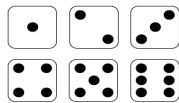



Probability, Random Variables, and Distributions

Probability Topics

- Basic Probability Questions
- Laws of chance
- An Application – DAWN Survey
- Random Variables
- An Application – Repair times

Roll a fair die



1. Roll once. What's the chance it lands  ?
 2. Roll twice. What's the chance the sum is 7?
 3. Roll twice. What's the chance the max is 5?
1. The die is equally likely to fall on any of its 6 sides: $1/6$.
 2. The 2 rolls can land in 36 possible ways (1,1), (1,2), ... (6,6). Of these 6 combinations sum to 7: $6/36 = 1/6$
 3. $9/36$

Formalize Rules of Probability

Ω = set of all possible outcomes from the chance process

A = a collection of outcomes, AKA an event

B = another collection of outcomes

- $P(\Omega) = 1$
- $0 \leq P(A) \leq 1$
- If A and B disjoint, then $P(A \text{ or } B) = P(A) + P(B)$

From these 3 Rules

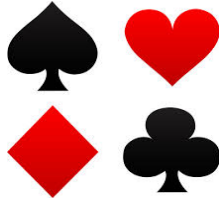
1. $P(\Omega) = 1$
 2. $0 \leq P(A) \leq 1$
 3. If A and B disjoint, then $P(A \text{ or } B) = P(A) + P(B)$
- If B is contained in A, then $P(B) \leq P(A)$
 - $P(A^c) = 1 - P(A)$
 - $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Toss Die 2 times

- $\Omega = 36$ ordered pairs $\{(1,1), \dots, (6,6)\}$
- A = sum is 7
 $= \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\}$
- B = max is 5
 $= \{(1,5), (2,5), (3,5), (4,5), (5,5), (5,4), (5,3), (5,2), (5,1)\}$
- Are A and B disjoint?
- $P(A \cup B) = P(\{(2,5), (5,2)\}) = 2/36$
- $P(A \cap B) = P(A) + P(B) - P(A \cup B)$
 $= 6/36 + 9/36 - 2/36 = 13/36$

Deal from a deck of playing cards

- 52 cards
- 26 red, 26 black
- 4 suits – diamonds, hearts, spades, clubs
- 13 in each suit – ace, 2, 3, 4, ..., 10, jack, queen, king

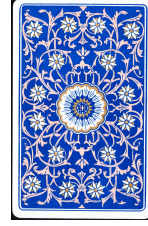
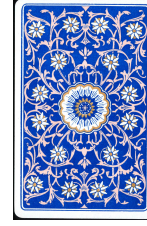


Deal 3 cards from a well-shuffled deck

1

2

3



A = 1st card
is the Ace of
diamonds

B = 2nd card is
a Diamond

C = 3rd card
is Red

Chance 1st card Ace of diamonds

- 52 cards
- Only 1 of them is the Ace of diamonds
- Deck is well-shuffled so all cards are equally likely
- $1/52$

Chance 2nd card is a diamond

- 52 cards
- 13 diamonds
- Deck is well-shuffled so all cards are equally likely
- The chance the 1st card is a diamond is $13/52$ or $1/4$
- By symmetry, the 2nd card should have the same chance of being a diamond as the 1st

Chance 3rd card is red

- 52 cards
- 26 red cards
- Deck is well-shuffled so all cards equally likely
- The chance the 1st card is red is $26/52$ or $1/2$
- By symmetry, the 3rd card should have the same chance of being red as the 1st

Deal 3 cards from a well-shuffled deck

1

2

3

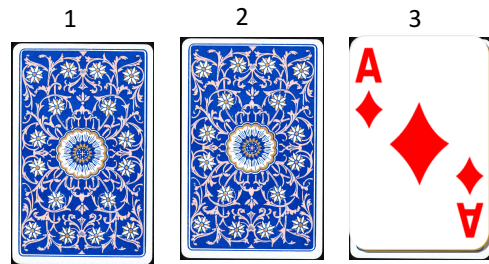


D = At least one of them is red?

Chance at least 1 red

- Chance they are all black?
- Chance the 1st card is black is $1/2$
- But we want the chance all three are black
- $26 * 25 * 24$ combinations of cards where all 3 are black
- $52 * 51 * 50$ combinations of 3 cards
- $(26*25*24)/(52*51*50)$ chance all black
- $1 - (26*25*24)/(52*51*50) =$ chance at least 1 red

Deal 3 cards from a well-shuffled deck



What's the chance the 1st card is red given the 3rd is the Ace of diamonds?

Conditional Probability

- $P(A|B)$ means the chance of A given B occurs
- $P(A|B) = P(A \text{ and } B)/P(B)$
- Independence of A and B means $P(A|B) = P(A)$
- $P(A \text{ and } B \text{ and } C) = P(A)P(B|A)P(C|A,B)$
- If independent, then

$$P(A \text{ and } B \text{ and } C) = P(A)P(B)P(C)$$

Chance all 3 cards are black

- Chance 1st and 2nd and 3rd cards are black
- Chance the 1st card is black is $26/52$
- Given the 1st card is black, the chance the 2nd is black is $25/51$ (51 cards left and 25 are black)
- Given the 1st and 2nd cards are black, the chance the 3rd is black is $24/50$
- ANS: $(26/52)*(25/51)*(24/50)$

Ways to solve probability problems

- Equally likely outcomes –

$$P(A) = \# \text{ outcomes in } A / \# \text{ outcomes in } \Omega$$
- Complement rule – $P(A) = 1 - P(A^c)$
- Symmetry – A and B are equivalent events so $P(A) = P(B)$
- Sequence of outcomes $P(A \text{ and } B \text{ and } C)$

Return to 3 cards from the deck

- According to the equally likely principle, what's the chance a particular combination of 3 cards occurs?
- $P\{\text{Ace, 2, 3 of diamonds as 1st, 2nd, 3rd cards}\}$

$$= 1/52 * 1/51 * 1/50$$

$$= 1/(52 * 51 * 50)$$
- $P\{\text{Ace, 2, and 3 of diamonds any order}\}$

$$= 6 / (52 * 51 * 50)$$

$$= \frac{1}{\binom{52}{3}}$$

$$\binom{n}{k} \text{ is } n!/k!(n-k)!$$

$$k! = k \times (k-1) \times \dots \times 1!$$

Simple Random Sample

- Like dealing cards from a well-shuffled deck
 - Deck with N cards, deal n
 - Each subset of n cards is equally likely
- $P(n \text{ particular individuals in sample}) =$

$$\frac{1}{\binom{N}{n}} = \frac{n!(N-n)!}{N!}$$

Simple Random Sample

The SRS is a chance process

- We can calculate the chance of an individual or group of individuals being selected
- A Statistic computed from the sample tends to be close to the population statistic
- We can use the sampling scheme to measure the typical error in our sample statistic

Drug Abuse Warning Network Survey

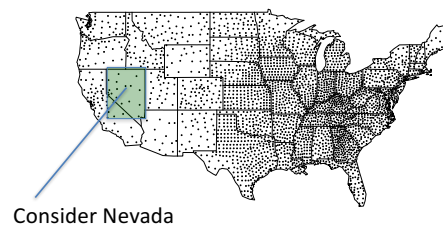


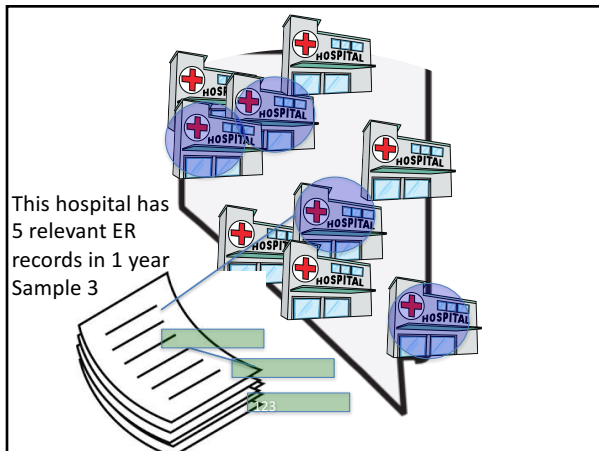
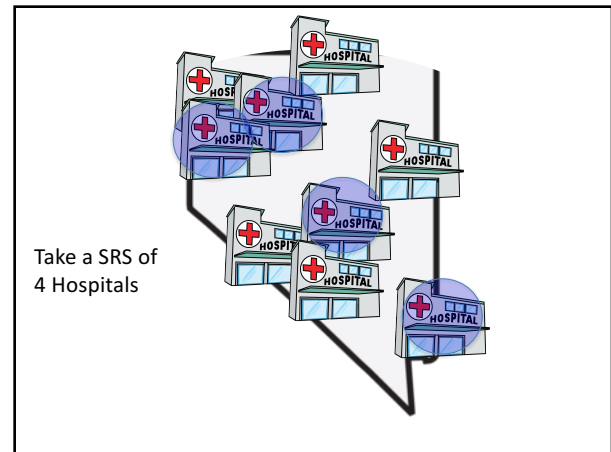
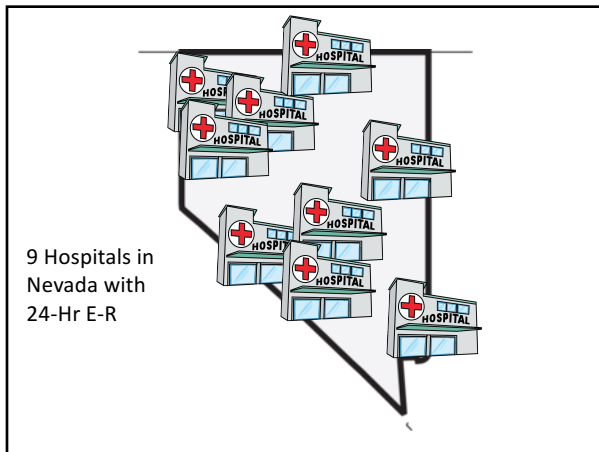
DAWN is a public health surveillance system for monitoring drug-related hospital emergency department visits in order to report on the impact of drug use, misuse, and abuse

Design: Divide the country into geographic regions (states)



In each state, Take a SRS of Hospitals with 24-hour ER





Probability Model for 2-Stage Cluster Sample

$P(\text{record 123 in ABC-ER in Las Vegas is in sample})$

$= P(\text{ABC-ER is in the Nevada sample}) \times$

$P(\text{record 123 in sample} | \text{ABC-ER in sample})$

$= 4/9 \times 3/5$

Probability Sample

- Is every substance-related ER record in the country equally likely to be in the sample?
- Is every hospital with a 24-HR ER equally likely to be in the 1st stage of the sample?
- Is every hospital in one geographic region equally likely to be in the region's sample?
- Is every ER-record in a hospital equally likely to be in the sample?

DAWN Design: 2-stage cluster

1. Divide US into geographic regions. Take a Simple Random Sample of hospitals with a 24-hr emergency room within region


Supplement with additional samples of hospitals from selected Metro areas, e.g. LA

2. Take a Simple Random sample of records from ER visits for substance-related reasons

Probability Sample

- Can we find the chance that ER record A from hospital X in geographic region G is in the sample?

$$\begin{aligned} P(\text{ER record in sample}) &= \\ &= P(\text{hospital } X) \times P(\text{ER } A \mid \text{hospital } X) \\ &= \frac{\# \text{ hospitals sampled from } G}{\# \text{ hospitals in } G} \times \frac{\# \text{ records sampled from } X}{\# \text{ records in } X} \end{aligned}$$



DAWN
DRUG ABUSE
WINNING NETWORK

0406 (Rev. 05/02-07/10) Name: _____
 12/10/2008

Emergency Department Case Report

U.S. Department of Health and Human Services • Substance Abuse and Mental Health Services Administration

1. Facility <input type="text"/> <input type="text"/> <input type="text"/>			
2. Date of Visit MONTH DAY YEAR 20 0 0	3. Time of Visit HOUR MINUTE : :	<input type="checkbox"/> a.m. <input type="checkbox"/> p.m. <input type="checkbox"/> military	4. Age <input type="text"/> <input type="text"/> <input type="checkbox"/> Less than 1 year <input type="checkbox"/> Not documented
5. Patient's Home ZIP Code <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> Otherwise, select one response: <input type="checkbox"/> No fixed address (e.g., homeless) <input type="checkbox"/> Institution (e.g., shelter/jail/hospital) <input type="checkbox"/> Outside U.S. <input type="checkbox"/> Not documented	6. Sex <input type="checkbox"/> Male <input type="checkbox"/> Female <input type="checkbox"/> Not documented	7. Race/Ethnicity Select one or more: <input type="checkbox"/> White <input type="checkbox"/> Black or African American <input type="checkbox"/> Hispanic or Latino <input type="checkbox"/> Asian <input type="checkbox"/> American Indian or Alaska Native <input type="checkbox"/> Native Hawaiian or Other Pacific Islander <input type="checkbox"/> Not documented	

8. Diagnosis List up to 4 diagnoses noted on the physician's chart. Do not list ICD codes.

1. 2.

3. 4.

9. Case Description Beginning with the presenting complaint, describe how the drug(s) was related to the ED visit. Copy verbatim from the patient's chart when possible.

10. Substance(s) Involved Using available documentation, list all substances that caused or contributed to the ED visit. Record substances as specifically as possible (i.e., brand [brand name preferred over generic name preferred over chemical name, etc.]). Do not record the same substance but two

Route of Administration

Ingest One

Mark if ☐ Inhalant ☐ Injection ☐ Intranasal ☐ Intravenous ☐ Other

Data – One record

12251082, 9426354082 3 4 1 2201141 2 865 105 1102005 1 2 1 2 0.000-7.000-
7.000-7.000001255 105 1142032 4 1 1 2 5.0 5.0100-7.0000-7.0000 -7 -7 -7 -7-7-7-
7.00-7.00-7.0000-7.0000-7.0000 -7 -7 -7 -7-7-7-7-7.00-7.00-7.0000-7.0000-7.0000 -7
-7 -7 -7-7-7-7-7.00-7.00-7.0000-7.0000-7.0000 -7 -7 -7-7-7-7-7.00-7.00-7.0000-
7.0000-7.0000 -7 -7 -7-7-7-7.00-7.00-7.0000-7.0000-7.0000-7.0000 -7 -7 -7-7-7-7-
7.00-7.00-7.0000-7.0000-7.0000 -7 -7 -7-7-7-7-7.00-7.00-7.0000-7.0000-7.0000 -7 -7
-7-7-7-7-7.00-7.00-7.0000-7.0000-7.0000 -7 -7 -7-7-7-7-7.00-7.00-7.0000-7.0000-
7.0000 -7 -7 -7-7-7-7-7.00-7.00-7.0000-7.0000-7.0000-7.0000-7.0000-7.00-7.00-7.00-
7.0000-7.0000-7.0000 -7 -7 -7-7-7-7-7.00-7.00-7.0000-7.0000-7.0000-7.0000 -7 -7 -7-7-
7.00-7.00-7.0000-7.0000-7.0000-7.0000 -7 -7 -7-7-7-7-7.00-7.00-7.0000-7.0000-7.0000
-7 -7 -7-7-7-7-7.00-7.00-7.0000-7.0000-7.0000 -7 -7 -7-7-7-7-7.00-7.00-7.00-7.00-
7.0000-7.0000 -7 -7 -7-7-7-7-7.00-7.00-7.0000-7.0000-7.0000-7.0000-7.0000-7.00-7.00-
7.00-7.00-7.0000-7.0000-7.0000 -7 -7 -7-7-7-7-7.00-7.00-7.0000-7.0000-7.0000-7.0000 -7
-7 -7-7-7-7-7.00-7.00-7.0000-7.0000-7.0000-7.0000-7.0000-7.0000-7.0000-7.0000 -7

Data – One record

[illegible]

1 year of ER visits
Representative sample

DAWN Codebook

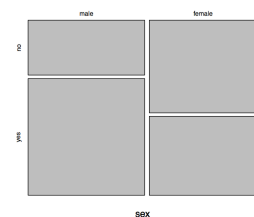
SEX	GENDER			
Location:	36-37 (width: 2; decimal: 0)			
Variable Type:	numeric			
Range of Missing Values (M):	-8			
Value	Label	Unweighted Frequency	%	Valid %
1	MALE:(1)	119111	52.0 %	52.0 %
2	FEMALE:(2)	110030	48.0 %	48.0 %
-8 (M)	NOT DOCUMENTED:(-8)	70	0.0 %	

Based upon 229141 valid cases out of 229211 total cases.

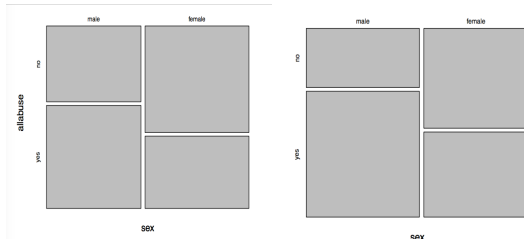
Missing Values
-7 not applicable
-8 undocumented
-9 missing

EDA

- Like most surveys, almost all variables are categorical
- Relationship between sex and whether the ER visit was solely for substance-reason
- Mosaic plot carves up the block according to proportions (conditional)



Account for the chance an ER Record appears in sample



How do the proportions of substance only records change for each gender?
How does the relationship change?

Random Variables and Probability Distributions

Random Variable

Numeric value for outcome
Use capital letter to denote the outcome, usually from the end of the alphabet

Example

U = number face up on roll of a fair die
 X = 1 if 1st card drawn is diamond, 0 otherwise

Probability Distribution Table

All possible outcomes of the random variable
Chance of each outcome

U = number face up on roll of a fair die

u	1	2	3	4	5	6
$P(u)$	1/6	1/6	1/6	1/6	1/6	1/6

We have all the information we need to compute the change of an event

Probability Distribution Table

U = number face up on roll of a fair die

u	1	2	3	4	5	6
$P(u)$	1/6	1/6	1/6	1/6	1/6	1/6

$$P(\text{odd}) = P(1, 3, 5) = 1/2$$

$$P(1 \text{ or } 6) = 1/3$$

Discrete Uniform Distribution

U = number face up on roll of a fair die

u	1	2	3	4	5	6
$P(u)$	1/6	1/6	1/6	1/6	1/6	1/6

This type of distribution is so common, we give it a name: discrete uniform

$$U \sim \text{discrete uniform}(k, m)$$

$$P(j) = 1/(m-k+1) \text{ for } k \leq j \leq m$$

$$= 0 \text{ otherwise}$$

Bernoulli Distribution

X = indicator for 1st card a diamond

x	0	1
$P(x)$	3/4	1/4

This probability distribution is so common it has a name too

Bernoulli(p)

$P(0) = 1-p$ and $P(1) = p$

Combining Random Variables

U_1 = number face up on 1st roll of a fair die

U_2 = number face up on 2nd roll of a fair die

$S_2 = U_1 + U_2$

s	2	3	4	5	6	7	8	9	10	11	12
$P(s)$						6/36					

Combining Random Variables

U_1 = number face up on 1st roll of a fair die

U_2 = number face up on 2nd roll of a fair die

$S_2 = U_1 + U_2$

s	2	3	4	5	6	7	8	9	10	11	12
$P(s)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Combining Random Variables

$S_2 = U_1 + U_2$

s	2	3	4	5	6	7	8	9	10	11	12
$P(s)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

$M = \max(U_1, U_2)$

m	1	2	3	4	5	6
$P(m)$				9/36		

Combining Random Variables

$S_2 = U_1 + U_2$

s	2	3	4	5	6	7	8	9	10	11	12
$P(s)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

$M = \max(U_1, U_2)$

m	1	2	3	4	5	6
$P(m)$	1/36	3/36	5/36	7/36	9/36	11/36

Probability Distribution

- Probability Distribution Table:

- All Possible Values/Outcomes
- Chance of each value/outcome

- Function or Rule for calculating the chance of each possible outcome

Bernoulli

Roll a die 3 times

$B_1 = 1^{\text{st}}$ roll a 1 or 6

$B_2 = 2^{\text{nd}}$ roll a 1 or 6

$B_3 = 3^{\text{rd}}$ roll a 1 or 6

Each is a Bernoulli(1/3)

They are independent of one another

$$P(B_1 = 1 \mid B_2 = 0) = P(B_1 = 1) = 1/3$$

$$\text{Sum: } S_3 = B_1 + B_2 + B_3$$

$$P(3) = P(111)$$

$$= (1/3)^3 = 1/27$$

$$P(2) = P(110 \text{ or } 101 \text{ or } 011) \quad \text{each has same chance}$$

$$= 3(1/3)^2 (2/3) = 6/27 \quad \text{disjoint}$$

$$P(1) = P(100 \text{ or } 010 \text{ or } 001)$$

$$= 3 (1/3) (2/3)^2 = 12/27$$

$$P(0) = P(000)$$

$$= (2/3)^3 = 8/27$$

General Formulation

$$S_n = B_1 + B_2 + \dots + B_n$$

where $B_i \sim \text{Bernoulli}(p)$ independent

$$P(k) = \text{number of ways to get } k \text{ 1s} \times p^k (1-p)^{n-k}$$

For $k = 0, 1, \dots, n$

This distribution is called the Binomial(n, p)

Binomial Distribution

- Many problems reduce to observing a particular characteristic (nominal or ordinal data)
- Of interest is how many of a particular type/level did we see in our data?
- A question that often arises is how likely is it that we saw so many (or so few) of that type?

Binomial Distribution

- n trials
- p chance of success on any one trial
- trials are independent
- What is the chance of k successes?

Binomial Distribution

- n independent trials, p chance of success

$$\mathbb{P}(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\binom{n}{k} \text{ is } n! / k!(n-k)!$$

$$k! = k \times (k-1) \times \dots \times 1!$$

Probability & Statistics

- In probability, we work with known distributions and study their properties
 - Sometimes the chance process is so complex we use simulation to help us study it

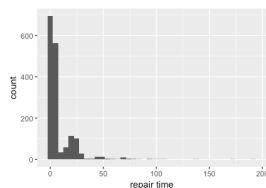
Probability & Statistics

- In statistics, we work with data that come from some distribution
 - Often we don't know the parameters of the distribution, we want to estimate them from the data, these estimates are random variables
 - Other times, we are unsure the distribution fits the data
 - We can use simulation to help us check the fit of a distribution and to understand the variability in our parameter estimates

Response Times for Network Service

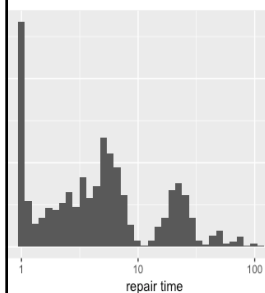
- NY Commission monitors response time for service repairs in the state
- The commission requests data on the repair times for a 3-month time period
- They are interested in average repair time

EDA



- Skewed right
- Some repairs took several days
- Possibly bimodal distribution – sharp peak around a few hours, smaller peak around 24 hours

EDA



- Now see a spike at 0, indicating many repairs are completed immediately
- Gap around 10, indicating that some repairs wait to the next day
- Large mode - Most repairs completed in a day

Average Repair Time

AVG = 9.4 hours

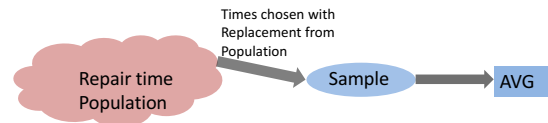
- Another time period will have a different average time
- Would like to give an interval estimate of the repair time
- The interval reflects the typical deviation in average repair time

Average Repair Time

- We expect the distribution of repair times will be similar to our sample
- We can use our sample as the population of repair times
- We generate new samples of repair times from this “population” and examine the variability in the average

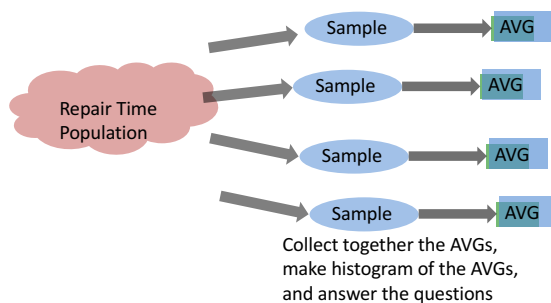
The Bootstrap

Probability Model for Repair Time



- Do we get the same average when we observe another time period?
- How does the average vary?
- What are the likely values?

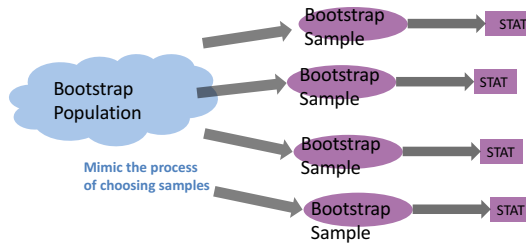
Ideally we would repeat the process many times



We don't know the Data Generator

- The sample should look the underlying population (data generator)
- Create a “Bootstrap Population” from the sample
- Mimic the process of generating samples – draws from the population

Imitate the Chance process with bootstrap population



Core Idea

- We generate bootstrap samples from the bootstrap population in a similar manner as the sample was generated from the population –
- To Do This We Need To Understand the Chance Process That Generated Our Data
- We compute the bootstrap statistic from the bootstrap sample, in the same way we compute the statistic from the sample

Core Idea

bootstrap sample relation to bootstrap population

≈ sample relation to the population

bootstrap statistic's distribution shape & variability

≈ statistic's distribution shape and variability

We often studentize our statistic θ before we bootstrap:

Bootstrap the studentized statistic : $\frac{\hat{\theta} - \theta}{SE(\hat{\theta})}$

Bootstrap version $\frac{\hat{\theta}^* - \theta^*}{SE(\hat{\theta}^*)}$

* Mean bootstrap version

Estimating the SE may require a second bootstrap

Then find the 2.5% and 97.5%tiles of the studentized distribution and use them to create an interval for θ

Example: Studentized AVG

θ is the population average

$\hat{\theta}$ is the sample average

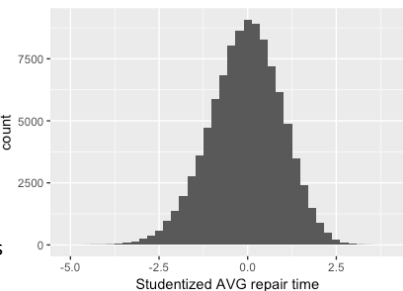
$SE(\hat{\theta})$ is the standard error the of sample average, which is $\approx SD(sample) / \sqrt{n}$

Bootstrapped Sampling Distribution of the studentized sample average

2.5% is -2.16
97.5% is 1.81

$9.4 - 2.16 * 0.36$
 $9.4 + 1.81 * 0.36$

Or
(8.6, 10.1) hours



Fun Probability Problems

- Andrew and Sam are playing a game with dice
- Each player stakes \$32 to play the game
- Andrew picks a number between 1 and 6
- Sam picks another number between 1 and 6
- One round of the game: roll a die until Sam or Andrew's number shows
- Who ever wins 3 rounds gets the \$64
- BUT – The game is interrupted at the point where Sam has won 2 rounds and Andrew 1. How should they divide the \$64?

- Three cards are marked as follows:
 - one card has a square on each side
 - One card has a circle on each side
 - One card has a square on one side and circle on the other
 - The cards are otherwise indistinguishable
- I shuffle the 3 cards behind my back and hold one up so that you see one side and I see the other side
- Give you see a circle, what's the chance that I also see a circle?