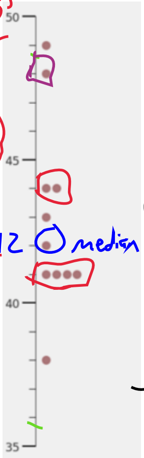
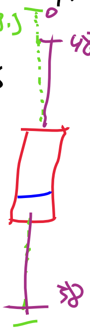
CSCI 3022-002 Intro to Data Science
Intro to Probability

"best" → max
→ min

Draw the box-whisker plot for the data to the left.



units →
6 pts
42 median
44 = Q_3
41 = Q_1
• 11 points



personal: • start w/ histogram (default ok!)

• box plots used for comparing data sets (hard to overlay histogram)

whisker length:

$$1.5(IQR)$$
$$= 1.5(44-41) = 9/2 = 4.5$$

Announcements and To-Dos

Announcements:

1. HW 1 Posted, due Friday!
2. Another nb day this Friday! (nb03 contains useful stuff for the HW: will post it after class *today*!)

↳ 1) cleaning: string → numeric

type( )

2) boxplot usages

Last time we learned:

1. Drawing pictures out of our data.

To do:

1. Start that HW! Ensure you can load the data and work with it. Practice your TeX/markdowns.

Overview: Probability

Many aspects of the world seem random and unpredictable.

1. Are we tall or short?
2. Do we have Mom's eyes or Dad's?
3. Is the hurricane going to hit Alabama?
4. Which team will win the NFC North?
5. How long until the Stampede bus shows up?
6. Which grocery store line should I get in?

One main objective of statistics/data science is to help make good decisions under conditions of uncertainty.

Overview: Probability

Many aspects of the world seem random and unpredictable.

1. Are we tall or short?
2. Do we have Mom's eyes or Dad's?
3. Is the hurricane going to hit Alabama?
4. Which team will win the NFC North?
5. How long until the Stampede bus shows up?
6. Which grocery store line should I get in?

One main objective of statistics/data science is to help make good decisions under conditions of uncertainty.

Overview: Definitions

Definition: *Set*
A *set* is a collection of objects. – finite, interval.

Definition: *Probabilistic Process*
A *probabilistic process* is system/experiment whose outcomes are uncertain.

Definition: *Outcome* → could be 20m! could be 2 weeks!
An *outcome* is a possible result of a probabilistic process .

Definition: *Sample Space*
A *sample space* (denoted Ω) of a probabilistic process is the set of all possible outcomes of that process.

Discrete vs. Continuous

Sets can contain many types of objects, both discrete and continuous. Our associated mathematics will shift accordingly.

$\Sigma \longleftrightarrow \int$

Discrete (Structures)

1. Math: summation, counting, sorting
2. Sets: times, ~~intervals~~ $\{$counting $\mathbb{N}$ or integers$\}$ $\{0, 1, 2, \ldots\ldots\}$
   rounded?

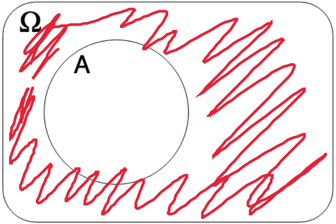
Continuous

1. Math: integrals, derivatives, smooth functions
2. Sets: times, intervals $[0, 1]$

Basic Set Operations

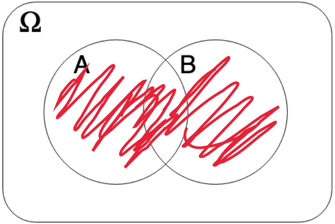
For sets A, B: in Sample space Ω

$A \subseteq \Omega$; $B \subseteq \Omega$

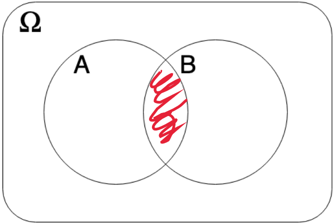


Complement;
A^C;
"Not"

$\bar{A}$

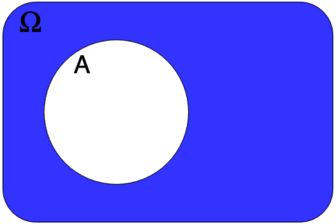


Union;
$A \cup B$;
inclusive "Or"

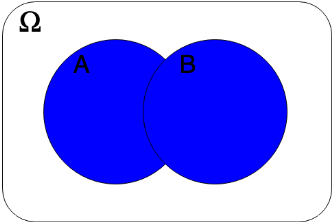


Intersection;
$A \cap B$;
"And"

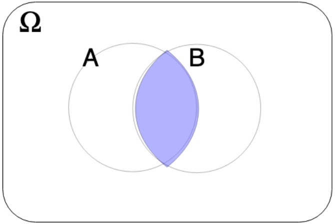
Basic Set Operations



Complement;
A^C;
"Not"



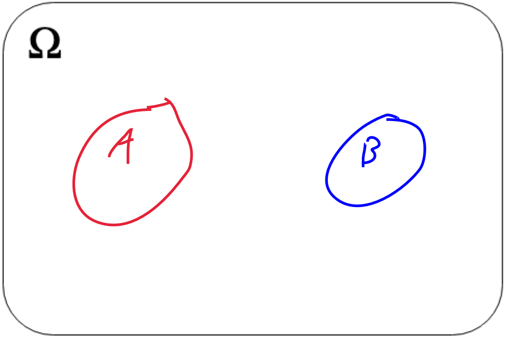
Union;
$A \cup B$;
"Or"



Intersection;
$A \cap B$;
"And"

Basic Set Definitions

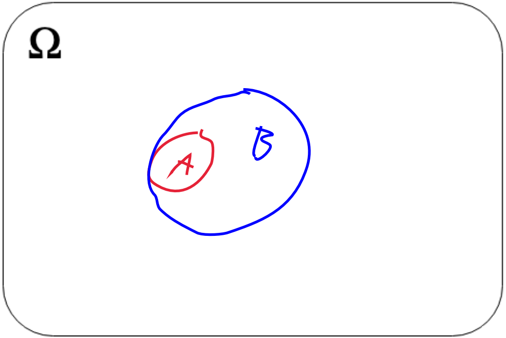
$A \subseteq B$



Mutually Exclusive

A & B don't overlap

$(A \cap B) = \emptyset$

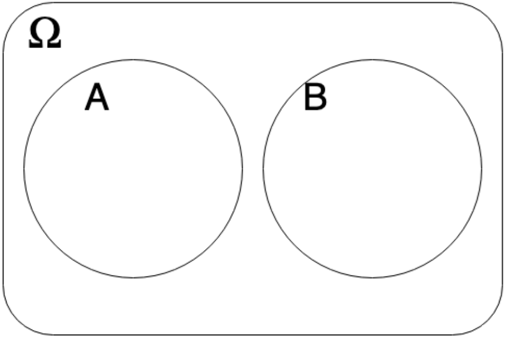


Subset

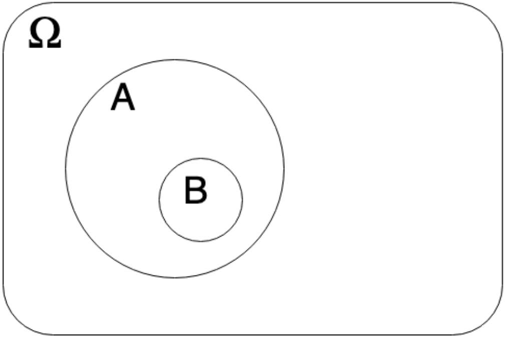
$A \subseteq B$

$A \subset B$ "strict"

Basic Set Definitions



Mutually Exclusive;
$(A \cap B) = \emptyset$;
"If A, not B; If B, not A."

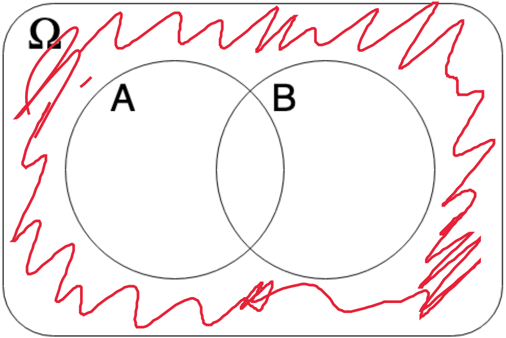


Subset;
$A \supseteq B$;
$A \supset B$;

1/100 of years we get 20" rain.

DeMorgan's Laws

$(A)^c$ & $\cup$ vs. $\cap$.



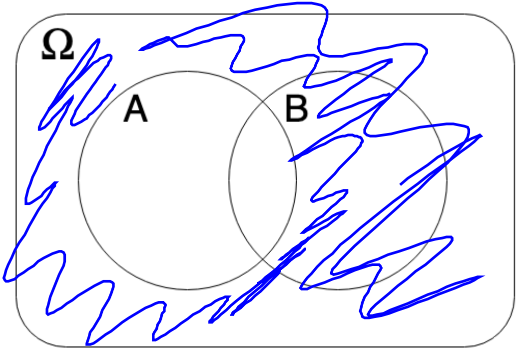
"neither A nor B"

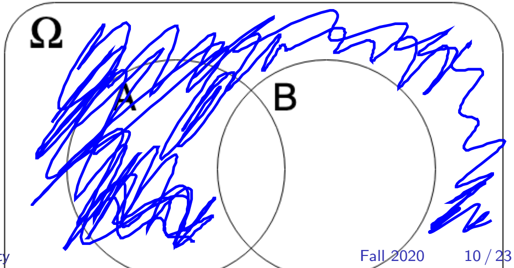
compared to:

"not A"

"not B"

not A AND also not B

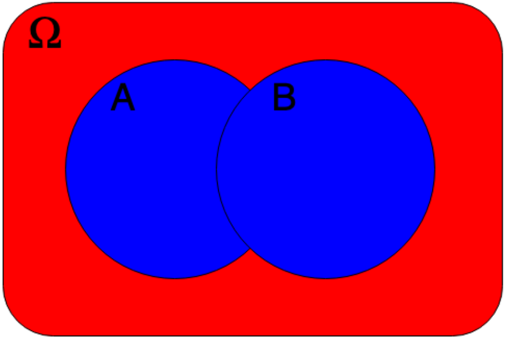




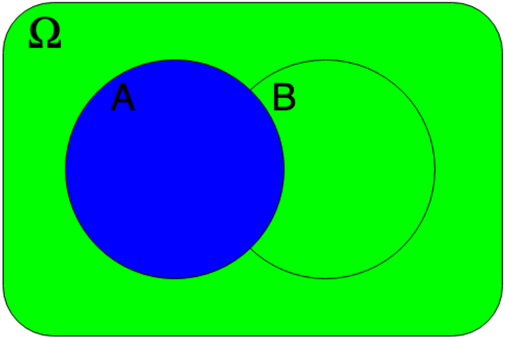
neither A nor B $(A \cup B)^c = A^c \cap B^c$ $\qquad \neg(A \vee B) \equiv \neg A \wedge \neg B$

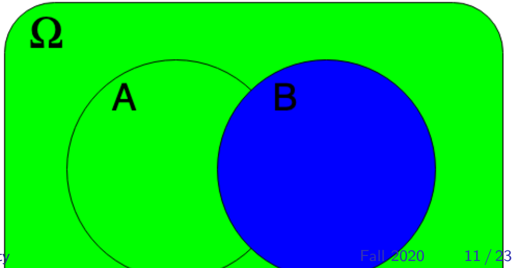
DeMorgan's Laws

not both $(A \cap B)^c = A^c \cup B^c$



"neither A nor B;"
$(A \cup B)^C =$
$A^C \cap B^C$;
"not A AND not B"





Some Sample Spaces

Describe sample spaces for:

{00, 01, 10, 11}

1. Tossing a coin twice = {HH, HT, TH, TT}

2. Selecting a card from a deck

52-card: spade

3. Measuring your commute time on a particular morning

buff bus bad! [1 hr, 7000 hr]

(0, 24 hr) ? continuous

Some Sample Spaces

Describe sample spaces for:

1. Tossing a coin twice
   $\{HH, HT, TH, TT\}$
2. Selecting a card from a deck
   $\{2\clubsuit, 2\spadesuit, 2\diamondsuit, 2\heartsuit, 3\clubsuit \ldots\}$
3. Measuring your commute time on a particular morning
   $\{t : t \in (0, T]\}$ where T is... infinity? The maximum reasonable time it *could* take?

Event

Definition: *Event*
An *event* is any collection (subset) of outcomes from the sample space.

{HT}

An event is simple if it consists of exactly one outcome and compound if it consists of more than one outcome.

Specific :

→ { {HT} and {TH} }

When an experiment is performed, an event A is said to *occur* if the resulting experimental outcome is contained in A.

Events

binary outcome!

flip 1 flip 2

Example: Suppose that we flip a coin 3 times.

Sample space: $\{HHH, \ldots\}$

HHH $\subseteq E_1$
HHT
HTH
1 HTT
THH
2 THT
3 TTH
TTT $\subseteq E_1$

3 flips; each 2 possibilities

$2 \cdot 2 \cdot 2 = 8$ possibilities

Some Possible Event(s):

E_1 : the event that we see the same flip all 3 times.

E_2 : the event that flip # 2 is heads.

What outcomes or elements(s) of Ω are in $E_1 \cap E_2$?

$\{HHH\}$ simple event

Events

Example: Suppose that we flip a coin 3 times.

Sample space:
$\{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$

Some Possible Event(s):
E_1 : the event that we see the same flip all 3 times.
E_2 : the event that flip $\#$ 2 is heads.

What outcomes or elements(s) of Ω are in $E_1 \cap E_2$? : just $\{HHH\}$

Probability Axioms

sets & events

Definition: *Probability*
Probability is a function that takes in sets (and later, we'll see, random variables) and outputs numbers according to the following rules:

1. *Non-negativity:* $P(A) \geq 0$
2. *Unity:* $P(\Omega) = 1$
3. σ*-additivity:* $P(A) + P(B) = P(A \cup B)$ if A & B are exclusive $(A \cap B = \emptyset)$

add mutual exclusive things

Probability Axioms

Definition: *Probability*
Probability is a function that takes in sets (and later, we'll see, random variables) and outputs numbers according to the following rules:

1. *Non-negativity:* For every $A \in \Omega$, $P(A) \geq 0$.
2. *Unity:* Given a sample space Ω, $P(\Omega) = 1$.
3. *σ-additivity:* If A and B are disjoint (mutually exclusive) sets, $P(A \cup B) = P(A) + P(B)$.

Probability Theorems

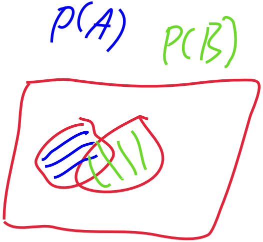
The axioms of probability give us a couple of important results.

$(A)\cup(A^c)=\Omega$; A & A^c are exclusive!

1. *Complementation:*

$$P(A)+P(A^c)=1$$

$$P(A)=1-P(A^c)$$



2. *Inclusion/Exclusion:* What is $P(A \cup B)$?

inclusive

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)$$

A not B
A and B

B not A
B and A

Probability Theorems

The axioms of probability give us a couple of important results.

1. *Complementation:* $P(A^C) = 1 - P(A)$.
   Proof:

2. *Inclusion/Exclusion:* What is $P(A \cup B)$? $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
   Proof:

Probability Theorems

The axioms of probability give us a couple of important results.

1. *Complementation:* $P(A^C) = 1 - P(A)$.
   Proof: From unity, $P(\Omega) = 1$, and $\Omega = A \cup A^C$, which are disjoint sets. So $P(\Omega) = P(A \cup A^C) = P(A) + P(A^C) = 1$.

2. *Inclusion/Exclusion:* What is $P(A \cup B)$? $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
   Proof: $A \cup B$ is "A or B," which can happen 3 disjoint ways:

   0.1 "A not B;" or $A \cap B^C$, with probability $P(A) - P(A \cap B)$;

   0.2 "B not A;" or $B \cap A^C$, with probability $P(B) - P(A \cap B)$;

   0.3 "both;" with probability $P(A \cap B)$;

   Summing these 3 probabilities gives the desired result.

Probability Theorems

The axioms of probability give us a couple of important results.

1. *Complementation:*

2. *Inclusion/Exclusion:* What is $P(A \cup B)$? $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
   Proof:

   This idea works for more than 2 sets!

Probabilities on Random Variables

$P(H) = .5$
$P(T) = .5$

Let $X = \#$ of heads in three tosses of a fair coin. X is a *random variable:* it maps events (a count of heads) into real numbers through probabilities.

What is the underlying probabilistic process?

flipping

What is the sample space? all 8 flip-orders

What are the possible values for X? $\{0, 1, 2, 3\}$

What is the probability that X is equal to 1: $P(X = 1)$?

Count! $\frac{|E|}{|\Omega|} = \frac{|\{X=1\}|}{|\Omega|} = \frac{|\{HTT, THT, TTH\}|}{|\Omega|} = \frac{3}{8}$

Probabilities on Random Variables

Let $X = \#$ of heads in three tosses of a fair coin. X is a *random variable:* it maps events (a count of heads) into real numbers through probabilities.

What is the underlying probabilistic process?
Flipping a fair coin.

What is the sample space?
The same 8 flip-outcomes as before.

What are the possible values for X?
$X \in \{0, 1, 2, 3\}$

What is the probability that X is equal to 1: $P(X = 1)$?
If all outcomes are equally likely in a set, we can arrive at this by counting the elements of Ω in X compared to all of Ω: or $\frac{|X|}{|\Omega|} = 3/8$

Probabilities on Random Variables

(binary process)

$P(H) \neq .5$

Why stop at fair coins? What if our coin is *unfair*, and comes up heads p proportion of the time, so $P(\{H, T\}) = \{p, q\}$? note: $p+q=1$ or $q=1-p$; $\{p, (1-p)\}$

What is the probability that I flip a biased coin twice and both flips come up heads?

Sample space for one flip: $\{H\}$ $\{T\}$

Sample space for both flips (a product of sample spaces!): $\{HH, HT, TH, TT\}$

Should the probability of the second flip change based on the result of the first?

Probabilities on Random Variables

Why stop at fair coins? What if our coin is *unfair*, and comes up heads p proportion of the time, so $P(\{H, T\}) = \{p, q\}$? Note: $q = 1 - p$.
What is the probability that I flip a biased coin twice and both flips come up heads?

Sample space for one flip: $\{H, T\}$

Sample space for both flips (a product of sample spaces!): $\{HH, HT, TH, TT\}$

Should the probability of the second flip change based on the result of the first?
Not usually: we call these *independent*... Not everything is independent!

Probabilities on Random Variables

Our coin is *unfair*, and comes up heads p proportion of the time. What is the probability that I flip a biased coin twice and both flips come up heads?

Sample space for one flip:

Sample space for both flips (a product of sample spaces!):

Should the probability of the second flip change based on the result of the first?

Probabilities on Random Variables

Our coin is *unfair*, and comes up heads p proportion of the time. What is the probability that I flip a biased coin twice and both flips come up heads?

Sample space for one flip: $\{H, T\}$

Sample space for both flips (a product of sample spaces!): $\{HH, HT, TH, TT\}$

Should the probability of the second flip change based on the result of the first?
Not usually: we call these *independent*... Not everything is independent! (**Idea:** two or more trials are *independent* if they don't affect each other)

Independence and Probabilities

Our coin is *unfair*, and comes up heads p proportion of the time. What is the probability that I flip a biased coin twice and both flips come up heads?

If two outcomes are independent, probabilities on their intersection ("and") becomes a product.

$$P(\{T_1 T_2\}) = (P(T))^2 = (1-p)^2$$

Result: What are $P(\{HH\})$ and $P(\{TT\})$?

$$P(\{H_1 \text{ AND } H_2\}) = P(H_1) \cdot P(H_2) = (P(H))^2 = (p)^2$$

If two outcomes are disjoint, probabilities on their union ("or") becomes a sum.
Result: What is $P(\{HT\}$ **OR** $\{HT\})$?

$$P(HT) + P(TH) = P(H) \cdot P(T) + P(T)P(H) = 2p(1-p)$$

Sanity check: did we just add up to 1?

all added: $(p^2 + 2p(1-p) + (1-p)^2) = (p + (1-p))^2$

Independence and Probabilities

Our coin is *unfair*, and comes up heads p proportion of the time. What is the probability that I flip a biased coin twice and both flips come up heads?

If two outcomes are independent, probabilities on their intersection ("and") becomes a product.

Result: What are $P(\{HH\})$ and $P(\{TT\})$?
A: $p \cdot p$ and $q \cdot q = (1-p)^2$

If two outcomes are disjoint, probabilities on their union ("or") becomes a sum.
Result: What is $P(\{HT\}$ **OR** $\{HT\})$?
A: $P(\{HT\} = pq$ PLUS $P(\{TH\} = qp$

Sanity check: did we just add up to 1?

Counting outcomes

Finally, what is the probability of I flip our biased coin ~~five~~ 3 times and get *exactly* one heads?

outcomes: $P(HTT) + P(THT) + P(TTH)$

and

$$\underline{P(H)\cdot P(T)\cdot P(T)} + \underline{P(T)\, P(H)\cdot P(T)} + \underline{P(T)\cdot P(T)\cdot P(H)}$$

$$= \boxed{3}\, \underbrace{P(H)\cdot (P(T))^2}$$

# of ways

mean prob of 1 outcome

Counting outcomes

Finally, what is the probability of I flip our biased coin five times and get *exactly* one heads?

This is the set of events $\{HTTTT, THTTT, TTHTT, TTTHT, TTTTH\}$.

Counting outcomes

Finally, what is the probability of I flip our biased coin five times and get *exactly* one heads?

This is the set of events $\{HTTTT, THTTT, TTHTT, TTTHT, TTTTH\}$.

Each is composed of 5 independent flips, so the probability of any one of these events is the product pq^4

Counting outcomes

Finally, what is the probability of I flip our biased coin five times and get *exactly* one heads?

This is the set of events $\{HTTTT, THTTT, TTHTT, TTTHT, TTTTH\}$.

Each is composed of 5 independent flips, so the probability of any one of these events is the product pq^4

Each outcome is disjoint/exclusive, so the full cumulative probability is the sum of 5 of these: $5pq^4$

Moving Forward

Suppose we have a coin and we don't know if it's biased... what could we do? (nb04, lecture next week to come!)

$$\frac{\sum(x_i-\bar{x})^2}{n-1}$$

Daily Recap

n - ddof

np.median(x)

sd(x, ddof = 1)

def mymedian(x)

sort(x)

Today we learned

1. A review of probability
2. Think about when we can use "all outcomes equally likely" and then just *count* those outcomes. This is a big part of *independence*.

if x is odd (x mod 2 == 1?)

return x[n/2]

if x even

$$\text{return } \frac{x[\frac{n}{2}] + x[\frac{n}{2}+1]}{2}$$

Moving forward:

- No class Monday for Labor Day.
- Friday: making some histograms, boxplots, and playing around with data frames: scrubbing data!

Next time in lecture:

- We probably talk even more about probability!