Sampling and Long-Run Probabilities

School of Information Studies
Syracuse University

Creative Video

From a demonstration by Yihui Xie:

```
library(animation)
ani.options(interval = 0.1, nmax = 250)
op = par(mar = c(3, 3, 1, 0.5), mgp = c(1.5, 0.5, 0), tcl = -0.3)
clt.ani(type = “s”)
```

Learning Topics for This Week

Define the meaning of population and sample.

Describe the process of sampling.

Generate a simulated population with R.

Use R to generate example distributions.

Describe the law of large numbers and the central limit theorem.

Demonstrate the ability to reason about events in a sampling distribution.

Use R to graph distributions with a histogram and mark quantiles.

Use marked distributions to reason about probability.

Define the Meaning of Population and Sample

Population and Sample

A population is the entirety of a phenomenon that we hope to study.

Populations often have a conceptual definition but are "unreachable" as a whole: for example, all iPhones in use globally—easy to define, virtually impossible to study as a complete set.

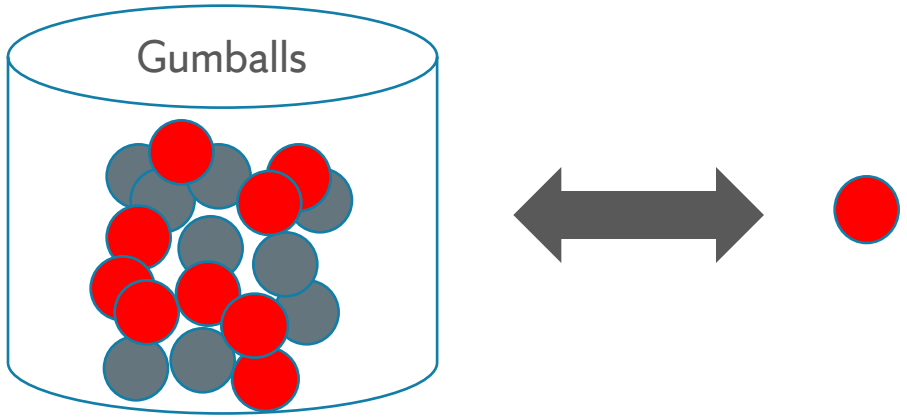
Because nontrivial populations are often difficult to access, we use "sampling" instead. Sampling is any systematic process for selecting individual cases from a population.

Sampling is both a science and an art. Statisticians have established the mathematical basis for many forms of sampling, but as applied researchers, it is up to us to make sound and thoughtful decisions on how to obtain our data.

Describe the Process of Sampling

Sampling Processes

Sampling is the process of drawing elements from a population. Random sampling is when every element has an equal chance of being drawn. Replacement is the curious idea that we put the element back after we have drawn it but before we draw the next random element.



Generate a Simulated Population With R

Toast Angle Data: A Simulated Population

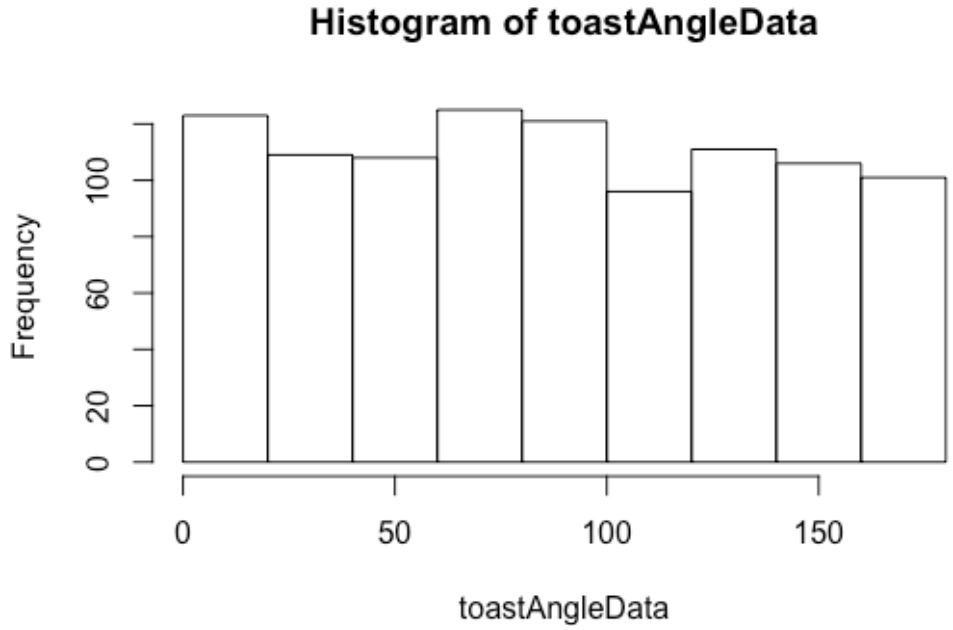
In the commands below, we create a fake population consisting of "angle data" from dropping toast.

Our toast angle data follows the uniform distribution.

```
> toastAngleData <- runif(1000,0,180)  # Random numbers: uniform distribution
> head(toastAngleData)                 # Look at the first few numbers in the list
[1] 60.77672 99.40319 97.76814 105.08108 177.08945 48.84156
> tail(toastAngleData)                 # Look at the last few numbers in the list
[1] 74.38751 161.30393 115.12818 163.97210 151.53179 50.21762
> mean(toastAngleData)                 # Report the population mean
[1] 92.45193
```

Toast Angle Data: A Simulated Population

hist(toastAngleData) produces a histogram of all 1,000 of our data points. If this is a uniform distribution, why are all of the bar heights slightly different from each other?



Use R to Generate Example Distributions

Repetitious Sampling With R

```
> sample(toastAngleData, size=14, replace=TRUE)
[1] 152.07620 102.20549  89.35385  42.75709  21.13263
158.35032 141.95377
[8]  16.31910 136.32748 171.54875  44.39738  36.60672
87.92560 149.97864
> mean(sample(toastAngleData,size=14,replace=TRUE)) # Mean
of one sample
[1] 93.88385
```

Repetitious Sampling With R

```
# Summarize the 14 elements by taking the mean;
# Use replicate() to repeat that process
samplingDistribution <-
replicate(10000,mean(sample(toastAngleData,
size=14,replace=TRUE)), simplify=TRUE)
```

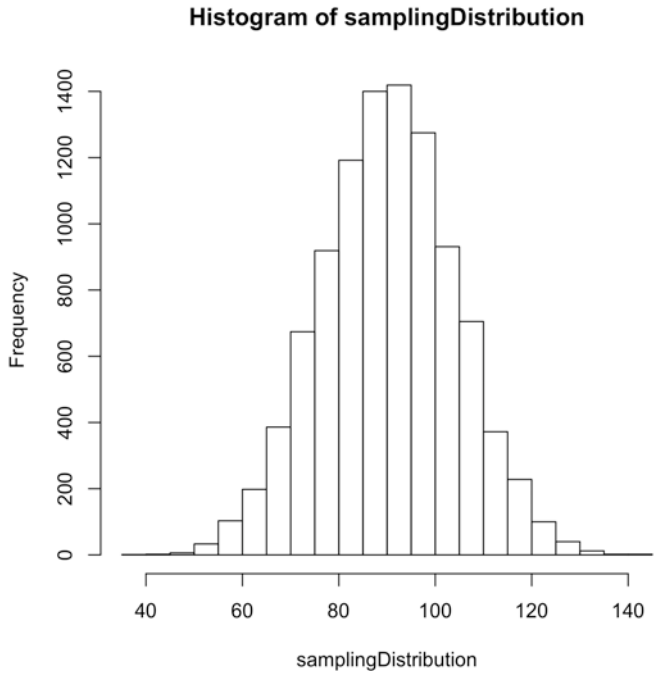
hist(samplingDistribution)

Features of this histogram:

The data plotted here consist of 10,000 means sampled from toastAngleData, where each sample is n=14 observations

The mode (the highest bar) is right near 90, which also happens to be near the mean of the original uniform distribution of toastAngleData

The distribution is forming a symmetric bell shape

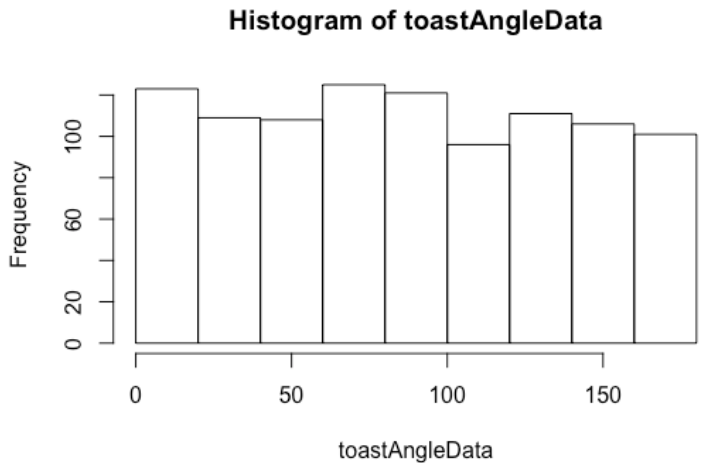


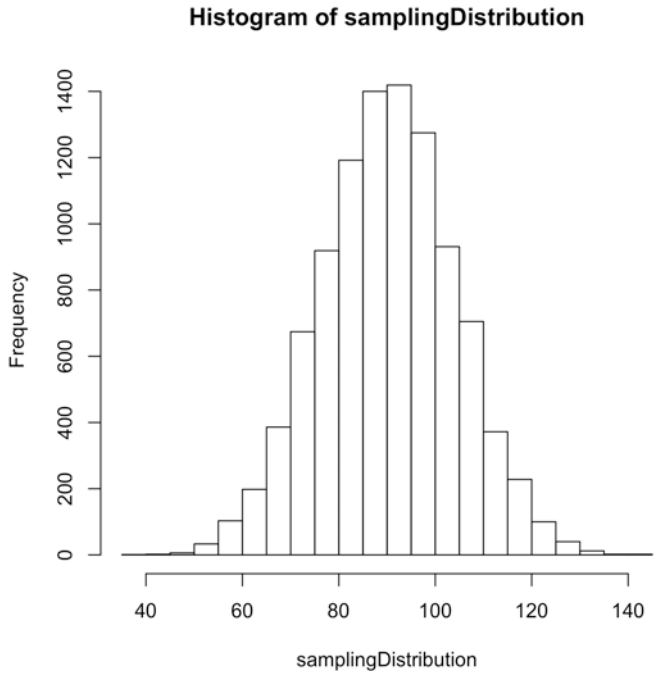
Important Distinction to Remember

The mean of the raw data above (a uniform distribution ranging from 0 to 180) is about 90.

The mean of the sampling distribution over on the right (10,000 individual means calculated from samples of n=14) is also about 90.

What makes the distribution shapes so different?





Describe the Law of Large Numbers and the Central Limit Theorem

The Law of Large Numbers

If you run a statistical process like sampling a large number of times, it will generally converge on a particular result.

Swiss mathematician and astronomer Jakob Bernoulli suggested this idea in a book called *The Art of Conjecturing*.

For example, if you keep track of the number of heads when tossing a fair coin, after a large number of trials the proportion will converge on 50% heads.

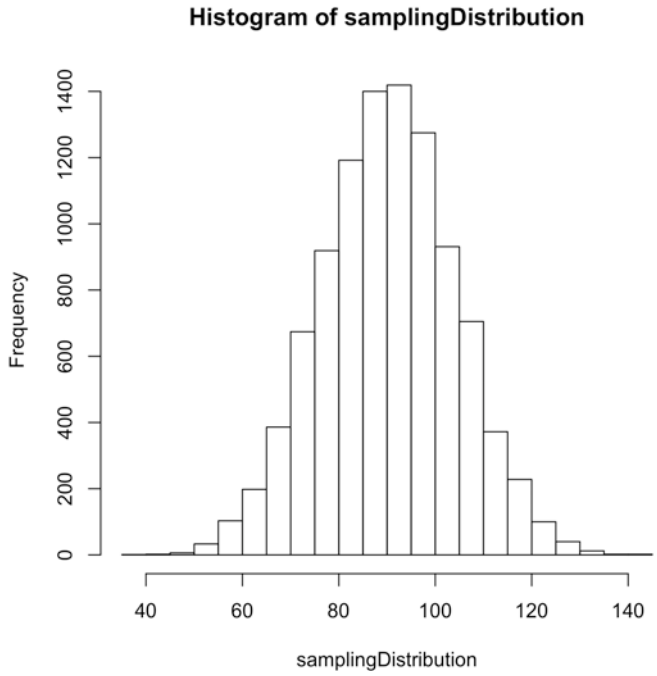
Jakob Bernoulli, 1645-1705 public domain image

The Central Limit Theorem

When independent variables are combined, the resulting distribution will be normal, even if the variables themselves are not normally distributed.

We are combining variables by calculating a mean of samples drawn from a population.

In this case, the central limit theorem also states that, over the long run, the mean of the sampling distribution will match the mean of the underlying population.



Marking Quantiles on a Distribution

Quantile

Many people have heard the term percentile, or perhaps quartile. These are both terms that mark a position within a collection of values.

For example, the median, which you learned about last week, is the 50th percentile and also the 2nd quartile.

A more general term is quantile. Quantile is like percentile but expressed as a decimal instead of a percent.

The median is the 0.50 quantile.

You can use the quantile command in R to probe any collection of numbers. For example:

```
quantile(0:100,probs=0.75)
```

This command looks for the 0.75 quantile within the digits 0 through 100.

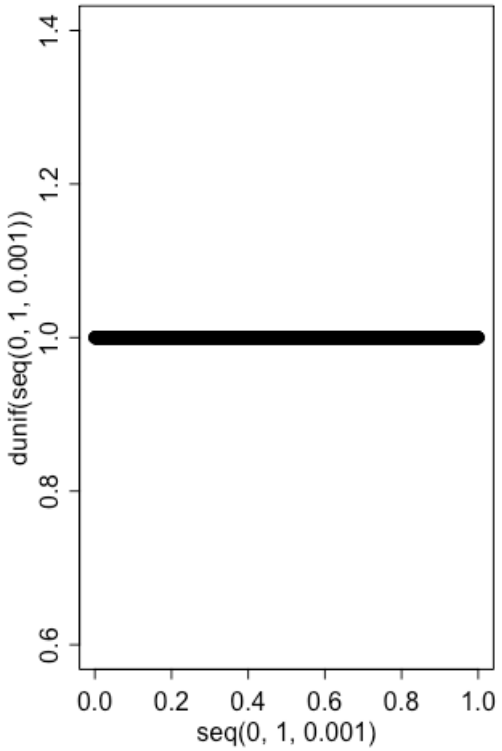
Area Under a Curve

This is a concept from calculus that allows us to estimate or calculate the amount of space underneath a diagram like this one.

This is a graph of the uniform distribution in the range 0 to 1, and it is easy to see that the area under the line is exactly 1 because the measurements are 1×1.

You can create this plot with:

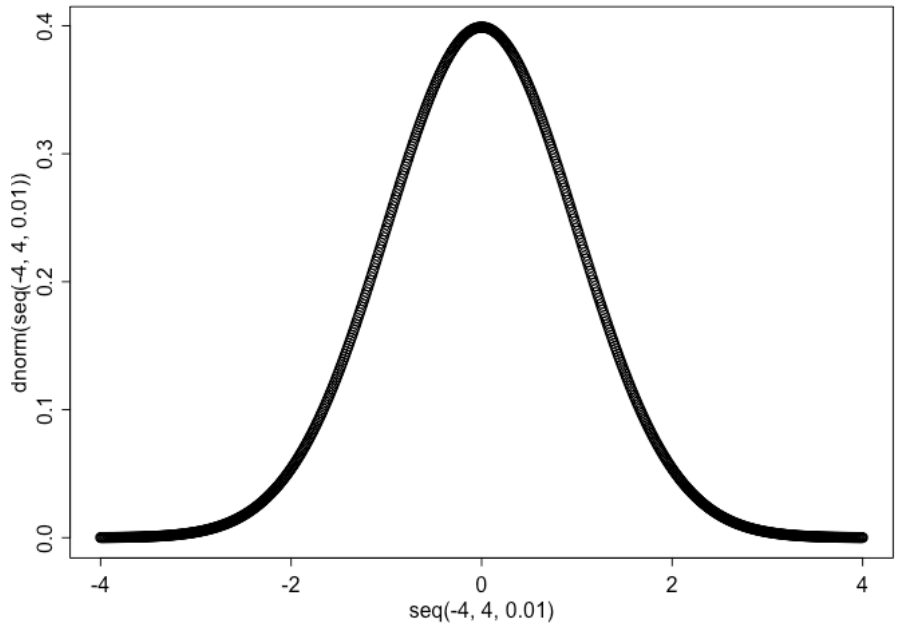
plot(seq(0,1,.001), dunif(seq(0,1,.001)))



Area Under a Normal Curve

Figuring the area under something that is actually curvy is trickier, but the plot below helps visualize the "unit curve" for a normal distribution.

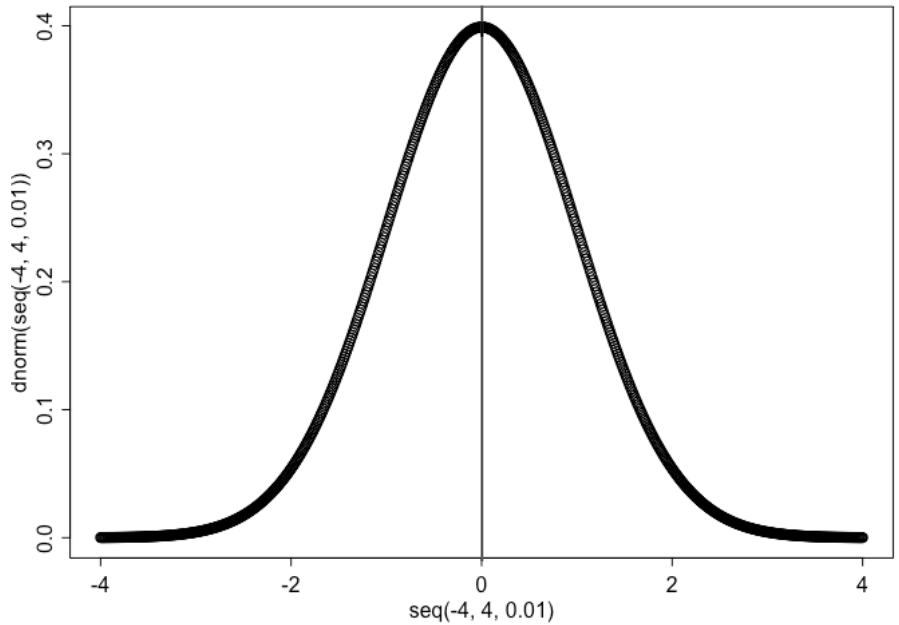
plot(seq(-4,4,.01), dnorm(seq(-4,4,.01)))



Marking the Curve

We can use the abline() command to draw a line, and the qnorm() command to mark any quantile on the normal curve.

abline(v=qnorm(0.50))

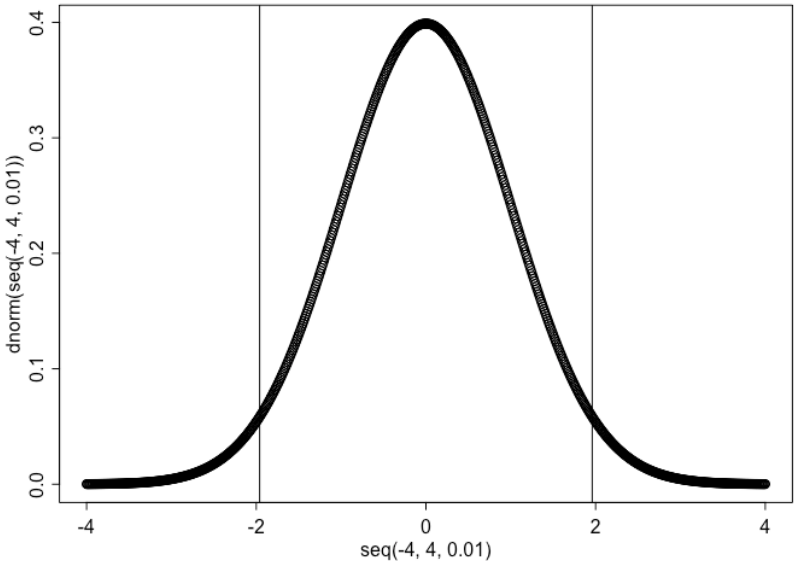


Putting It All Together

In this diagram, we plotted a normal curve with vertical lines for the 0.025 and 0.975 quantiles:

```
plot(seq(-4,4,.01),
          dnorm(seq(-4,4,.01)))
abline(v=qnorm(0.975))
abline(v=qnorm(0.025))
```

So, 95% of the area under the curve is in the central region

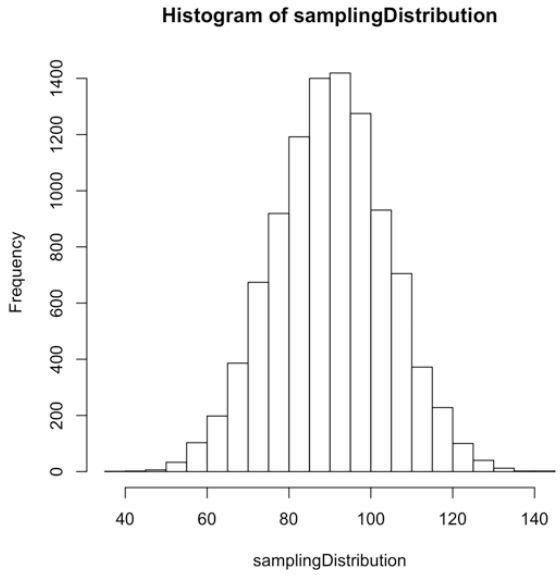


Use R to Graph Distributions With a Histogram and Mark Quantiles

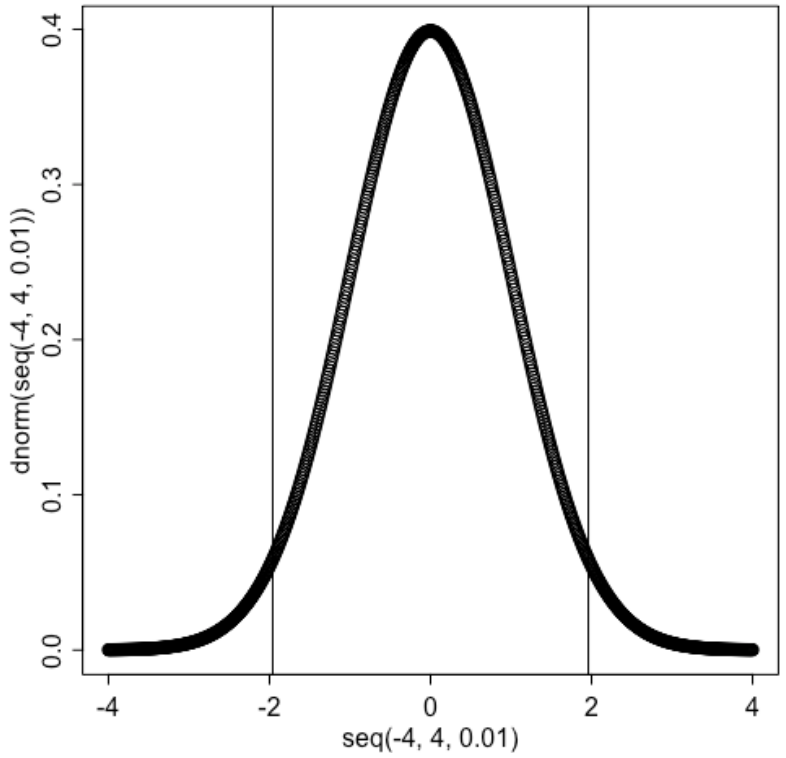
Now With Real Data!

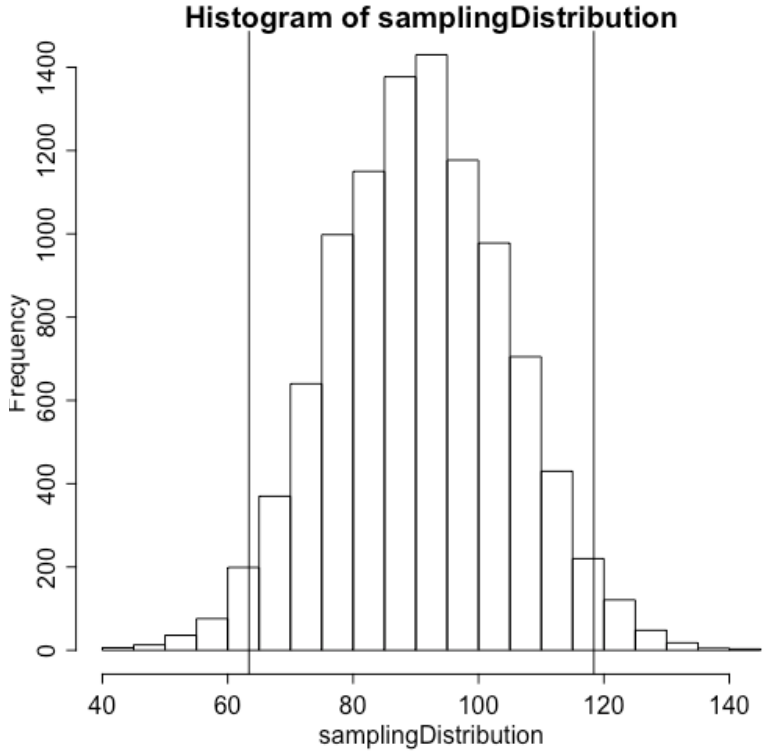
This code resamples our toast angle data, plots it, and marks off the 0.025 and 0.975 quantiles, leaving 95% of the area under the curve in the central region

```
samplingDistribution <- replicate(10000,
        mean(sample(toastAngleData,
        size=14,replace=TRUE)),simplify=TRUE)

hist(samplingDistribution)

abline(v=quantile(samplingDistribution,0.025))

abline(v=quantile(samplingDistribution,0.975))
```



Comparing the Ideal With the Reality





Use Marked Distributions to Reason About Probability

Reasoning About Distributions

Now we have a bird's-eye view of a sampling process that comprises 10,000 means of samples drawn from our toast angle data.

First, in line with the central limit theorem, the center of this distribution is about 90—the same as the underlying raw population data. It is also a pretty normal looking curve.

Second, we now know a lot about common and unusual sampling results: 95% of all sample means of toastAngleData fall between 63 and 118. Any samples more extreme than that are rare.