BUAD621: Representing Uncertainty

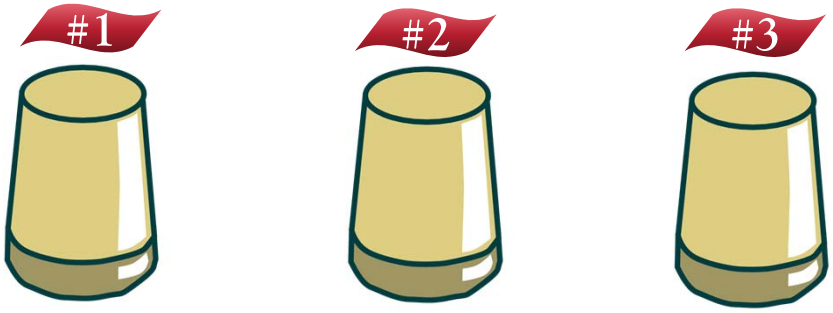
Intro to RV's and Probability Calibration

UNIVERSITY OF DELAWARE

Complete Uncertainty (i.e. ignorance) is Unsettling

Would you pay $2 for the right to choose a cup and receive its contents?

If so, which cup will you choose?



CERTAINTY IS EASIER

Would you pay $2 for the right to choose a cup and receive its contents?

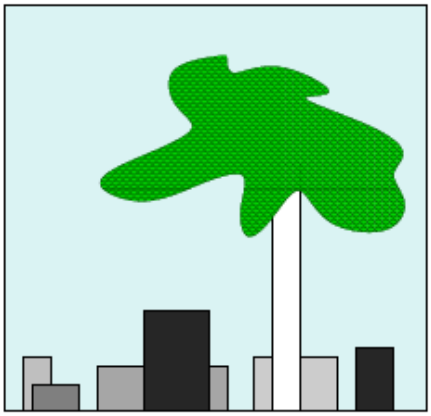
If so, which cup will you choose?

Real-Life is Between Certainty And Ignorance

Would you pay $2 for the right to choose a cup and receive its contents?

If so, which cup will you choose?

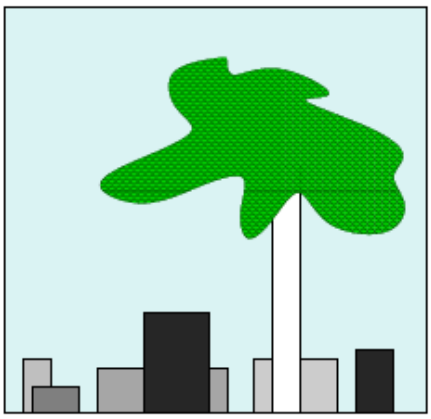
In Real-Life, We Combine Expectations And Data

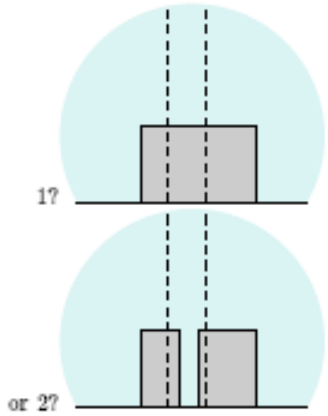


How many boxes are shown in this picture?

In Real-Life, We Combine Expectations And Data

Which is behind the tree?





Data Helps Us To Better Calibrate Our Uncertainty

Beliefs About The World + **Data** = **Updated Beliefs**

1. Beliefs that are consistent with the data become more plausible
2. Beliefs that are inconsistent with the data become less plausible

An Example

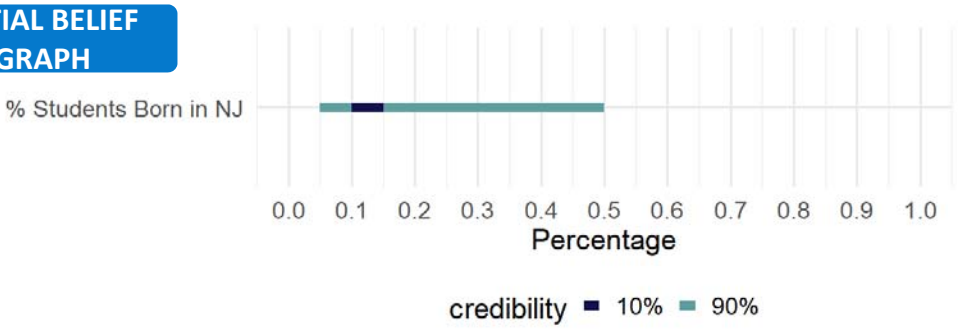
Beliefs About The World + **Data** = **Updated Beliefs**

CURRENT BELIEFS:
What is your belief about the percentage of students in this class born in New Jersey?

An Example

Beliefs About The World + **Data** = **Updated Beliefs**

INITIAL BELIEF GRAPH



An Example

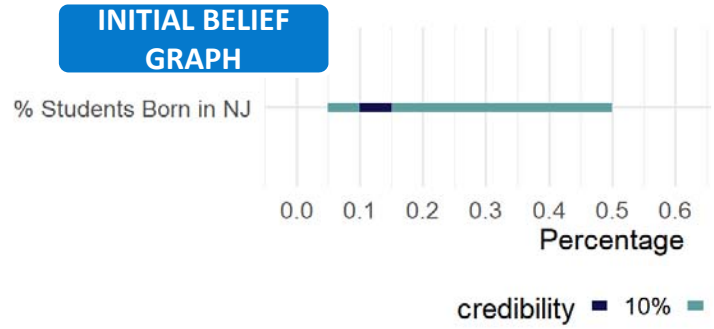
Beliefs About The World + **Data** = **Updated Beliefs**

DATA:
IF YOU ARE SITTING IN THE FIRST ROW OF THE CLASSROOM,
raise your hand and keep it raised if you were born in New Jersey?

An Example

Beliefs About The World + Data = Updated Beliefs

INITIAL BELIEF GRAPH



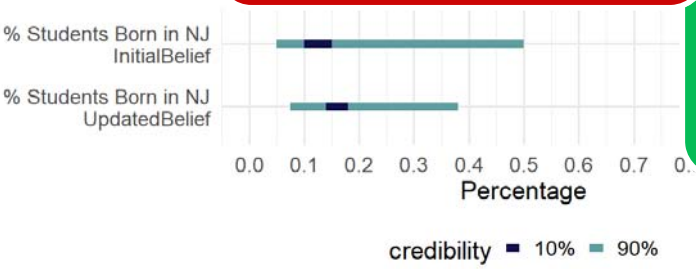
UPDATED BELIEFS: Interval to the left should get smaller than shown and reflect what you learned from the data.

UNIVERSITY OF DELAWARE

An Example

Beliefs About The World + Data = Updated Beliefs

DATA:
Assume 1 out of 6 polled students were born in NJ



UPDATED BELIEFS: What is your new belief about the percentage of students in this class born in New Jersey?

UNIVERSITY OF DELAWARE

Data Helps Us To Better Calibrate Our Uncertainty

1. Beliefs that are consistent with the data become more plausible
2. Beliefs that are inconsistent with the data become less plausible

Analytics is Simply Using Data to Inform What "Guesses" Seem to Be Better Than Others

All Data (i.e. measurements) Are Random Variables

Drives A Honda
X

$X \equiv Randomly\ Chosen\ Person\ In\ This\ Class\ Drives\ a\ Honda$

All Data (i.e. measurements) Are Random Variables

What values of X are possible?

Drives A Honda
X

$X \equiv Randomly\ Chosen\ Person\ In\ This\ Room\ Drives\ a\ Honda$

Random Variables Map Outcomes to Real Numbers

$X \equiv Randomly\ Chosen\ Person\ In\ This\ Class\ Drives\ a\ Honda$

What values of X are possible?

$$X \equiv \begin{cases} 0, & if\ person\ does\ not\ drive\ a\ Honda \\ 1, & if\ person\ does\ drive\ a\ Honda \end{cases}$$

Probability Describes Uncertainty in Random Variables

$X \equiv Randomly\ Chosen\ Person\ In\ This\ Class\ Drives\ a\ Honda$

$$X \equiv \begin{cases} 0, & if\ person\ does\ not\ drive\ a\ Honda \\ 1, & if\ person\ does\ drive\ a\ Honda \end{cases}$$

Outcome	Realization (x)	$P(X = x)$
No Honda	0	80%
Honda	1	20%

A Probability Distribution Tells Us Everything

Outcome	Realization (x)	$P(X = x)$
No Honda	0	80%
Honda	1	20%

What values of X are possible?

What is the probability someone drives a Honda?

What is the probability $X = 0$?

What is the probability $X = 2$?

What is the probability $X \leq 2$?

Some Probability Distributions Get Special Names

$$X \sim Bernoulli(\theta)$$

$$X \equiv \begin{cases} 0, & failure \\ 1, & success \end{cases}$$

Outcome	Realization (x)	$P(X = x)$
Failure	0	$1 - \theta$
Success	1	θ

A Probability Distribution Tells Us Everything (revisited)

$$X \sim Bernoulli(\theta = 60\%)$$

What values of X are possible?	What is the probability $X = 0$?	What is the probability $X = 0.6$?	What is the probability $X \leq 1$?

Same questions, just answer without knowing θ

$$X \sim Bernoulli(\theta)$$

What values of X are possible?	What is the probability $X = 0$?	What is the probability $X = 0.6$?	What is the probability $X \leq 1$?

Two Ways of Mathematically Specifying a Bernoulli Random Variable

$$X \sim Bernoulli(\theta)$$ Math Notation

Outcome	Realization (x)	$P(X = x)$
Failure	0	$1 - \theta$
Success	1	θ

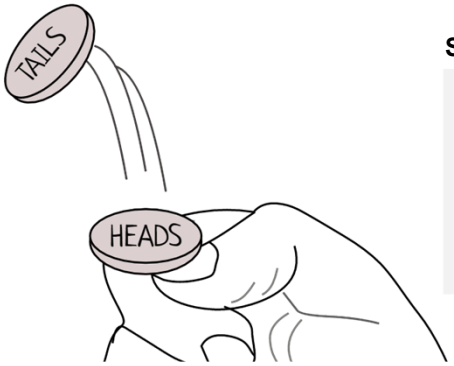
A Lookup Table

Examples of Bernoulli Random Variables

- Will the user click my ad?
- Will the drug lower a patient's cholesterol?
- Will the new store layout increase sales?
- Will the well yield oil?
- Will the customer pay back their loan?
- Will the passenger show up for their flight?
- Is this credit card transaction fraudulent?

The Real World Lacks Mathematical Precision

$$X \equiv Outcome\ of\ Coin\ Flip\ is\ Heads$$



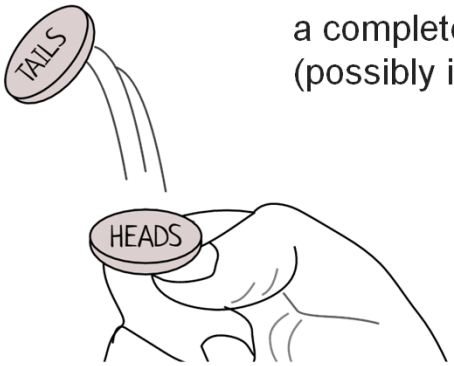
Simulating seven flips of a mathematically perfect coin

```
library(causact)
set.seed(123)
rbern(n=7,prob=0.5)
```

Based only these seven flips, what might you estimate for $P(X = 1)$?

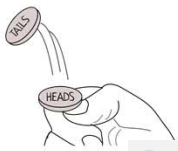
Representative Samples

A **representative sample** is an incomplete collection or subset of data that exhibits similarity to a complete collection of data from an entire (possibly infinite) population.



Do you think seven coin flips:
0 1 0 1 1 0 1
gives a representative sample?

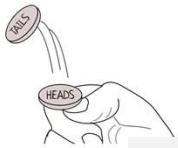
How many flips do you think you need to get a representative sample of coin flips?



$$X \equiv Outcome\ of\ Coin\ Flip\ is\ Heads$$

```
library(tidyverse)
set.seed(123)

numFlips = 7
df = data.frame(flipNum = 1:numFlips,
                coinFlip = rbern(n=numFlips,prob=0.5)) %>%
  mutate(headsProportion = cummean(coinFlip))

ggplot(df, aes(x = flipNum, y = headsProportion)) +
  geom_point() +
  geom_line() +
  geom_hline(yintercept = 0.5, color = "red")
```


How many flips do you [illegible] representative sample of coin flips?



$$X \equiv Outcome\ of$$

1) Open `numFlips.R`

2) Change the value of:

`numflips`

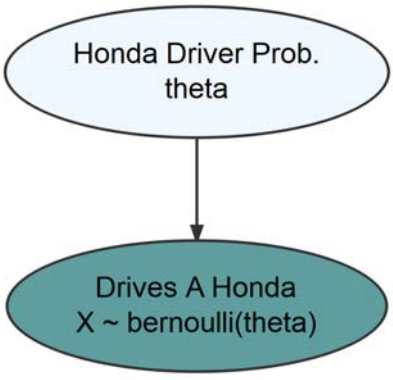
and see how many flips you need to get a representative sample?

```
library(tidyverse)
set.seed(123)

numFlips = 7
df = data.frame(flipNum = 1:numFlips,
                coinFlip = rbern(n=numFlips,prob=0.5)) %>%
  mutate(headsProportion = cummean(coinFlip))

ggplot(df, aes(x = flipNum, y = headsProportion)) +
  geom_point() +
  geom_line() +
  geom_hline(yintercept = 0.5, color = "red")
```

All Data (i.e. measurements) Are Random Variables

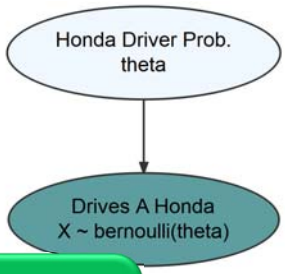
$X \equiv Randomly\ Chosen\ Person\ In\ This\ Class\ Drives\ a\ Honda$

Often Interested in Parameters of Probability Dist



$X \equiv Randomly\ Chosen\ Person\ In\ This\ Class\ Drives\ a\ Honda$

Do we know anything about θ?

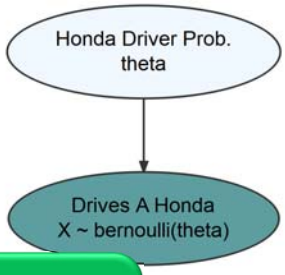


Beliefs About The World + **Data** = **Updated Beliefs**

$X \equiv Randomly\ Chosen\ Person\ In\ This\ Class\ Drives\ a\ Honda$

Do we know anything about θ?

What is your 90% confidence interval for θ? Give me a lower bound that has a 5% chance that the true proportion is less than it and an upper bound with a 5% chance that the true proportion is greater than it?



Beliefs About The World + **Data** = **Updated Beliefs**

$X \equiv Randomly\ Chosen\ Person\ In\ This\ Class\ Drives\ a\ Honda$

Calibrating Your Beliefs

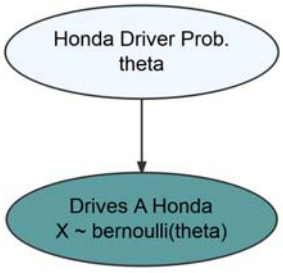
- Create a 90% confidence interval
  - Provide both an upper bound and a lower bound that you believe there is a 90% chance that the answer will be between your bound

1. You will win $1,000 if the true answer is between the numbers you gave for the upper and lower bound

2. You win $1,000 if you draw a red marble from a sack of marbles with 9 red marbles and 1 white marble. (i.e. there is a 90% chance you win $1,000).

Let's Get Calibrated

Poll for true value of θ



See CANVAS for calibration activity