Approximate Confidence Intervals for Means of Non-Normal Probability Distributions

When we have independent observations $X_1, X_2, \ldots, X_n$ with common mean $E[X]$ and standard deviation $\sqrt{Var(X)}$, the sample mean $\overline{X}_n$ has expected value $E[\overline{X}_n] = E[X]$ and standard deviation $\sqrt{Var(X)/n}$.

We know from the CLT that

$$\frac{\sqrt{n}\left(\overline{X}_n - E[X]\right)}{\sqrt{Var[X]}} \overset{approx}{\sim} N(0,1)$$

for `large' n.

If the X's are normal, the standardized mean has a standard normal probability distribution. And if the X's have some other probability distribution, the standardized mean still has an approximate standard normal distribution when the sample is large.

This enables us to determine an approximate 95% confidence interval for the expected value, $E[X]$, when the sample size is `large' regardless of the underlying distribution:

$$\overline{X}_n \pm 1.96\sqrt{Var(X)/n}$$

Why? Because

$$P\left(-1.96 < \frac{\sqrt{n}\left(\overline{X}_n - E[X]\right)}{\sqrt{Var[X]}} < 1.96\right) \approx 0.95$$

or

$$P\left(\overline{X}_n - 1.96\sqrt{\frac{Var[X]}{n}} < E[X] < \overline{X}_n + 1.96\sqrt{\frac{Var[X]}{n}}\right) \approx 0.95$$

This is a powerful and amazing result, because all we need to assume is that we know

1) Var[X] and
2) that n is `large' enough

But number 2 is empirically verifiable, and number 1 is not necessary if we use a t-distribution, replacing Var[X] with S^2 !!

Before we go on, lets examine the CLT.

Example:

To take a familiar extreme case, suppose the *X*'s are i.i.d. Bernoulli(θ) random variables. The Central Limit Theorem ensures that for large n

$$\frac{\sqrt{n}\left(\overline{X}_n - \theta\right)}{\sqrt{\theta(1-\theta)}} \overset{approx}{\sim} N(0,1)$$

So,

$$P\left(\frac{\sqrt{n}\left(\overline{X}_n - \theta\right)}{\sqrt{\theta(1-\theta)}} > 1.645\right) \approx 0.05$$

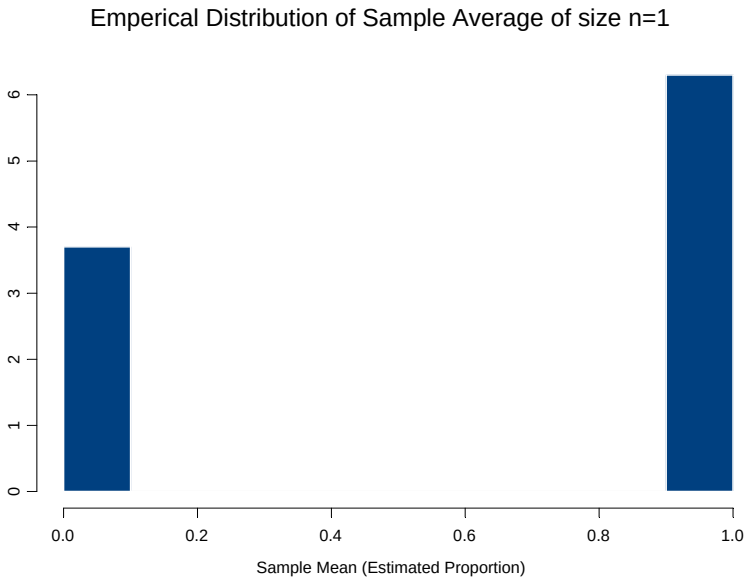
$$P\left(\frac{\sqrt{n}\left(\overline{X}_n - \theta\right)}{\sqrt{\theta(1-\theta)}} < -1.96\right) \approx 0.025$$

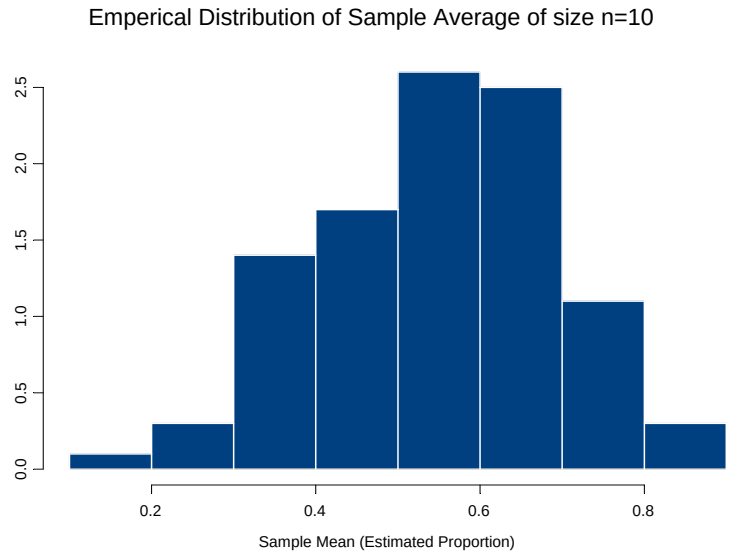
and so forth.

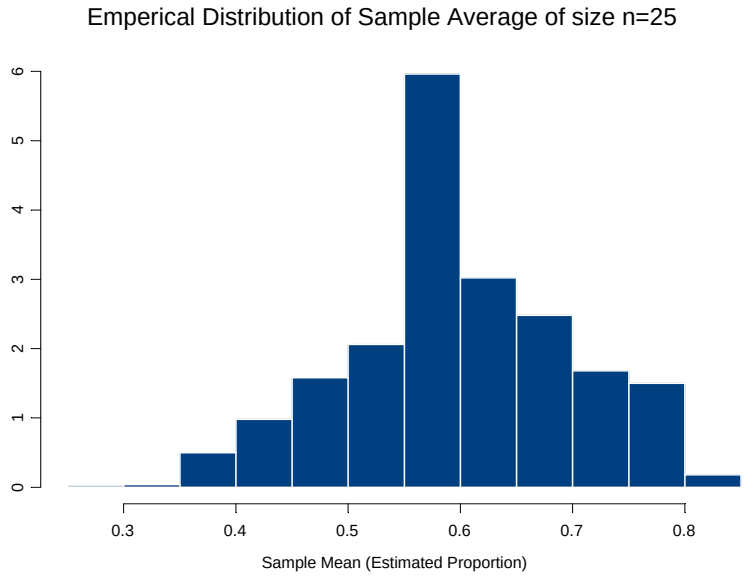
Now this is really quite remarkable: The Bernoulli probability distribution is not at all like the normal — instead of the normal's continuous, symmetric, bell-shaped density, the Bernoulli has all of its probability concentrated in two (usually unequal) blobs, on the points 0 and 1.

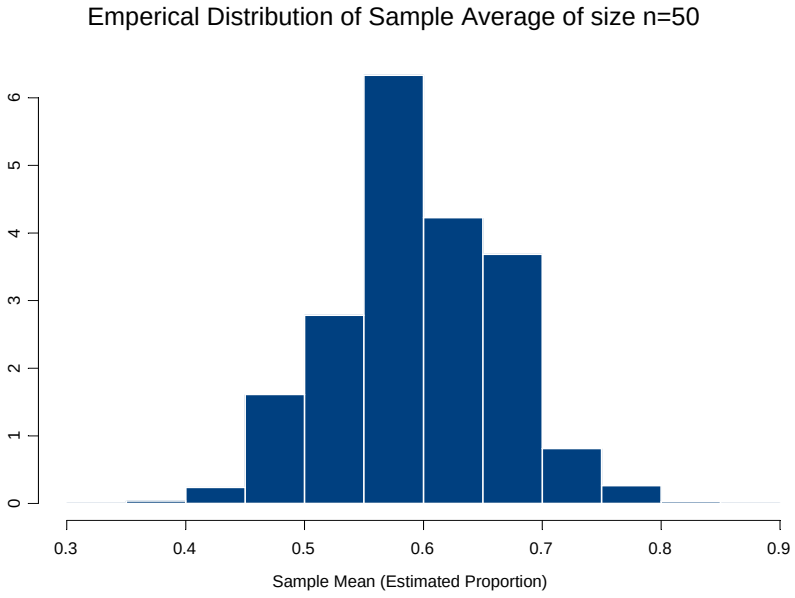
Yet, the standardized average of many independent Bernoulli random variables has approximately the same (standard normal) probability distribution as the standardized average of independent normal random variables.

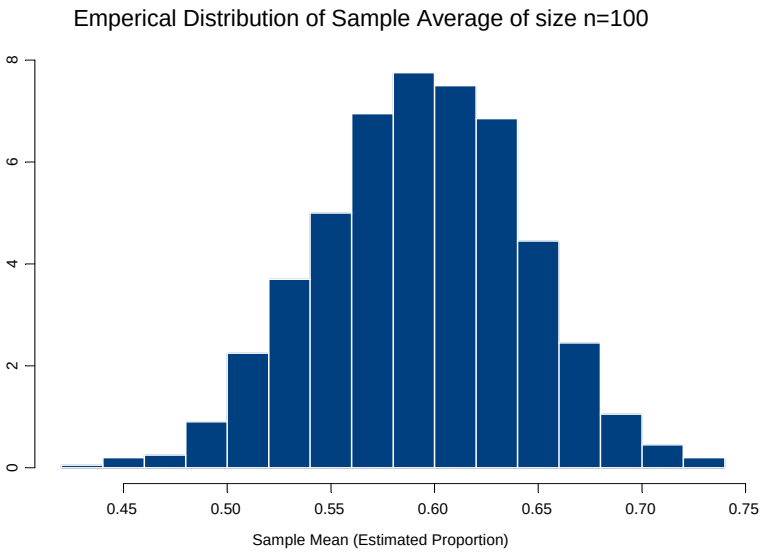
We've already seen this process in action. For the Bernoulli(0.6) distribution the histogram for $\overline{X}_{25}$ already looks much like the normal density with the same mean and variance, and for $n = 100$, the similarity is even stronger.

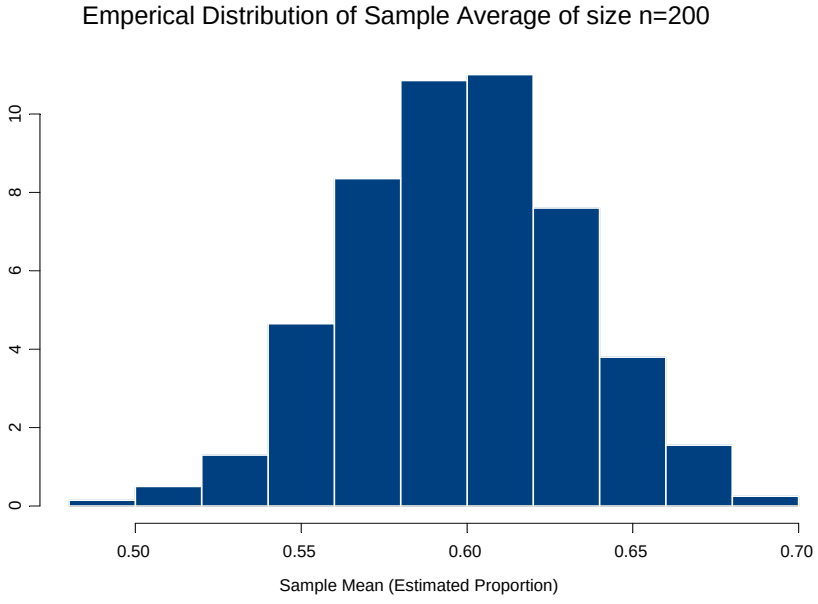

Emperical Distribution of Sample Average of size n=1
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
1.0
Sample Mean (Estimated Proportion)


Emperical Distribution of Sample Average of size n=10
0.0
0.5
1.0
1.5
2.0
2.5
0.2
0.4
0.6
0.8
Sample Mean (Estimated Proportion)


Emperical Distribution of Sample Average of size n=25
0
1
2
3
4
5
6
0.3
0.4
0.5
0.6
0.7
0.8
Sample Mean (Estimated Proportion)


Emperical Distribution of Sample Average of size n=50
0
1
2
3
4
5
6
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Sample Mean (Estimated Proportion)


Emperical Distribution of Sample Average of size n=100
0
2
4
6
8
0.45
0.50
0.55
0.60
0.65
0.70
0.75
Sample Mean (Estimated Proportion)


Emperical Distribution of Sample Average of size n=200
0
2
4
6
8
10
0.50
0.55
0.60
0.65
0.70
Sample Mean (Estimated Proportion)

Let's compare some of the probabilities: For the Bernoulli(0.6),

$$0.05 \approx P\left(\frac{\sqrt{n}\left(\overline{X}_n - \theta\right)}{\sqrt{\theta(1-\theta)}} > 1.645 \right) = P\left(\frac{\sqrt{n}\left(\overline{X}_n - 0.6\right)}{\sqrt{0.6(1-0.6)}} > 1.645 \right)$$

$$= P\left(\overline{X}_n > 0.6 + 1.645\sqrt{\frac{0.6 \times 0.4}{n}} \right)$$

$$= P\left(\sum_{i=1}^{n} X_i > 0.6n + 1.645\sqrt{n}\sqrt{0.6 \times 0.4} \right)$$

The last expression is the probability that a **Binomial(n, 0.6)** random variable (the sum of x's) will exceed the quantity on the right hand side.

For any value of n we can calculate this probability exactly and see how close it is to the probability that the standard normal will exceed 1.645 (which is 0.05). For n = 25, 100, and 1000, the exact probabilities are 0.029, 0.040, and 0.050.

The normal distribution plays a central role in statistics. This is not because many of the random variables that we observe have normal distributions. (Most do not.) It is because the normal can be used to approximate the distributions of averages (and of many other summary quantities as well) of samples from all type of distributions. The reason for this is the Central Limit Theorem.

What makes this <u>really</u> important is that it remains true when Var(X) is replaced by the sample variance, S_n^2:

$$\frac{\sqrt{n}\left(\overline{X}_n - E[X]\right)}{S_n} \overset{approx}{\sim} N(0,1)$$

(for large n). This means that

$$\overline{X}_n \pm 1.96\ SE\left(\overline{X}_n\right)$$

is an approximate 95% CI, not only when the individual X's are normal and the standard error is given, but also when the X's have some other probability distribution, and the standard error is estimated from the sample. (WOW!!)

Example:

In the Bernoulli(θ) case, the average, $\overline{X}_n$, is the proportion of successes, which is used to estimate the probability of success, θ. In this case, $\overline{X}_n$ is often represented by the symbol $\hat{\theta}$ ("theta hat"), to indicate its role as an estimator of θ. The standardized mean is

$$\frac{\sqrt{n}\,(\,\overline{X}_n - \theta\,)}{\sqrt{\,\theta(1-\theta\,)}} = \frac{\sqrt{n}\,(\,\hat{\theta}_n - \theta\,)}{\sqrt{\,\theta(1-\theta\,)}}$$

If we estimate the variance, which is θ(1-θ), by the sample variance S_n^2 the result

$$\frac{\sqrt{n}\left(\overline{X}_n - \theta\right)}{S_n} \overset{approx}{\sim} N(0,1)$$

Another natural estimate of the variance in this special case (Bernoulli) would be $\hat{\theta}(1-\hat{\theta})$ (i.e., $\overline{X}_n(1-\overline{X}_n)$).

$$\frac{\sqrt{n}\left(\overline{X}_n - \theta\right)}{\sqrt{\overline{X}_n\left(1-\overline{X}_n\right)}} = \frac{\sqrt{n}\left(\hat{\theta} - \theta\right)}{\sqrt{\hat{\theta}\left(1-\hat{\theta}\right)}} \overset{approx}{\sim} N(0,1)$$

When n is large we can use either variance estimate -- they are essentially the same:

$$S_n^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2}{n-1} = \frac{\sum_{i=1}^{n} X_n^2 - n\left(\overline{X}_n\right)^2}{n-1}$$

$$= \frac{\sum_{i=1}^{n} X_n - n\left(\overline{X}_n\right)^2}{n-1}$$

$$= \frac{n\overline{X}_n - n\left(\overline{X}_n\right)^2}{n-1}$$

$$= \overline{X}_n\left(1-\overline{X}_n\right)\left(\frac{n}{n-1}\right)$$

Thus for a Bernoulli sample with large *n*, the sample variance S_n^2 is essentially just the same as

$$\overline{X}_n\left(1-\overline{X}_n\right) = \hat{\theta}\left(1-\hat{\theta}\right)$$

because for large n the ratio n/(n-1) is approximately 1.

What does all this tell us about the Student's t confidence interval?

Suppose we observe i.i.d. random variables $X_1, X_2, ..., X_n$ and want a 95% confidence interval for their expected value. If the distribution is normal, then the appropriate confidence interval is the Student's t interval:

$$\overline{X}_n \pm t_{n-1} S_n / \sqrt{n}$$

If the distribution of the *X*'s is normal, the coverage probability

$$P\left(\overline{X}_n - t_{n-1}\sqrt{\frac{S_n^2}{n}} < E[X] < \overline{X}_n + t_{n-1}\sqrt{\frac{S_n^2}{n}} \right) = 0.95$$

is exactly 0.95,

but even if the *X*'s have some other distribution the coverage probability is still approximately 0.95.

Why? The reasoning goes like this:

$$
\begin{aligned}
& P\left(-t_{n-1} < \frac{\sqrt{n}\left(\overline{X}_n - E[X]\right)}{S_n} < +t_{n-1} \right) \\
& \approx P\left(-1.96 < \frac{\sqrt{n}\left(\overline{X}_n - E[X]\right)}{S_n} < +1.96 \right) \\
& \approx P\left(-1.96 < \frac{\sqrt{n}\left(\overline{X}_n - E[X]\right)}{\sqrt{Var[X]}} < +1.96 \right) \\
& = 0.95
\end{aligned}
$$

This approximation gets better because

1) the t-distribution $\rightarrow$N(0,1)

2) $S_n^2 \rightarrow Var[X]$

3) $\frac{\sqrt{n}\left(\overline{X}_n - E[X]\right)}{\sqrt{Var[X]}} \rightarrow N(0,1)$

as the sample size increases.

The Student's t confidence interval whose coverage probability is exactly 0.95 when the *X*'s are normal is still approximately 0.95 when they are not normal (as long as the sample is not too small).

This is terribly important in practice, because we never really know what distribution we're actually sampling from. Because they remain valid (covering the true mean with approximately the probability that they are supposed to) under quite general conditions, Student's t CI's are said to be "robust."

Statistics "works" in the real world because many of its procedures, like the Student's t CI, have this property of "robustness," or approximate validity under a wide range of probability models, i.e. under a wide range of assumptions about the probability distribution.