Central Limit Theorem

The Deepest Thought Ever Thunk

Paul E. Johnson[1,2]

[1]Department of Political Science
University of Kansas

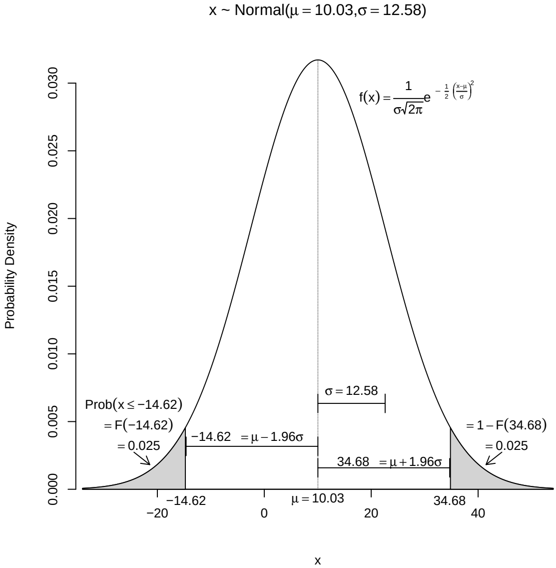
[2]Center for Research Methods and Data Analysis
University of Kansas

2011

Outline

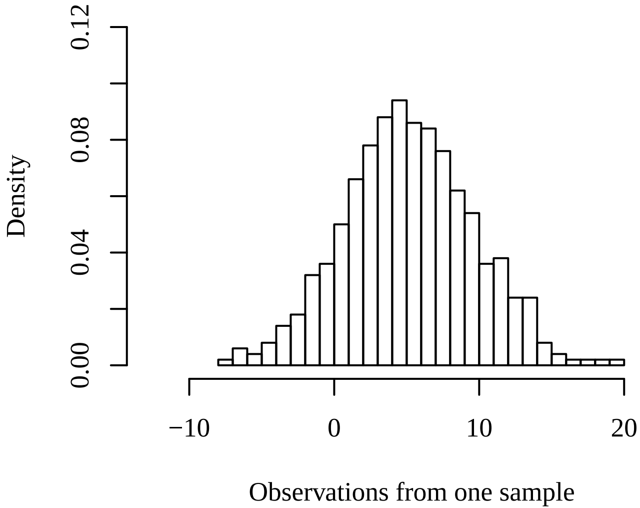
1. The Difference Between A Sample and The Truth
2. Sampling Distribution
3. Asymptotic Properties
4. The Central Limit Theorem

Do you remember your friend, the Normal Distribution?

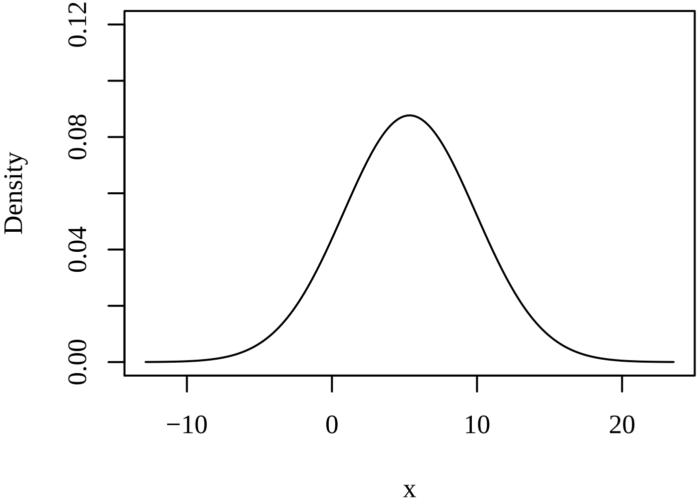


- Single Peaked
- Symmetric
- $E[x] = \mu$
- $Var[x] = \sigma^2$
- $SD[x] = \sigma$

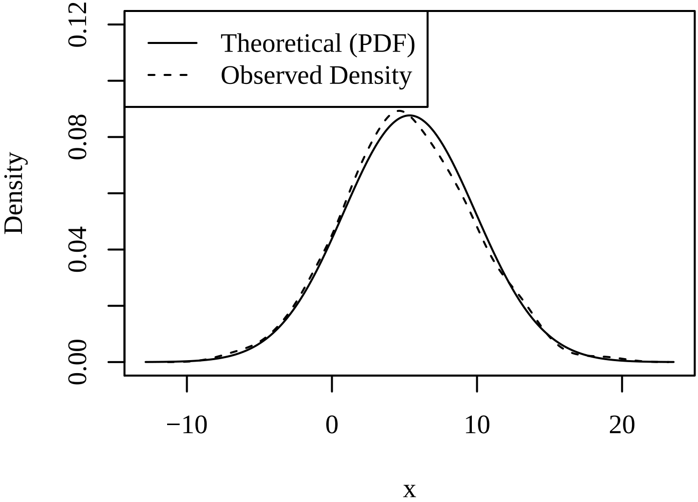
Draw one Normal Sample from $N(5.353, 4.55^2)$



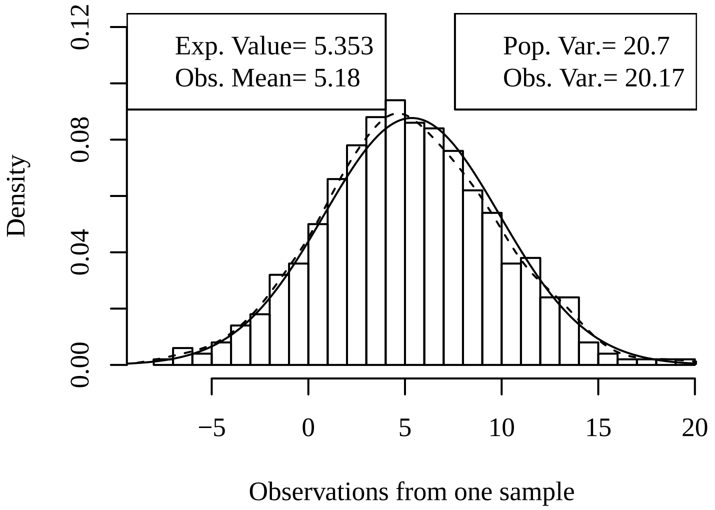
The Theoretical PDF Is This:



But the Observed Density Differs



But the Observed Density Differs



The Basic Idea

- Draw a lot of samples
  - Collect M samples of size N
  - Calculate the mean for each sample
- What distribution will be observed among all of those means?
- Do you expect the distribution of means will be different from the distribution of x itself?

Important Term: Sampling Distribution

- Definition: Sampling Distribution is the PDF of the "true" distribution for an estimator like $\bar{x}$
- Drawing 500, or 5000, or 100,000 samples, and then creating a histogram of the estimates, approximates the sampling distribution.
- This histogram (or observed density) will not be exactly the same as the sampling distribution, but it might get very close!

General Claims about the Sampling Distribution of $\bar{x}$

This is the first set of facts I need to establish

- If $E[x] = \mu$, then $E[\bar{x}] = \mu$
- If $Var[x] = \sigma^2$, then $Var[\bar{x}] = \frac{Var[x]}{N}$
- Which implies $SD[\bar{x}] = \frac{SD[x]}{\sqrt{(N)}}$

In Other Words...

The distribution of $\bar{x}$

- Is Centered on the same spot as x_i
- But $\bar{x}$ is clusterd much more "tightly' than the distribution of x_i itself.

That's impossibly easy to see

- Algebraically.
- By simulation.

Let's define terms.

The mean of a sample $x_1, x_2, x_3, \ldots, x_N$ is:

$$\bar{x} = \frac{1}{N}\sum_{i}^{N} x_i \tag{1}$$

If we have data on the frequency of each possible score x_j, calculate proportions

$$Prop.(x_j) = \frac{Frequency(x = x_j)}{N} \tag{2}$$

$$Mean(x_i) = \bar{x} = \sum_{j=1}^{m} Prop(x_j)x_j \tag{3}$$

where $Prop(x_j)$ is the proportion of observations that have value x_j. (sums across possible values of x_j, rather than summing across all individuals observed).

The Expected Value of x, $E[x]$

- EV=Another term for the "population mean" or "true mean"
- Recall, population=the random process that generates x_i.
- discrete distribution makes it easiest to compare formulae for $\bar{x}$ and $E[x]$
  - f is a "probability mass function"

$$Expected\ Value(x) = E[x] = \sum f(x_j)x_j \tag{4}$$

  - Similar to sample mean formula, except replace the "observed proportion" ($Prop(x_j)$) with the "theoretical probability" $f(x_j)$.
- Similar for a continuous distribution with pdf $f(x)$

$$E[x] = \int_{-\infty}^{+\infty} f(x)\,x\,dx. \tag{5}$$

One Little Tricky Bit Needs explaining First

- Think of a "variable" as one single observation from a distribution

$$x_i \tag{6}$$

- We were comfortable discussing a variable x as a collection of observations.
- We said x is normally distributed, usually thinking of a collection
- Now think of x_1, x_2 and so forth as separate variates from the same distribution.
- Appeal to Intuition. $E[x] = E[x_1] = E[x_2] = \ldots E[x_N]$
- To me, that was the only really surprising idea in all of this.

Calculate the Expected Value of $\bar{x}$

$$
\begin{aligned}
E[\bar{x}] = & \quad E\left[\frac{x_1+x_2+x_3+\ldots+x_N}{N}\right] \\
& = \frac{1}{N}\left\{E[x_1]+E[x_2]+E[x_3]+\ldots+E[x_N]\right\} \\
& = \frac{1}{N}\left\{N \cdot E[x]\right\} \\
& = E[x]
\end{aligned}
$$

Conclusion: The expected value of the mean is the same as the expected value of one draw from a given distribution.
Implication: $\bar{x}$ is an **unbiased estimator** of $E[x]$

Variance

- Recall Variance in a sample is the average of squared errors (aka "mean square error")

$$Variance(x_i) = \frac{1}{N}\sum(x_i - \bar{x})^2 \tag{7}$$

- Maybe you divide by $N-1$ in order to make this a 'consistent' estimator. Not a huge issue at this point.
- With frequency data:

$$Variance(x_i) = \sum Prop.(x_j)(x_j - \bar{x})^2 \tag{8}$$

where $Prop(x_j)$ is the proportion of observations that have value x_j.

Population Variance, same as Theoretical Variance

The "population variance" of the random process that generates x_i.
For discrete variable, use the PMF in place of $Prop.(x)$:

$$Theoretical\ Variance(x_i) = \sum f(x_i)(x_i - \bar{x})^2 \tag{9}$$

For a continuous variable f, use the PDF instead of proportions:

$$Theoretical\ Variance(x_i) = \int f(x_i)(x_i - \bar{x})^2 dx_i \tag{10}$$

Recall the Variance of A Sum

The variance of a sum of two variables $x1$ and $x2$ can be found:

$$Var[x1 + x2] = Var[x1] + Var[x2] + 2Cov[x1, x2] \tag{11}$$

And

$$Var[ax1 + bx2] = a^2 Var[x1] + b^2 Var[x2] + 2abCov[x1, x2] \tag{12}$$

Here a and b are constants.
We want a simple result, so we often assume the $Cov[x1, x2] = 0$ on the grounds that the observations are "statistically independent."

Calculate the Variance of the Mean

What is the variance of the mean itself?

$$Var[\bar{x}] = Var[\frac{1}{N}x_1 + \frac{1}{N}x_2 + \ldots + \frac{1}{N}x_N] \quad (13)$$

Invoking the "statistical independence" principle to eliminate the Covariance terms, we apply the "Variance of a sum" rule

$$Var(\frac{1}{N}x_1 + \frac{1}{N}x_2 + \ldots + \frac{1}{N}x_N) = \quad (14)$$

$$\frac{1}{N^2}Var(x_1) + \frac{1}{N^2}Var(x_2) + \ldots + \frac{1}{N^2}Var(x_N) \quad (15)$$

If all the observations were drawn from the same random process–the same population–then they all have the same variance, which is just $Var(x_i)$. So the previous instantly reduces to this:

$$Var(\bar{x}) = \frac{1}{N^2}\frac{NVar(x_i)}{1} \tag{16}$$

$$= \frac{1}{N} Var(x_i) \tag{17}$$

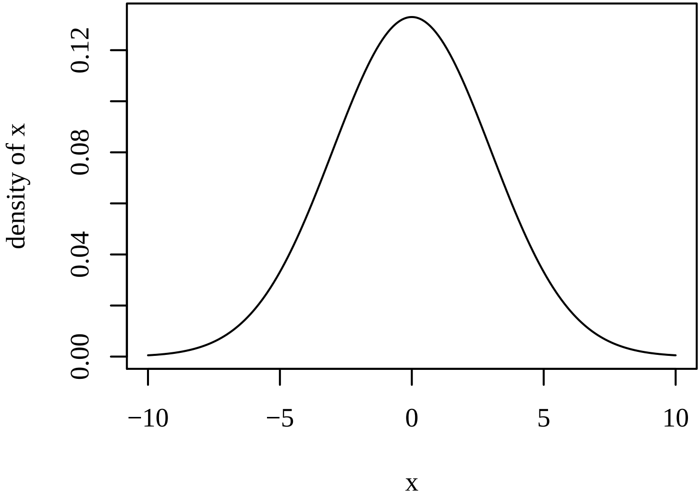
In words, the variance of the mean of x_i is the variance of x_i divided by N, the sample size upon which the mean is calculated.
That must mean the standard deviation of the means is

$$Standard\ Deviation(\bar{x}) = \frac{Standard\ Deviation(x_i)}{\sqrt{N}}$$

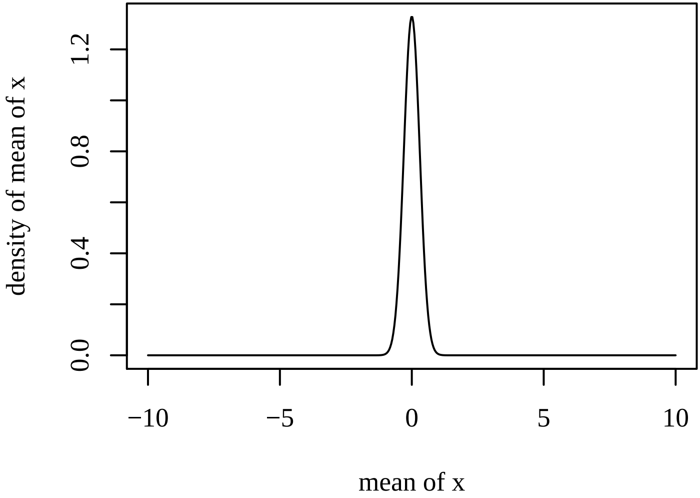
The Distribution of the Mean is "Spike-ish"

Please observe the illustration of the effect of sample size on the variance of $\bar{x}$.

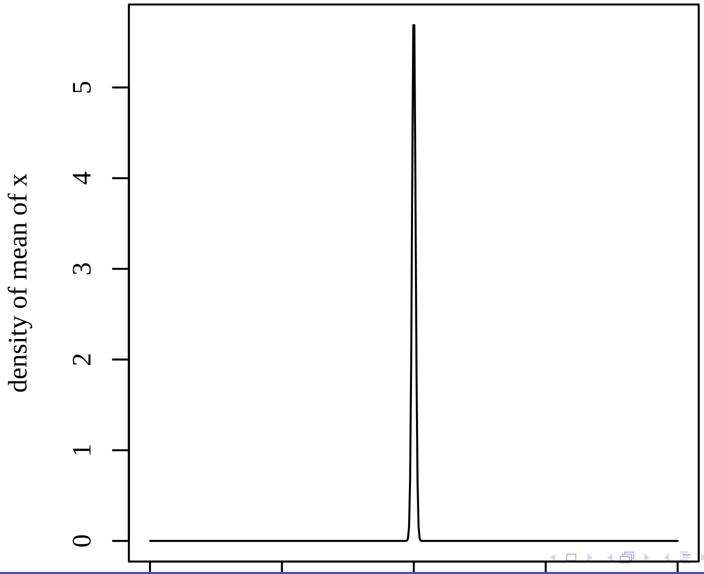
Distribution of $x \sim Normal(0, 3^2)$



Distribution of Mean, Sample=100 ($Normal(0, 3^2/100)$)



Distribution of Mean, Sample=2000 ($Normal(0, 3^2/2000)$)



Terms

- Asymptotic: related to very large (tending to infinite) sample sizes
- Consistency: an estimator (formula's result) 'tends to' the correct value as sample size tends to infinity

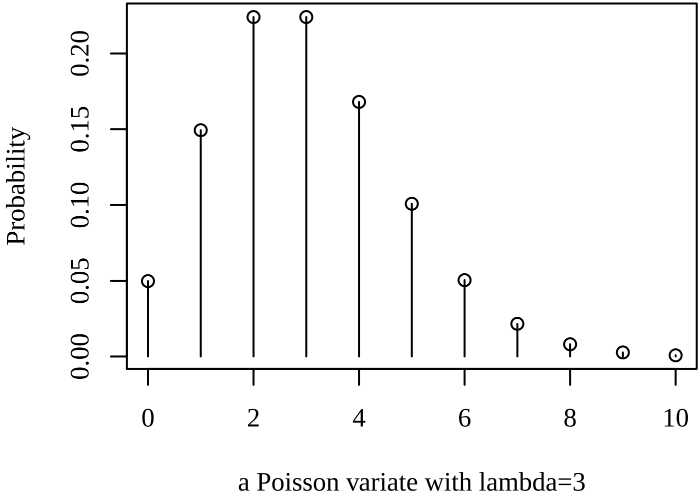
Law of Large Numbers

As the Sample Size Increases, $\bar{x}$ tends to the Expected Value (The True Mean)
This is the "law of large numbers".

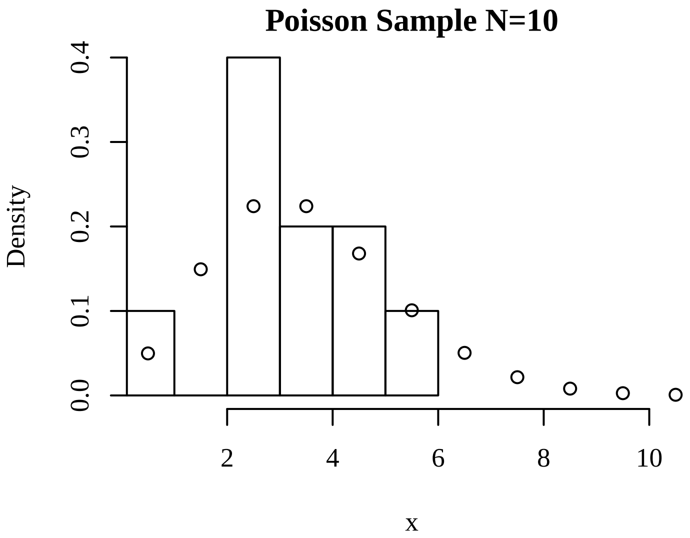
The Basic Idea of the CLT

- For ANY DISTRIBUTION (not just the normal) of x, the distribution of $\bar{x}$ approaches a normal distribution as the size of the sample upon which $\bar{x}$ is calculated tends to infinity.
- This one is difficult to prove algebraically, but it is quite easy to demonstrate with simulation

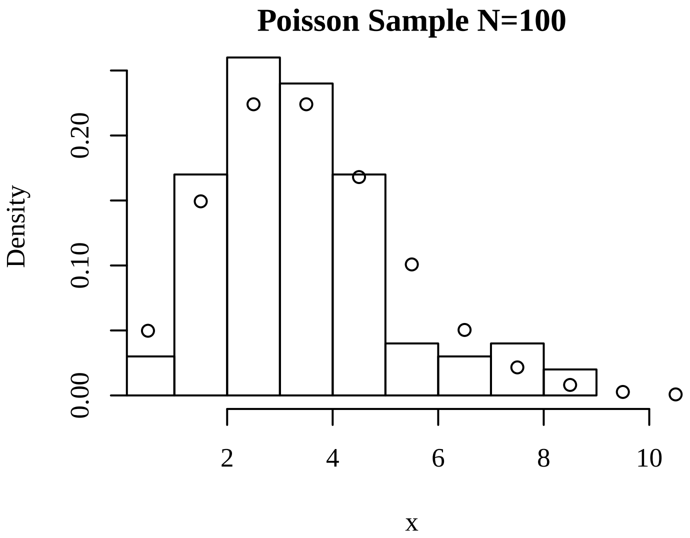
Take, for example, the Poisson Distribution



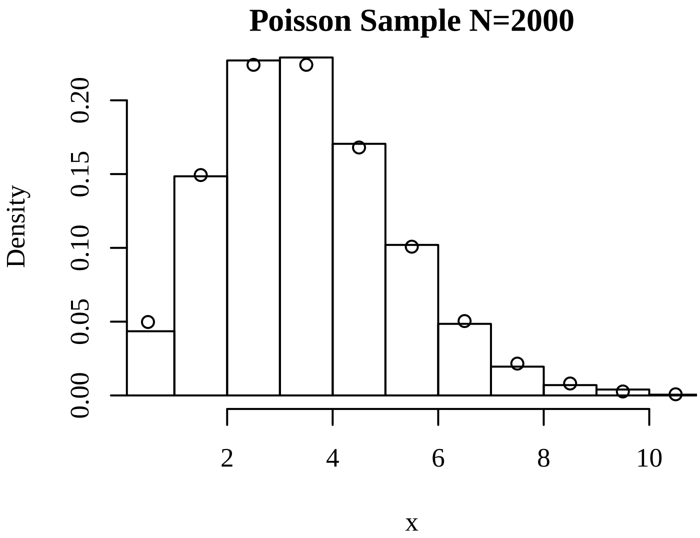
Poisson(3), SampleSize=10



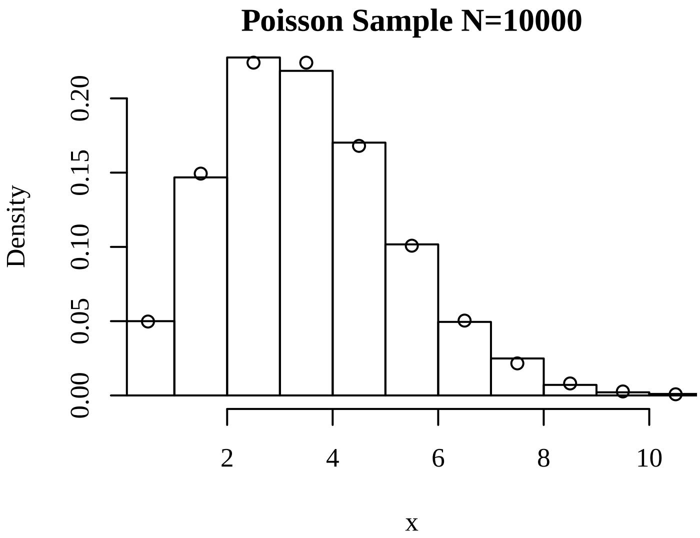
Poisson(3), SampleSize=100



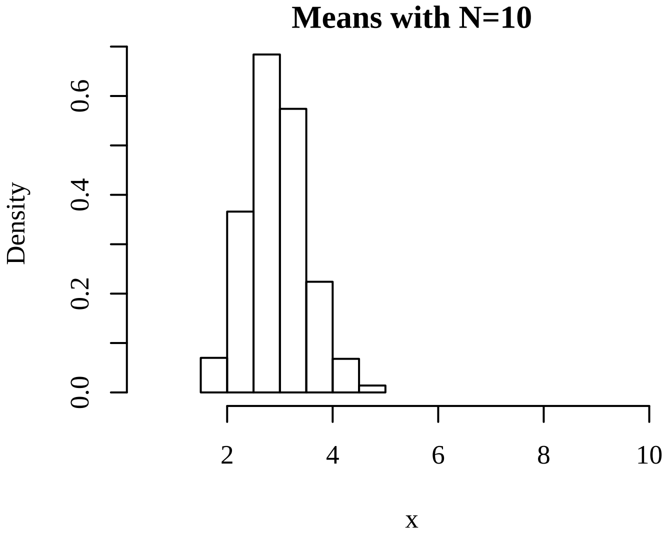
Poisson(3), SampleSize=2000



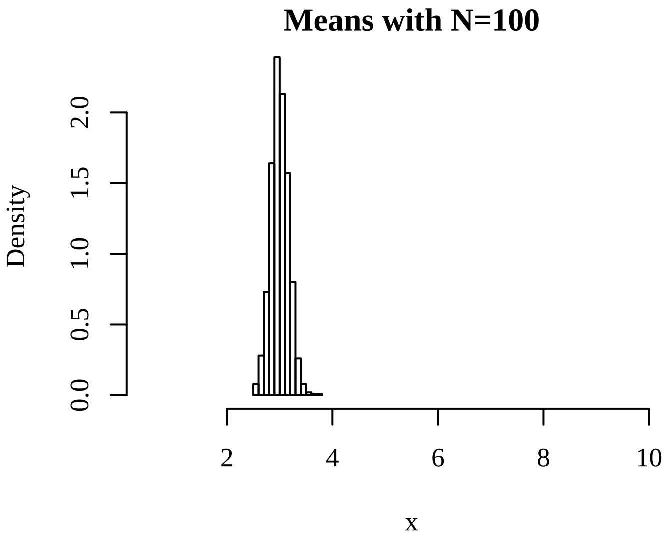
Poisson(3), SampleSize=10000



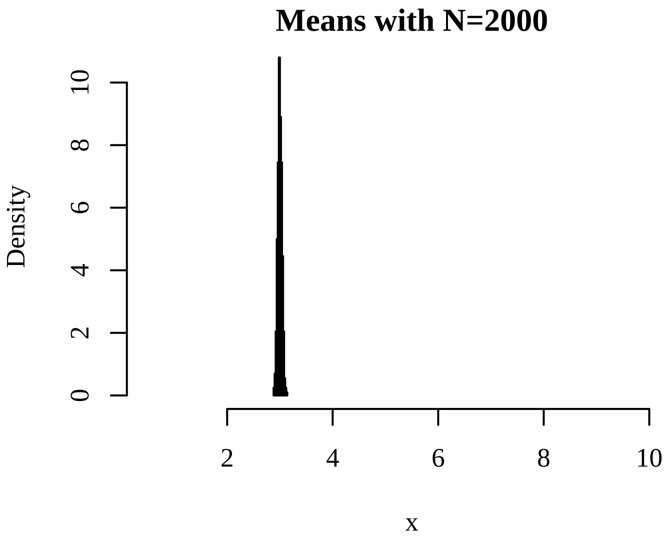
Means of 1000 Poisson Samples, Sample Size 10.



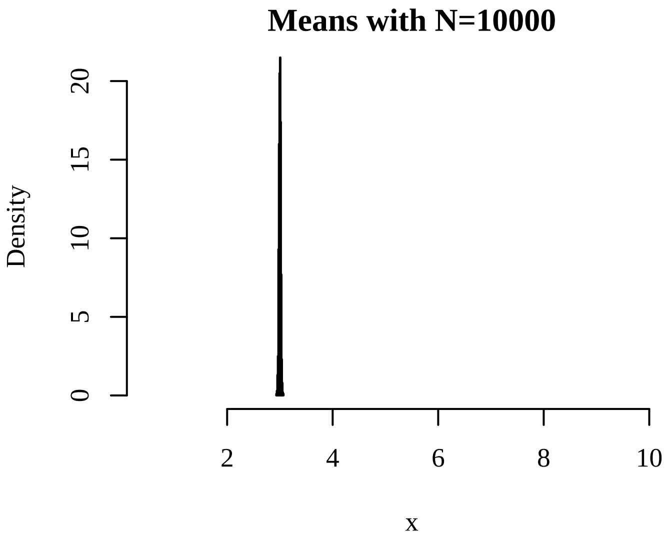
Means from 1000 Poissons, Sample Size=100



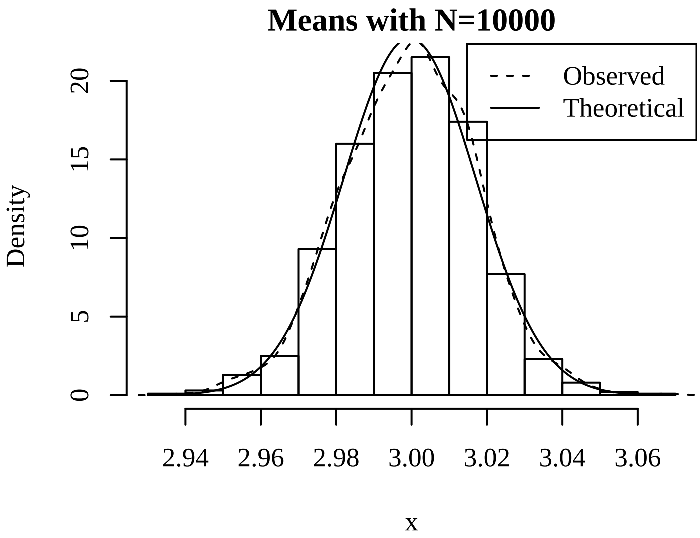
Means from 1000 Poisson samples, Sample Size=2000



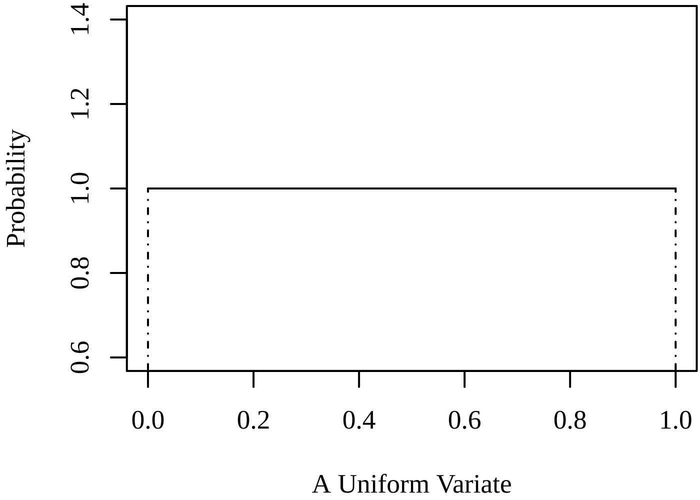
Means from 1000 Poisson samples, Sample Size=10000



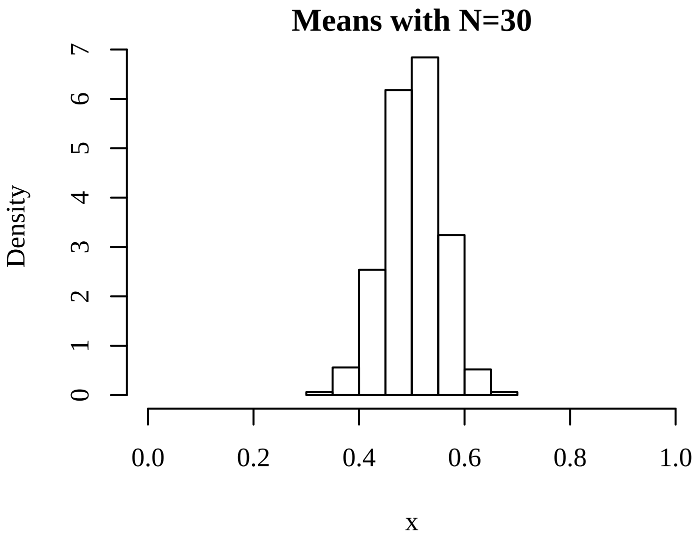
Same thing, bigger picture (N=10000)



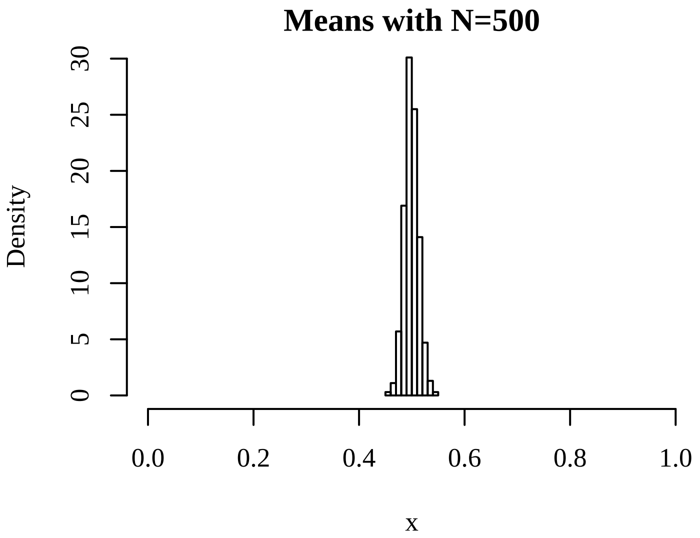
Consider the Uniform Distribution



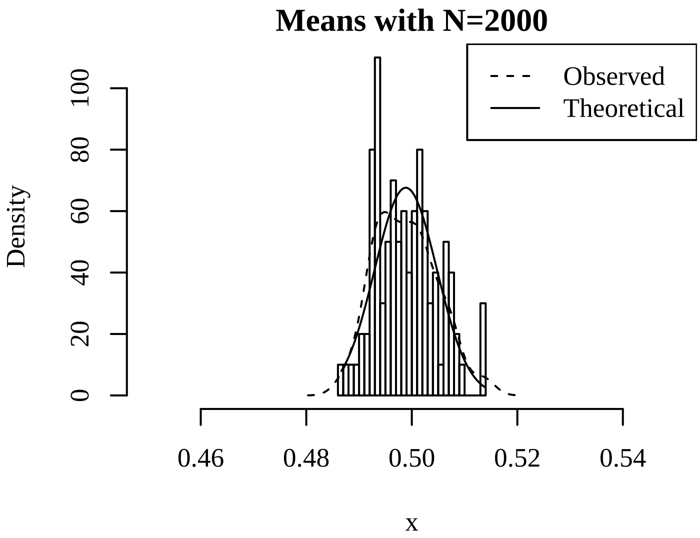
Means from 1000 Uniform samples, Sample Size=30



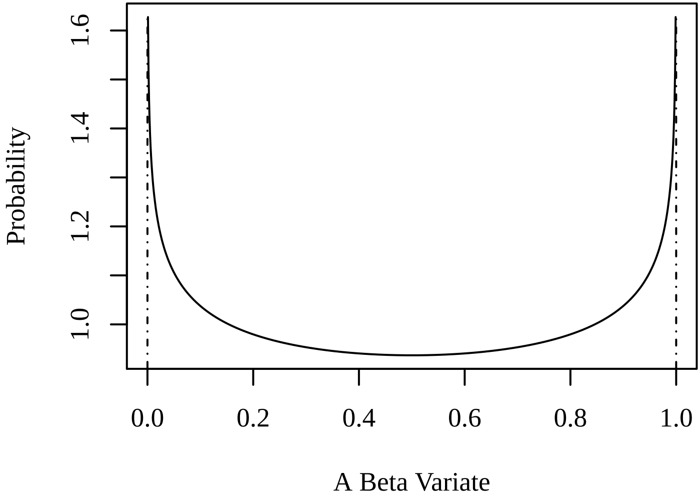
Means from 1000 Uniform samples, Sample Size=500



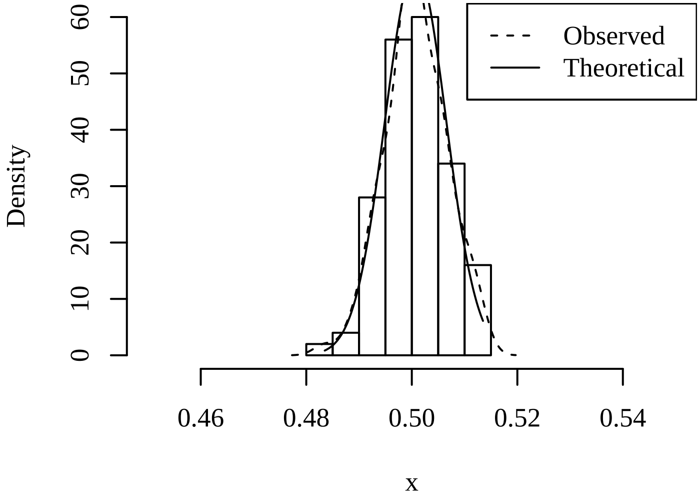
Means from 1000 Uniform samples, Sample Size=2000



OK, Challenge Me With Your Beta(0.9,0.9)



Means from 1000 Beta Samples, Sample Size=2000



My Mantra

From whatever distribution you pick, the Central Limit Theorem (CLT) says the "Sampling Distribution of the Mean is Normal".