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

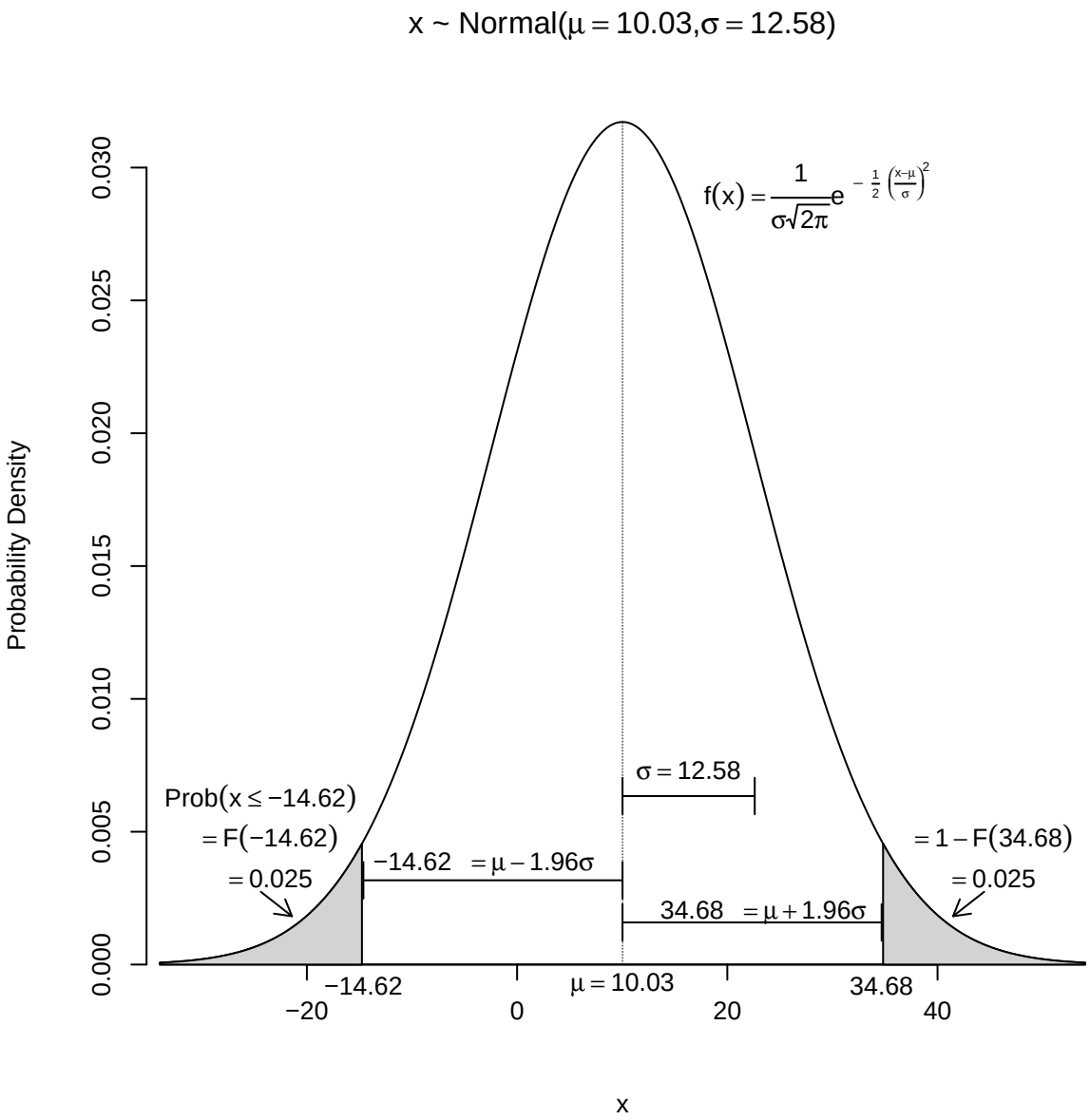
September 14, 2020

Outline

We think of one survey, one estimate

- Nature has a "data generating mechanism:"
- Nature's probability density function is never fully revealed to us, we only get samples.
- Samples fluctuate: no two samples are the same.
- From that one sample, we try to want to answer a LOT of questions.
  - we calculate an **estimate**: a single number that represents something.
  - We estimate a distribution's Expected Value, Variance, or other parameters
  - Develop a model of the PDF of the estimator. Almost NEVER are we interested in estimating Natures PDF that generates the data. Almost Always, we want to know the PDF of the estimator.

Normal Distribution PDF depends on μ and σ^2



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

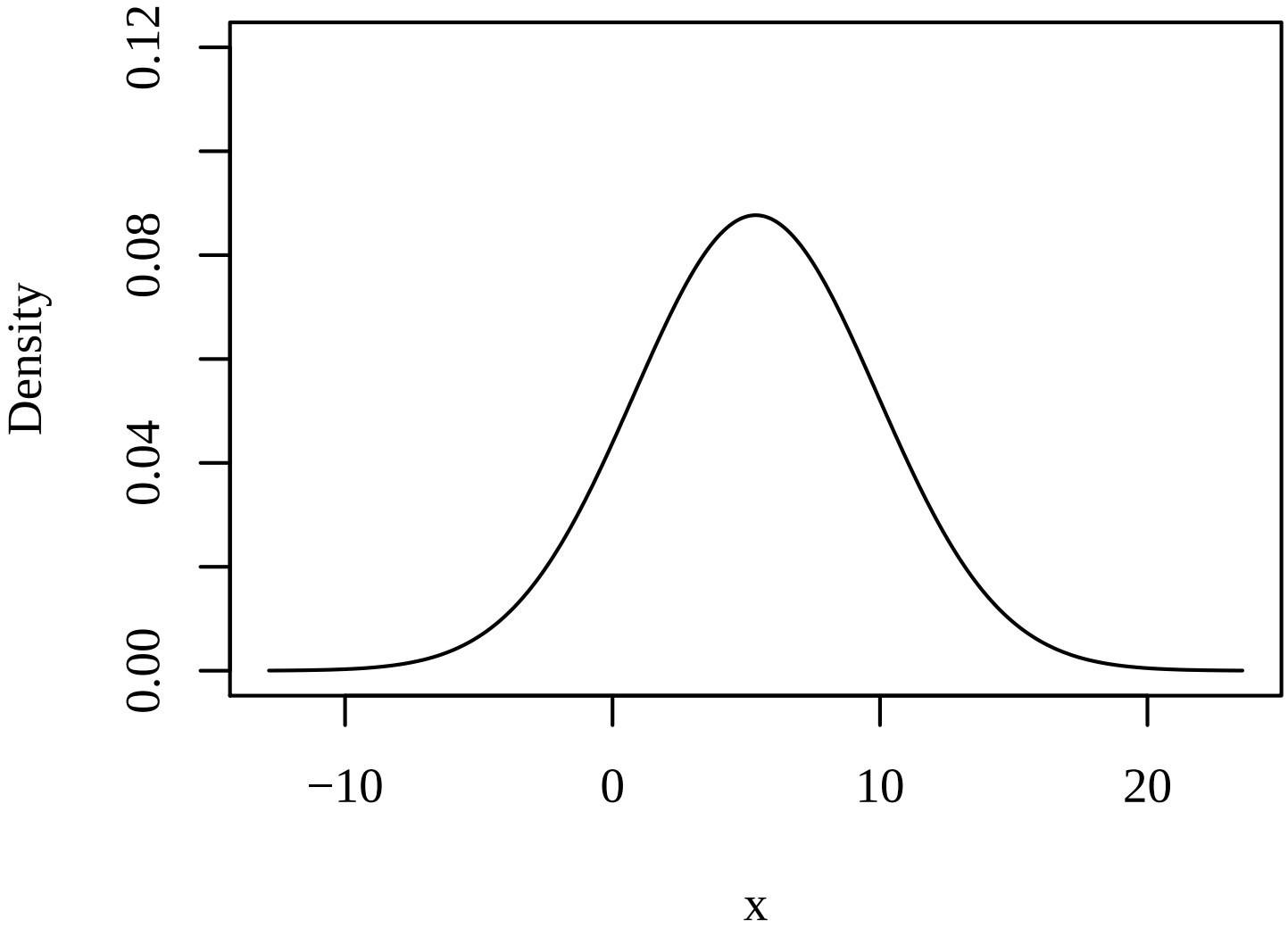
μ and σ^2 are Parameters

- Every distribution can have its "own letters" for parameters
- For generality, refer to them as θ
- I say: The estimates from sample data have hats:

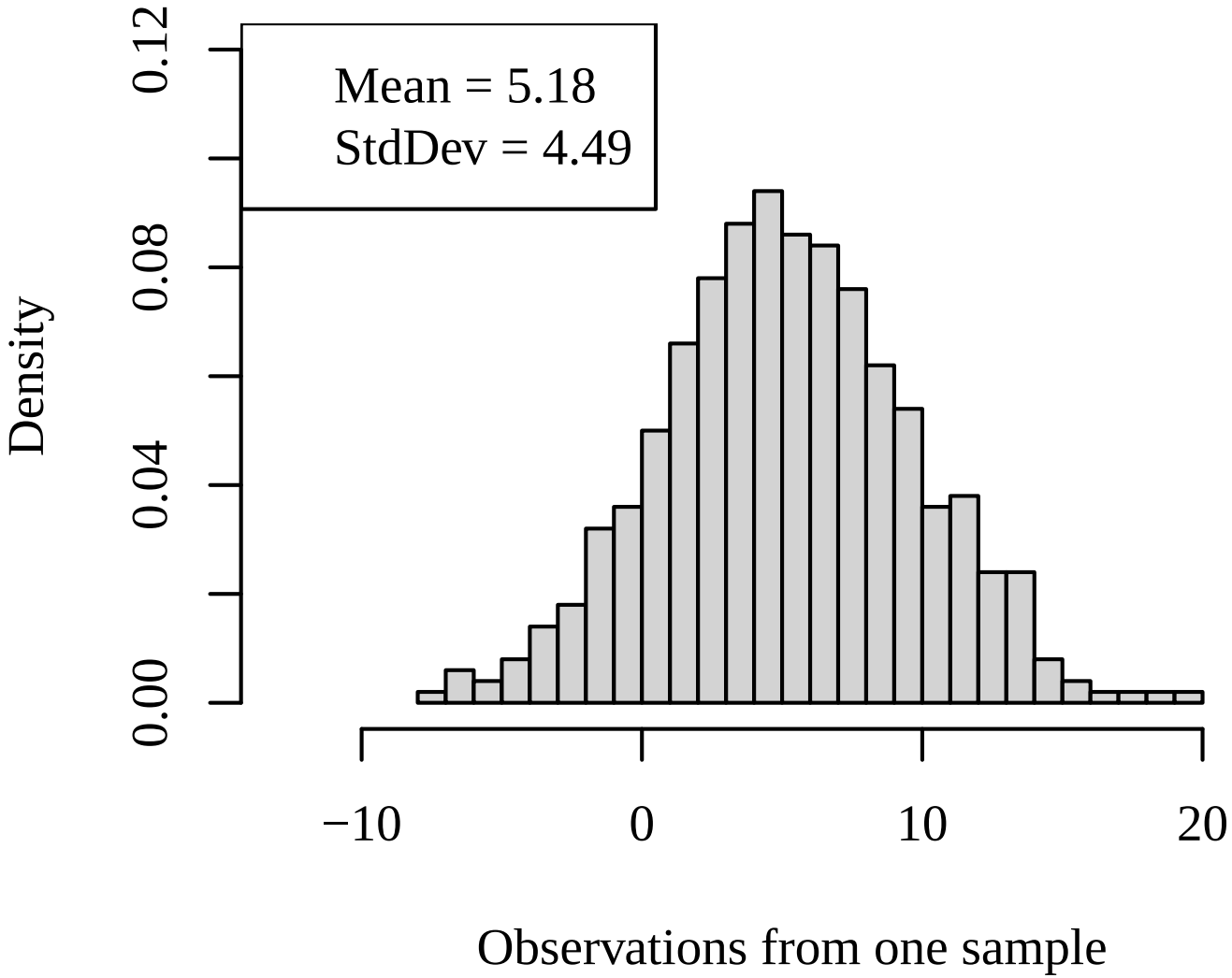
$$\hat{\mu} \qquad \widehat{\sigma^2}$$

- Some people prefer notation like: $\bar{x}$ and s^2.

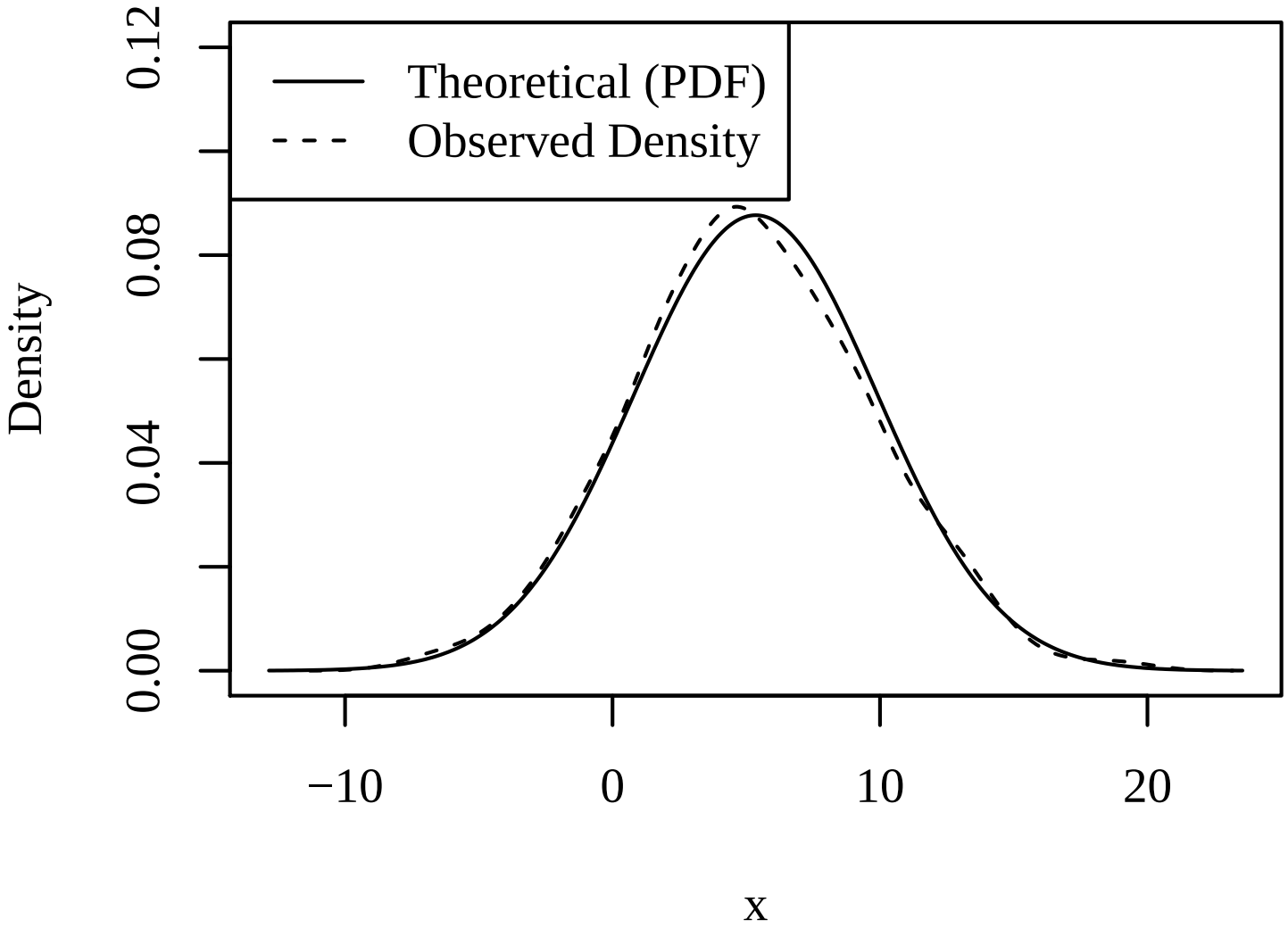
The Theoretical PDF Is This:



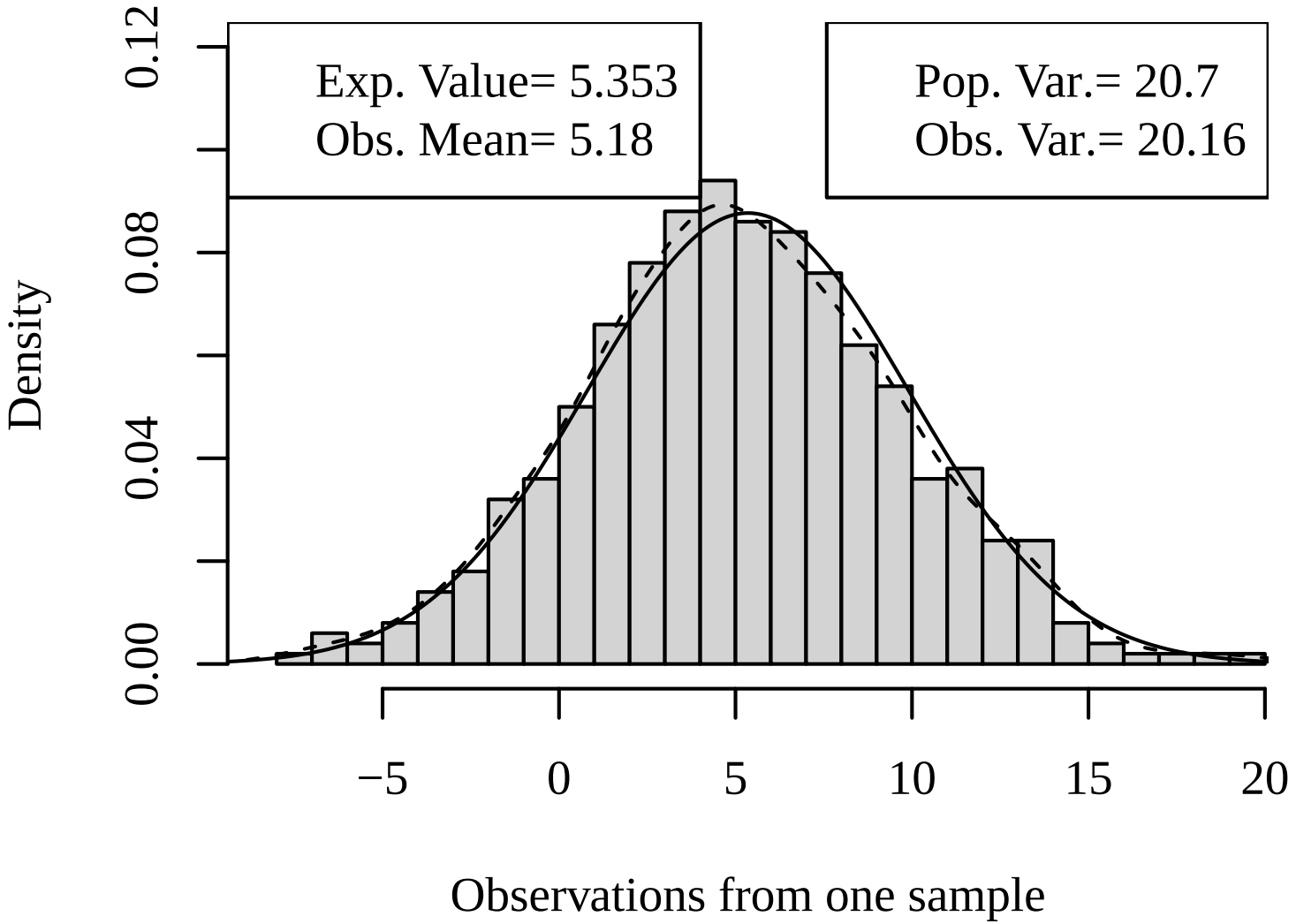
Draw one Normal Sample from $N(5.353, 4.55^2)$



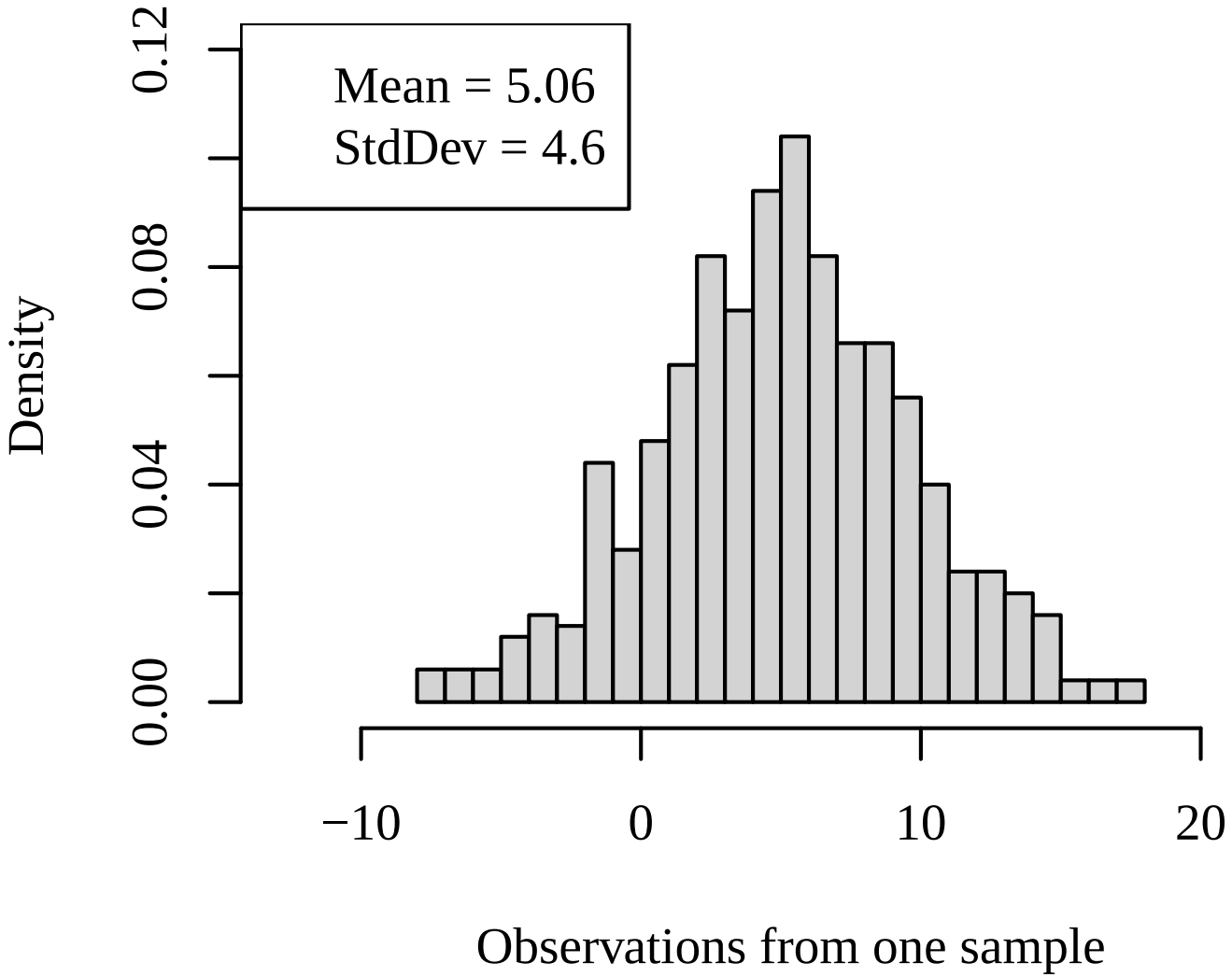
But the Observed Sample (Kernel) Density Differs



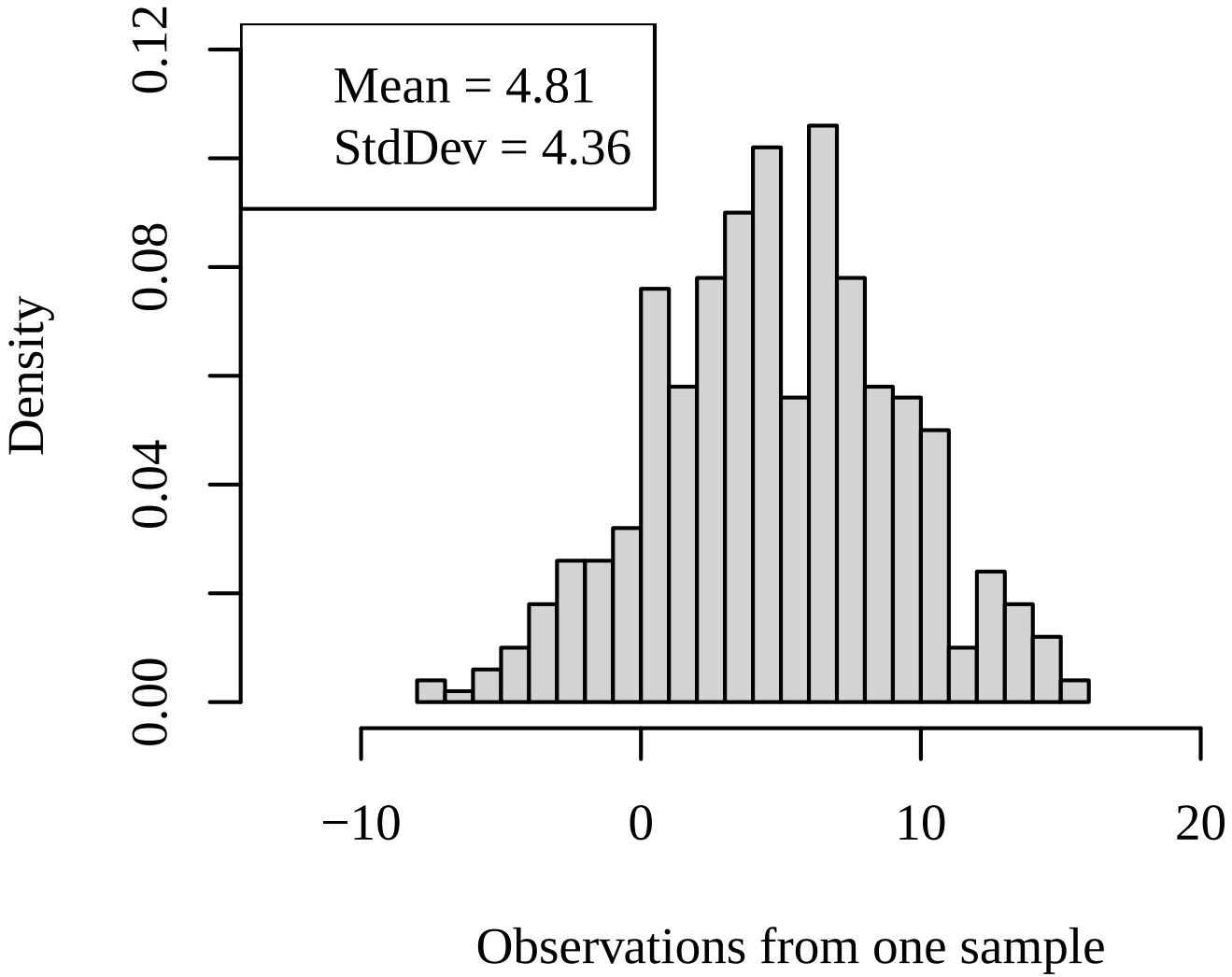
Histogram with PDF and KDE Superimposed



Draw another



Draw another



Important Term: Sampling Distribution

- Definition: Sampling Distribution is the PDF of an estimator. like $\bar{x}$
- The "true" sampling distribution is a theoretical "thing" (process?). Like the data generating process, it is never observed.
- Math can give us some "exact" characterizations about sampling distributions
- We can simulate repeated-sampling to visualize sampling distributions.

Visualize a Sampling Distribution Via Simulation

- Suppose you could repeatedly draw samples
- Calculate an estimate from each sample
  - Perhaps $\hat{\mu} \equiv \bar{x}$ (the mean) for each sample
  - More generally, any $\hat{\theta}$
  - Create a histogram of those observed estimates
  - We want to know
    - Are the estimates close to the "true" value?
    - Are the estimates symmetrically distributed?
    - Are there any abstract patters worth finding in these distributions of estimates?

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

This is true for the Normal distribution, AND ALL OTHER DISTRIBUTIONS we will work with!

- If the expected value is of x is μ, the expected value of the mean of a sample is also μ.

$$If\ \ E[x] = \mu, then E[\bar{x}] = \mu$$

- If the variance of x is σ^2, the Variance of the sampling distribution of the mean is $\frac{1}{N} Var[x]$

$$If Var[x] = \sigma^2, then Var[\bar{x}] = \frac{Var[x]}{N}$$

- 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 clustered much more "tightly' than the distribution of x_i itself.

That's impossibly easy to see

- Algebraically.
- By simulation.

I've moved the algebraic proof to the end of these notes, but have just one comment about it on the next 2 slides.

Spotlight on one Tricky Bit: One Observation x_i Has Expected Value and Variance!

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

$$x_i \tag{1}$$

- In past, we discussed $x = x_1, x_2, \ldots, x_N$ as a collection of observations. Easy to think of the "mean" or "variance" of sample and expected value $E[x]$ and variance $Var[x]$
- We said x is normally distributed, thinking of x as a variety of outcomes. In your mind, a "histogram" with PDF curve.
- We can't calculate a sample mean from a single observation, but that one observation still has an "expected value" because its drawn from a data generating process.

Spotlight on one Tricky Bit: One Observation x_i Has Expected Value and Variance!

- Now think of x_1, x_2 and so forth as separate variates from the same distribution.
- Appeal to Intuition. Each individual draw has the same expected value. So $E[x] = E[x_1] = E[x_2] = \ldots E[x_N]$
- Similarly, each draw has same variance.
- It should be obvious how we derive the claim that the expected value of an average is the expected value.

$$E[\bar{x}] = E[\frac{x_1 + x_2 + \ldots + x_N}{N}]$$

And the Variance of the Estimated Mean is Manageable as well

- Again, we are supposing we know $Var(x) = \sigma_x^2$.
- If we calculate the average of a sample,

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

- Apply the variance operator to both sides

$$Var(\bar{x}) = Var(\frac{x_1 + x_2 + \ldots + x_N}{N})$$

its pretty easy to get the result we want.

The Variance of a Sampling Distribution is REALLY Important I

- We don't care about means, in particular. We care about all kinds of parameter estimates, $\hat{\theta}$
- We want to have precise estimates
- The route to a small variance of the estimate is especially clear in the case of the estimated mean:

$$Var(\bar{x}) = \frac{1}{N} Var(x).$$

  But not all estimators have such a clear, simple formula that makes it easy to see how to reduce the estimator's variance
- For just a few kinds of parameter estimates, we can actually know $Var(\hat{\theta})$

The Variance of a Sampling Distribution is REALLY Important II

- Most often, we have to estimate $Var(\hat{\theta})$. I'm entertained by the two hat notation:

$$\widehat{Var(\hat{\theta})}$$

an estimate of our uncertainty about of an estimate.

- In Regression analysis, we won't write $\widehat{Var(\hat{\theta})}$ very often. We instead talk about its square root, which by custom is called

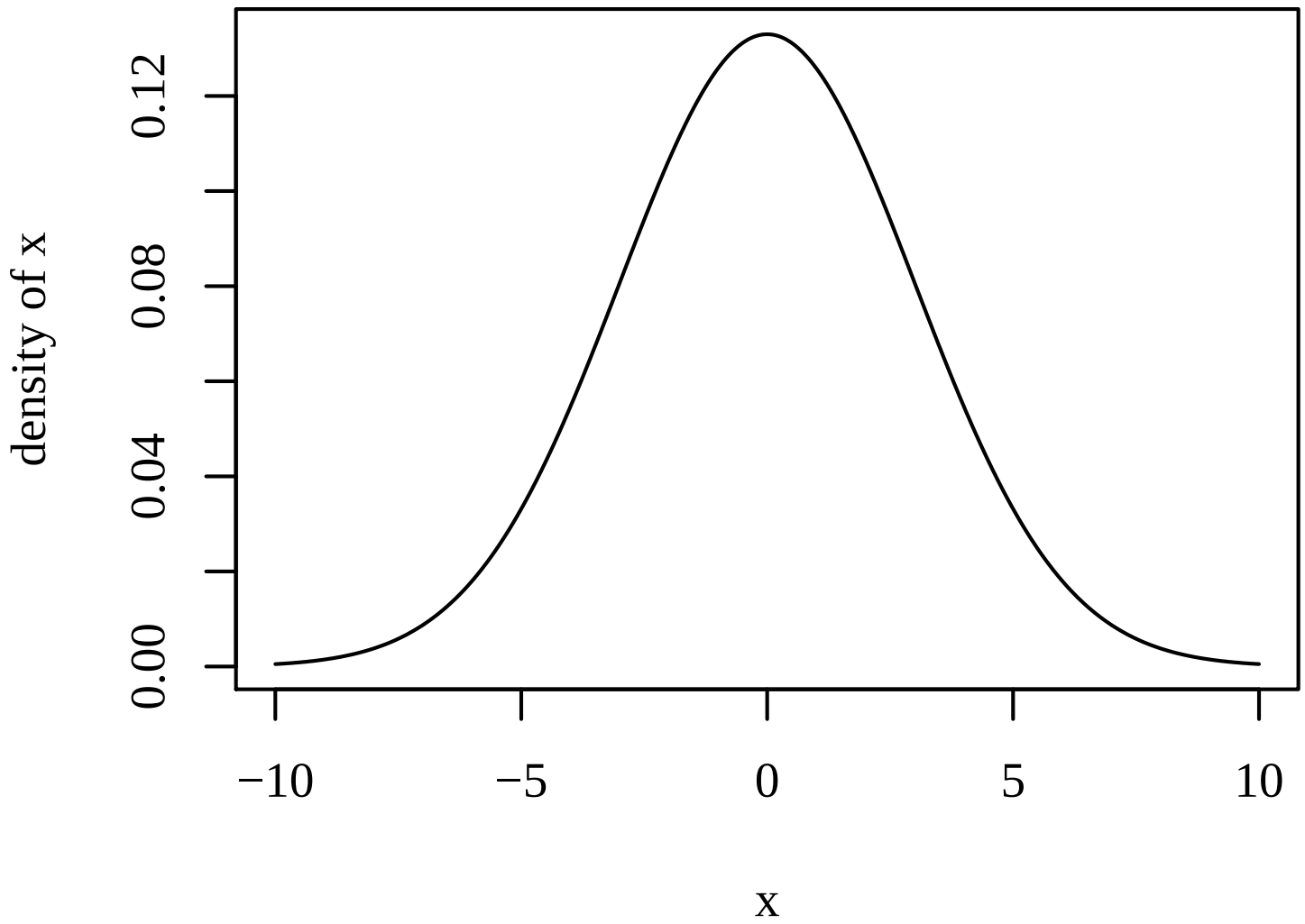
$$s.e.(\hat{theta}) : the standard\ error\ of\ \hat{\theta}$$

i.e., standard error is an estimate of the standard deviation of a sampling distribution

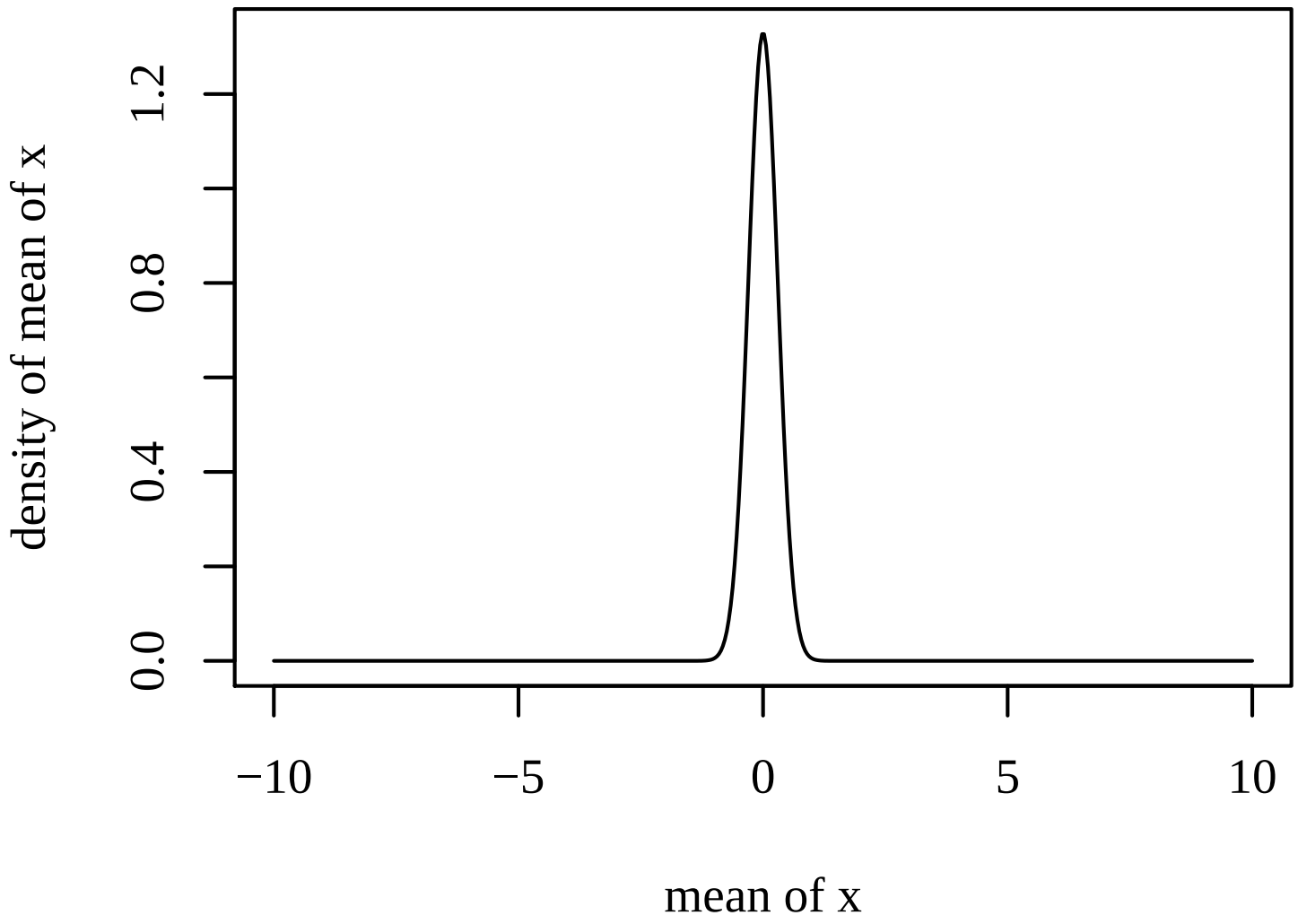
The Distribution of the Mean is "Spike-ish"

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

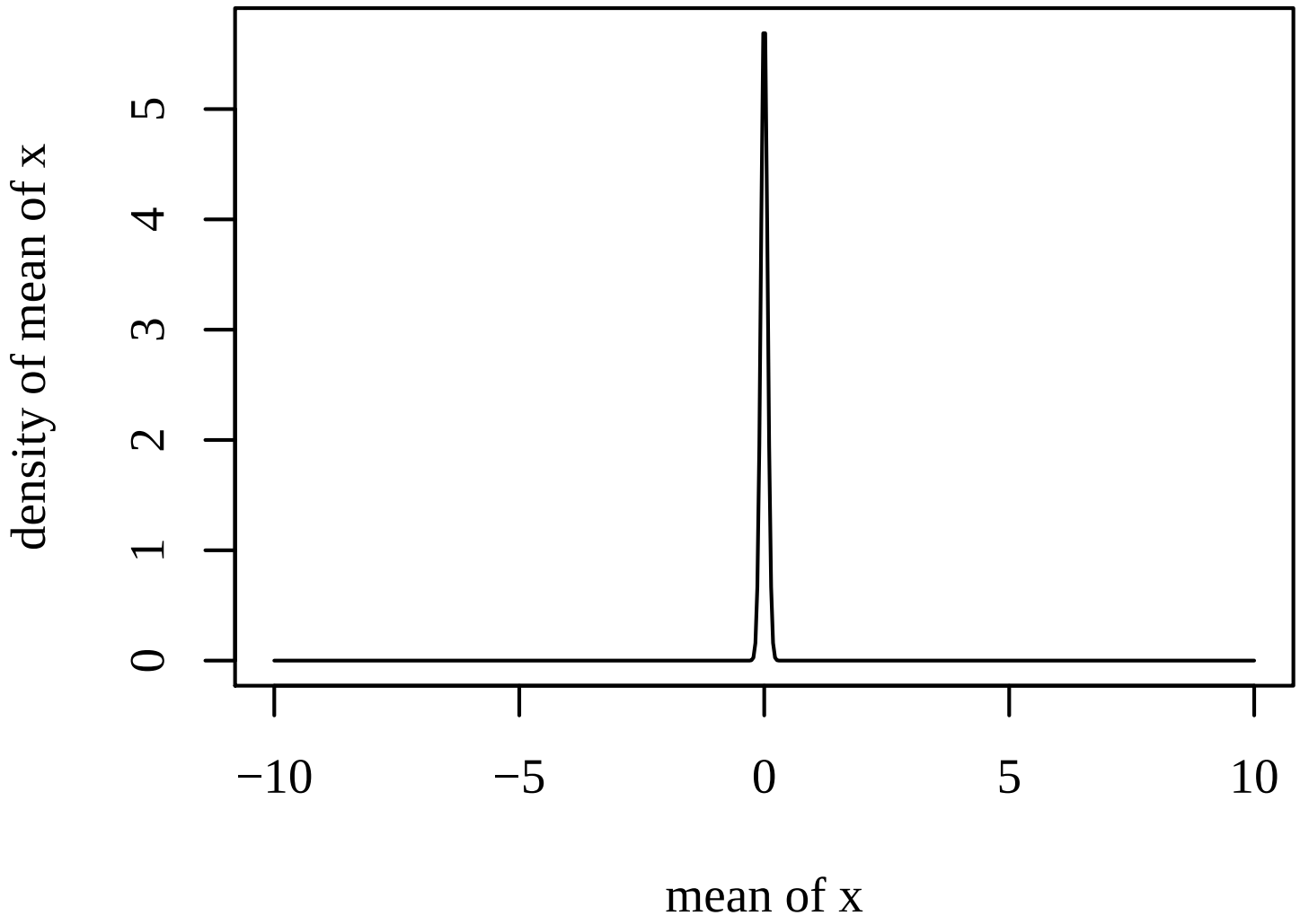
Distribution of $x \sim Normal(0, 3^2)$



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

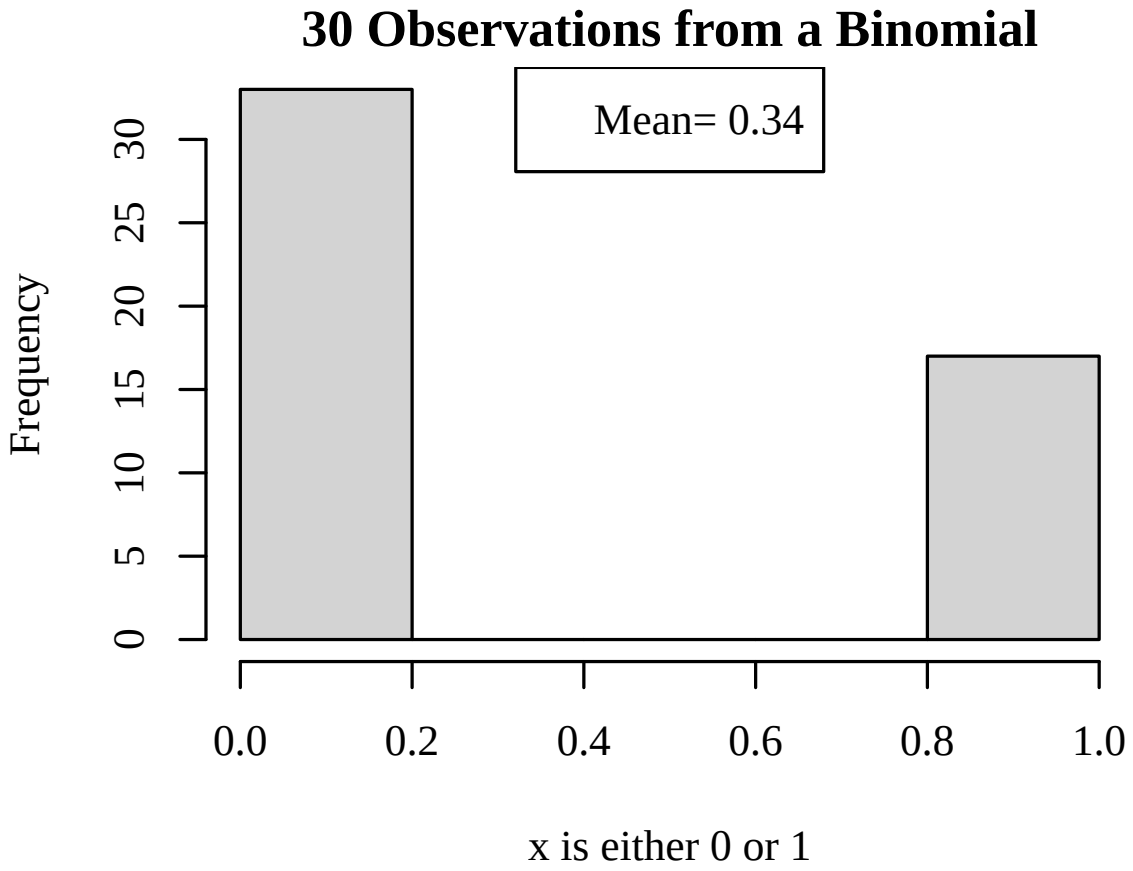
As the Sample Size Increases, $\bar{x}$ tends to the Expected Value (The True Mean)
This is the "law of large numbers".

The Basic Idea of the CLT

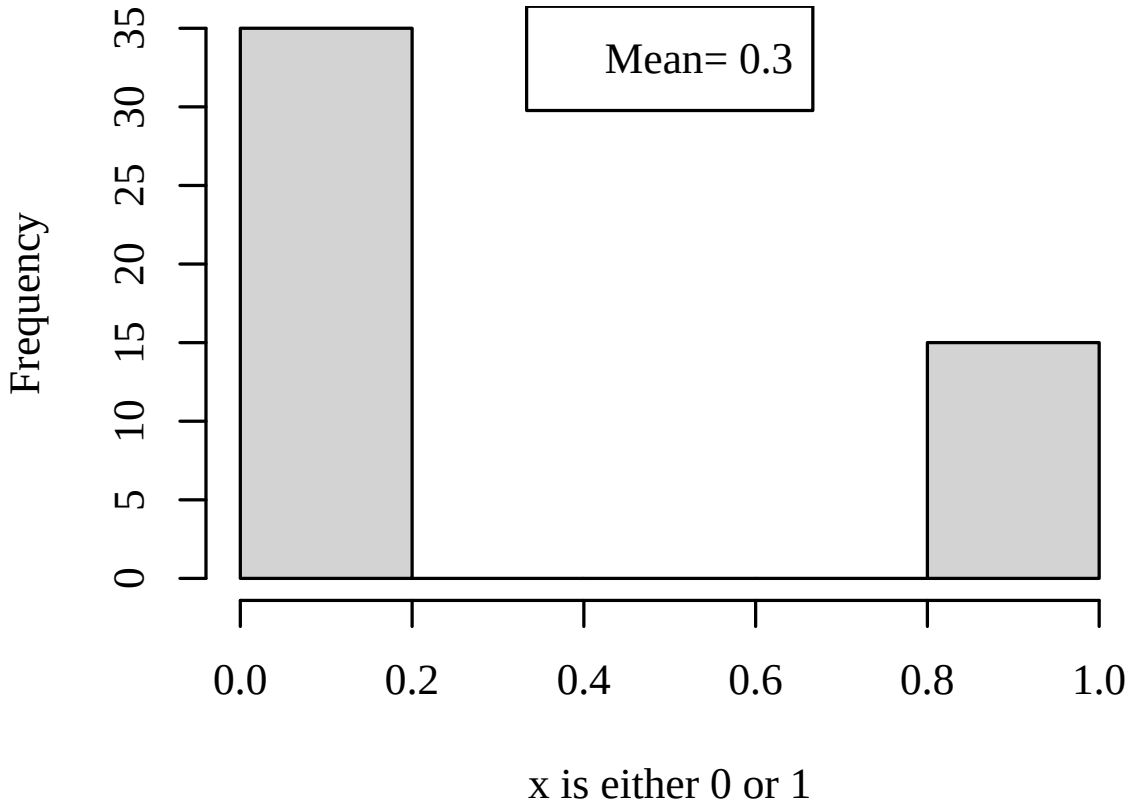
- For ANY DISTRIBUTION (not just the normal) of x, the distribution of $\bar{x}$ approaches a normal distribution as the size of the sample upon which $\bar{x}$ is calculated tends to infinity.
- This one is difficult to prove algebraically, but it is quite easy to demonstrate with simulation

The CLT with 0,1 data

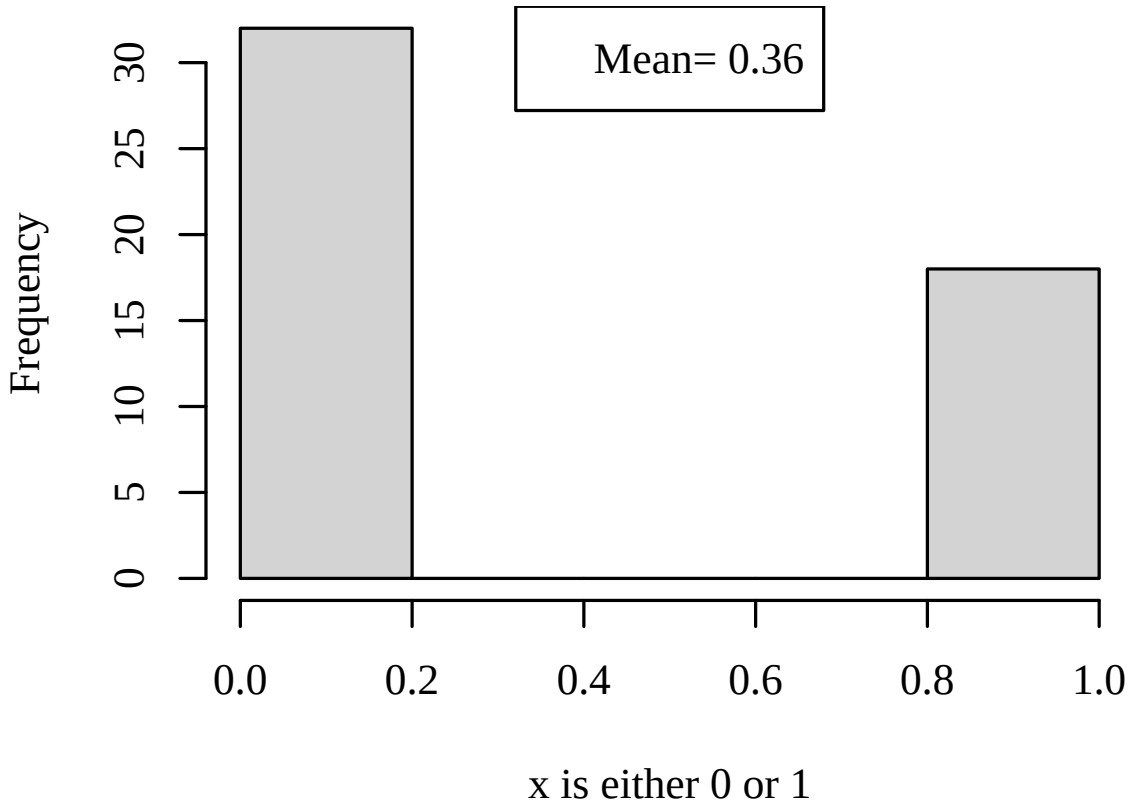
Draw 50 observations where the probability of success on each one is 0.30.



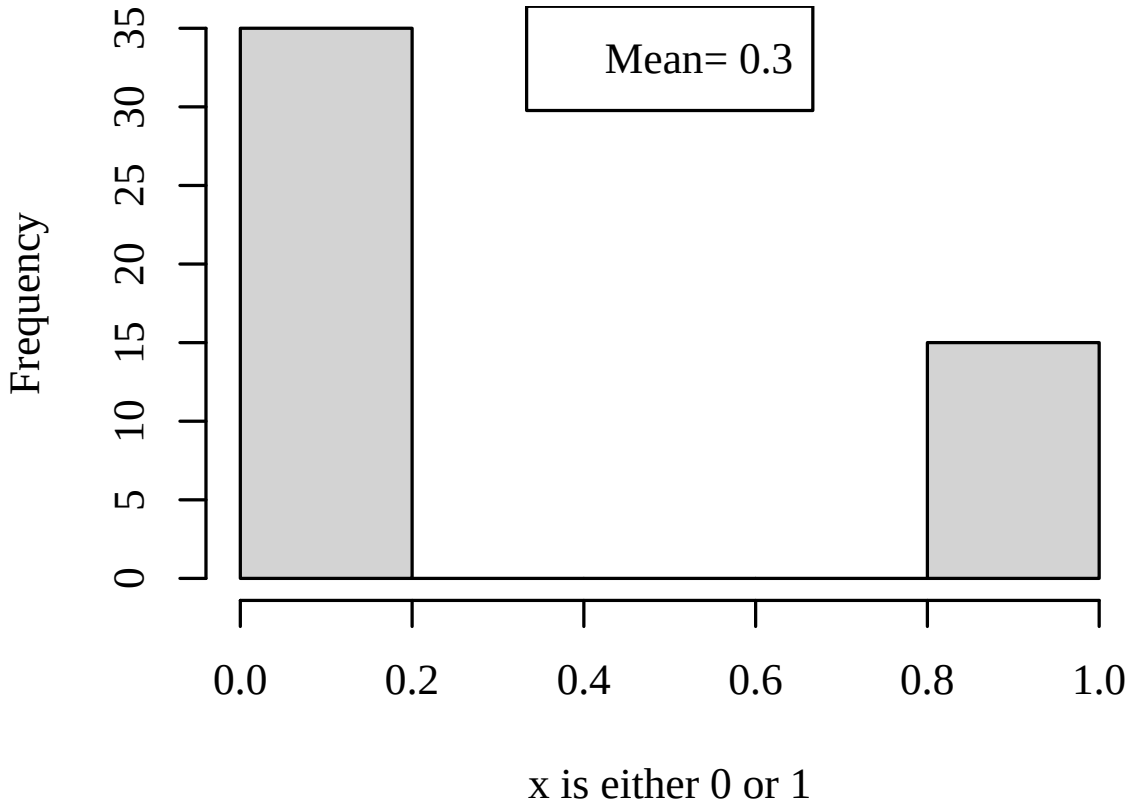
Draw Another Sample



Draw Another Sample



Draw Another Sample

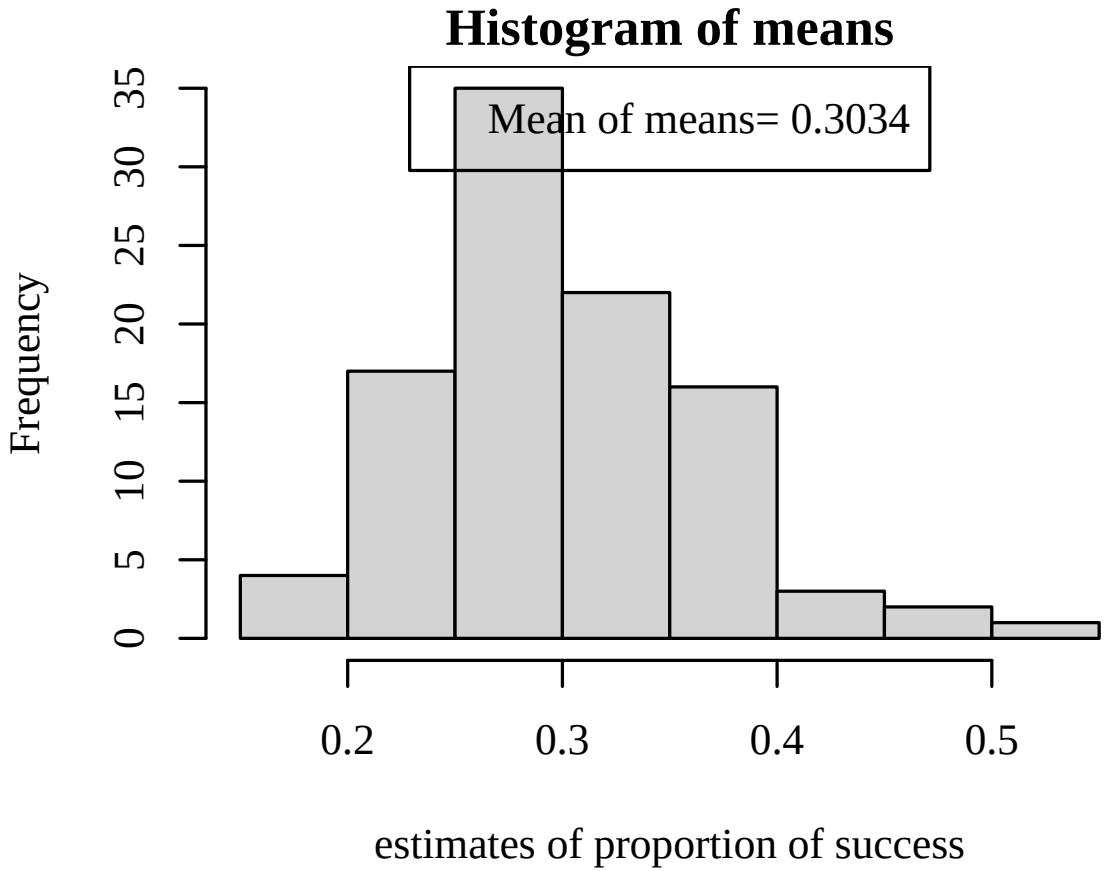


What about those means?

- Do that over and over again.
- what do you guess the distribution of the means would look like?
- I'll make a guess. It will be tightly clustered around "0.30" and it will be normally distributed

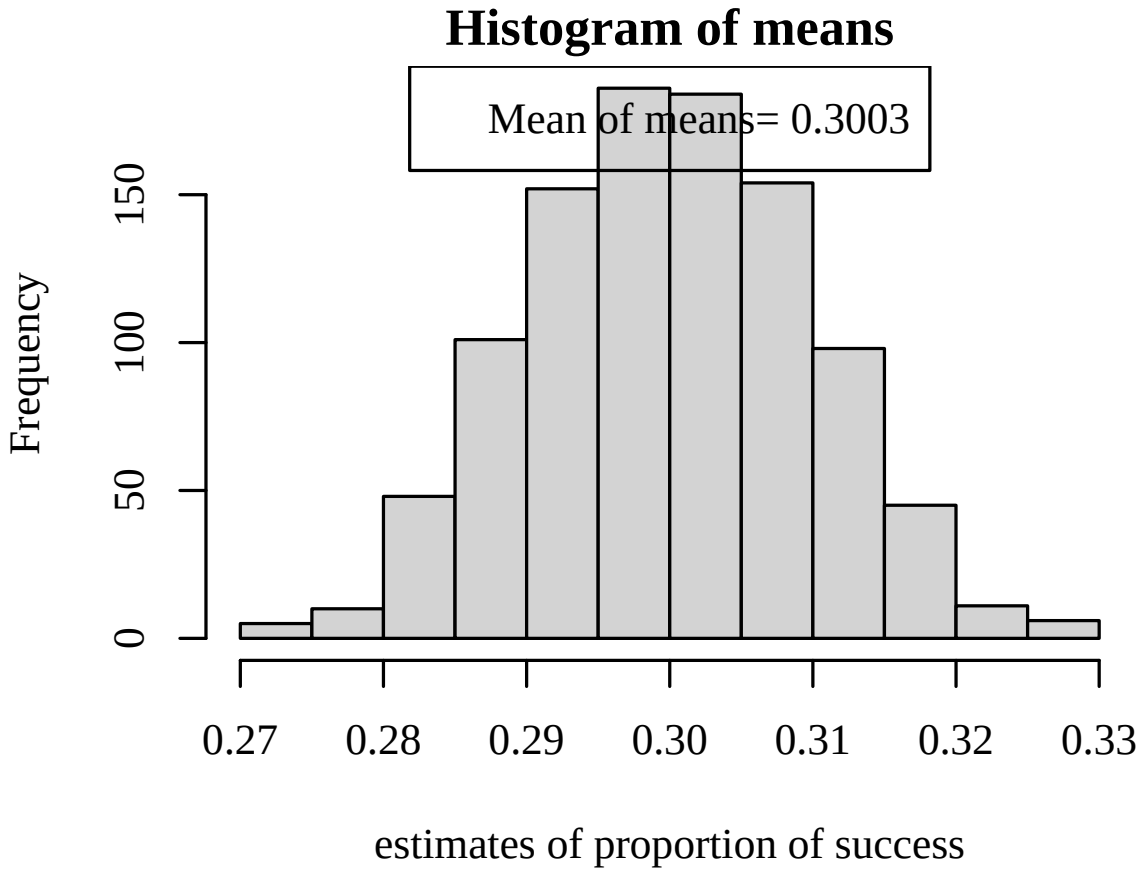
Lots of means from 0,1 data

100 samples, each including 50 random draws

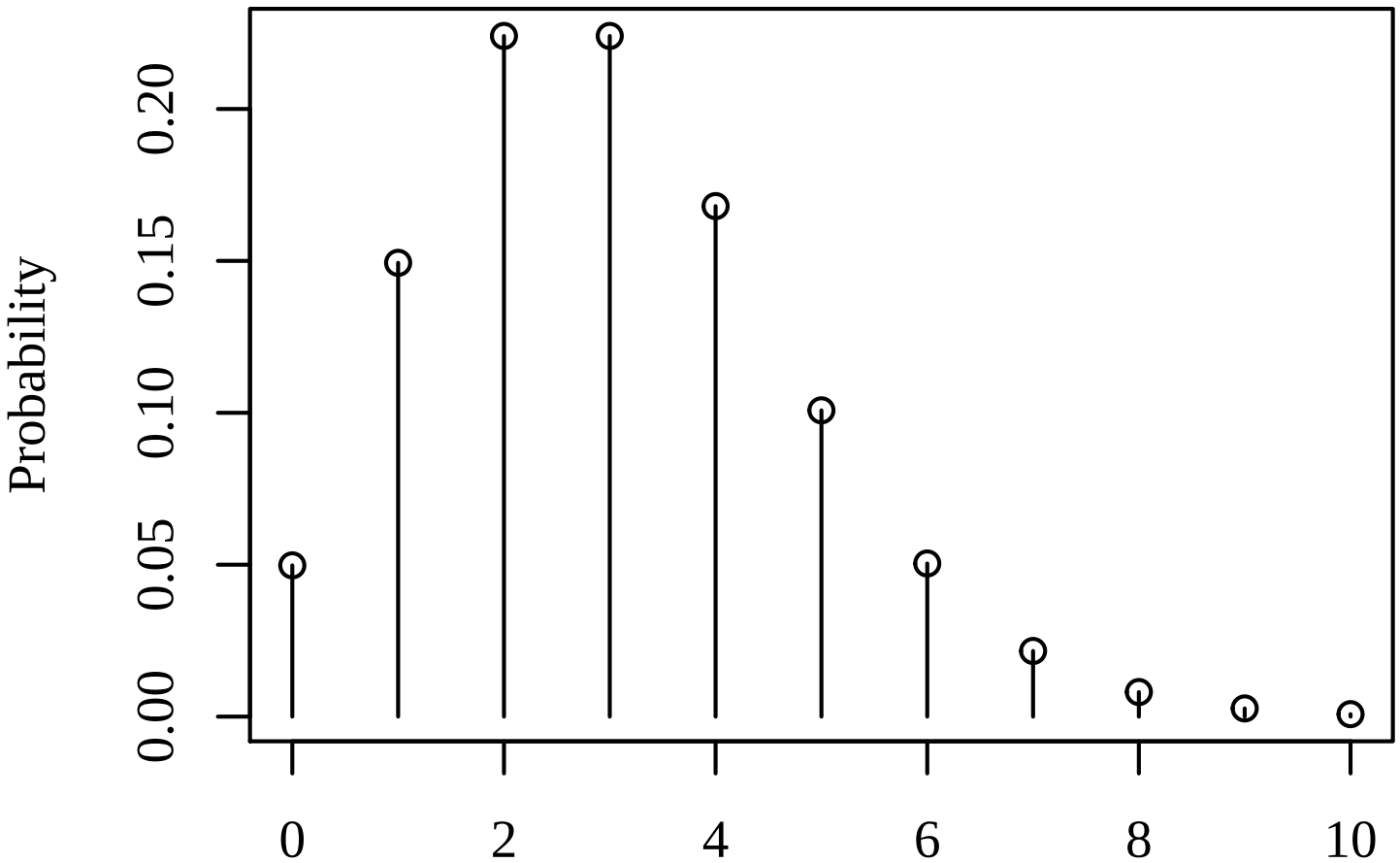


Lots of means from 0,1 data

1000 samples, each including 2000 random draws

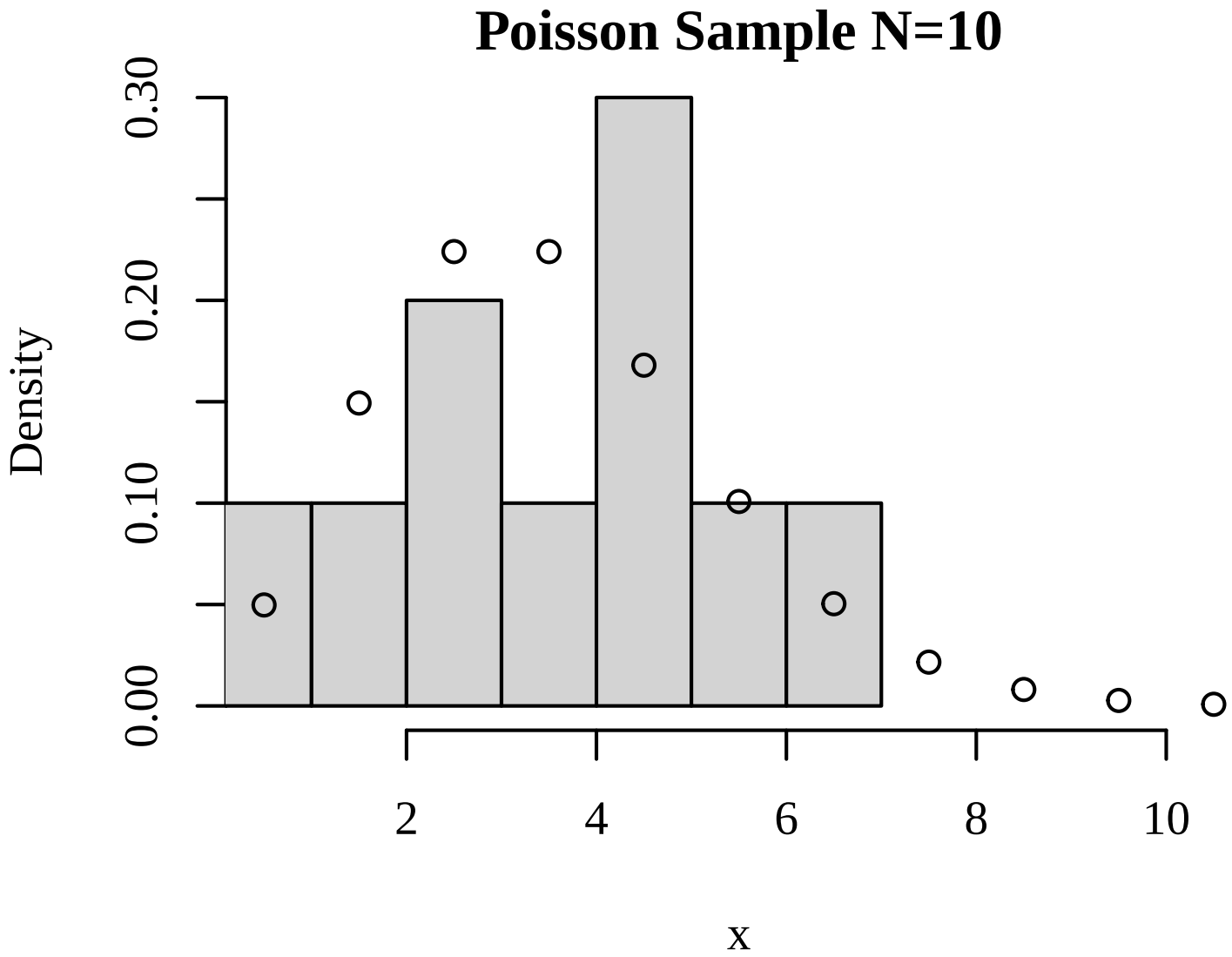


Take the Poisson Distribution for another example

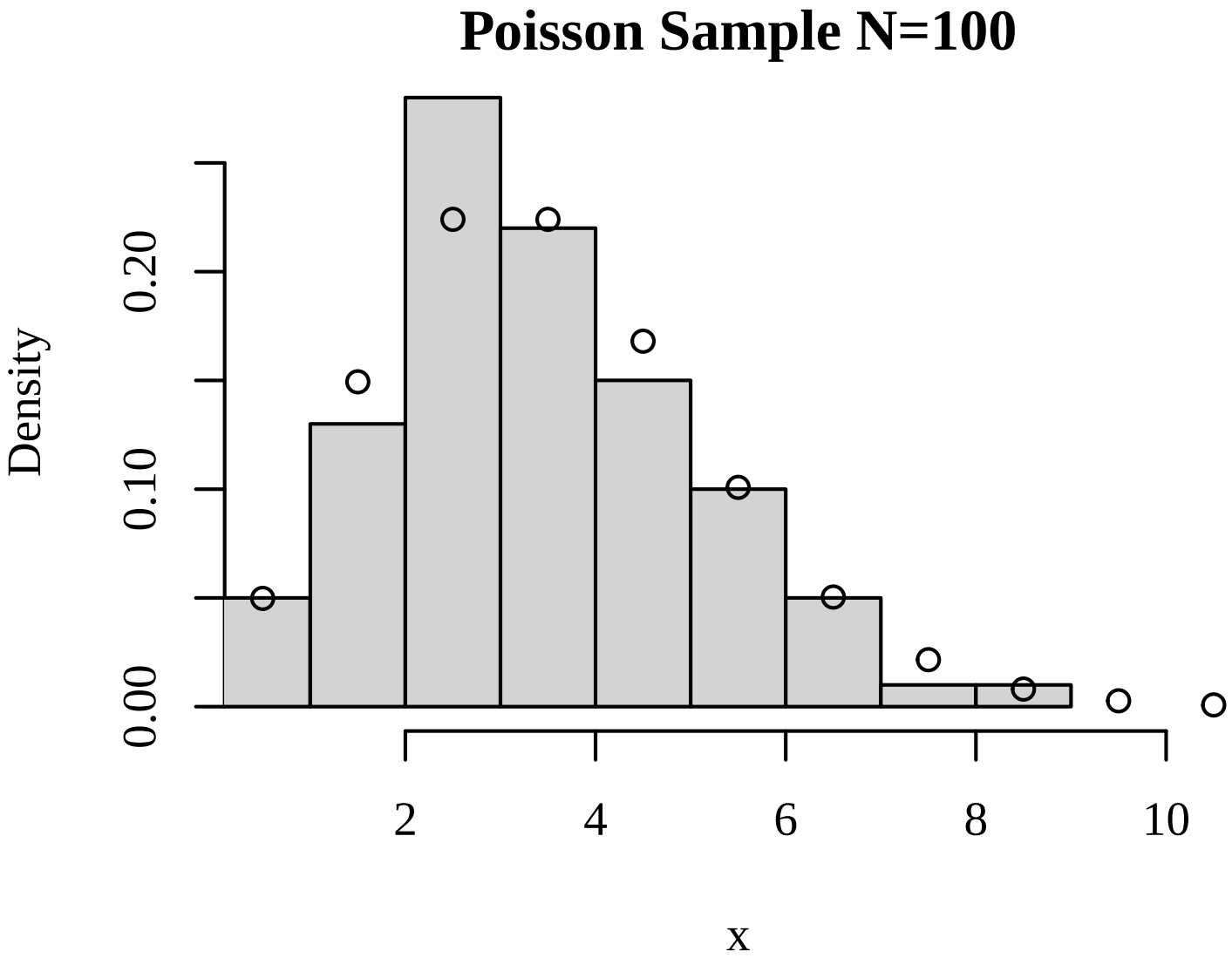


a Poisson variate with lambda=3

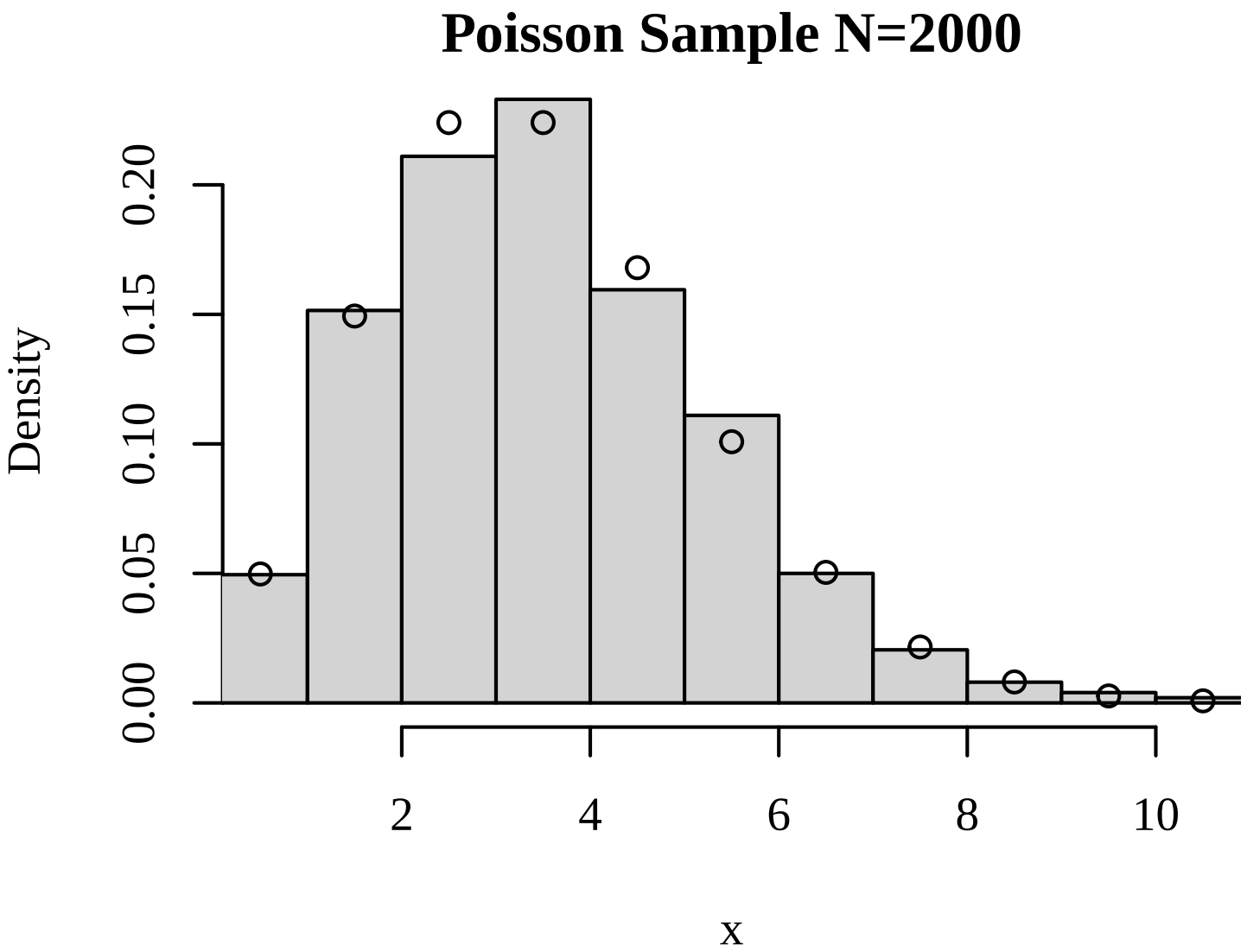
Poisson(3), SampleSize=10



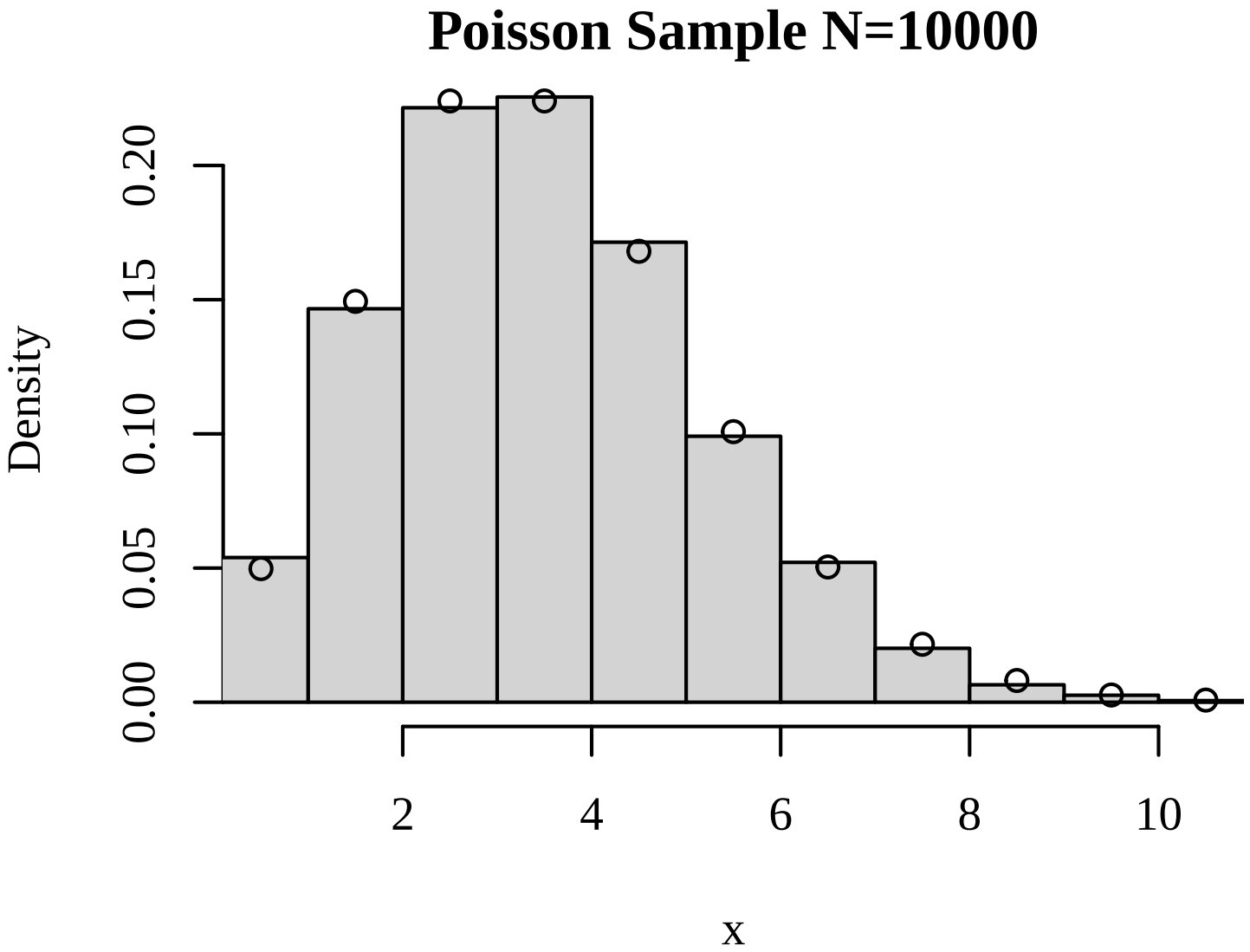
Poisson(3), SampleSize=100



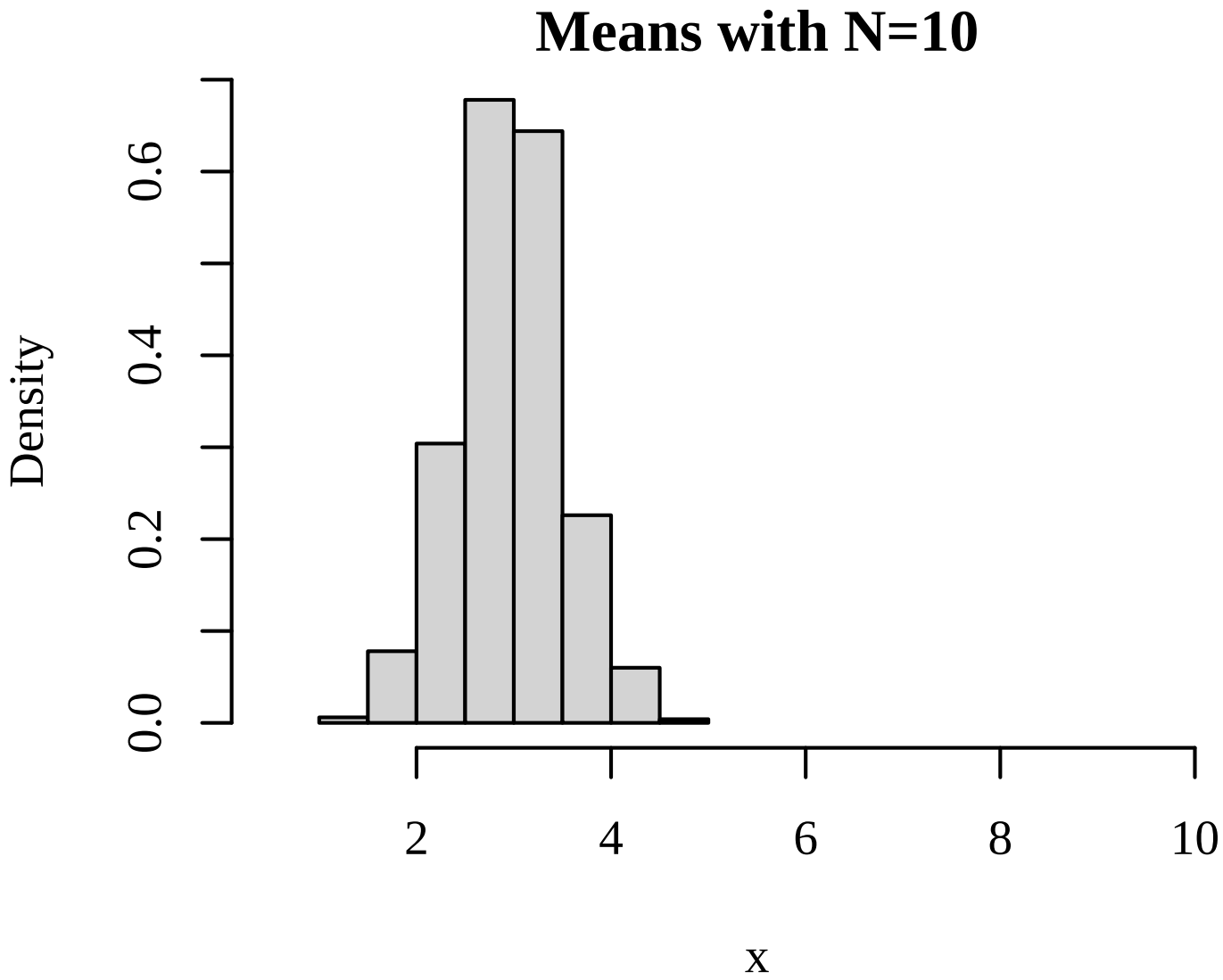
Poisson(3), SampleSize=2000



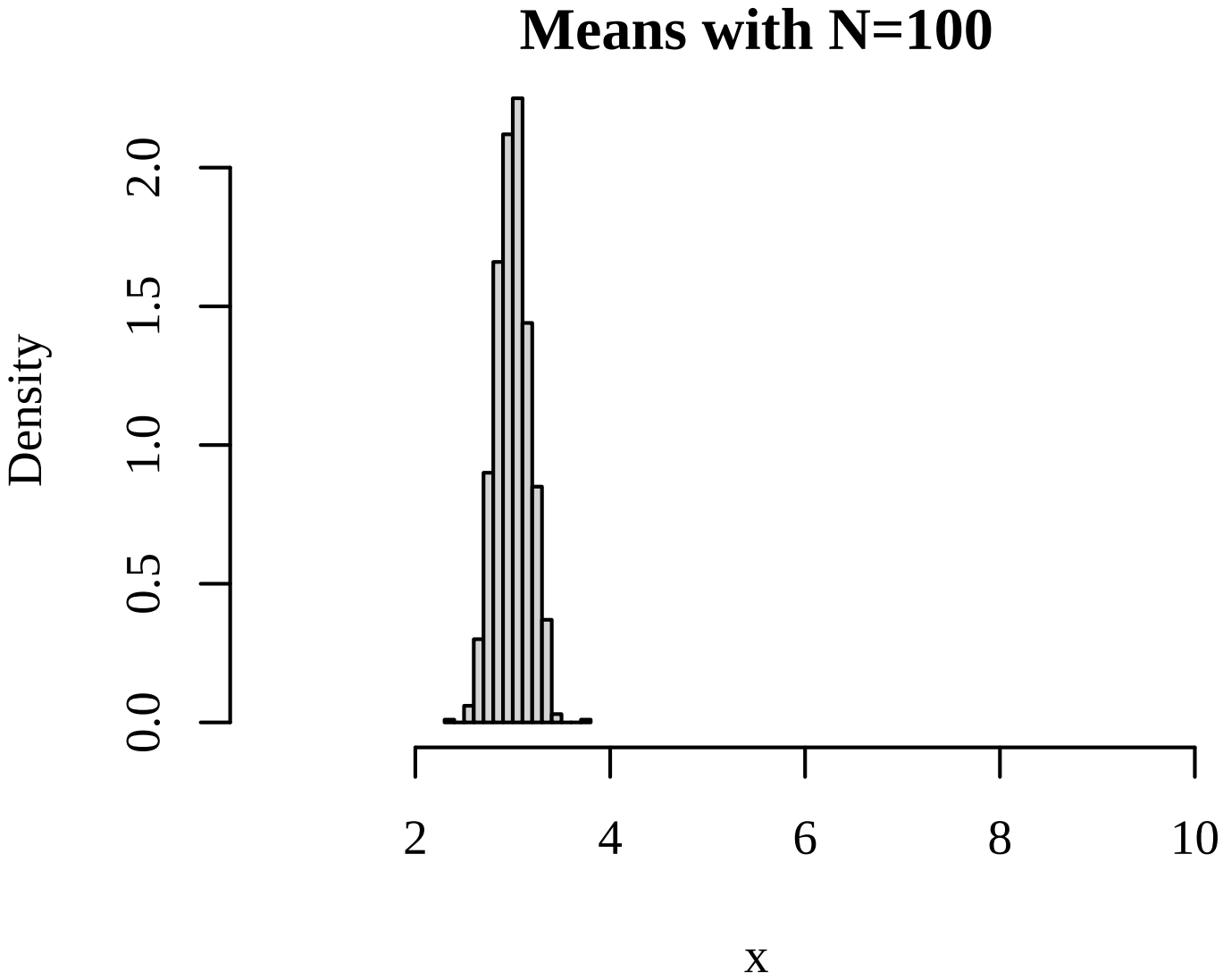
Poisson(3), SampleSize=10000



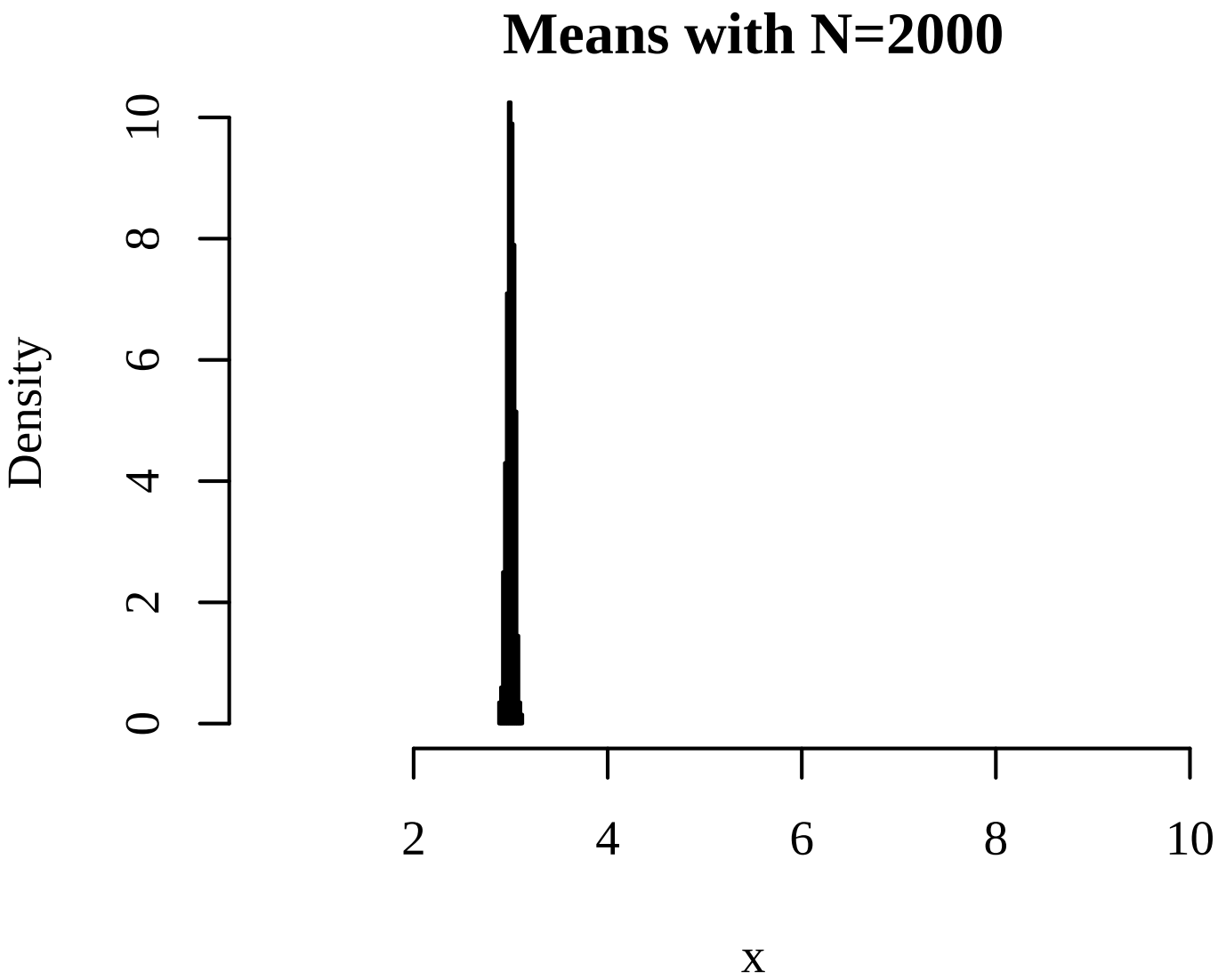
Means of 1000 Poisson Samples, Sample Size 10.



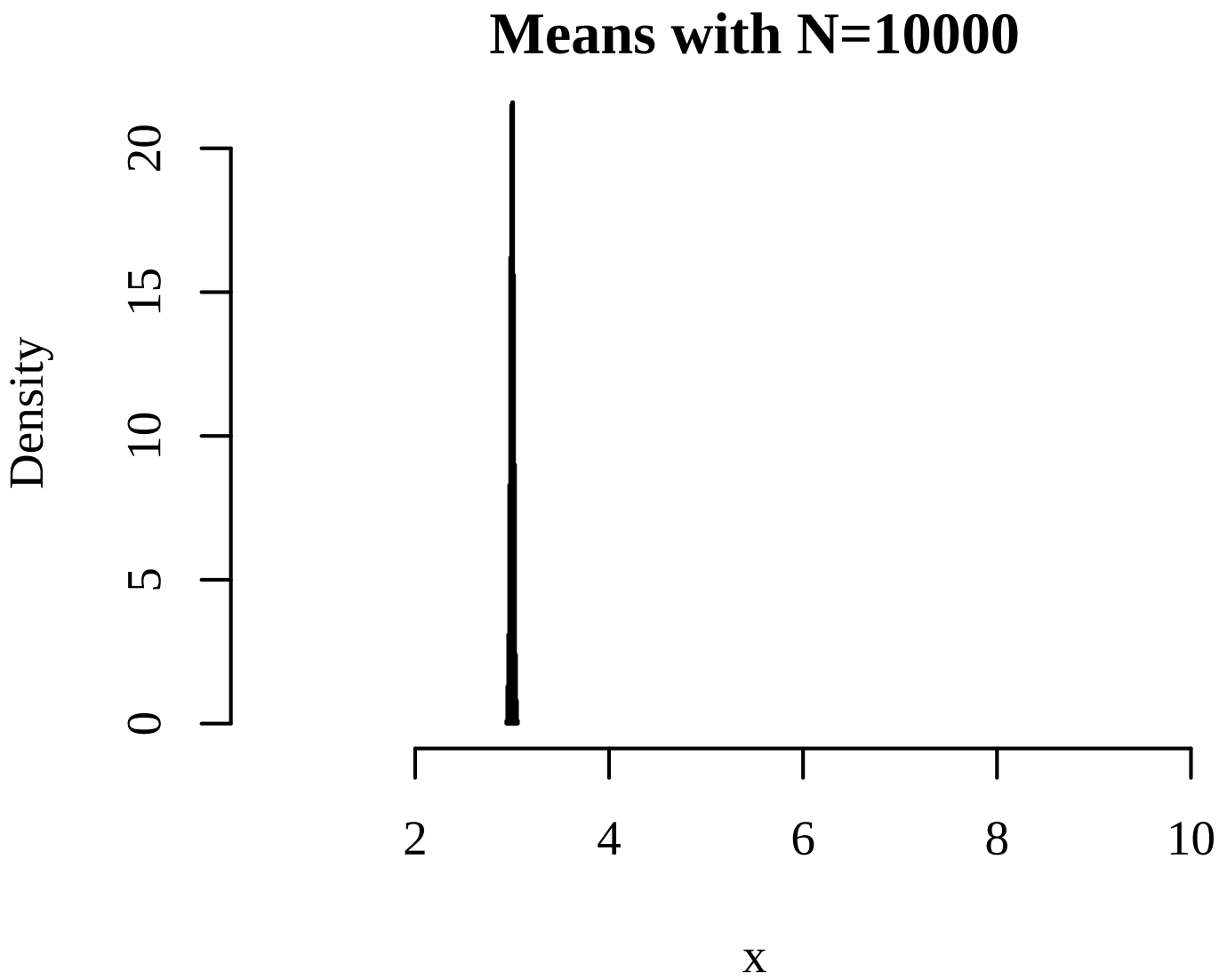
Means from 1000 Poissons, Sample Size=100



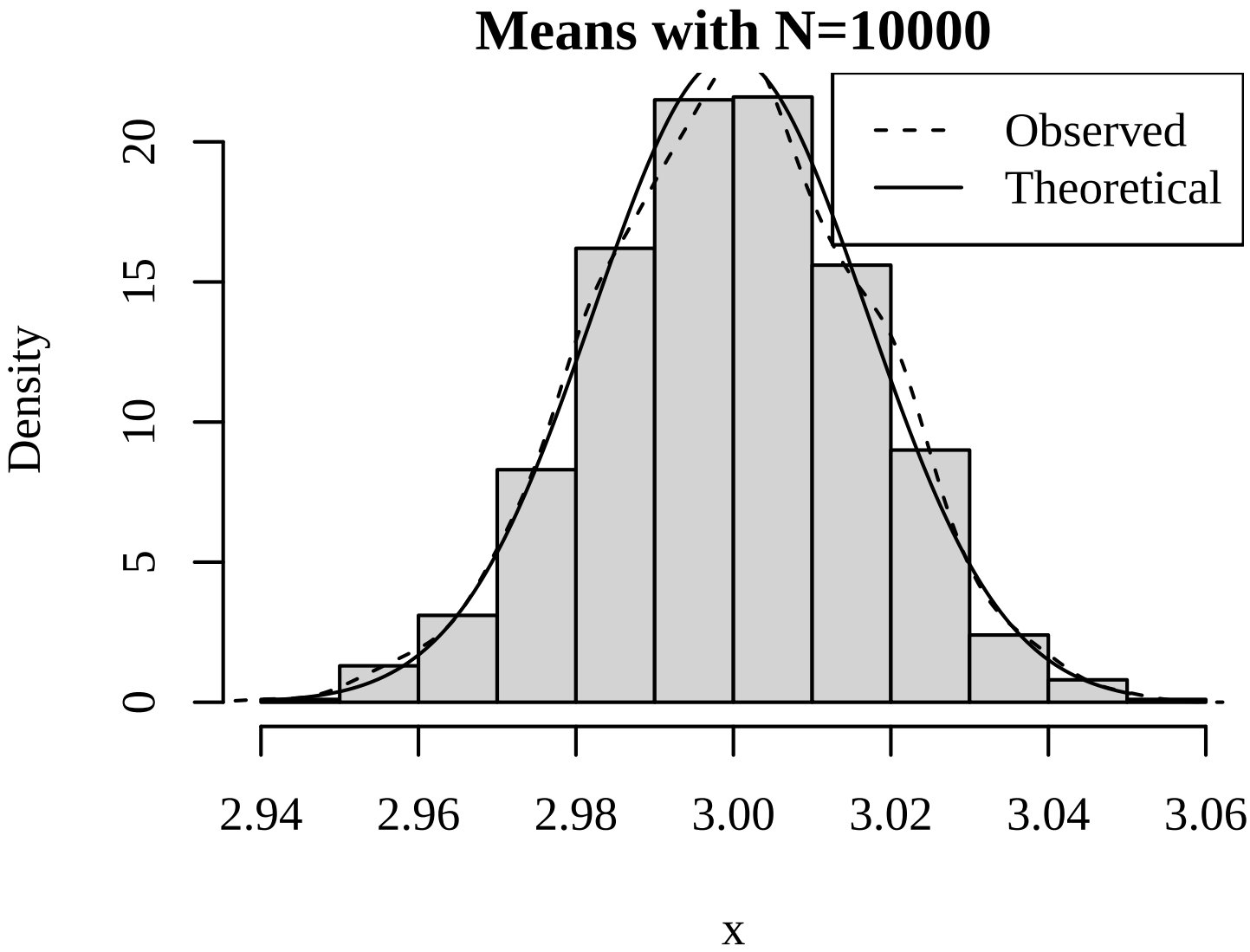
Means from 1000 Poisson samples, Sample Size=2000



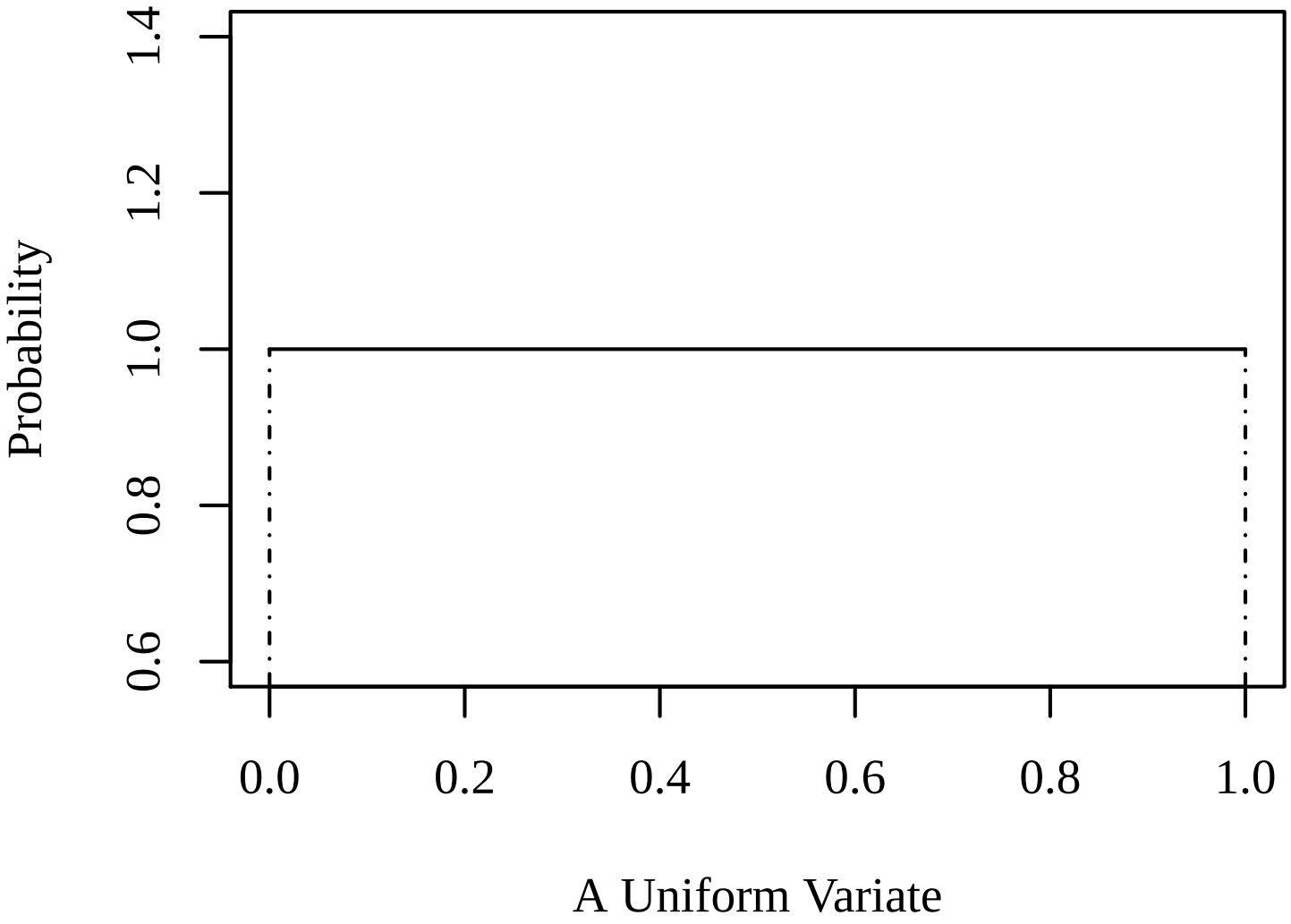
Means from 1000 Poisson samples, Sample Size=10000



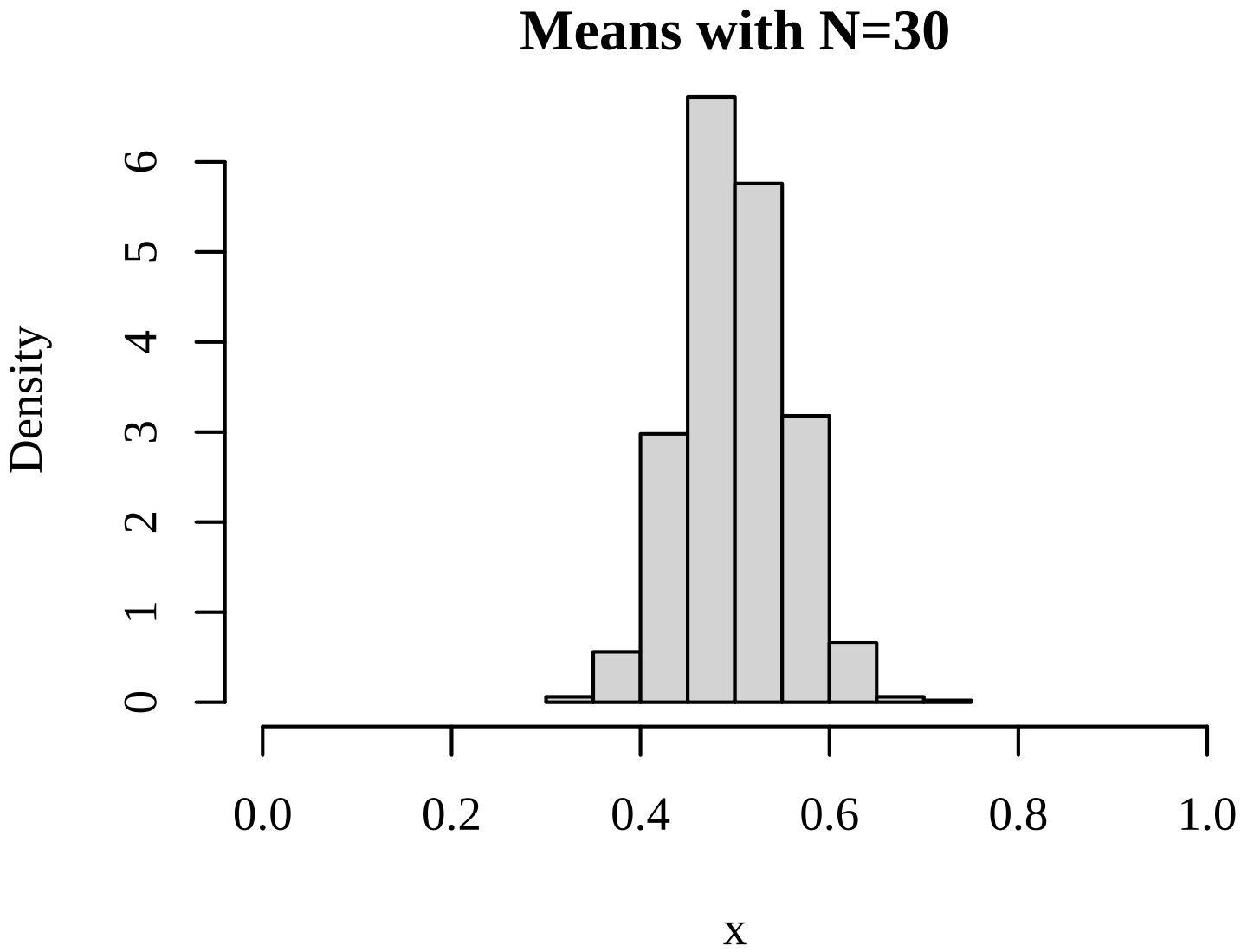
Same thing, bigger picture (N=10000)



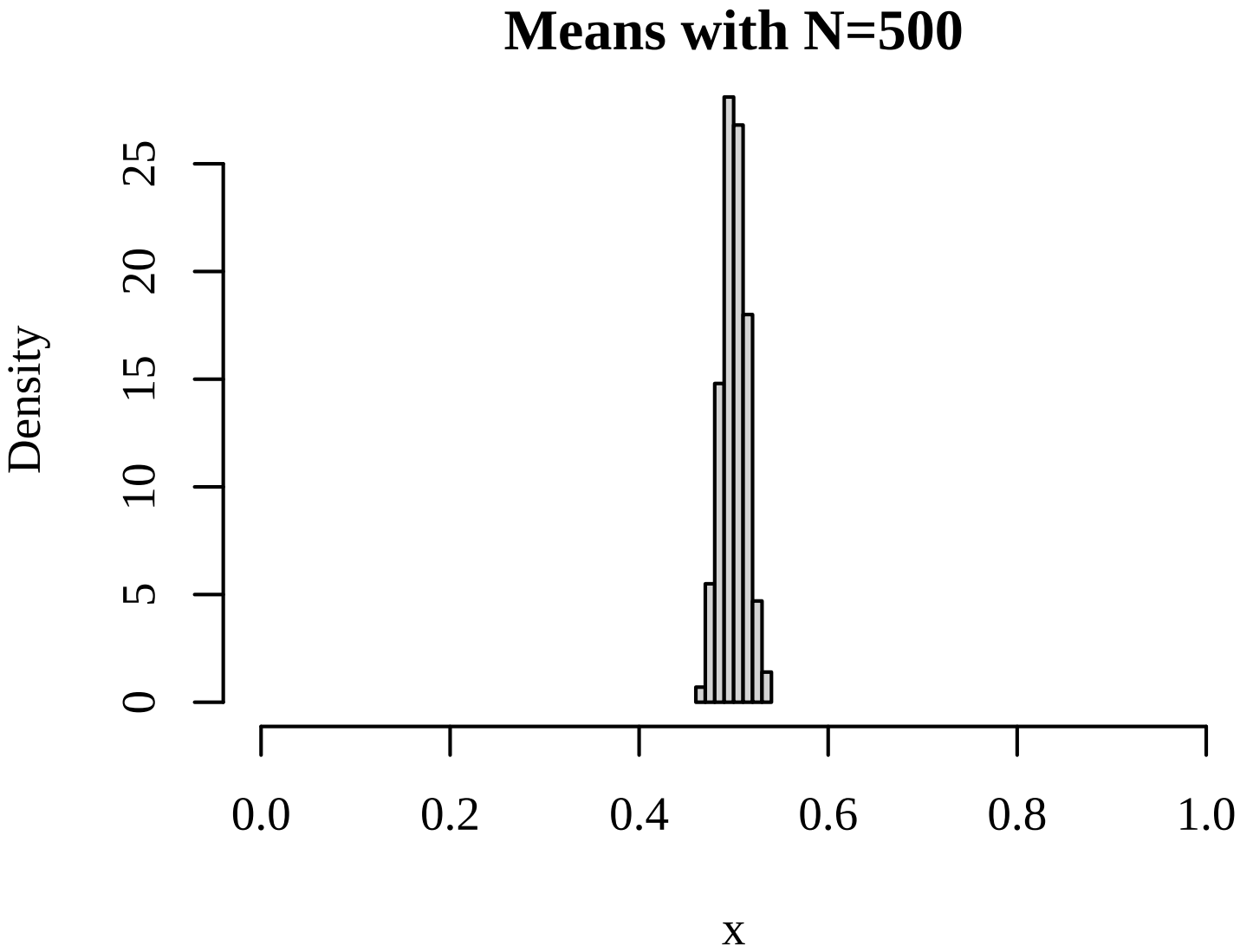
Consider the Uniform Distribution



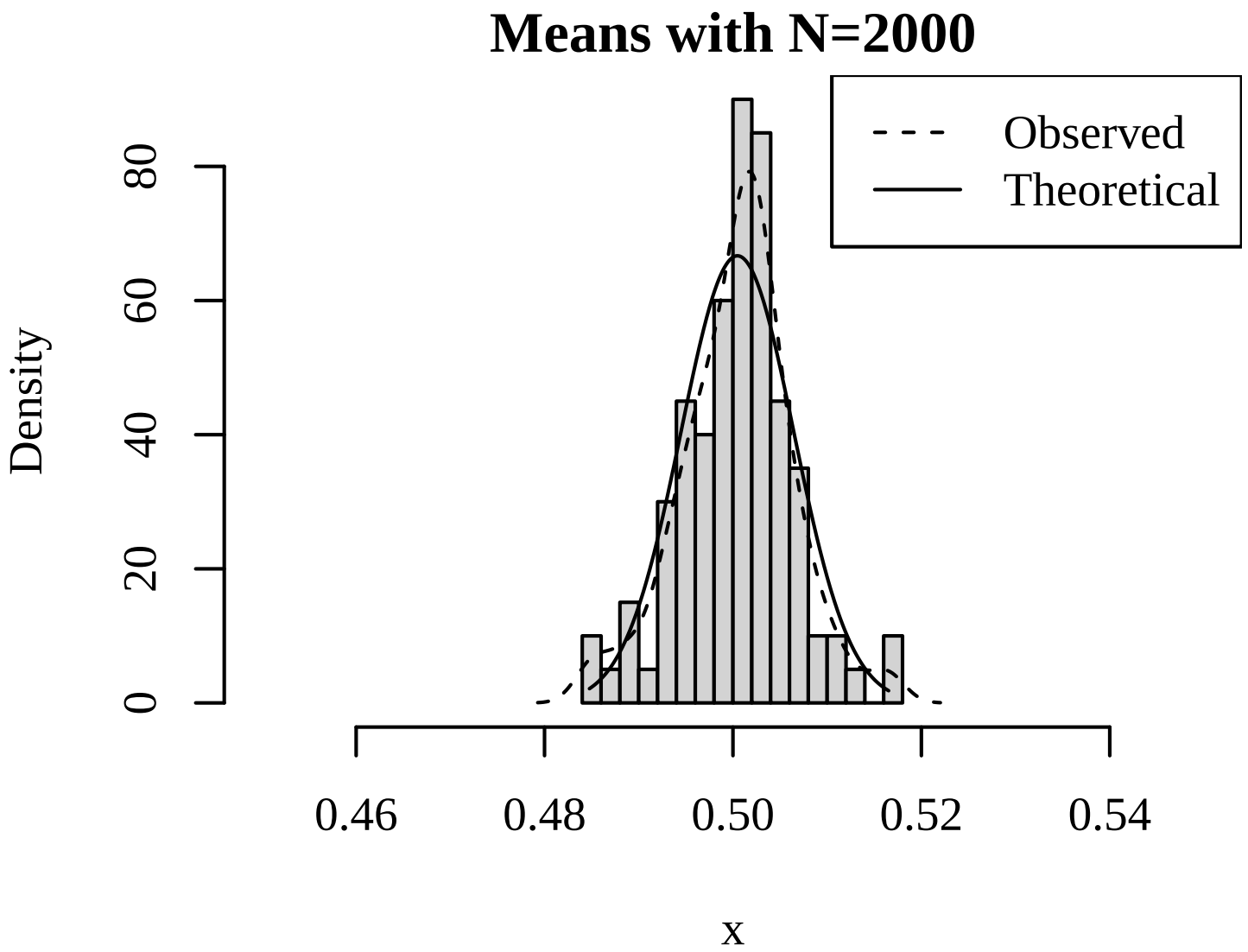
Means from 1000 Uniform samples, Sample Size=30



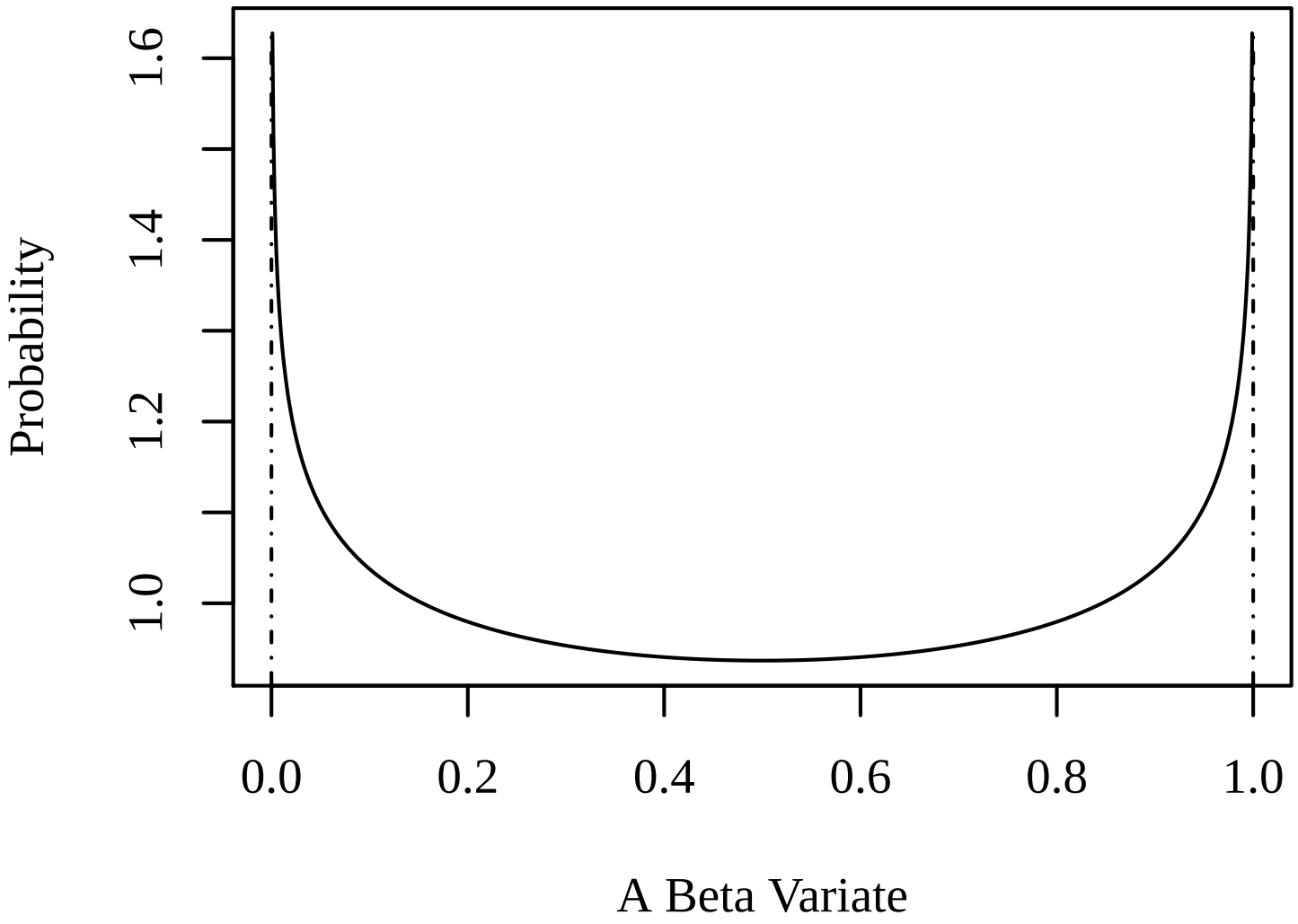
Means from 1000 Uniform samples, Sample Size=500



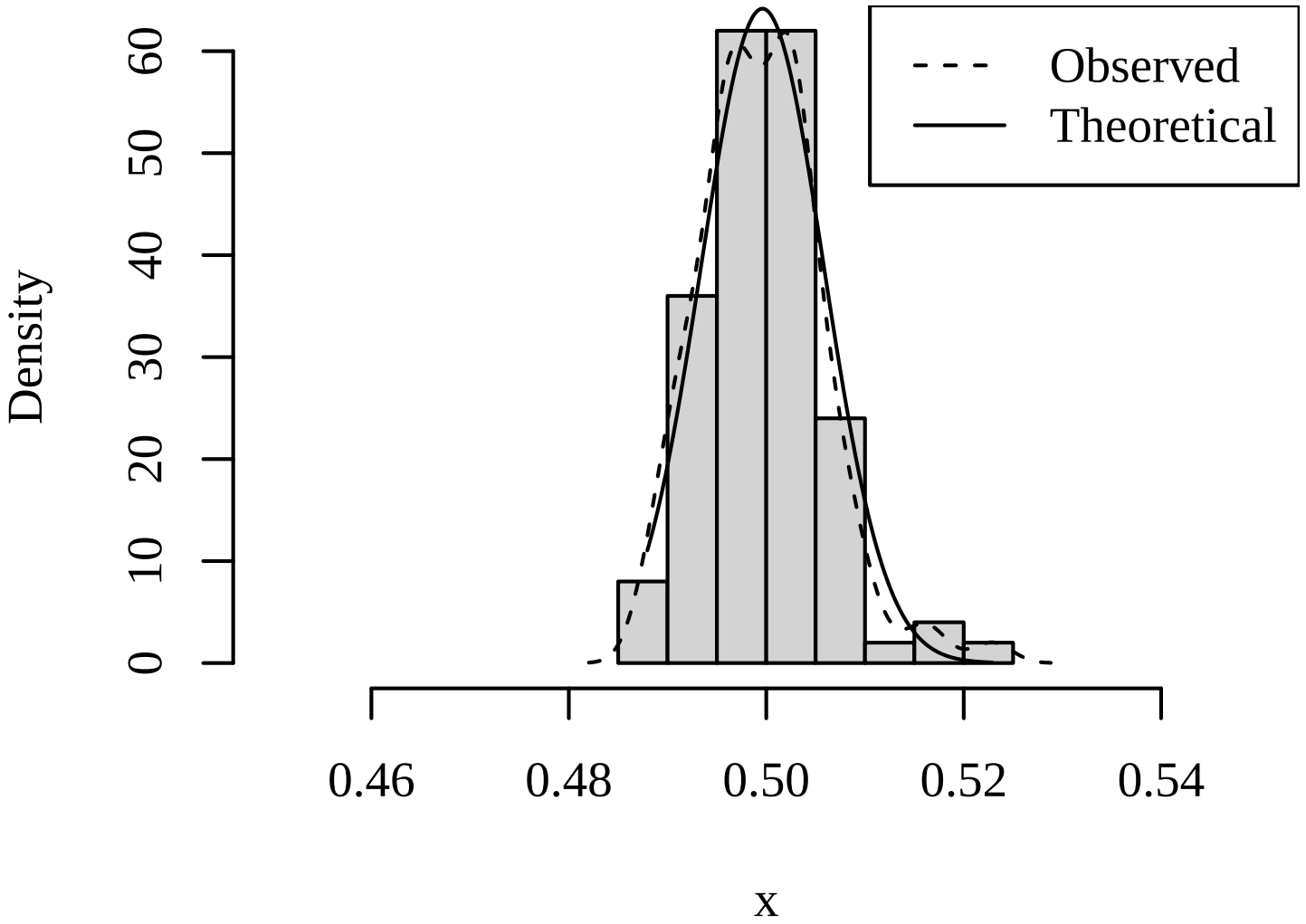
Means from 1000 Uniform samples, Sample Size=2000



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".

The CLT generalizes to any sum of random variables

- We are not interested primarily in the distribution of the mean
- But we have many other estimators that are weighted sums of random variables.
- Example: the estimated slope of a regression line

What is the Benefit of the CLT?

- Consider an estimator $\hat{\theta}$ that follows the CLT.
- Suppose the $E[\hat{\theta}] = 0$.
  - Why? We usually think of $\hat{\theta}$ as fluctuations around an estimator around the true value. $\hat{\theta}$ is "unbiased".
- IF (IF IF) we knew the standard deviation of the sampling distribution, then this would be a "standardized Normal variable"

$$\frac{\hat{\theta}}{true\, std.dev.(\hat{\theta})} \tag{2}$$

- I mean, it would be $N(0, 1)$, a VERY manageable quantity.

T Distribution Fix: If we don't know the true standard deviation of $\hat{\theta}$

- Divide this (which is $N(0,1)$)

$$\frac{\hat{\theta}}{true\ std.dev(\hat{\theta})} \tag{3}$$

- By this

$$\frac{\widehat{std.dev(\hat{\theta})}}{true\ std.dev.(\hat{\theta})} \tag{4}$$

- Gosset proved that the ratio follows a distribution that we now call T. T depends on the number of cases used to calculate the estimate, a number we call degrees of freedom.

Division cancels the unknown true $std.dev.(\hat{\theta})$

$$\frac{\frac{\hat{\theta}}{true\ std.dev(\hat{\theta})}}{\frac{\widehat{std.dev}(\hat{\theta})}{true\ std.dev.(\hat{\theta})}} \tag{5}$$

- After division, it is

$$\frac{\hat{\theta}}{\widehat{std.dev}(\hat{\theta})} \tag{6}$$

- We give a special name to the estimated standard deviation of the sampling distribution. It is called *standard error.*

$\hat{\theta}/std.err.(\hat{\theta})$ is Everpresent in Stats

- the T distribution is almost like the Normal(0,1).
- If the degrees of freedom is large (more than 1000), T and N(0,1) are virtually identical.
- Thus, the range $\hat{\theta} \pm 1.96 * std.err.(\hat{theta})$ contains about 95% of the distribution
- By implication, outcomes outside that 95% region are deemed "unusual" (2.5% of cases at either tail of distribution)
- If degrees of freedom is smaller, we just replace 1.96 with a slightly larger magic number (see "T table").

For Other Estimators, Much Detailed Research is Required

- We (applied social scientists) usually don't have training or interest in developing new math for sampling distributions.
- We do have some simulation tools to approximate unknown sampling distributions
- Simulation based ideas
  - Bootstrap: draw many samples from the data sample, re-calculate the estimate for each. The resulting distribution may approximate the sampling distribution.
  - Markov Chain Monte Carlo (MCMC): a computer simulation model developed during WWII as a way to explore complex probability models. Described in my review essay.

The Algebraic Argument for $E[\bar{x}] = E[x]$

The average (estimate of mean) of a sample $x_1, x_2, x_3, \ldots, x_N$ is:

$$\bar{x} = \frac{1}{N} \sum_{i}^{N} x_i \tag{7}$$

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

$$Prop.(x_j) = \frac{Frequency(x = x_j)}{N} \tag{8}$$

$$Mean(x_i) = \bar{x} = \sum_{j=1}^{m} Prop(x_j) x_j \tag{9}$$

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 sometimes thought of as the "population mean" or "true mean"
- Recall, population=the random process that generates x_i.
- Consider a discrete distribution f. Note $\bar{x}$ and $E[x]$
  - f is a "probability mass function"

$$Expected\ Value[x] = E[x] = \sum f(x_j)x_j \tag{10}$$

  - Same as 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 $f(x)$

$$E[x] = \int_{-\infty}^{+\infty} f(x)\, x\, dx. \tag{11}$$

Proof of claim that Expected Value of $\bar{x}$ equals Expected Value of x

Calculate the expected value of $\bar{x}$

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

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 $E[x]$

Recall the Variance of A Sum

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

$$Var[x1 + x2] = Var[x1] + Var[x2] + 2Cov[x1, x2] \tag{12}$$

And

$$Var[ax1 + bx2] = a^2 Var[x1] + b^2 Var[x2] + 2abCov[x1, x2] \tag{13}$$

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?

$$Var[\bar{x}] = Var[\frac{1}{N}x_1 + \frac{1}{N}x_2 + \ldots + \frac{1}{N}x_N] \tag{14}$$

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

$$Var(\frac{1}{N}x_1 + \frac{1}{N}x_2 + \ldots + \frac{1}{N}x_N) = \tag{15}$$

$$\frac{1}{N^2}Var(x_1) + \frac{1}{N^2}Var(x_2) + \ldots + \frac{1}{N^2}Var(x_N) \tag{16}$$

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:

$$Var(\bar{x}) = \frac{1}{N^2} \frac{NVar(x_i)}{1} \tag{17}$$

$$= \frac{1}{N} Var(x_i) \tag{18}$$