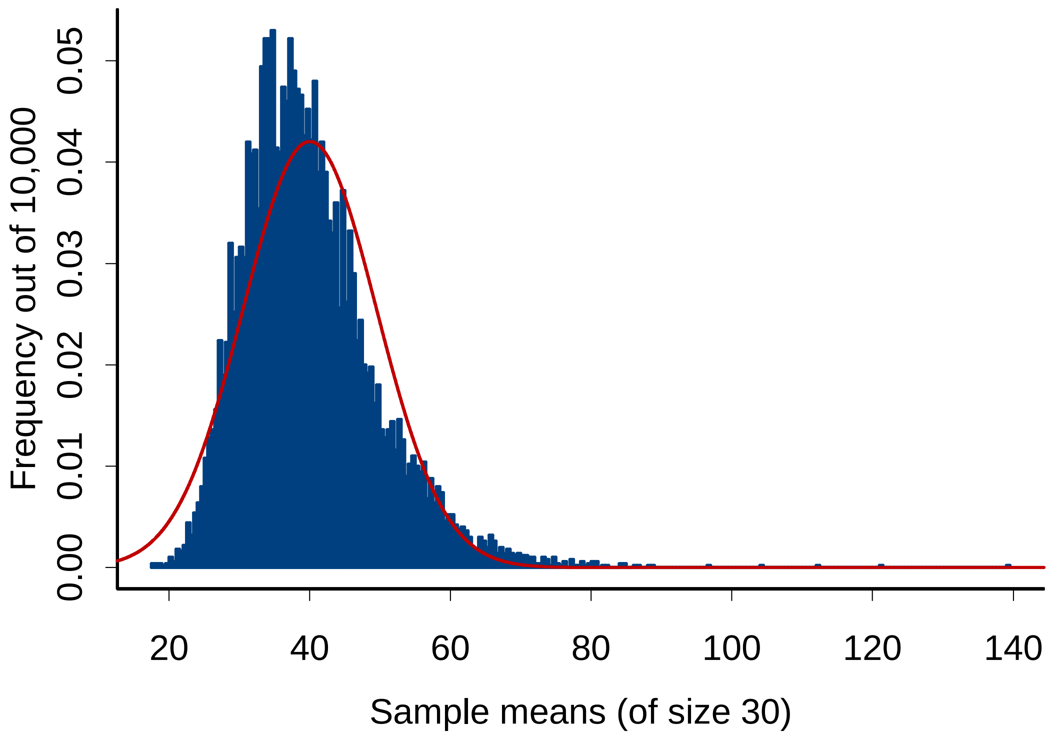
And if I repeat this for the case when I only take 5 observations, I get (average was 39.74; standard deviation was 23.05):

5 samples	**Percentile (%)**						
Salary(K)	**2.5**	**5**	**10**	**50**	**90**	**95**	**97.5**
Empirical	13.1	15.1	18.0	34.3	67.1	81.3	96
Normal	-5.2	2.1	10.5	40.0	69.5	77.9	85.2



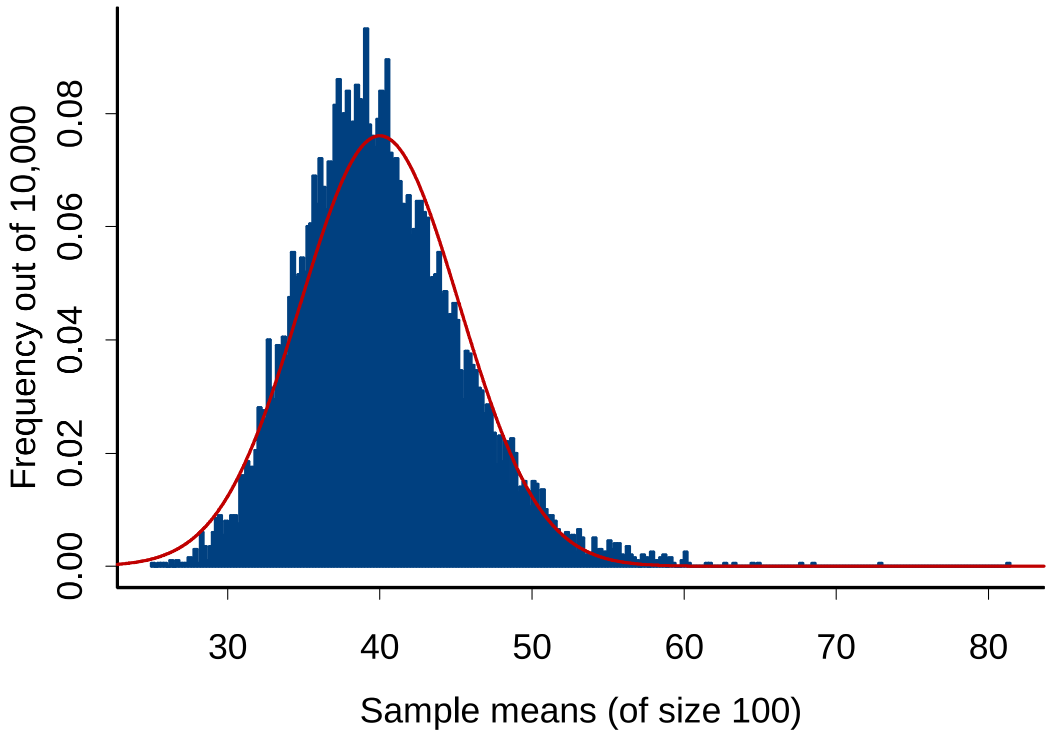
But when I take 30 observations, I do better (average was 39.95; standard deviation was 9.49):

30 samples	**Percentile (%)**						
Salary(K)	**2.5**	**5**	**10**	**50**	**90**	**95**	**97.5**
Empirical	25.6	27.3	29.3	38.6	52.6	57.5	62.0
Normal	21.4	24.4	27.8	40.0	52.2	55.6	58.6



And when I have 100, I do great (average was 40.09; standard deviation was 5.24):

30 samples	**Percentile (%)**						
Salary(K)	**2.5**	**5**	**10**	**50**	**90**	**95**	**97.5**
Empirical	31.2	32.4	33.8	39.6	46.9	49.3	51.4
Normal	29.7	31.4	33.3	40.0	46.7	48.6	50.3

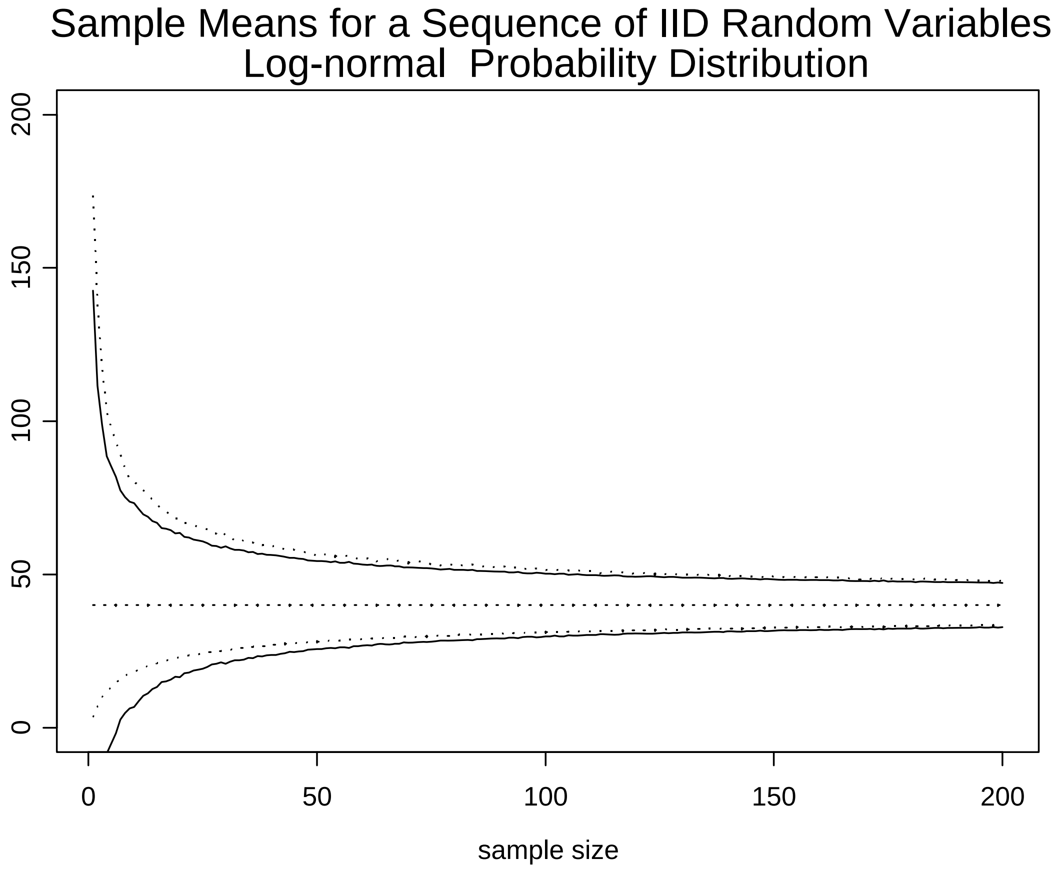


Notice how the standard deviation of the sample means (the standard error!) gets smaller as the sample size get larger.

We also see that as the sample size gets larger, the normal approximation gets better. The only suspect part of this approximation is in the tail areas where the approximation is almost always “rough”.

Below shows the empirical (true) distribution and the normal approximation for the true 95% interval for the sample mean.

Here the dotted lines are the empirical (true) distribution and the solids lines are the normal approximation.



Remember that our confidence intervals for our initial sample (`Mean = 41.62; sd = 44.59`) were

(9.72, 73.52) or (13.98, 69.23)

And the graph above indicates that these intervals, which are approximating the solid lines at 10 (19.4 and 77.35), are probably underestimates (but the width looks ok, 57.99 versus 63.8 and 54.25).