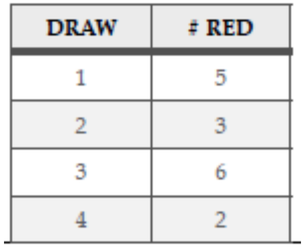
Introduction to Sampling

Sample in a Jar

Sampling distributions are the conceptual key to statistical inference. Many approaches to understanding sampling distributions use examples of drawing marbles or gumballs from a large jar to illustrate the influence of randomness on sampling. Using the list of U.S. states illustrates how a non-normal distribution nonetheless has a normal sampling distribution of means.

Data Science



DRAW	# RED
1	5
2	3
3	6
4	2

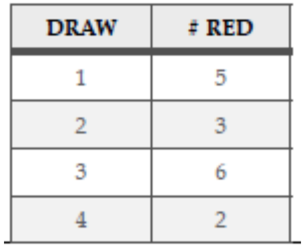
- Red gum balls, blue gum balls
- Same ratio of each in a jar
  - – Draw a sample of eight (one draw)
  - – What mix of red and blue gum balls will we get with one draw? → Really don't know.

- Forces of "randomness" driving uncertainty

- "Long run test" → multiple draws

Data Science

- “Drawing” process
  - Population (gum balls)
  - Sample
  - Distribution



DRAW	# RED
1	5
2	3
3	6
4	2

- Use R to draw samples
  - sample(USstatePops$april10census, size=16, replace=TRUE)
  - 12702379 19378102 8791894 19378102 9535483 6346105 4533372 5029196 25145561 6392017 19378102 6483802 8001024 8001024 12830632 814180

- Use R to calculate mean of the sample
  - mean(sample(USstatePops$april10census, size=16, replace=TRUE))
  - 5513472

Data Science

Question:

- Why is sampling from a population important?
- What are some key things to think about when sampling?

Replicating Samples

Data Science

- Not interested in one sample but what happens over time, i.e., many samples

- Replication
  - replicate(4,mean(sample(USstatePops$april10census,size=16,replace=TRUE)),simplify=TRUE)
  - 5234752   5978035   5876217   4222350

Data Science

-
  ```
  mean(replicate(400,mean(sample(
          USstatePops$april10census, size=16,
          replace=TRUE)), simplify=TRUE))
  ```

  ```
  > 6014258
  ```

- Interpretation
  - Draw 400 samples of size 16 from our state population.
  - Calculate the mean from each sample and keep it in a list.
  - Calculate the mean of the 400 sample means.
  - Calculated mean of means is off by 39,577.
    - 6,053,835 (mean of 51 states) – 6,014,258 (mean of means) = 39,577
    - 39,577 / 6,053,835 = .65% error

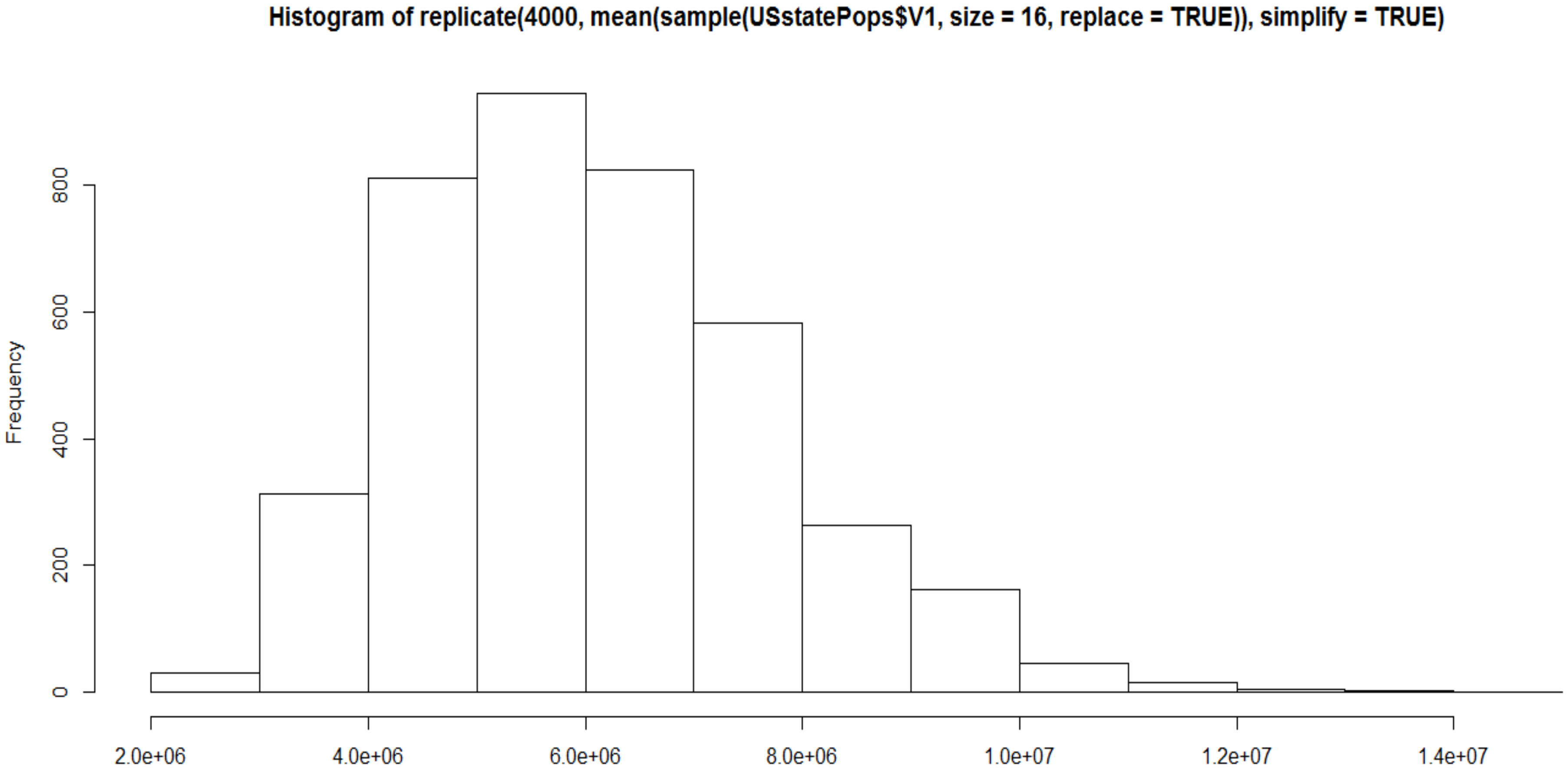
Data Science

- mean(replicate(4000,mean(sample(USstatePops $april10census,size=16,replace=TRUE)), simplify=TRUE))
  - 6053534

    6,053,835 (mean of 51 states) – 6,053,534 (mean of means) = 301

- Display distribution of 4000 means via a histogram as frequencies
  - hist(replicate(4000,mean(sample(USstatePops $april10census,size=16,replace=TRUE)),simplify=TRUE))

Data Science



Data Science

- Law of large numbers
  - If you run a statistical process a large number of times, it will converge on a stable result.
- Central limit theorem
  - When we look at sample means and take into account the "law of large numbers," the distribution of sampling means starts to create a bell-shaped or normal distribution, and the center of that distribution, the mean of the sample means, gets close to the population mean.

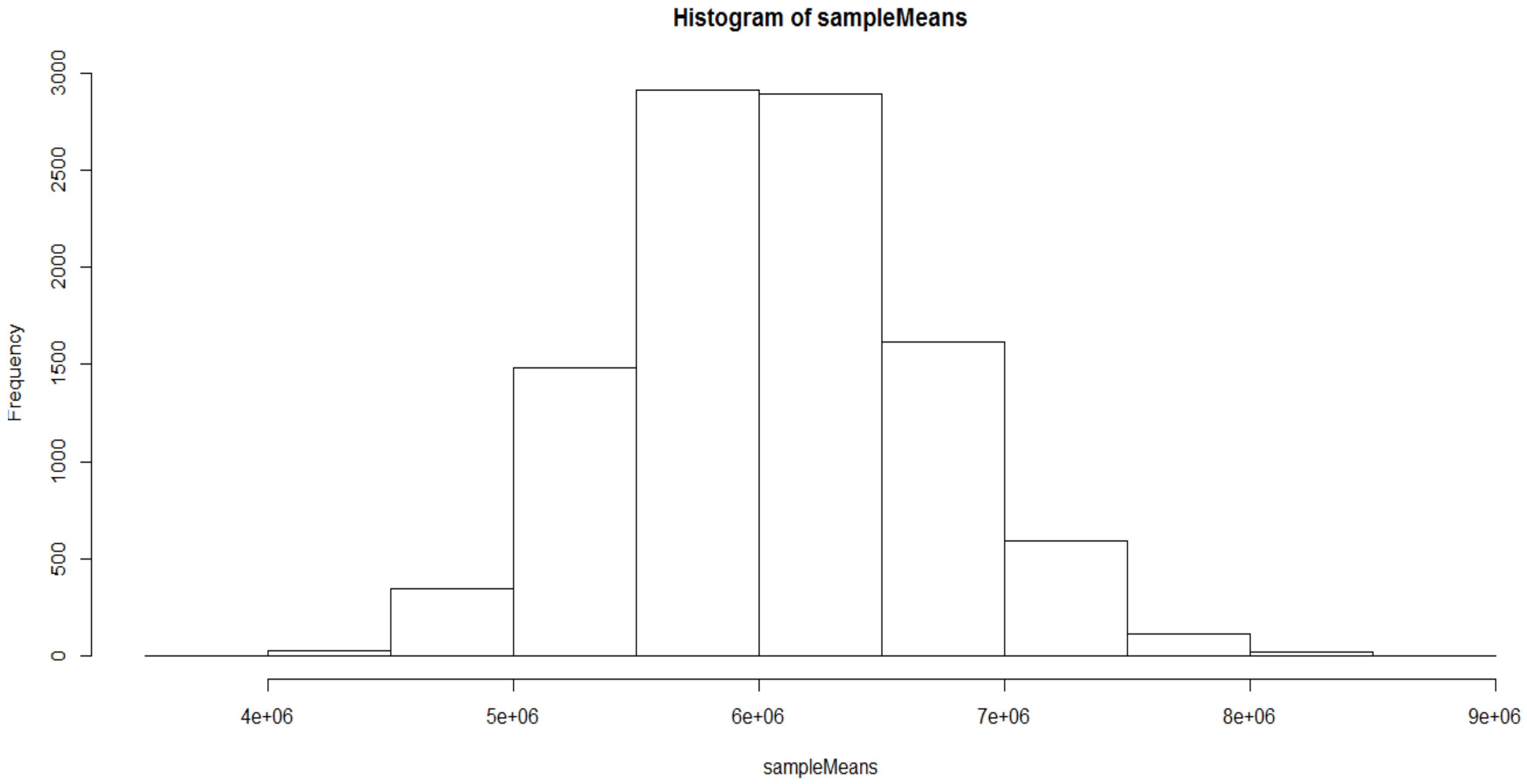
Data Science

- Sampling distribution
  - Save one distribution sample

```
SampleMeans <-replicate(10000, mean(sample(USstatePops$april10census,
size=120, replace=TRUE)), simplfy=TRUE)
```

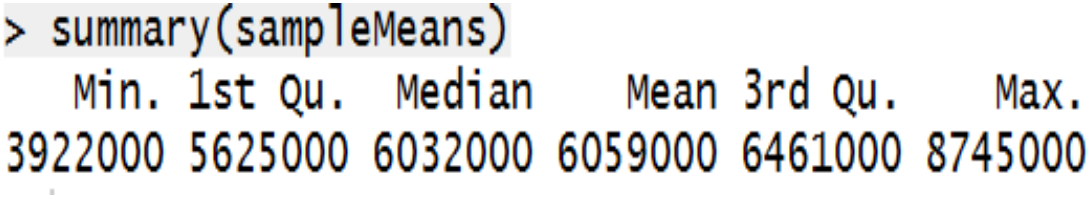
  - Length(SampleMeans)
    - 10000
  - mean(SampleMeans)
    - 6058734
  - Histogram
    - Hist(SampleMeans)

Data Science



Data Science

- Sampling distribution
  - Summary(SampleMeans)

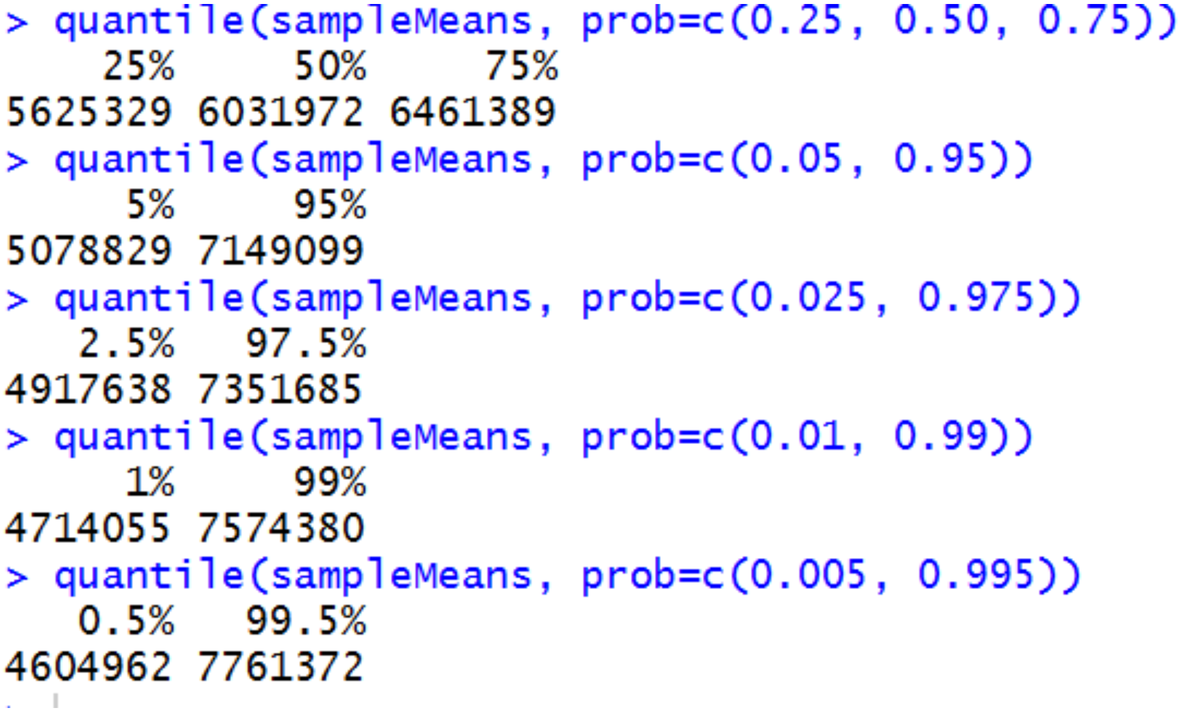


```
> summary(sampleMeans)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
3922000 5625000 6032000 6059000 6461000 8745000 
```

  - Interpretation
    - Min vs. mean
    - Max vs. mean
    - Median vs. mean
    - Quartiles

Data Science

- Sampling distribution
  - Quantile function, similar to summary function



```
> quantile(sampleMeans, prob=c(0.25, 0.50, 0.75))
    25%     50%     75% 
5625329 6031972 6461389 
> quantile(sampleMeans, prob=c(0.05, 0.95))
     5%     95% 
5078829 7149099 
> quantile(sampleMeans, prob=c(0.025, 0.975))
   2.5%   97.5% 
4917638 7351685 
> quantile(sampleMeans, prob=c(0.01, 0.99))
     1%     99% 
4714055 7574380 
> quantile(sampleMeans, prob=c(0.005, 0.995))
   0.5%   99.5% 
4604962 7761372 
```

Mystery Samples

Data Science

- MysterySample

```
> MysterySample <- c(3706690, 159358, 106405,
                                      55519, 53883)

> mean(MysterySample)
 816731
```

Data Science

- MysterySample: sample of U.S. states or something else?
- Basis of comparison
  - Subsequent USstatePops$april10census analysis
  - Sampling distribution of means: USstatePops$april10census represented in vector SampleMeans
  - Compare MysterySample mean to SampleMeans: "quantile analysis"
    - mean(SampleMeans)
    - quantile(SampleMeans, probs=c(0.25, 0.50, 0.75))
    - quantile(SampleMeans, probs=c(0.05, .095))
    - quantile(SampleMeans, probs=c(0.01, .099))
    - is MysterySample mean below 5% mark or above 95% mark
    - is MysterySample mean below 1% mark or above 99% mark

Data Science

- MysterySample: sample of U.S. states or something else?
  - 1% of all the SampleMeans are lower than 4,710,455.
  - Therefore, MysterySample mean of 816,731 would be a rare event.
  - We can infer, tentatively but based on good statistical evidence, that the MysterySample is not a sample of states.
  - Key takeaway
    - The mean of the MysterySample was sufficiently different from a known distribution of means such that we could make an inference that the sample was not drawn from the original population of data.

Data Science

- Basis for most all statistical inference
  - – Construct a comparison distribution.
  - – Identify a zone of extreme values.
  - – Compare new sample of data to the distribution relative to the "extreme" zones.
  - – If new sample does fall in the "extreme zone," you can tentatively conclude that the new sample was obtained from some other source than what you used to create the comparison distribution.

Data Science

- Brief recap
  - – Mean() of sampling distribution
  - – Sampling distribution shape via hist()

- Need to quantify the distribution spread via sd()
  - – sd(SampleMeans)
    - 621569

Data Science

- Standard deviation of the distribution of sampling means, also known as "standard error of the mean"

  sd(population)/sqrt (# of samples)

- Alternative to sd(SampleMeans) relative to calculating the "standard error of the mean"
  - sd(USstatePops$april10census)/sqrt (120)
    - 622941
  - Differences due to randomness of the distribution

Data Science

- Alternative to quantile() cut points
  - Use mean and standard error.
  - Two standard deviations down from the mean is the 5% cut point.
  - Two standard deviations up from the mean is the 95% cut point.
    - StdError <-sd(USstatePops$april10census)/sqrt(120)
    - CutPoint5<-mean(USstatePops$april10census)-(2 * StdError)
    - CutPoint95<-mean(USstatePops$april10census)+(2 * StdError)
    - CutPoint5
      - 4807951
    - CutPoint95
      - 7299717

Data Science

- Summary
  - Data set with 51 data points, numbers of people in 51 states
  - Use R to construct distribution of sampling means
  - Highlighted the process of statistical inference
    - Construct a comparison distribution.
    - Identify a zone of extreme values.
    - Compare new sample of data to the distribution relative to the “extreme” zones.
    - If new sample falls in the extreme zone, you can tentatively conclude that the new sample was obtained from some other source than what you used to create the comparison distribution.

Data Science

- Question:

  Why is it useful (or when is it useful) to compare two samples?