General Standardization

We have learned through properties of the normal distribution that the distribution of the sample average is also normal.

One special case is:

⇨ If $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu,\sigma^2)$

then $\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

This nice fact allows us to construct probability statements concerning the sample mean.

(a) P($\overline{X}_n$ > high limit)
(b) P($\overline{X}_n$ < low limit)
(c) P(low limit < $\overline{X}_n$ < high limit)
(d) P($\overline{X}_n$ > ?) = 0.05

etc....

To calculate these probabilities we need only standardize and look up the corresponding probability from the standard normal table in Pagano.

Remember:

We can standardize any **normal** random variable by subtracting its mean and dividing by its standard deviation. The sample mean is no exception:

$$\frac{\overline{\mathrm{X}}_n - \mu}{\sigma / \sqrt{n}} = \frac{\sqrt{n}\left(\overline{\mathrm{X}}_n - \mu\right)}{\sigma} = Z \sim N(0,1)$$

In addition, we see that more general standardization formula is:

$$\Rightarrow \qquad \frac{\overline{\mathrm{X}}_n - E\left(\overline{\mathrm{X}}_n\right)}{\sqrt{Var\left(\overline{\mathrm{X}}_n\right)}} = Z \sim N(0,1)$$

Example

In town Z, an average of 200 people per day visit the emergency room with a standard deviation of 15 people. What is the probability that the sample average over a 36 day period will exceed 204?

So,

$X_1, X_2, \ldots, X_{36}$ are i.i.d. $N(200,15^2)$

$$P\left(\overline{X}_n > 204\right) = P\left(\frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} > \frac{204 - \mu}{\sigma / \sqrt{n}}\right)$$

$$= P\left(Z > \frac{204 - 200}{15 / \sqrt{36}}\right)$$

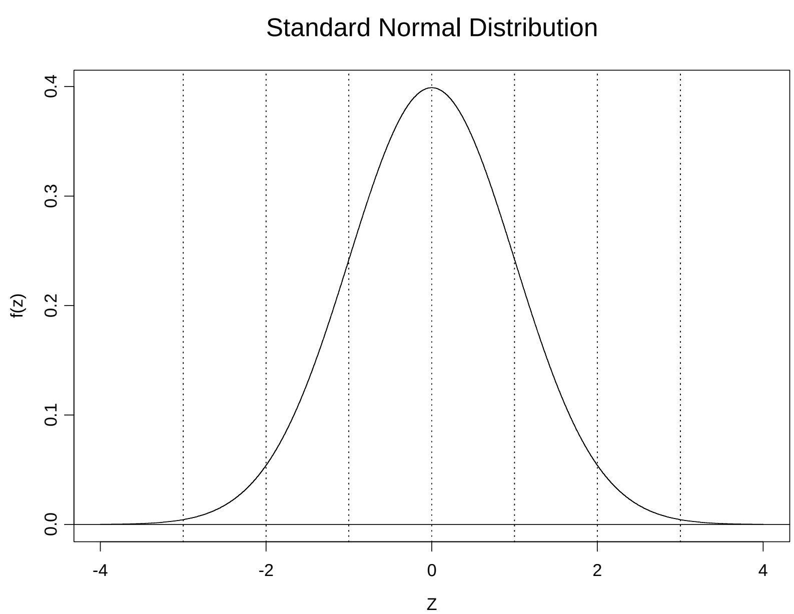
$$= P\left(Z > 1.6\right) = 0.0548$$

Back to the Law

We know that if X_1, X_2, ..., X_n are i.i.d. $N(\mu,\sigma^2)$

$$\text{then } \overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

And for normal random variables, almost 100% of the distribution lies between 3 standard deviations about the mean.



In fact, P(-3 < z < 3) = 0.9973

This means that the sample mean will be within 3 standard errors of the population mean with probability 0.9973 (because the standard error is the standard deviation of the sample mean).

Mathematically we write:

$$P(-3 < Z < 3) = P\left(-3 < \frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} < 3\right)$$

$$= P\left(\mu - 3\frac{\sigma}{\sqrt{n}} < \overline{X}_n < \mu + 3\frac{\sigma}{\sqrt{n}}\right)$$

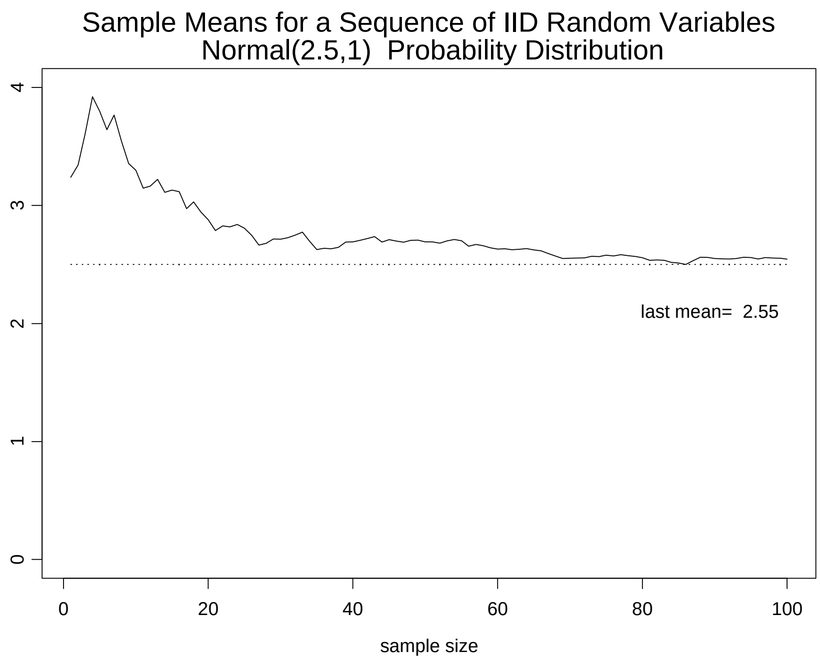
$$= 0.9973$$

So we expect that 99.73% of the time, the sample mean will fall between $\left[\mu - 3\frac{\sigma}{\sqrt{n}}, \mu + 3\frac{\sigma}{\sqrt{n}}\right]$.

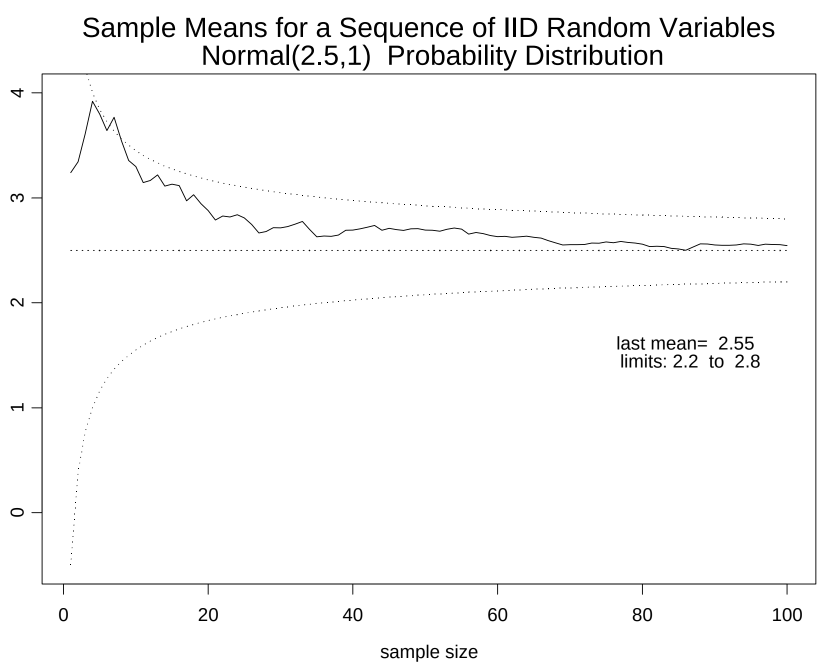
To demonstrate let's take another look at those great plots that demonstrated the Law of Large numbers!

Central Limit Theorem

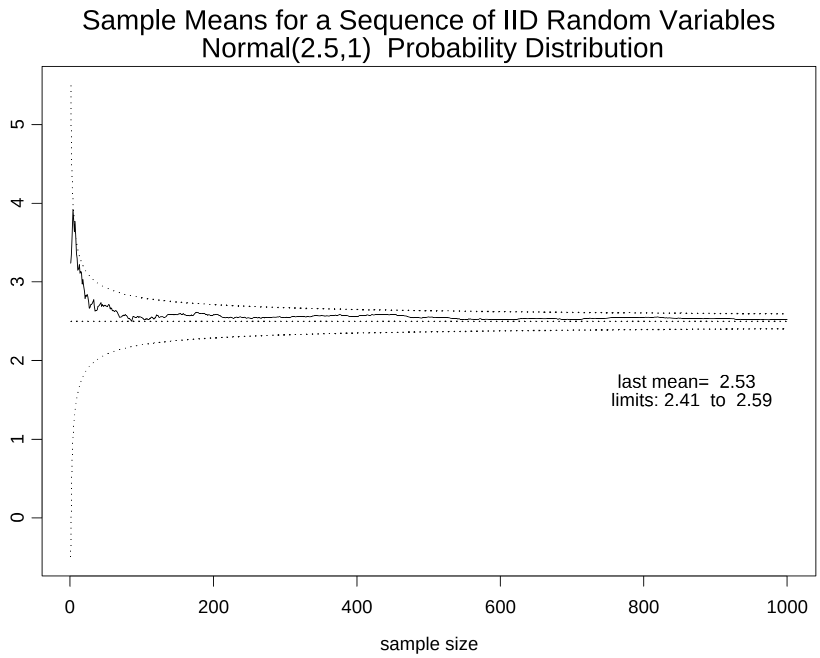
Suppose we collect 100 observations from a Normal (2.5,1) distribution. We see that



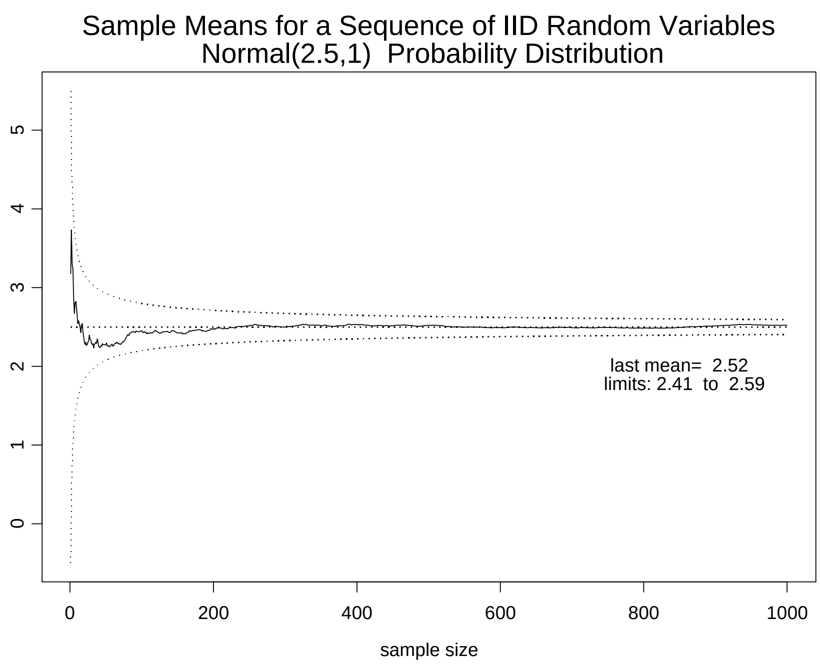
And with the limits we have:



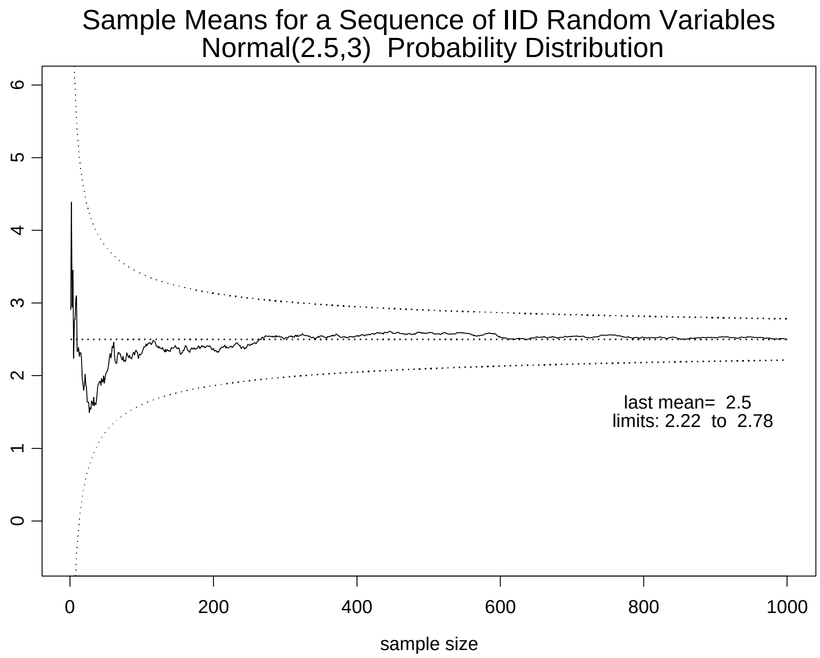
Here is the same sequence until 1,000:



And another sequence:



Here is a sequence from N(2.5,3):



In practice we never know μ and we can only estimate μ with $\overline{X}_n$. Thus our interval

$$\left[\mu - 3\frac{\sigma}{\sqrt{n}}, \mu + 3\frac{\sigma}{\sqrt{n}}\right]$$ is estimated with

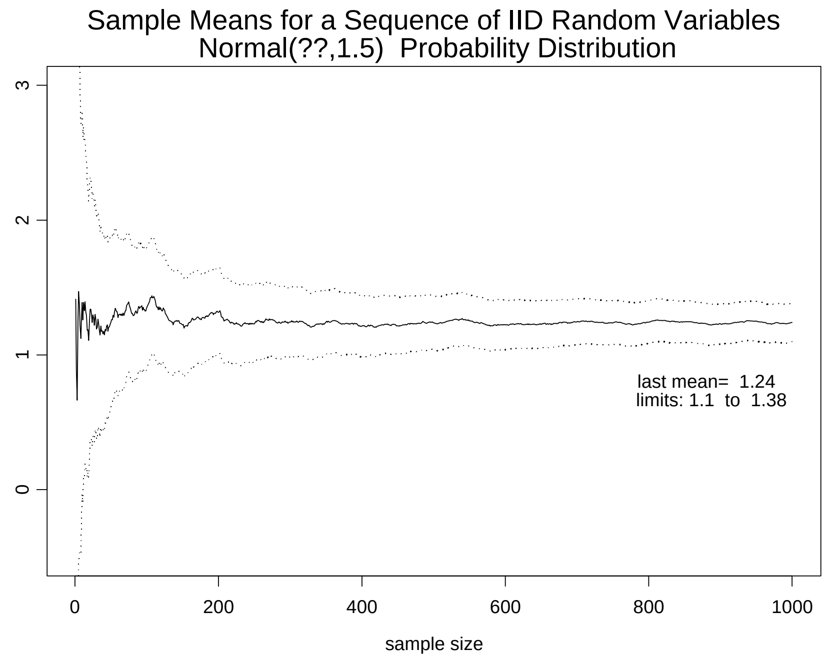
$$\left[\overline{X}_n - 3\frac{\sigma}{\sqrt{n}}, \overline{X}_n + 3\frac{\sigma}{\sqrt{n}}\right]$$

Interestingly enough:

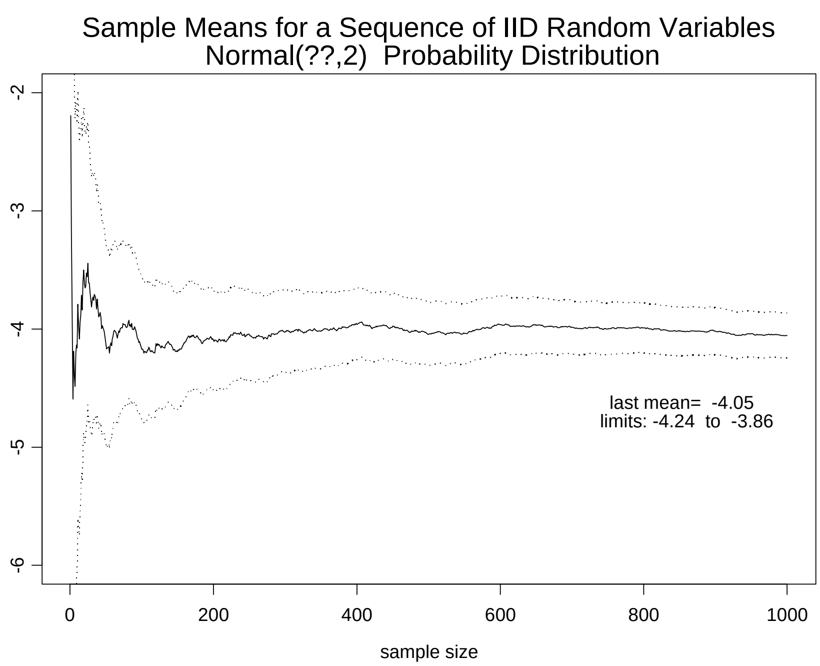
$$P\left(\overline{X}_n - 3\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}_n + 3\frac{\sigma}{\sqrt{n}}\right) =$$

$$P\left(\mu - 3\frac{\sigma}{\sqrt{n}} < \overline{X}_n < \mu + 3\frac{\sigma}{\sqrt{n}}\right) = 0.9973$$

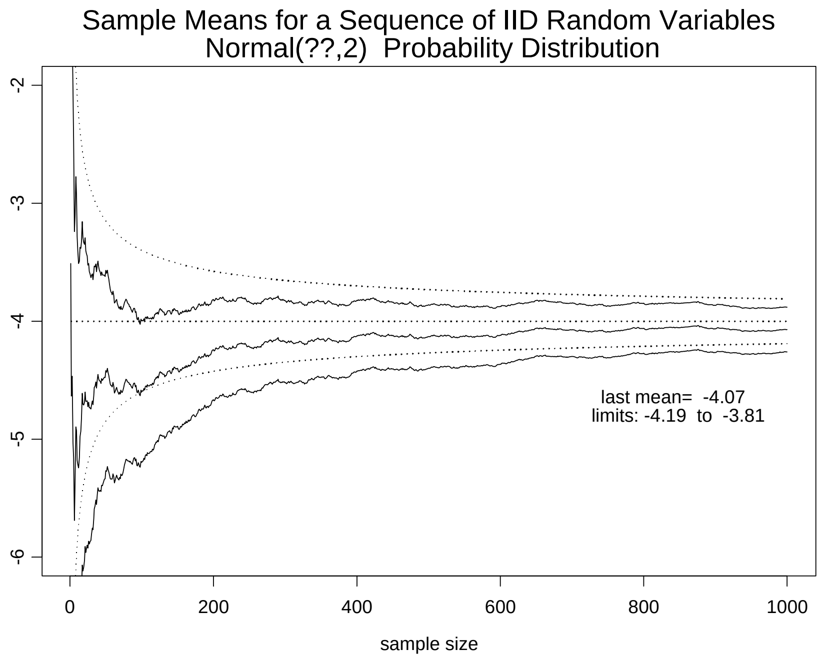
So 100 observations with variance 1.5 looks like:

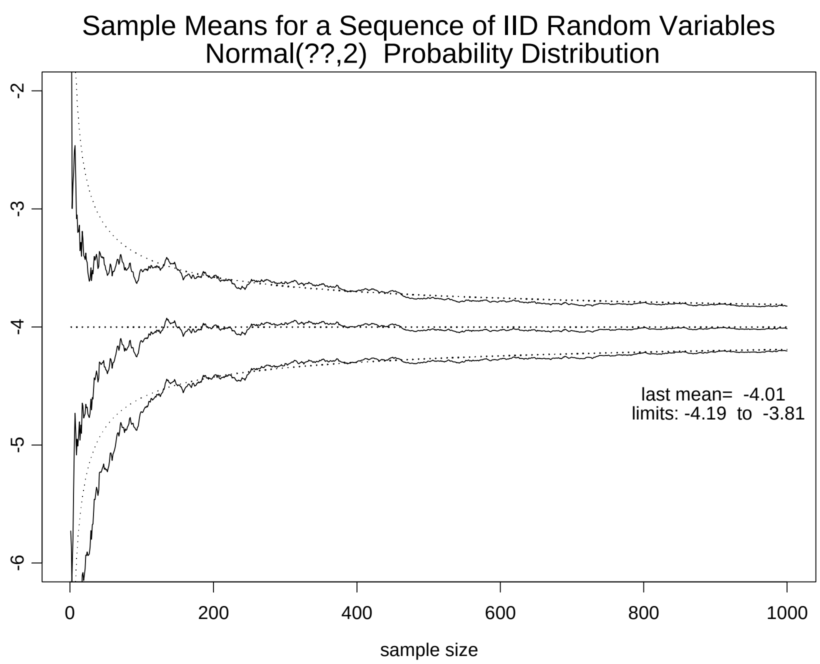


What is the mean here?



To see exactly what is going on, we can plot both the estimated interval and the true interval:





⇨ Notice how, as the sample size increases, the "spread" of the interval decreases, indicating that the variance (and the standard error) of $\overline{X}_n$ is decreasing.

For variables that are **<u>not</u>** normally distributed how can we describe the variability of the sample mean?

Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. Bernoulli(θ). We know that

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

But what can we say about

(a) P($\overline{X}_n$ > high limit)
(b) P($\overline{X}_n$ < low limit)
(c) P(low limit < $\overline{X}_n$ < high limit)
(d) P($\overline{X}_n$ > ?) = 0.05

We need to know the distribution of $\overline{X}_n$!

Central Limit Theorem:

Irrespective of the underlying distribution of the population (assuming E(X) exists), the distribution of the sample mean will be approximately normal in moderate to large samples.

Or

If X_1, X_2, ..., X_n are i.i.d. then

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

The central limit theorem tells us that we can approximate the distribution of the sample mean with a normal distribution. This implies that

$$\frac{\overline{X}_n - E(\overline{X}_n)}{\sqrt{Var(\overline{X}_n)}} = Z \text{ is approx } N(0,1)$$

in large distributions for any underlying probability model.

Example

Suppose X_1, X_2, ..., X_n are i.i.d. Ber(θ). Then in moderately large samples:

$$\frac{\overline{X}_n - E\left(\overline{X}_n\right)}{\sqrt{Var\left(\overline{X}_n\right)}} = \frac{\overline{X}_n - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} = Z \text{ is approx } N(0,1)$$

Question:
What is the probability that the sample proportion of success (out of 50 flips) is greater than 0.80, when the true probability of success is 0.75?

Answer:

$$P\left(\overline{X}_n > 0.80\right) = P\left(\frac{\overline{X}_n - \theta}{\sqrt{\theta(1-\theta)/n}} > \frac{0.80 - \theta}{\sqrt{\theta(1-\theta)/n}}\right)$$

$$= P\left(Z > \frac{0.80 - 0.75}{\sqrt{0.75(0.25)/50}}\right)$$

$$= P(Z > 0.8165) = 0.207$$

Remember that 20.7% is only an approximation!
(Called: Normal approximation to the Bernoulli)

Example

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. Poisson(λ). Then in moderately large samples:

$$\frac{\overline{X}_n - E(\overline{X}_n)}{\sqrt{Var(\overline{X}_n)}} = \frac{\overline{X}_n - \lambda}{\sqrt{\frac{\lambda}{n}}} = Z \text{ is approx } N(0,1)$$

Question:

What is the probability that the sample mean of 25 observations will be greater than 3.4, when the true event rate is 2.4?

Answer:

$$P(\overline{X}_n > 3.4) = P\left(\frac{\overline{X}_n - \lambda}{\sqrt{\lambda / n}} > \frac{3.4 - \lambda}{\sqrt{\lambda / n}}\right)$$

$$= P\left(Z > \frac{3.4 - 2.4}{\sqrt{2.4 / 25}}\right)$$

$$= P(Z > 3.22) = 0$$

Remember that this is only an approximation!
(Called: Normal approximation to the Poisson)

The Central Limit Theorem implies that the sample mean will be within approximately 3 standard errors of the population mean with probability 99.73 in moderate to large samples.

Mathematically we write (again an approximation):

$$P(-3 < Z < 3) = P\left(-3 < \frac{\overline{X}_n - E(\overline{X}_n)}{\sqrt{Var(\overline{X}_n)}} < 3\right)$$

$$= P\left(E(\overline{X}_n) - 3\sqrt{Var(\overline{X}_n)} < \overline{X}_n < E(\overline{X}_n) + 3\sqrt{Var(\overline{X}_n)}\right)$$

is approximately 99.73% in large samples.

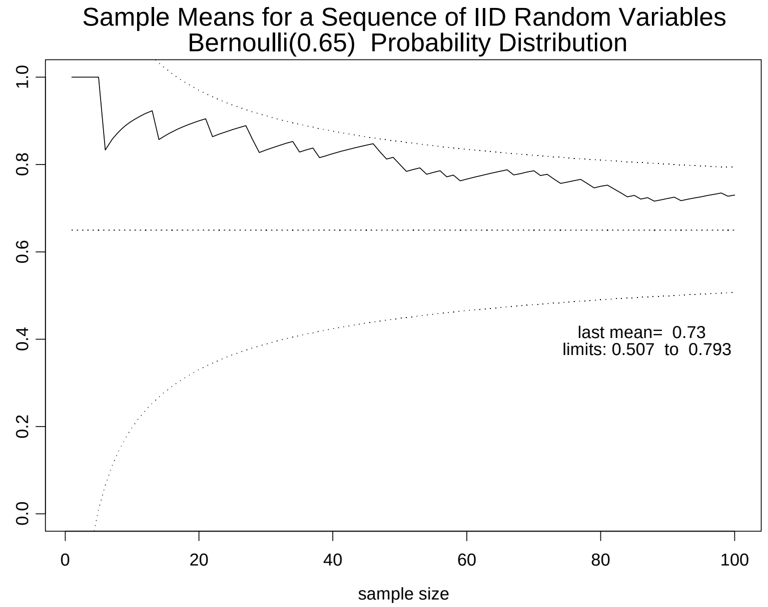
So we expect that, in large samples, 99.73% of the time, the sample mean of any sequence of independent observations will fall between.

$$[E(\overline{X}_n) - 3\sqrt{Var(\overline{X}_n)}, \quad E(\overline{X}_n) + 3\sqrt{Var(\overline{X}_n)}]$$

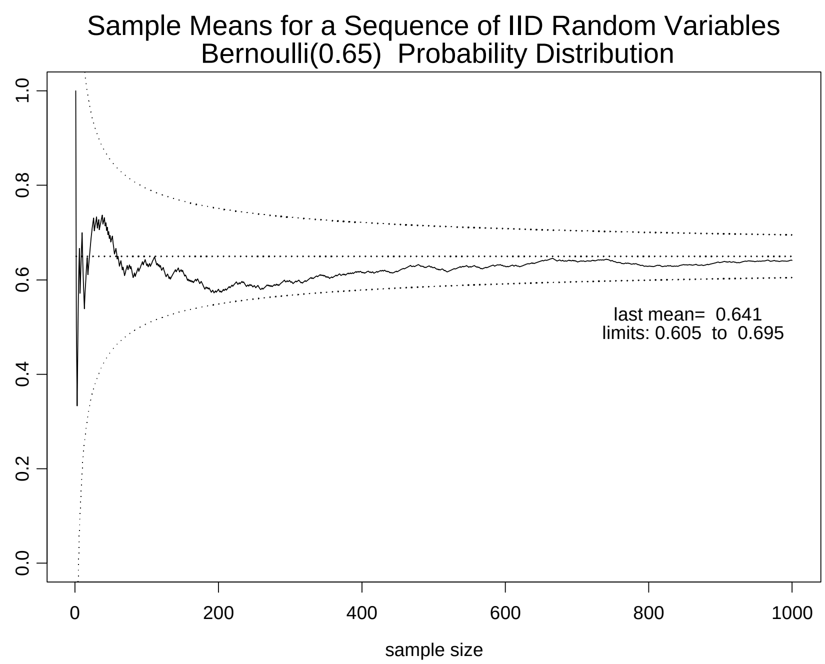
To demonstrate let's take another look at those great plots that demonstrated the Law of Large numbers!

Central Limit Theorem

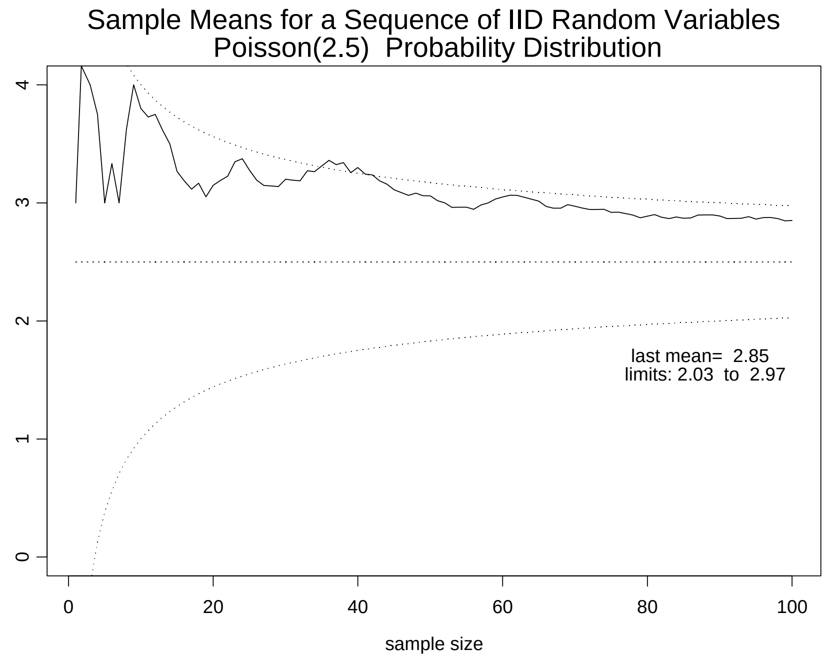
Suppose we collect 100 observations from a Bernoulli(0.65) distribution. We see that

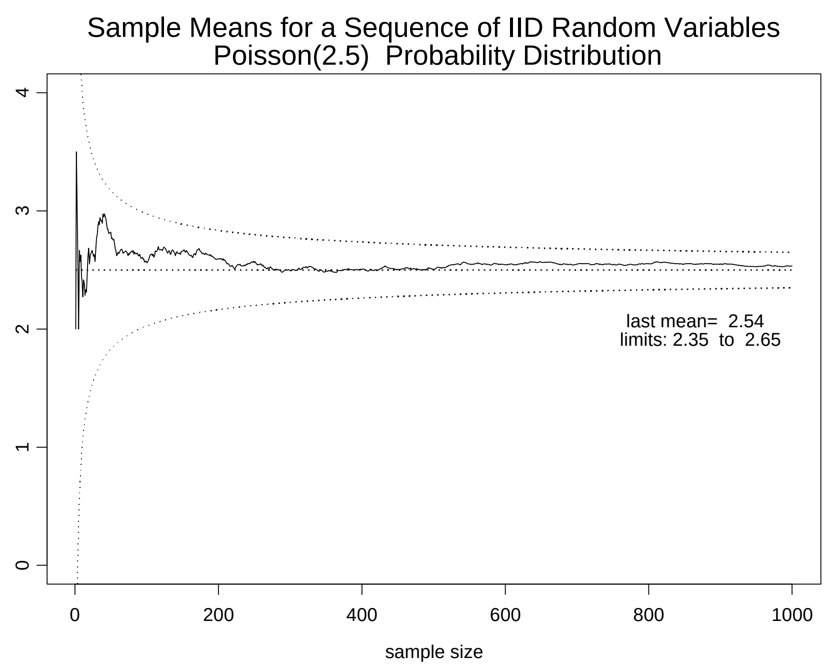


And for 1,000 observations



And for the Poisson:





Just like before, we never really know E(X), so we can only estimate it with $\overline{X}_n$. Thus our interval

$$\left[E(X)-3\sqrt{\frac{Var(X)}{n}},E(X)+3\sqrt{\frac{Var(X)}{n}}\right]$$

is estimated with

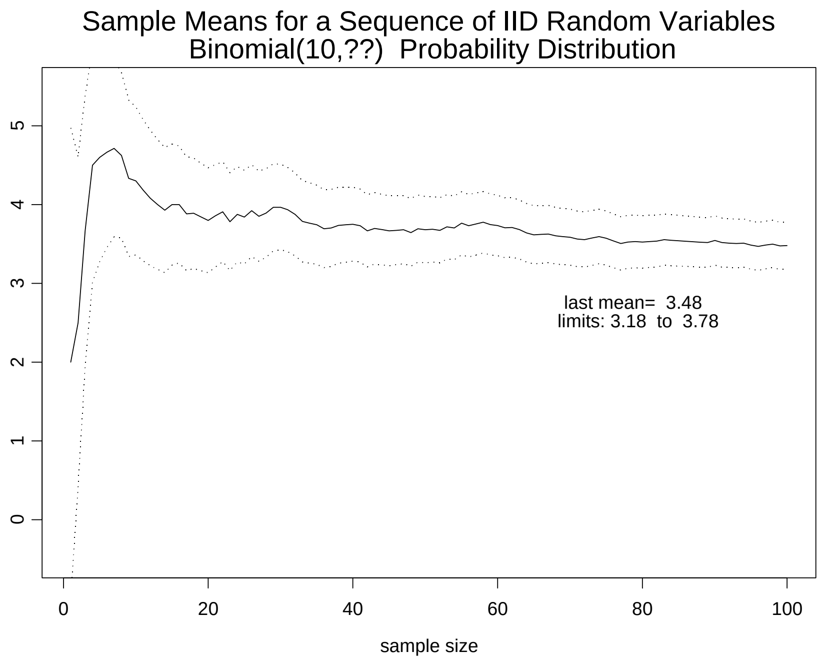
$$\left[\overline{X}_n-3\sqrt{\frac{Var(X)}{n}},\overline{X}_n+3\sqrt{\frac{Var(X)}{n}}\right]$$

Example:
Suppose that we observed 100 Binomial (10, 0.33) trials <u>without</u> knowing that $\theta=0.33$. To construct the above interval we would have a problem, because $Var(X)=10\theta(1-\theta)$, but we do not know idea what theta may be.

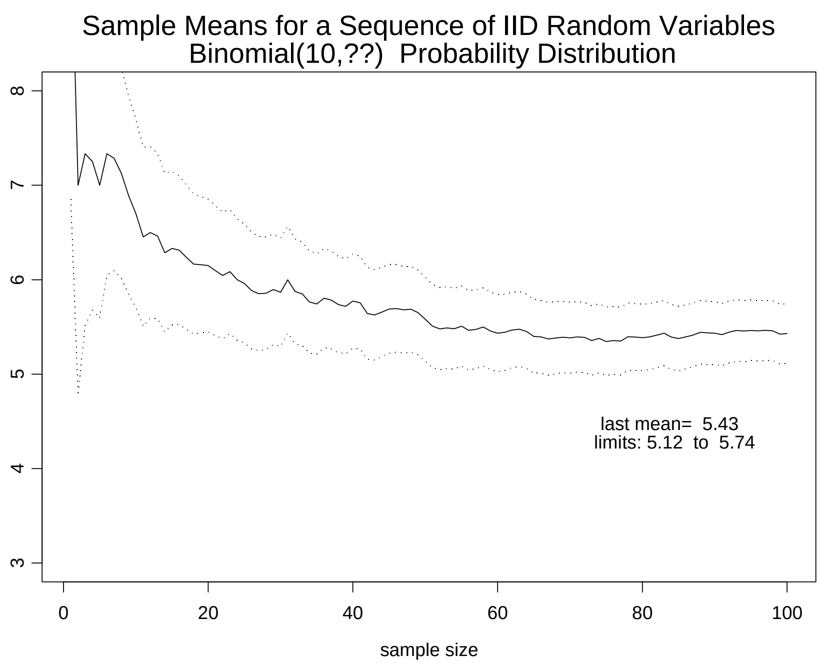
For our plots in this lecture I have just assume that we know θ. In practice we would simple replace θ with $\hat{p}$ (the sample proportion of successes) in the variance term.

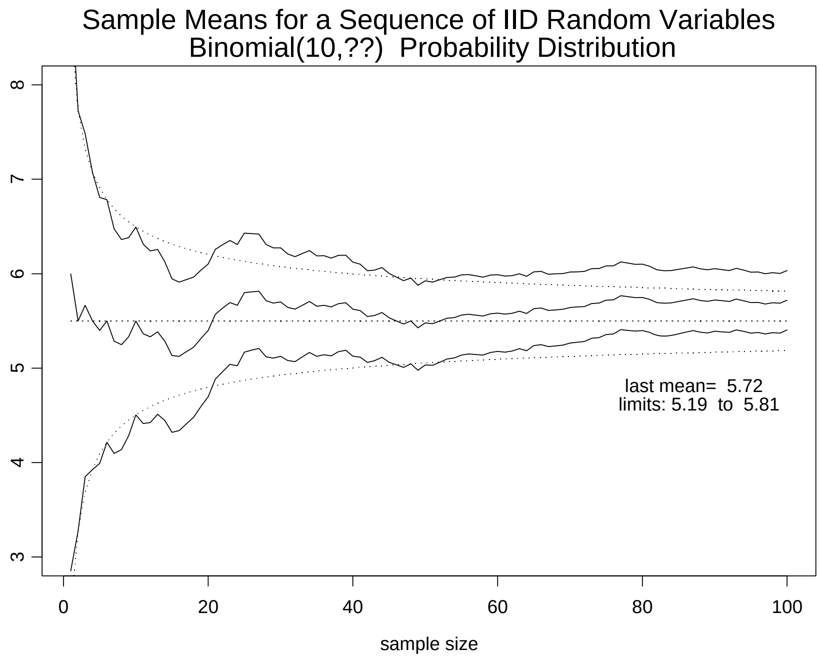
For now, we'll just assume we know the variance.

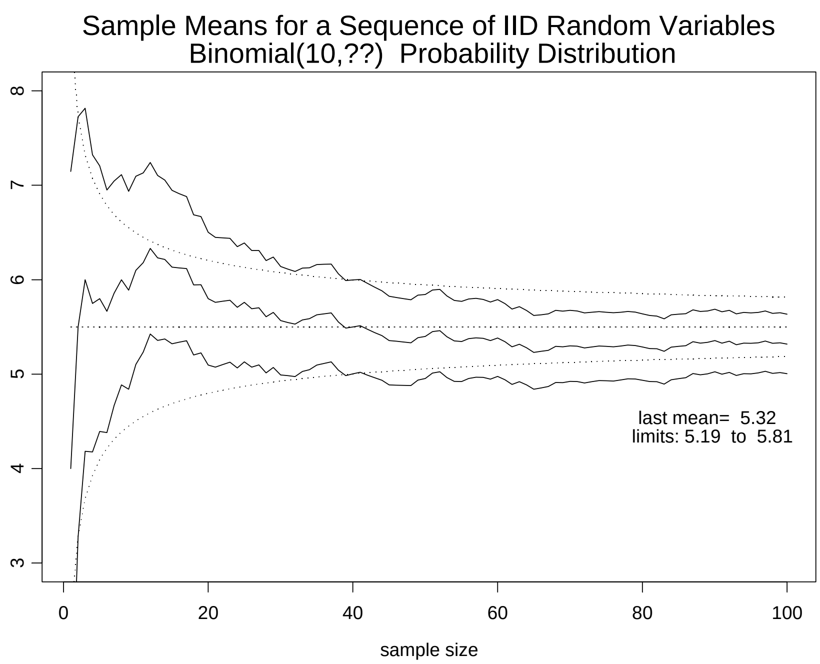
Can you guess E(X) and theta?



and now?




Sample Means for a Sequence of IID Random Variables
Binomial(10,??) Probability Distribution
last mean= 5.72
limits: 5.19 to 5.81
sample size
0
20
40
60
80
100
3
4
5
6
7
8


Sample Means for a Sequence of IID Random Variables
Binomial(10,??) Probability Distribution
last mean= 5.32
limits: 5.19 to 5.81
sample size
0
20
40
60
80
100
3
4
5
6
7
8

Everything settles down with a lot of observations: