Section 5.2 - The Sampling Distribution of a Sample Mean

Statistics 104

Autumn 2004

The Sampling Distribution of a Sample Mean

Example: Quality control check of light bulbs

Sample n light bulbs and look at the average failure time.

Take another sample, and another, and so on.

What is the mean of the sampling distribution? Variance? Standard Deviation?

What is $P[\bar{X}_n \geq \mu_x]$ or $P[\mu_x - c \leq \bar{X}_n \leq \mu_x + c]$?

We will consider the situation where the observations are independent (at least approximately). In the case of finite populations, we want $N \gg n$.

Under this assumption, we can use the same idea as we did to get the mean and variance of a binomial distribution.

$$
\begin{aligned}
\bar{X} &= \frac{1}{n}(X_1 + X_2 + \ldots + X_n) \\
\mu_{\bar{x}} &= \frac{1}{n}\underbrace{(\mu_x + \mu_x + \ldots + \mu_x)}_{n \text{ times}} = \mu_x \\
\sigma^2_{\bar{x}} &= \frac{1}{n^2}\underbrace{(\sigma^2_x + \sigma^2_x + \ldots + \sigma^2_x)}_{n \text{ times}} = \frac{\sigma^2_x}{n} \\
\sigma_{\bar{x}} &= \frac{\sigma_x}{\sqrt{n}}
\end{aligned}
$$

So the sampling distribution of $\bar{X}$ is centered at the same place as the observations, but less spread out as we should expect based on the law school example.

The smaller spread also agrees with the law of large numbers, which says $\bar{X} \to \mu_x$ as increases.

Note the sampling distribution of $\hat{p}$ is just a special case. $\hat{p}$ is just an average of n 0's or 1's. The formulas have the same form

$$\begin{aligned} \mu_{\hat{p}} &= p \\ \sigma^2_{\hat{p}} &= \frac{p(1-p)}{n} \\ \sigma_{\hat{p}} &= \sqrt{\frac{p(1-p)}{n}} \end{aligned}$$

Lets suppose that the life times of the light bulbs have a gamma distribution with $\mu_x = 2$ years and $\sigma_x = 1$ year.

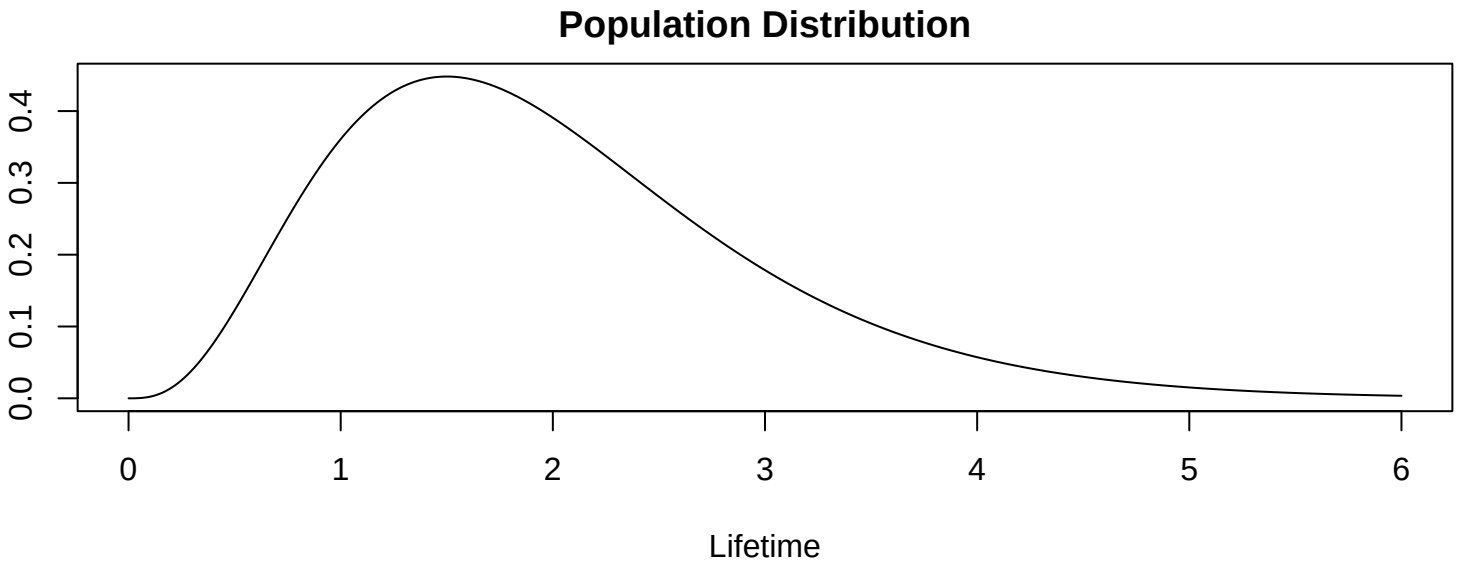
Sample n bulbs and calculate sample average

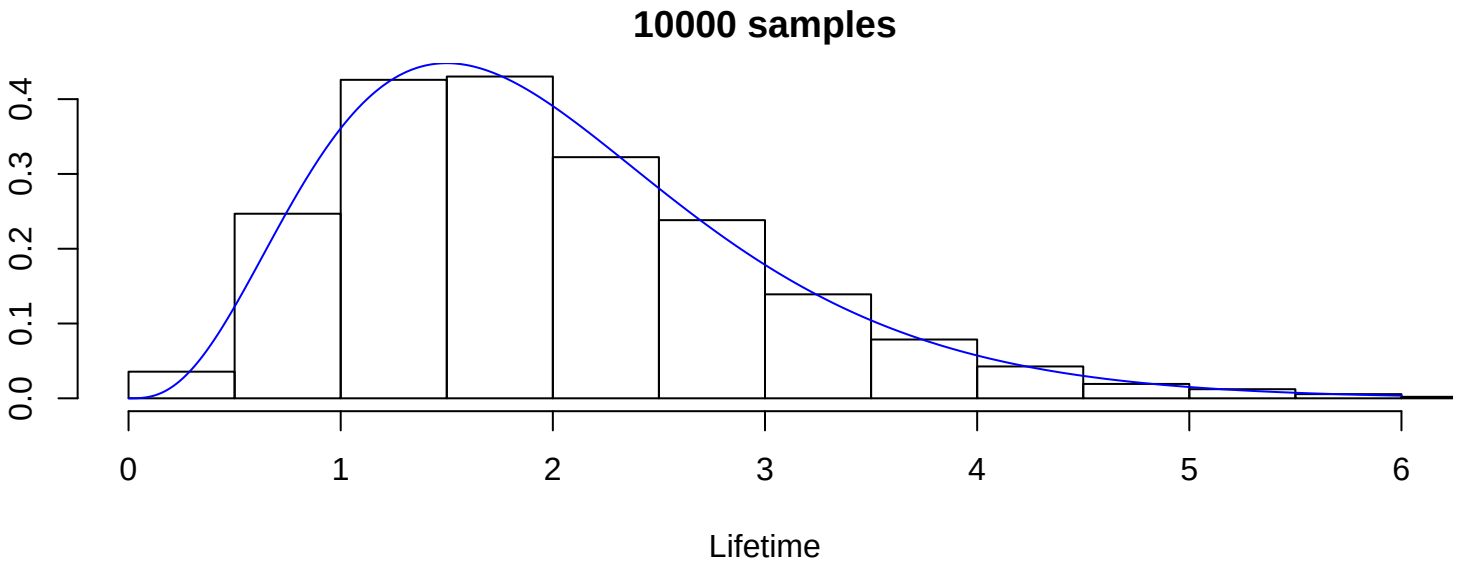
$$\begin{aligned} \mu_{\bar{x}} &= \mu_x = 2 \\ \sigma_{\bar{x}} &= \frac{\sigma_x}{\sqrt{n}} = \frac{1}{\sqrt{n}} \end{aligned}$$

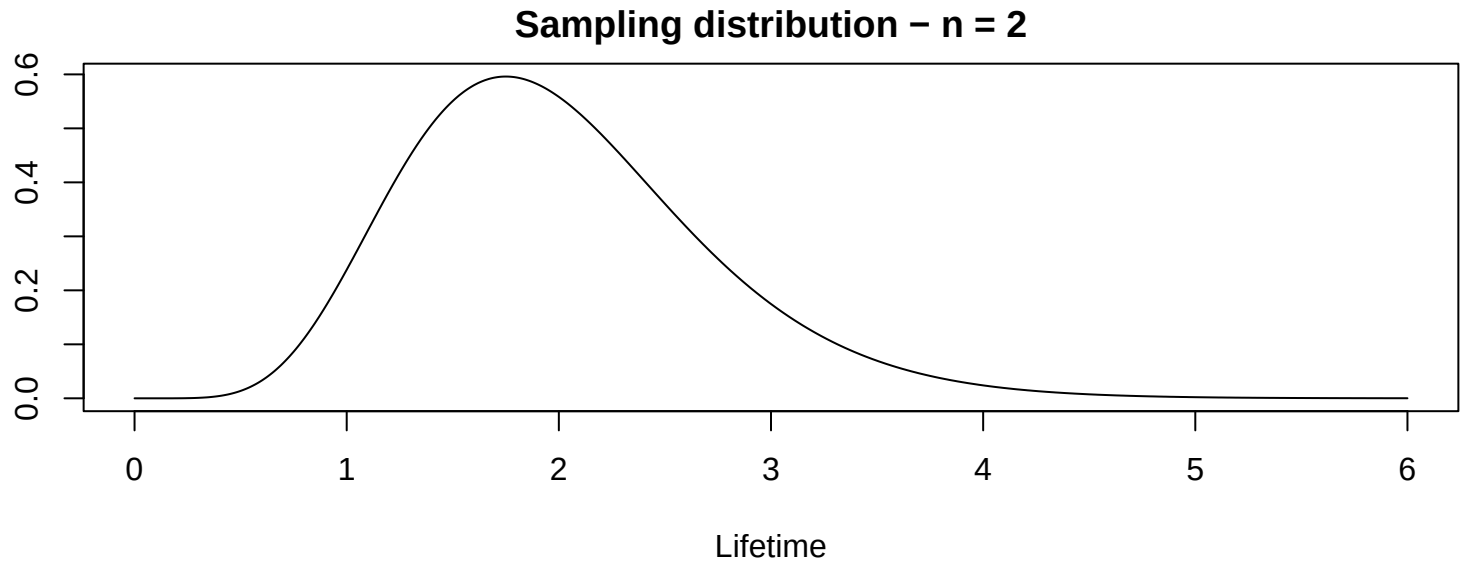
Lets see how the sampling distribution changes as n increases along the sequence $n = 2, 10, 50, 100$

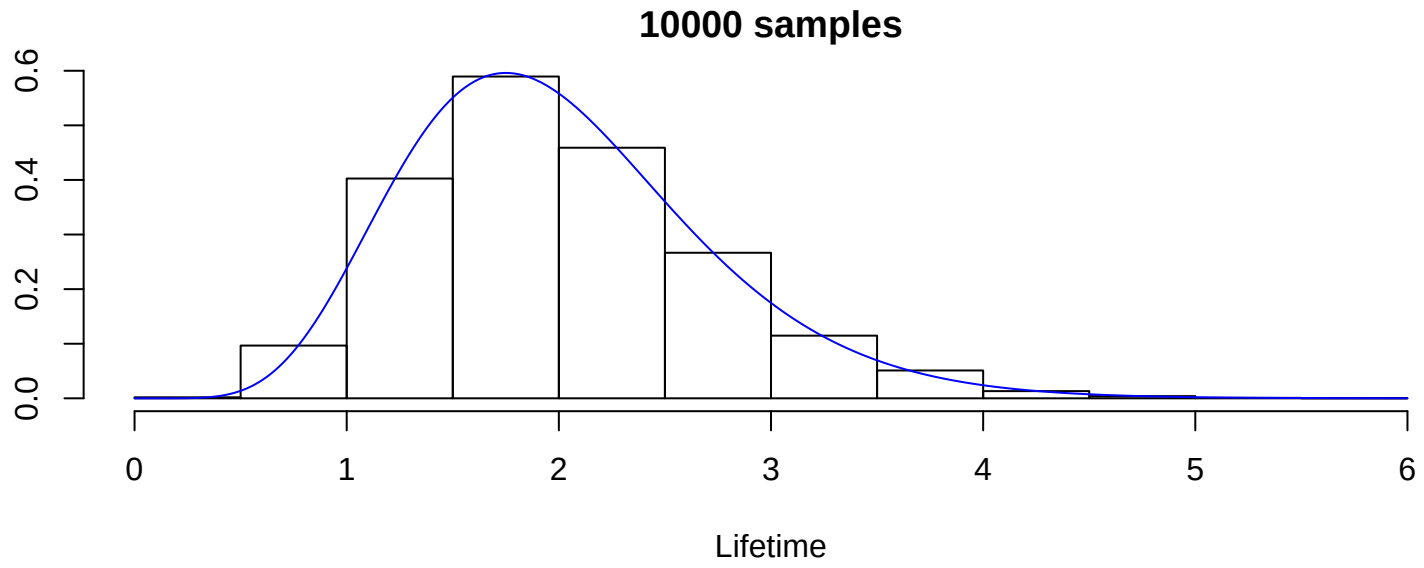
We will examine this two ways:

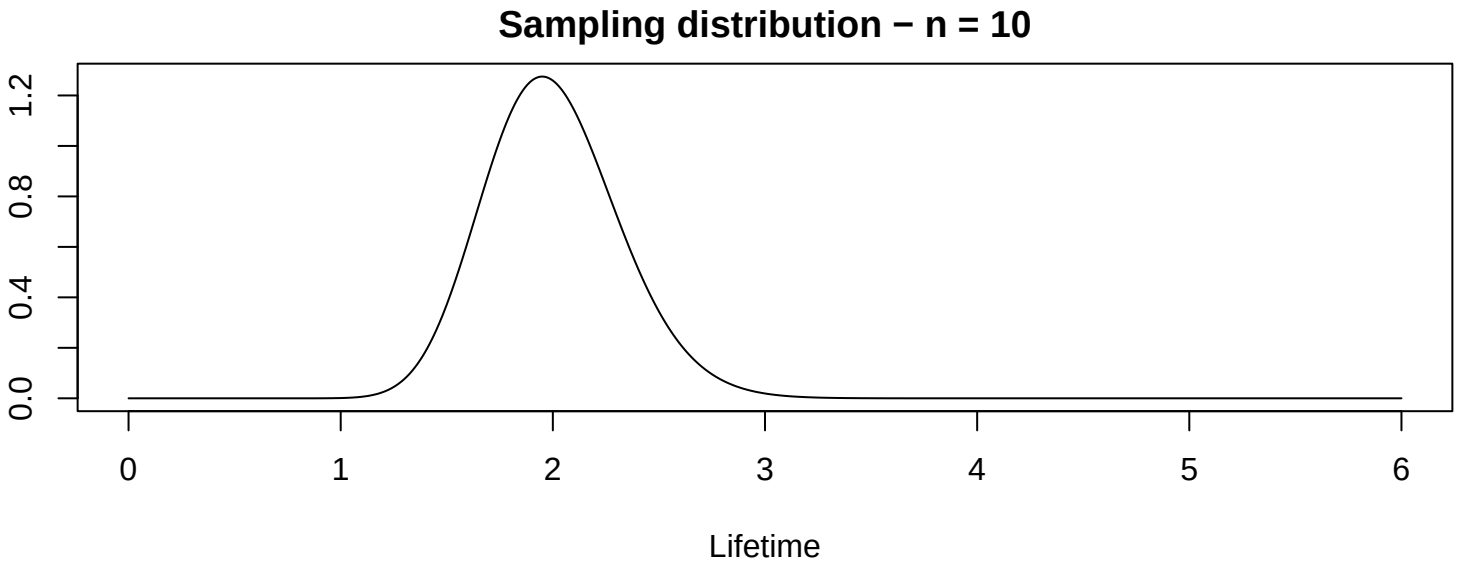
- The exact sampling distribution (which happens to be also a gamma distribution with the appropriate mean and standard deviation).
- Monte Carlo experiment with 10,000 samples of $\bar{X}$ for each n. The blue line is the exact sampling distribution

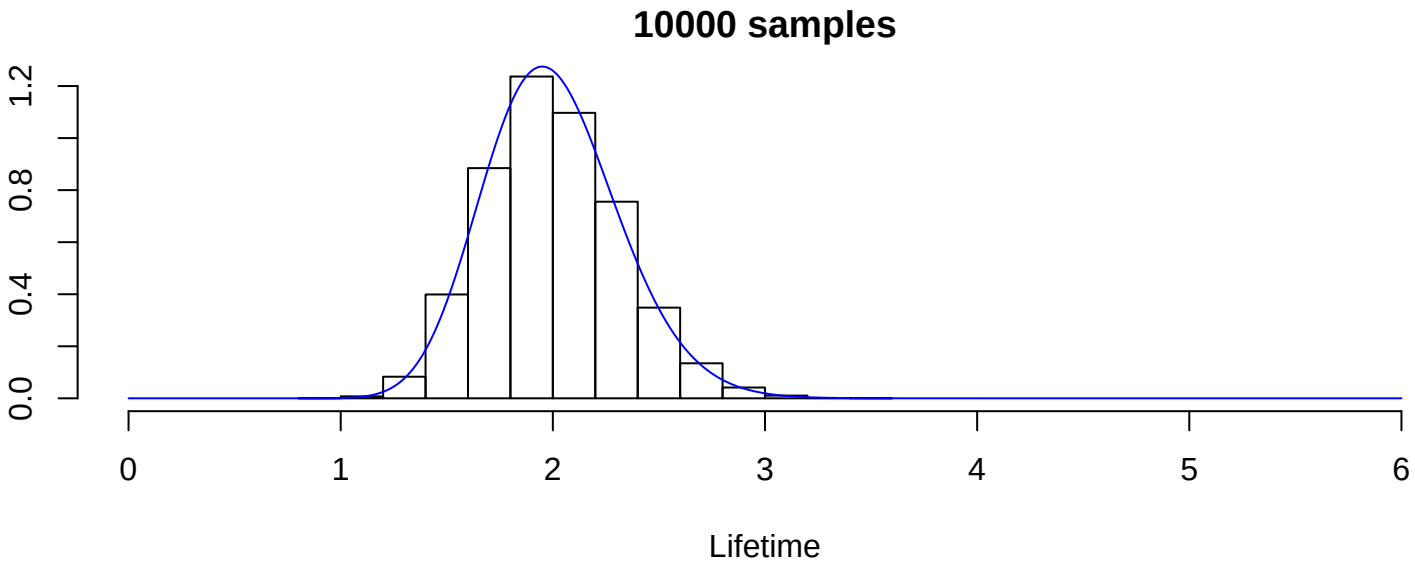

Population Distribution
0.0
0.1
0.2
0.3
0.4
0
1
2
3
4
5
6
Lifetime

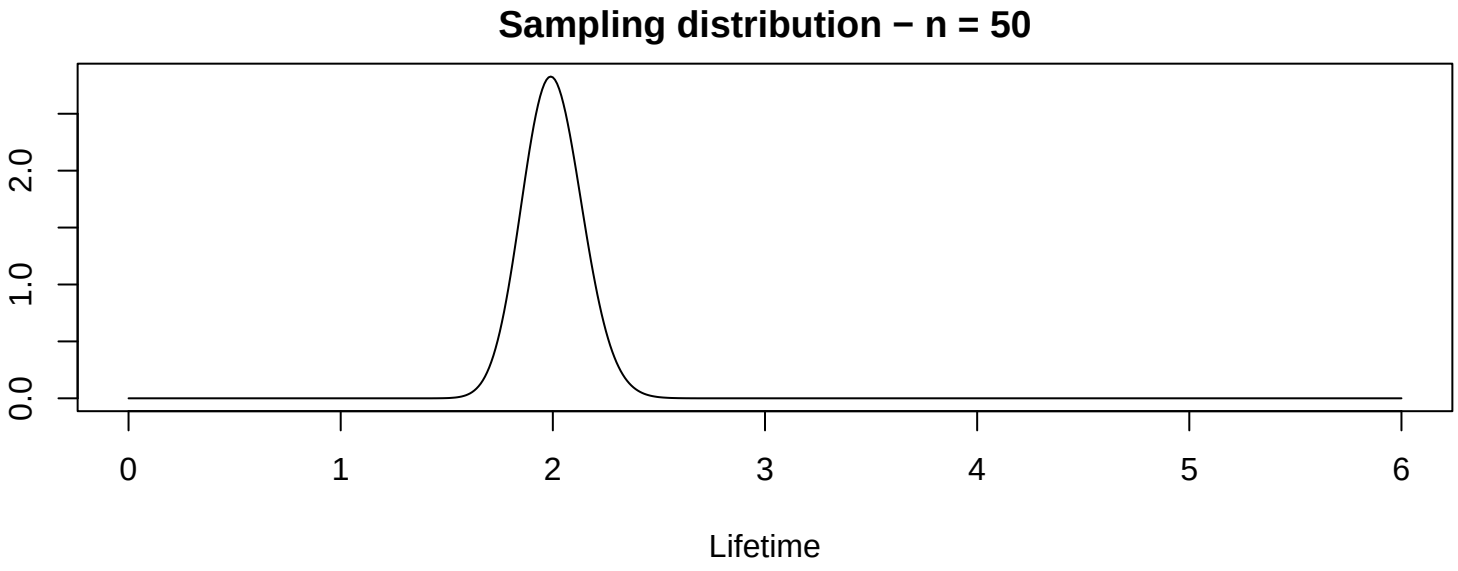

10000 samples
0.0
0.1
0.2
0.3
0.4
0
1
2
3
4
5
6
Lifetime

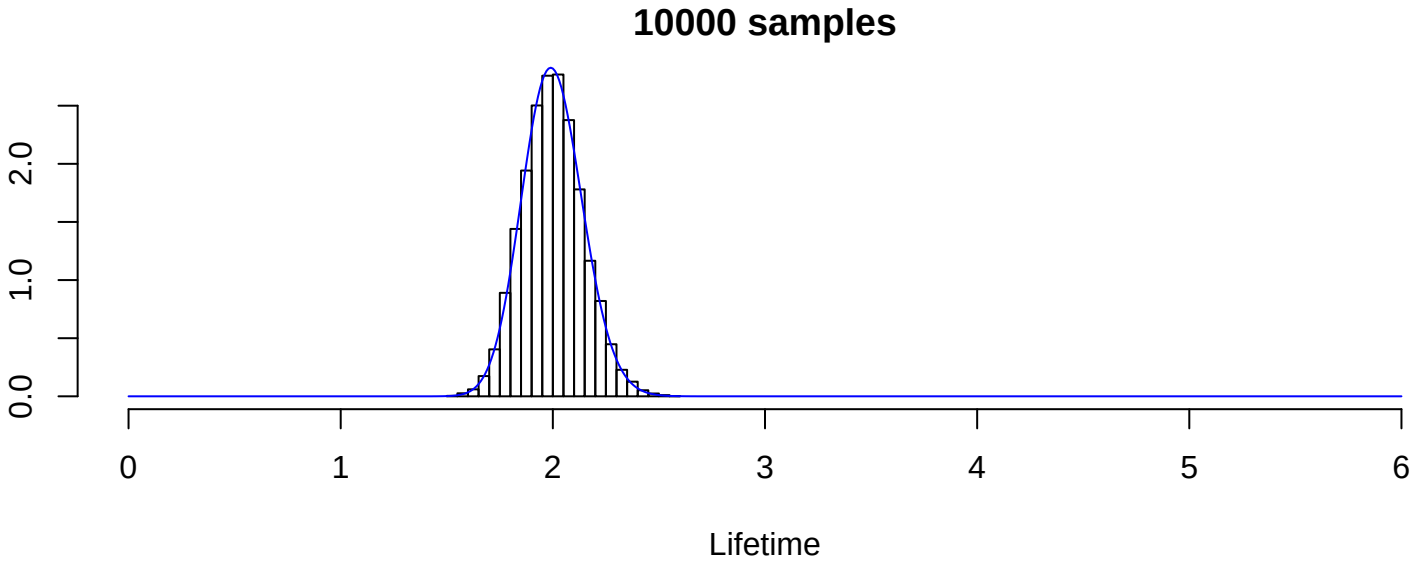

Sampling distribution − n = 2
0.0
0.2
0.4
0.6
0
1
2
3
4
5
6
Lifetime

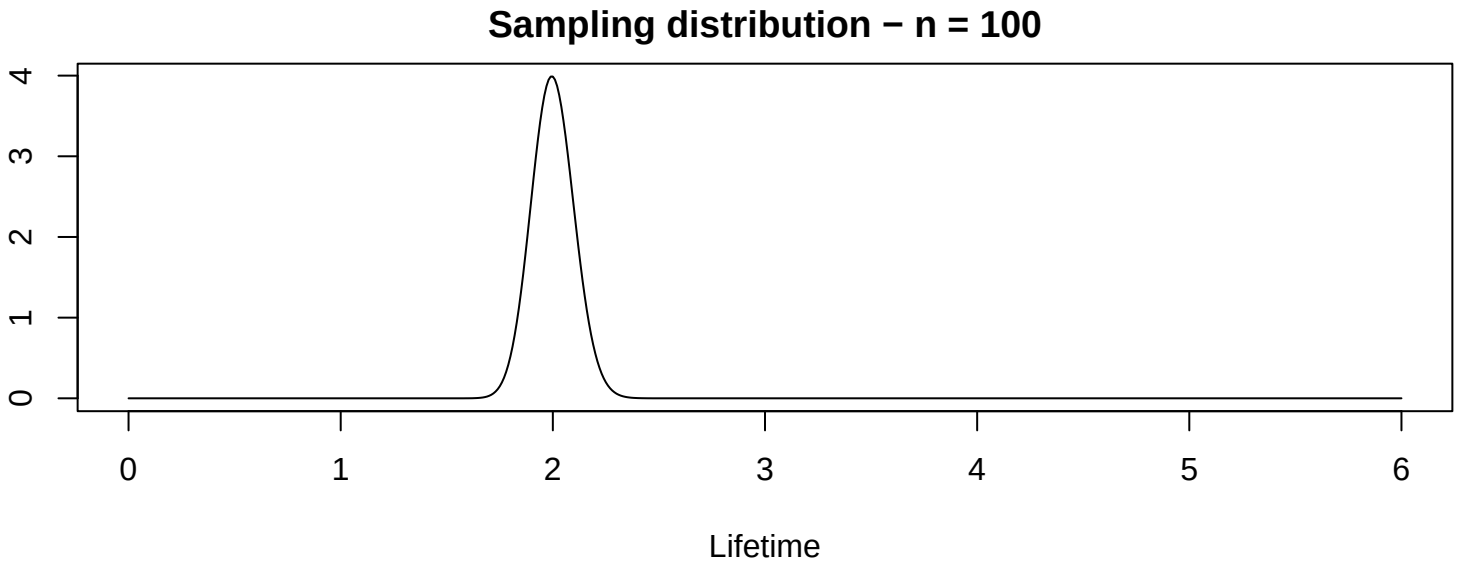

10000 samples
0.0
0.2
0.4
0.6
0
1
2
3
4
5
6
Lifetime

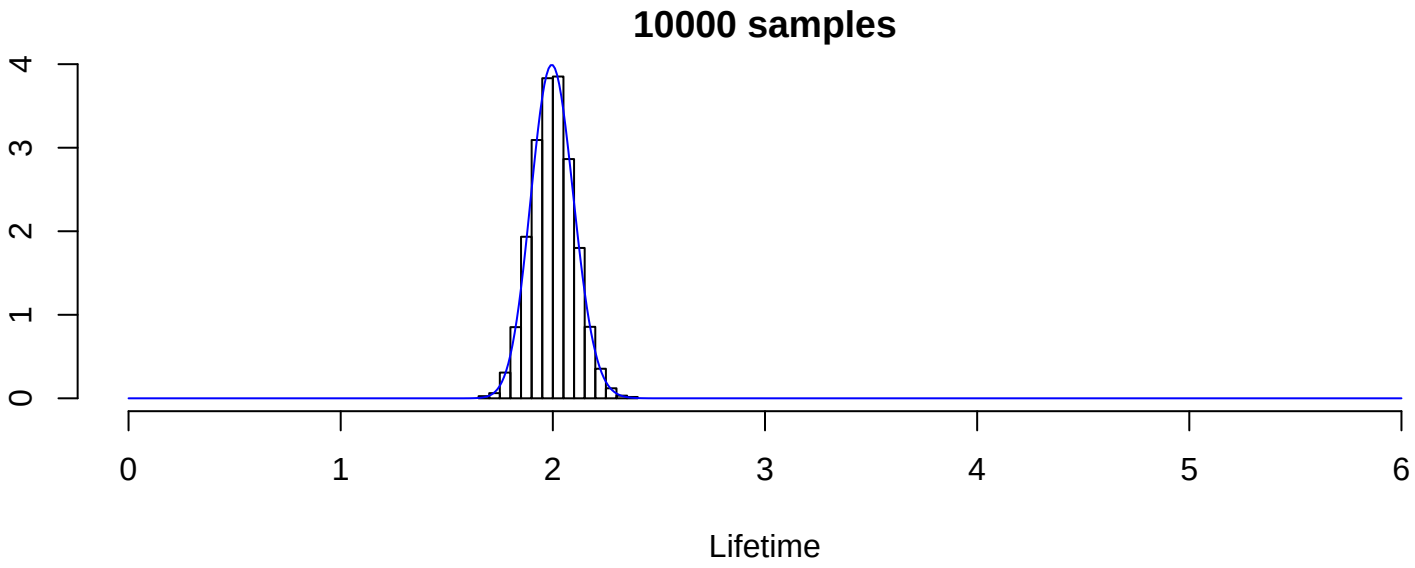

Sampling distribution − n = 10
0.0
0.4
0.8
1.2
0
1
2
3
4
5
6
Lifetime


10000 samples
0.0
0.4
0.8
1.2
0
1
2
3
4
5
6
Lifetime

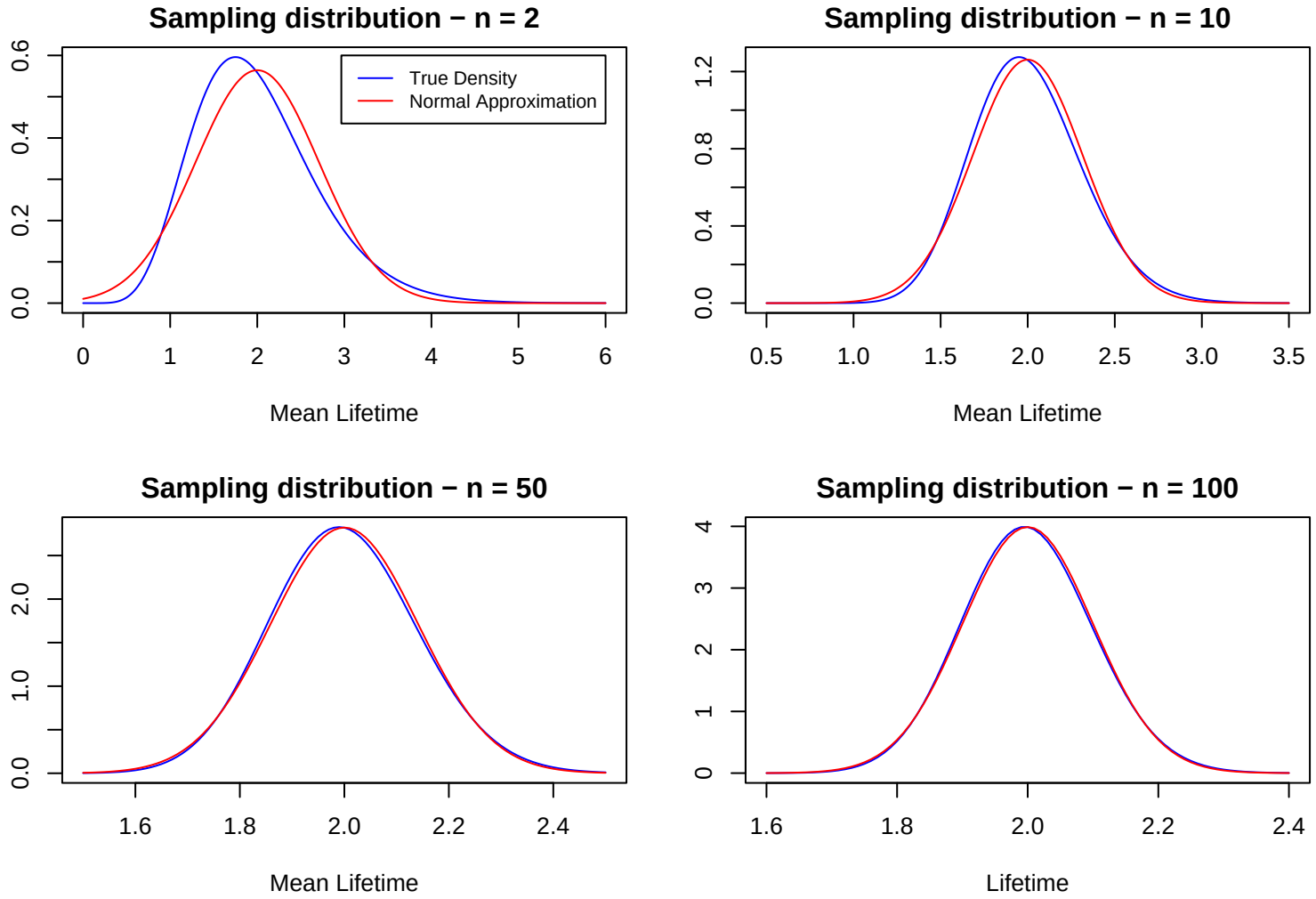

Sampling distribution − n = 50
0.0
1.0
2.0
0
1
2
3
4
5
6
Lifetime


10000 samples
0.0
1.0
2.0
0
1
2
3
4
5
6
Lifetime


Sampling distribution − n = 100
0
1
2
3
4
0
1
2
3
4
5
6
Lifetime


10000 samples
0
1
2
3
4
0
1
2
3
4
5
6
Lifetime

As the sample size n increases, the sampling distribution of $\bar{X}$ approaches a normal distribution.



Central Limit Theorem

Assume that $X_1, X_2, \ldots, X_n$ are independent and identically distributed with mean μ_x and standard deviation σ_x. Then for large sample sizes n, the distribution $\bar{X}$ is approximately $N(\mu_x, \frac{\sigma_x}{\sqrt{n}})$

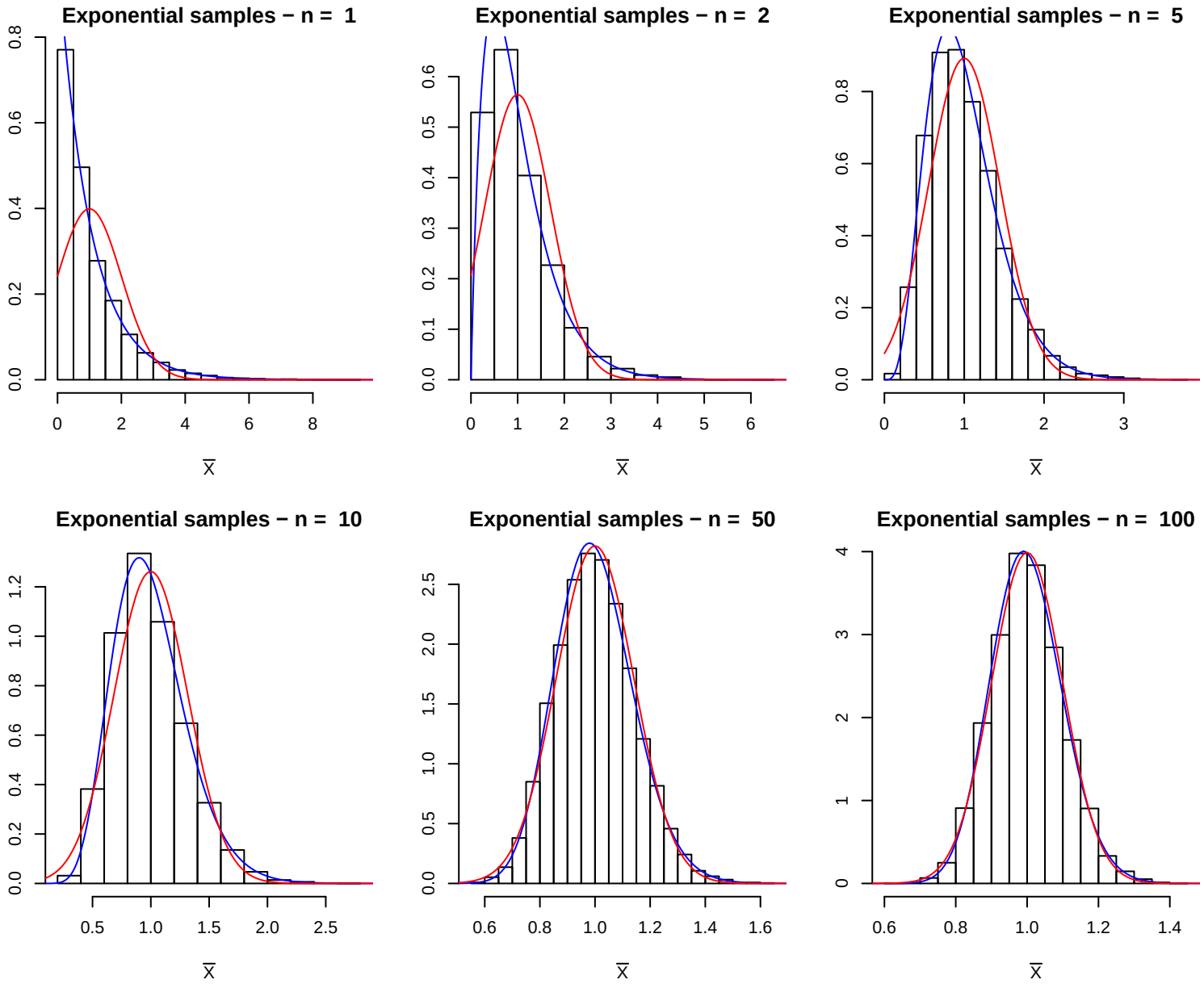
Note that the normal approximation for the distribution of $\hat{p}$ is just a special case of the Central Limit Theorem (CLT).

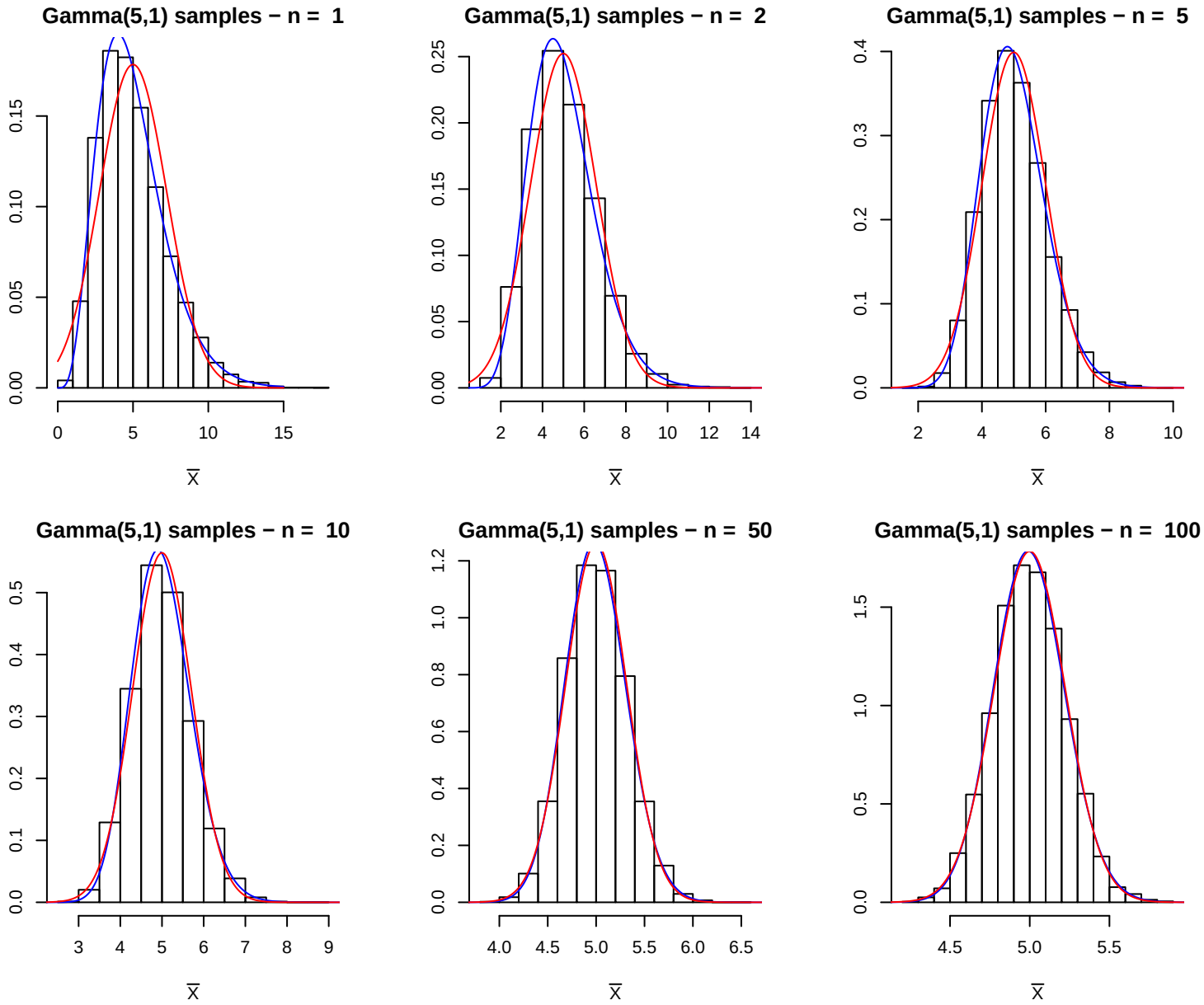
What is a large sample size?

As we saw with the binomial distribution, how well the normal approximation for $\hat{p}$ depended on how p, which influenced the skewness of the population distribution.

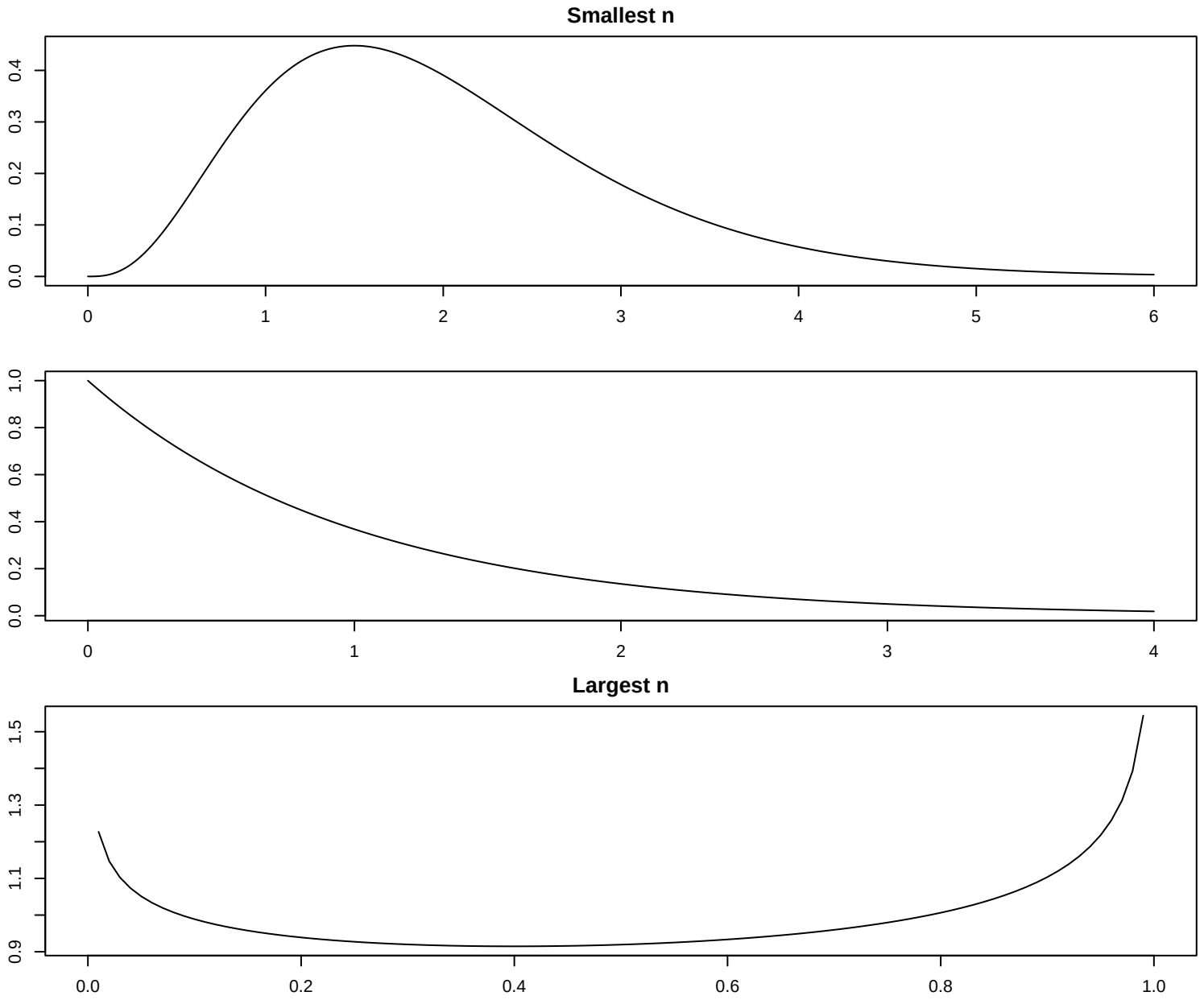
This same idea holds for the general for the CLT.

If the observations are normal, $\bar{X}$ has precisely a normal distribution for any n, since, as we've discussed before, sums of normals are normal. However the farther the density looks like a normal, the bigger that n needs to be for the approximation to do a good job.


Exponential samples − n = 1
Exponential samples − n = 2
Exponential samples − n = 5
Exponential samples − n = 10
Exponential samples − n = 50
Exponential samples − n = 100
X̄


Gamma(5,1) samples − n = 1
0.00
0.05
0.10
0.15
0 5 10 15
X̄
Gamma(5,1) samples − n = 2
0.00
0.05
0.10
0.15
0.20
0.25
2 4 6 8 10 12 14
X̄
Gamma(5,1) samples − n = 5
0.0
0.1
0.2
0.3
0.4
2 4 6 8 10
X̄
Gamma(5,1) samples − n = 10
0.0
0.1
0.2
0.3
0.4
0.5
3 4 5 6 7 8 9
X̄
Gamma(5,1) samples − n = 50
0.0
0.2
0.4
0.6
0.8
1.0
1.2
4.0 4.5 5.0 5.5 6.0 6.5
X̄
Gamma(5,1) samples − n = 100
0.0
0.5
1.0
1.5
4.5 5.0 5.5
X̄

The sampling distributions based on observations from the Gamma(5,1) distribution ($\mu_x = 5, \sigma_x = \sqrt{5}$) look more normal than the sampling distributions based on observations from the Exponential distribution ($\mu_x = 1, \sigma_x = 1$). For every sample size, the distribution of $\bar{X}$ is more normal for the Gamma distribution than the Exponential distribution.

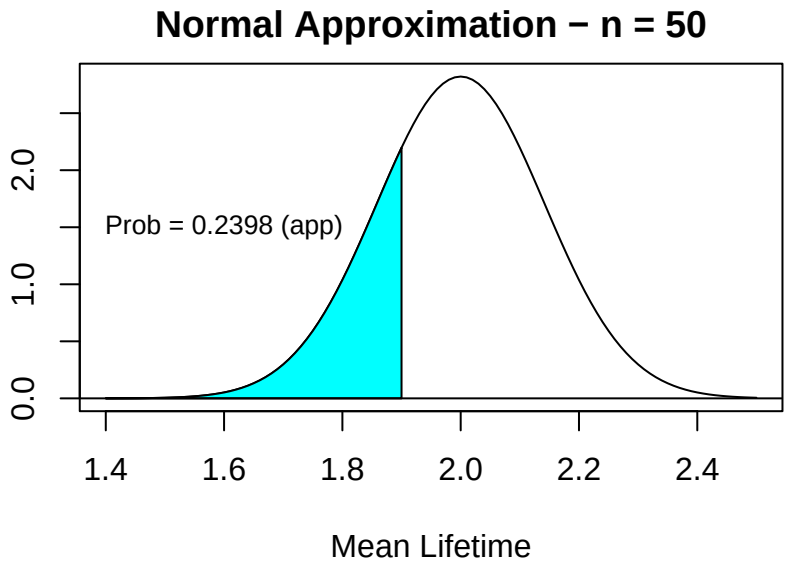

Smallest n
0.0
0.1
0.2
0.3
0.4
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
Largest n
0.9
1.1
1.3
1.5
0.0
0.2
0.4
0.6
0.8
1.0

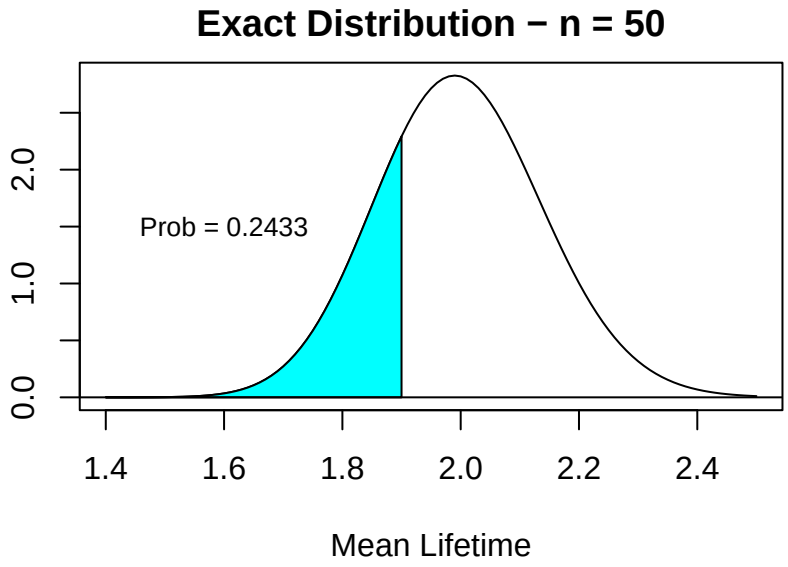
The Central Limit Theorem allows us to make approximate probability statements about $\bar{X}$ using a normal distribution, even though X is not normally distributed.

So for the light bulb example with $n = 50$, the $P[\bar{X} \leq 1.9]$ can be approximated by the normal distribution.

$$\begin{aligned} P[\bar{X} \leq 1.9] &= P\left[\frac{\bar{X} - 2}{\frac{1}{\sqrt{50}}} \leq \frac{1.9 - 2}{\frac{1}{\sqrt{50}}}\right] \\ &= P[Z \leq -0.707] \\ &\approx 0.2398 \end{aligned}$$

The true probability is 0.2433.





The CLT can also be used to make statements about sums of independently and identically distributed random variables. Let

$$S = X_1 + X_2 + \ldots + X_n = n\bar{X}$$

Then

$$\begin{aligned} \mu_S &= n\mu_{\bar{x}} = n\mu_x \\ \sigma_S^2 &= n^2\sigma_{\bar{x}}^2 = n^2\frac{\sigma_x^2}{n} = \sigma_x^2 n \\ \sigma_S &= \sigma_x\sqrt{n} \end{aligned}$$

So S is approximately $N(n\mu_x, \sigma_x\sqrt{n})$ distributed.

Relaxing assumptions for the CLT

The assumptions for the CLT can be relaxed to allow for some dependency and some differences between distributions. This is why much data is approximately normally distributed. The more general versions of the theorem say that when an effect is the sum of a large number of roughly equally weighted terms, the effect should be approximately normally distributed.

For example, peoples heights are influenced a (potentially) large number of genes and by various environmental effects. Histograms of adult men's and women's heights are both well described by normal densities.

Another consequence of this, is that $\bar{X}$ based on a simple random samples, with fairly large sampling fractions are also approximately normally distributed.

Sampling With and Without Replacement

Simple Random Sampling is sometimes referred to sampling without replacement. Once a member of the population is sampled, it can't be sampled again. As discussed before, the without replacement action of the sampling introduces dependency into the observations.

Another possible sampling scheme is sampling with replacement. In this case, when a member of a population is sampled, it is returned to the population and could be sampled again. This occurs if your sampling scheme is similar to repeated rolling of a dice. There is no dependency between observations in this case, as at each step, the members population that could be sampled are the same. This situation is also equivalent to drawing from an "infinite" population.

When SRS is used, the variance of the sampling distribution needs to be adjusted for dependency induced by the sampling.

The correction is based on the Finite Population Correction (FPC)

$$f = \frac{N - n}{N}$$

which is the fraction of the population which is not sampled.

Then the variance and standard deviation of $\bar{X}$ are

$$\begin{aligned} \sigma_{\bar{x}}^2 &= \frac{\sigma_x^2}{n} f \\ \sigma_{\bar{x}} &= \frac{\sigma_x}{\sqrt{n}} \sqrt{f} \end{aligned}$$

So when a bigger fraction of the population is sampled (so f is smaller), you get a smaller spread in the sampling distribution.

However when n is small relative to N, this correction has little effect. For sample, if 10% of the population is sampled, so $f = 0.9$, the standard deviation of the sampling distribution is about 95% ($\sqrt{0.9}$) of the standard deviation for that of with replacement sampling. If a 1% sample is taken, the correction on the standard deviation is 0.995.

Except when fairly large sampling fractions occur, the FPC is usually not used.