Intro to DS

Week 4: Samples and Populations

Summary of Previous Learning

Here's what you should know and be able to do at this point in the course:

1. List major skills needed by data scientists (e.g., communicate with client)
2. Describe a DS project and describe "domain analysis and decomposition" as well as "identification of SMEs"
3. Explain basic concepts of data modeling
4. Explain and use a data frame in R
5. Define the most common descriptive statistics and calculate them with R using appropriate functions
6. Demonstrate the development of a simple function in R

Data Science in the real world

The New Coke Machine

The New Coke Machine

1. Who is the customer?
2. Suggest a "focus area" for a data science project.
3. Give an example of data you want.
4. With the suggested focus and data, who is a domain expert to interview?

source: Bsquare

Coca-Cola Scores With Retro-Feel Creative Campaign Introducing Orange Vanilla Coke

Amy Barnes Contributor
Food & Drink
I cover the southern region, quirky foods and innovative packaging.

Orange Vanilla Coke and Orange Vanilla Coke Zero Sugar THE COCA-COLA COMPANY

https://www.forbes.com/sites/amybarnes/2019/02/25/coke-scores-with-retro-feel-creative-campaign-introducing-orange-vanilla-coke/

Skin Cancer: Detecting Melanoma

Skin cancer:

- Is the most common type of cancer.
- Occurs due to the abnormal growth of skin cells, usually on the areas exposed to sunlight.
- There are three major types of skin cancer
  - basal cell carcinoma, squamous cell carcinoma, and melanoma.

- **Melanoma is:**
  - Is the least common skin cancer.
  - Responsible for 75% of skin cancer deaths
  - if caught early, it can be cured with minor surgery

Detecting Melanoma

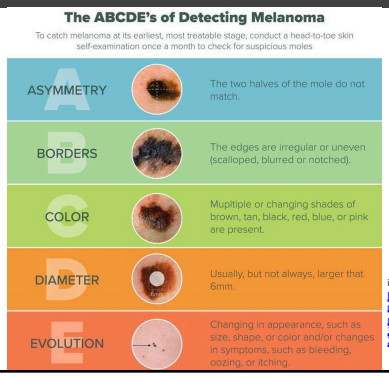


image from: https://www.facebook.com/SchifferHealthCenter/photos/a.10558673945392/2591280610251/?type=1&theater

Melanoma & Data Science?

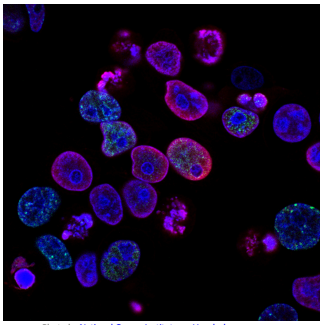


Photo by National Cancer Institute on Unsplash

Melanoma & Data Science?

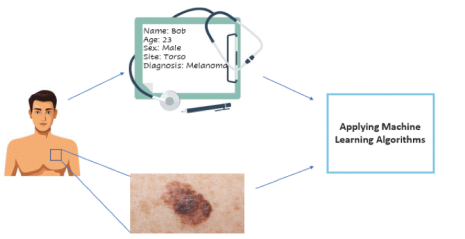


Image by https://towardsdatascience.com/machine-learning-in-healthcare-detecting-melanoma-70147e1b08de

Melanoma & Data Science?

Photo by National Cancer Institute on Unsplash

Sampling

Objectives for this Session

- Creating & interpreting **sampling distributions**
- Understanding the effects of **randomness**
- Discussion the "**law of large numbers**" and the "**central limit theorem**"
- Review **inferential & descriptive** statistics
- Standard deviation vs. standard error
- Demonstration of **R functions** - sampling, replication, and visualization via histogram

Population and Sample

- A population is the entirety of a phenomenon that we hope to study.
- Populations have a conceptual definition but are "unreachable" as a whole
  - ➢ All iPhones in use globally
    → impossible to study as a complete set
  - ➢ All Guinness beer – too expensive, nothing left

Population and Sample

- Use "Sampling" to access populations
  - Sampling is any systematic process for selecting cases from a population.
  - Sampling is science and art.
- Statisticians have established the math for sampling
- As data scientists, it is up to us to make sound and thoughtful decisions on how to obtain our data.

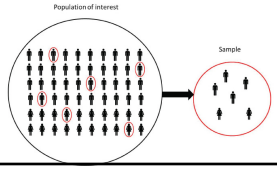
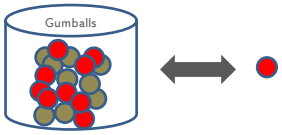


Image credit: NCES

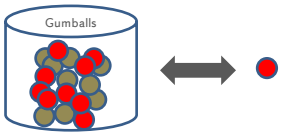
Sampling Process



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.

Sampling Process (continued)



Sampling Balls

- Same ratio of each in a jar
- Draw a sample (one draw)
- Red or blue ball with a draw → Really don't know.
  - Forces of "randomness" driving uncertainty
  - "Long run test" → multiple draws

Using rbinom()

- R has the capability to generate random numbers in a variety of configurations. Each configuration of random numbers reflects a specific underlying **distribution**.
- When working with events that have two states, such as heads and tails, the best distribution to use is the **binomial distribution**.
- In R, the rbinom() command generates random trials of the binomial distribution.
- Try this command:

```
rbinom(n=8, size=1, prob=0.5)
[1] 1 1 0 1 1 1 1 1
```

Label Results with factor()

factor(rbinom(n=8, size=1, prob=0.5), labels=c("red","blue"))

```
> set.seed(10)
>
> rbinom(n=8,size=1,prob=0.5)
[1] 1 0 0 1 0 0 0 0
>
> factor(rbinom(n=8,size=1,prob=0.5),labels=c("red","blue"))
[1] blue red  blue blue red  blue red  red
Levels: red blue
```

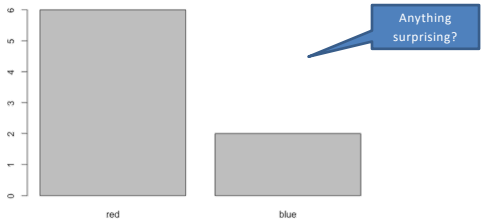
Label Results with factor()

table(factor(rbinom(n=8, size=1, prob=0.5), labels=c("red","blue")))

```
> table(
   factor(rbinom(n=8,size=1,prob=0.5),
          labels=c("red","blue"))
red blue
   4    4
```

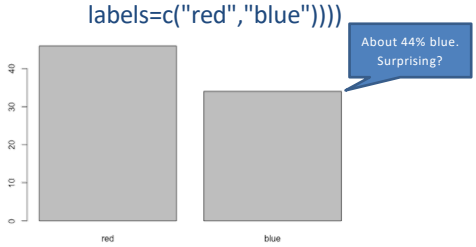
Tabulate and Plot

barplot(table(factor(
rbinom(n=8,size=1,prob=0.5),
labels=c("red","blue"))))



More Trials

```
barplot(table(factor(
        rbinom(n=80,size=1,prob=0.5),
                labels=c("red","blue"))))
```



States Populations

```
stateDF <- data.frame(state=state.name,
                      population=state.x77[,1])

realMean <- mean(stateDF$population)
realMean
[1] 4246.42
```

Samples States

- Use R to draw samples

  sample(stateDF$population, size=6, replace=TRUE)

  [1] 3100 2816 11197 590 2284 2212

- Use R to calculate mean of the sample

  mean(sample(stateDF$population, size=6, replace=TRUE))

  3090.667

Sampling Questions

- Why is sampling from a population an important process?
- What are some critical things to think about when obtaining a sample of data?

Replicating Samples

What is Replication?

- To understand the limitations of one sample we need to understand what happens over time (i.e., over many samples)
- *Doing the sampling process multiple times*

```
replicate(4,mean(sample(stateDF$population,
          size=16,replace=TRUE)),simplify=TRUE)
[1] 3859.250  3701.875  2854.375  3272.375
```

Mean of Means

- **mean(replicate(4,**
  mean(sample(stateDF$population, size=16,
  replace=TRUE)), simplify=TRUE))

  4095.422

- Interpretation
  - Draw 4 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 4 sample means.

- Calculated mean of means is off by 260

  4,246 (mean of 50 states) – 4,506 (mean of means) = -260

  -259 / 4246 = 6.12 % error

Mean of More Means

- **mean(replicate(400,**
  mean(sample(stateDF$population, size=16,
  replace=TRUE)), simplify=TRUE))

  4808.562

- 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 87.
  - 4,246 (mean of 51 states) – 4,159 (mean of means) = 87
  - 87 / 4,246 = 2% error

How Many Times to Replicate

- **mean(replicate(2000,**
  mean(sample(stateDF$population, size=16,
  replace=TRUE)), simplify=TRUE))
  - 4242

  4,246 (mean of 51 states) – 4,242 (mean of means) = 4

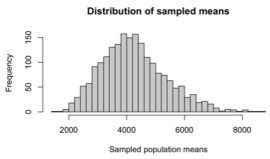
- Display distribution of 2000 means via a histogram as frequencies

  hist(replicate(2000,mean(sample(stateDF$population,
  size=16,replace=TRUE)),simplify=TRUE))

Histogram of Replication

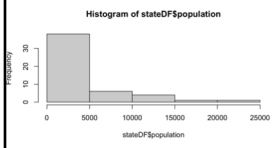
Features of this histogram:

- The data plotted consists of 2,000 means sampled from the census, where each sample is n=16 observations
- The mode (the highest bar) is right near 4000, which also happens to be near the mean of the original long-tailed distribution of the population
- The distribution is forming a symmetric bell shape

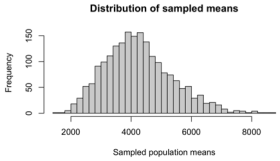



```
hist(replicate(2000,
     mean(sample(stateDF$population,
          size=16,replace=TRUE)),simplify=TRUE),
     breaks=30,
     main ="Distribution of sampled means",
     xlab="Sampled population means")
```

Important Distinction to Remember



The mean of the raw data is about 4000 (a long-tailed distribution)



The mean of the 2,000 means is also about 4000 (2,000 samples of n=16)

What makes the distribution shapes so different?

13

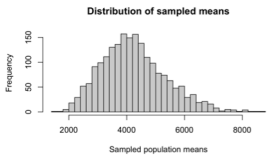
The Law of Large Numbers

- 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 the mean of values sampled 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.



Two Key Concepts

- 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
  - For sample means, and taking into account the "law of large numbers," the distribution of sampling means resembles a bell-shaped or normal distribution
  - The center of that distribution, the mean of the sample means, gets close to the population mean.

Percentile & Quantile

- **Percentile and Quartile**:
  - These terms mark a position within a collection of values. (ex. median is the 50th percentile and also the 2nd quartile)
- **A more general term is Quantile**:
  - Quantile is like percentile but expressed as a decimal
  - The median is the 0.50 quantile.
- **quantile command in R:**

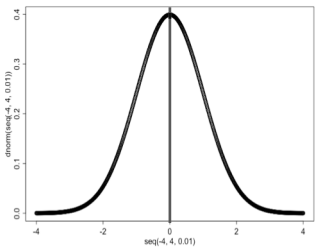
```
quantile(0:200,probs=c(0.25,0.75))
25% 75%
50  150
```

→ looks for the 0.25 and 0.75 quantiles within the digits 0 through 200

Marking the Curve

- We can use the abline() with qnorm() to draw a line at any quantile on the normal curve.

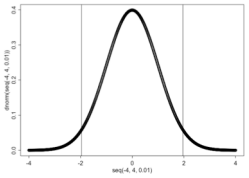
abline(v=qnorm(0.50))



Putting It All Together

- This diagram plots:
  - a normal curve
  - 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))
```

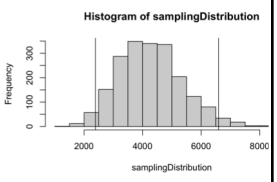


- So, 95% of the area under the curve is in the central region

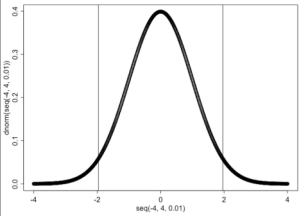
Now With Real Data!

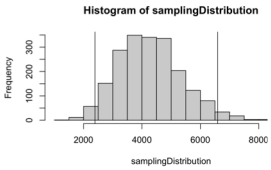
Resample the state data, plot it, and mark off the 0.025 and 0.975 quantiles (leaving 95% of the area under the curve in the central region)

```
samplingDistribution <-
      replicate(2000,
            mean(sample(stateDF$population,
                  size=16,replace=TRUE)))
hist(samplingDistribution)
abline(v=quantile(samplingDistribution,0.025))
abline(v=quantile(samplingDistribution,0.975))
```



Comparing the Ideal With the Reality

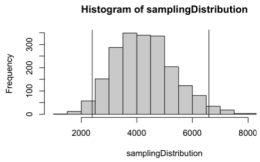




Reasoning About Distributions

Now we have a view of a sampling process that comprises 2,000 means of samples drawn from our state data

1. In line with the central limit theorem:
   - the center of the distribution is ~4,000 —the same as the mean of the underlying raw population data
   - It is also a pretty normal looking curve.
2. We now know a lot about common and unusual sampling results:
   - 95% of all sample means of state data fall between ~2,4000 and ~6,6000.
   - Any samples more extreme are rare.




```
quantile(samplingDistribution,
                probs=c(0.025, 0.975))
  2.5%      97.5%
2391.427.  6591.987
```

Mystery Sample

School of Information Studies
SYRACUSE UNIVERSITY
ischool.syr.edu

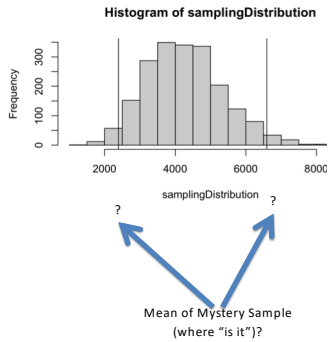
Exploring the Mystery Sample

- MysterySample:

```
> MysterySample <- c(333, 1088, 3699,
      3805,  3135, 3390, 99, 1049,
      235, 129, 4299, 3649, 2295,
      4663)
> mean(MysterySample)
[1] 2276
```

Is MysterySample a sample of states (from a simulated population) or something else?

Where is the MysterySample?



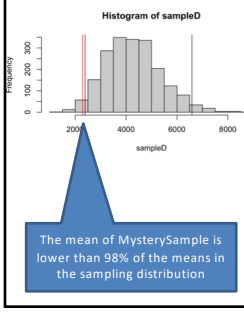
Comparing the Mystery Sample

- Is MysterySample a sample of states (from a simulated population) or something else?
- Compare MysterySample mean to our samplingDistribution "quantile analysis":

```
mean(samplingDistribution)  4,274
mean(MysterySample)         2,276   ← much smaller

quantile(samplingDistribution, probs=c(0.025, 0.975))
     2.5%      97.5%
2391.427   6591.987
```

Comparing Graphically




```
sampleD <- samplingDistribution

howMany <-
  sum(mean(MysterySample) > sampleD)

MysteryQuantileLocation <-
        howMany/length(sampleD)

hist(sampleD)
abline(v=quantile(sampleD,0.025))
abline(v=quantile(sampleD,0.975))
abline(v=quantile(sampleD,
         MysteryQuantileLocation),
         col="red")

MysteryQuantileLocation
[1] 0.0165
```

Analyzing the Results

- MysterySample:
  → *sample of U.S. states or something else?*
  - 98% of all the SampleMeans are higher than 2,276
  - Therefore, MysterySample mean (of 2,276) would be a rare event.
  - We can infer, based on solid statistical evidence (with greater than 98% confidence), that the MysterySample is not a sample from the simulated population 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.

Statistical Inference in a Nutshell

- Here's the basis for most all statistical inference
  - Construct a comparison distribution.
  - Identify a central zone containing 95% of the distribution: everything outside of this is considered an extreme value.
  - Compare a summary (e.g., mean) of a new sample of data to the distribution.
  - If new sample falls in the "extreme zone," tentatively conclude that the new sample was obtained from some other source than the comparison distribution.

Question

Describe an example of a situation where you would apply statistical inference in order to compare a sample mean to some benchmark value.