Probability

Overview

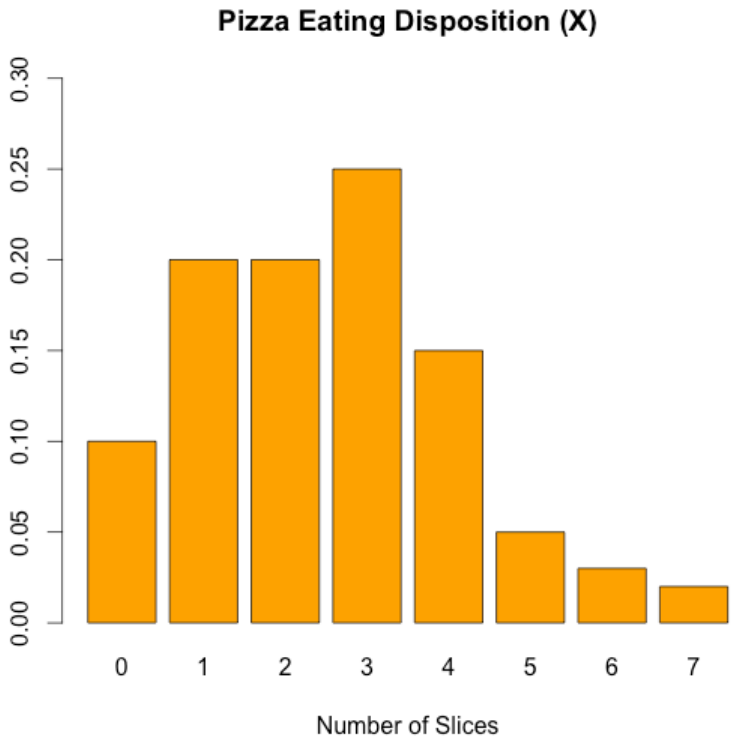
- Probability Theory
  - Sets
  - Laws: Bayes Theorem, Law of Total Probability, Chain Rule
  - Permutations, Combinations
  - PMF vs. PDF: E(X), Var(X), Cov(X,Y), Correlation(X,Y)
- Major Distributions
  - Bernoulli, Binomial, Poisson, Exponential, Uniform, Normal
- Joint, Marginal Distributions

X = # of pizza slices consumed

x	0	1	2	3	4	5	6	7
P(X = x)	0.1	0.2	0.2	0.25	0.15	0.05	0.03	0.02

X = # of pizza slices consumed

x	0	1	2	3	4	5	6	7
P(X = x)	0.1	0.2	0.2	0.25	0.15	0.05	0.03	0.02



Sums to 1!

X = # of pizza slices consumed

x	0	1	2	3	4	5	6	7
P(X = x)	0.1	0.2	0.2	0.25	0.15	0.05	0.03	0.02

$$E(X) = \sum_{x=0}^{7} x * P(X = x)$$

$$= 0 * 0.1 + 1 * 0.2 + 2 * 0.2 + 3 * 0.25 + ... + 7 * 0.02$$

$$= 2.52$$

$$Var(X) = E[(X - \mu)^2]$$

$$= \sum_{x=0}^{7} (x - \mu)^2 * P(X = x)$$

$$= (0 - 2.52)^2 * 0.1 + (1 - 2.52)^2 * 0.2 + ... + (7 - 2.52)^2 * 0.02$$

$$= 2.61$$

Suppose there are only two states of the stomach: "hungry" and "not hungry"

Y = # of pizza slices consumed, if hungry

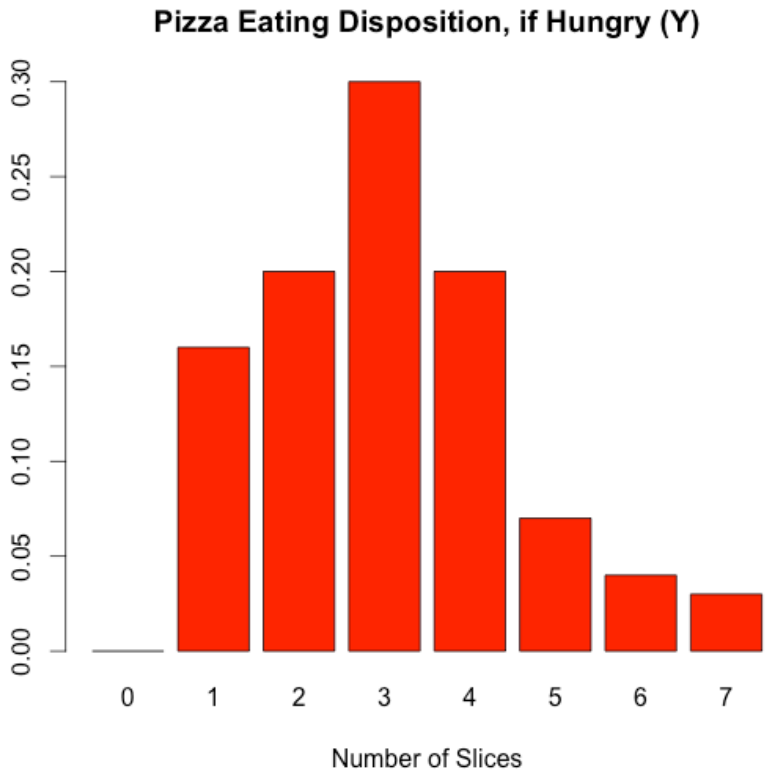
y	0	1	2	3	4	5	6	7
P(Y = y)	0	0.16	0.2	0.3	0.2	0.07	0.04	0.03

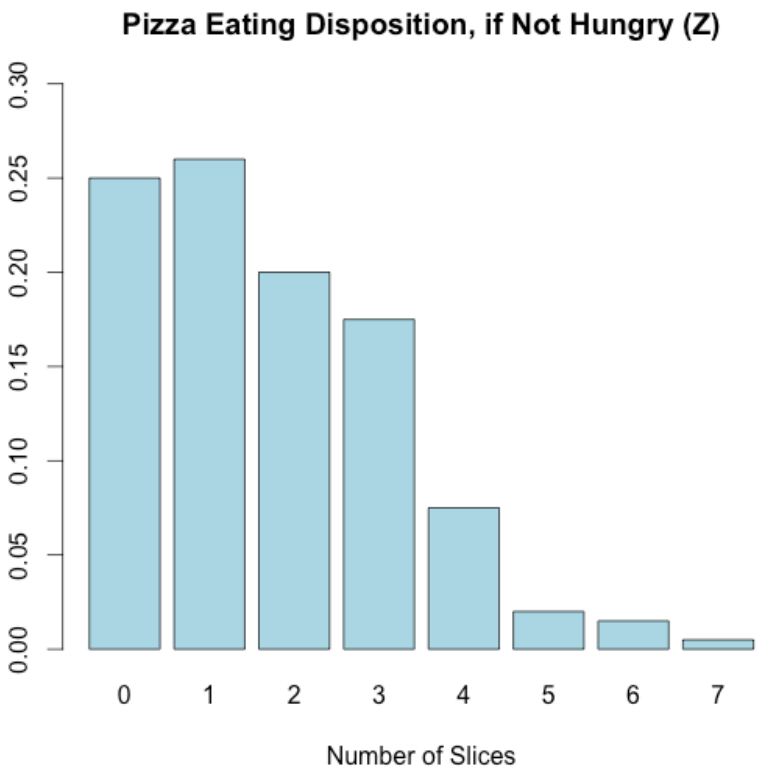
Z = # of pizza slices consumed, if not hungry

z	0	1	2	3	4	5	6	7
P(Z = z)	0.25	0.26	0.2	0.175	0.075	0.02	0.015	0.005

Y = # of pizza slices consumed, if hungry

Z = # of pizza slices consumed, if not hungry

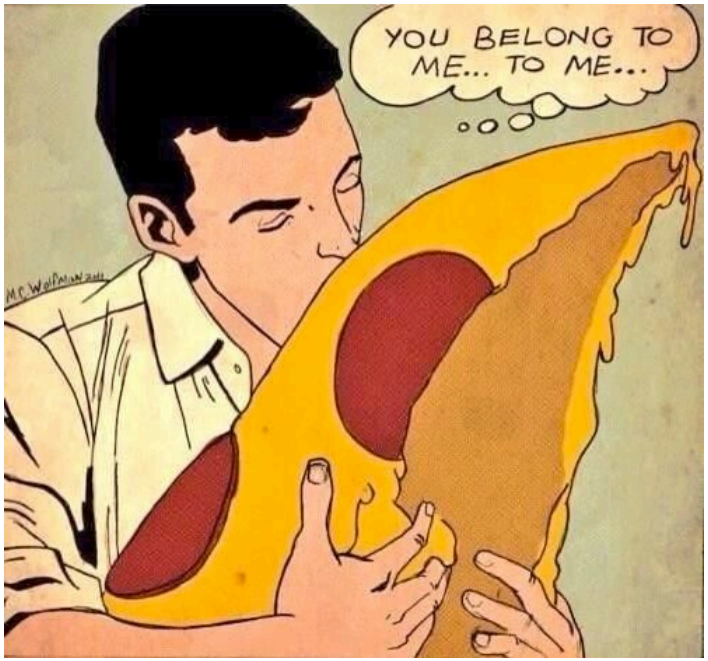




Only two states of the stomach:

"hungry" and "not hungry"

s	Hungry	Not Hungry
P(S = s)	0.6	0.4



X = # of pizza slices consumed

Y = X|S="hungry"

Z = X|S="not hungry"

X

S	0	1	2	3	4	5	6	7	
Hungry	0	0.096	0.12	0.18	0.12	0.042	0.024	0.018	0.6
Not Hungry	0.1	0.104	0.08	0.07	0.03	0.008	0.006	0.002	0.4
	0.1	0.2	0.2	0.25	0.15	0.05	0.03	0.02	1

X = # slices consumed

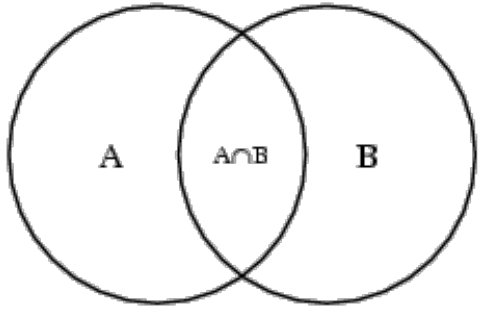
S

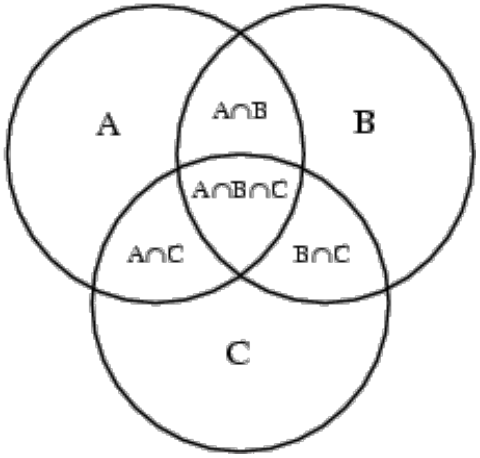
	0	1	2	3	4	5	6	7	
Hungry	0	0.096	0.12	0.18	0.12	0.042	0.024	0.018	0.6
Not Hungry	0.1	0.104	0.08	0.07	0.03	0.008	0.006	0.002	0.4
	0.1	0.2	0.2	0.25	0.15	0.05	0.03	0.02	1

Probability that…

- Bob eats 3 slices given that he is hungry?
- Bob was hungry given he ate 3 slices of pizza?
- Bob eats 4 slices of pizza when he's not hungry?

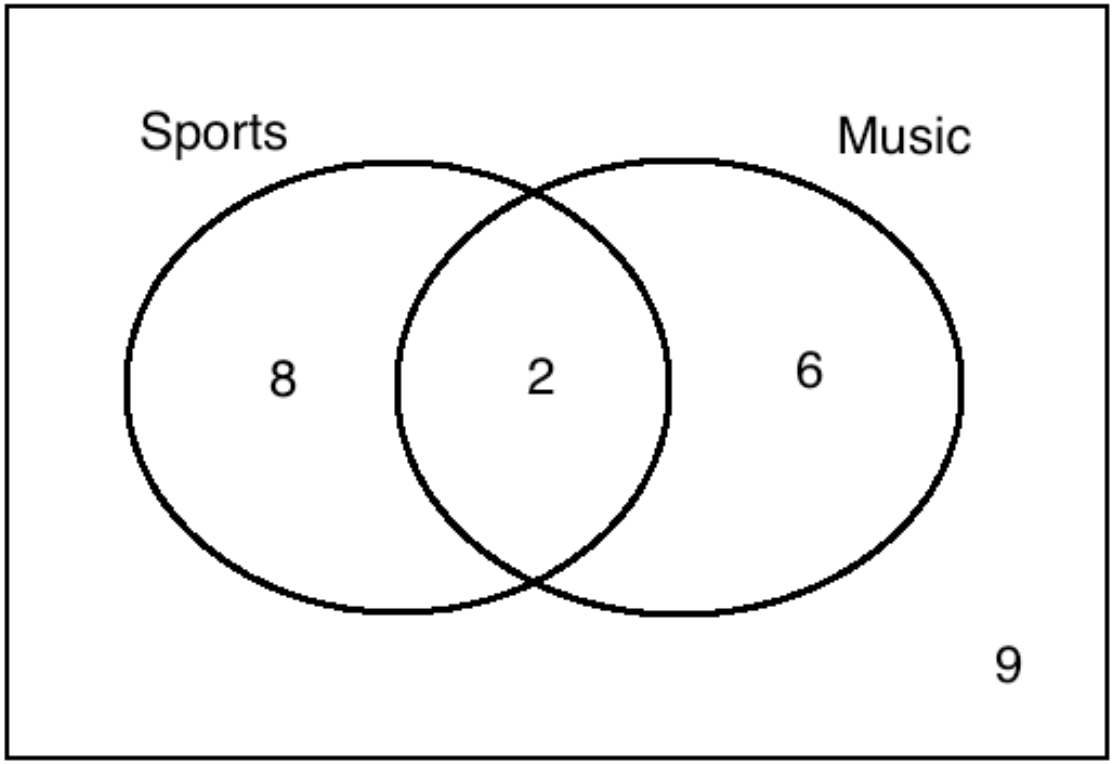
Sets



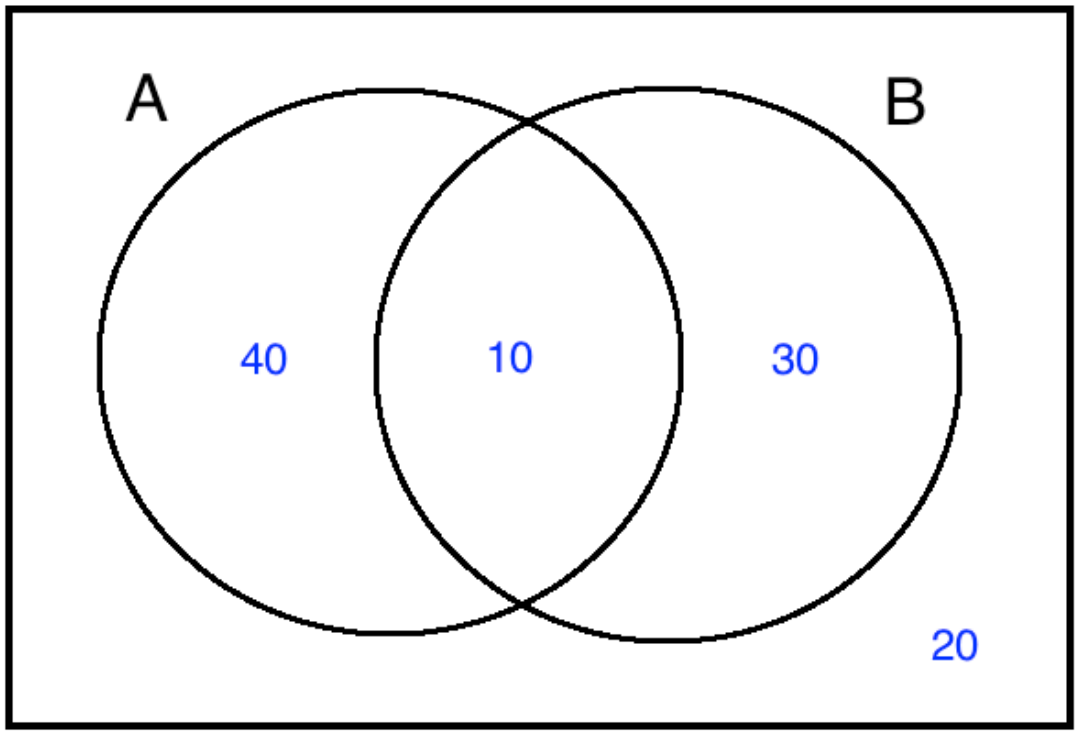


- Complement: A' denotes all elements of the universal set that are not in A
- Union: A U B denotes all elements that are in A or in B
- Intersection: A & B, also denoted A,B or A∩B denotes all elements in A and in B
  - Disjoint: Two sets are disjoint if their intersection is the null set

- In a class of 25 students, 10 are on a sports team, 8 are in a music program, 2 are both in sports and in music. How many students are in neither?



Just work inside out. 2 → 8, 6 → 9



Think about…

- A intersect B
- A | B
- A

Basic Probability

Suppose we're interested in the probability of **both** event A and B occurring

$$P(A \cap B) = P(A|B) * P(B)$$
$$= P(B|A) * P(A)$$

always!

$$P(A \cap B) = P(A) * P(B)$$

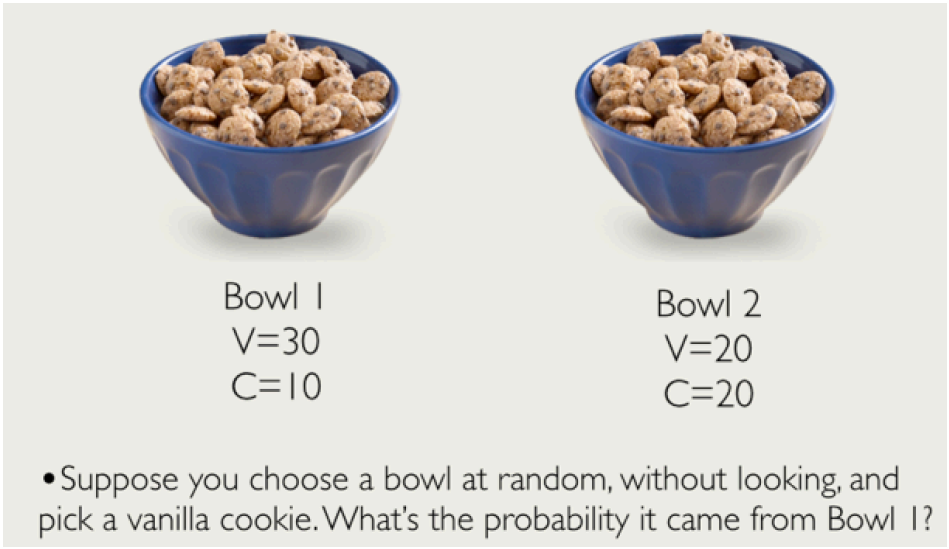
under independence

Cookies revisited…

$$P(bowl1 \cap vanilla)?$$

$$P(bowl1) = \frac{1}{2}$$

$$P(vanilla) = \frac{5}{8}$$



- There is a dependence between event bowl1 and event vanilla!

$$P(vanilla|bowl1) * P(bowl1) = \frac{3}{4} * \frac{1}{2} = \frac{3}{8}$$

- Also…

$$P(bowl1|vanilla) * P(vanilla) = ? * \frac{5}{8}$$

Bayes Theorem

Kind of hard…

Really easy!

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

The Cookie Problem

- Two bowls of cookies, Bowl 1 contains 30 vanilla and 10 chocolate cookies. Bowl 2 contains 20 of each.

Bowl 1
V=30
C=10

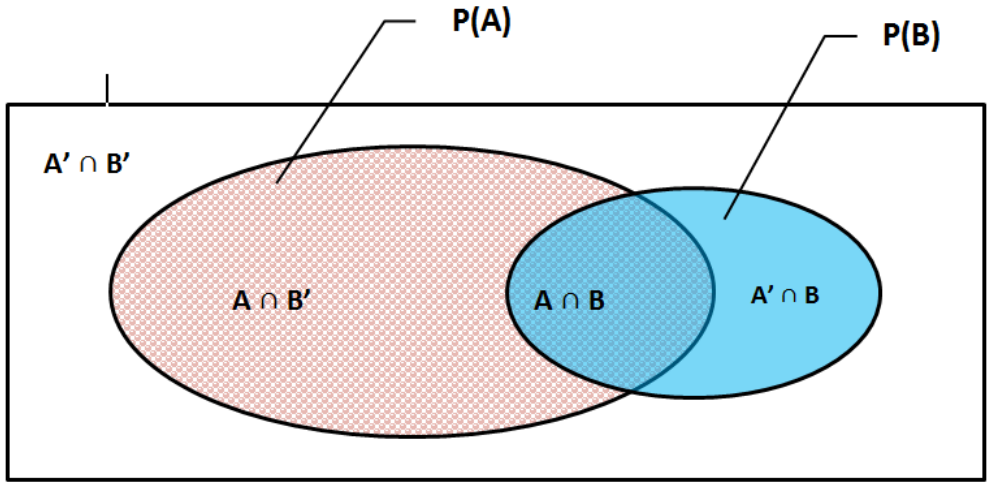
Bowl 2
V=20
C=20

- Suppose you choose a bowl at random, without looking, and pick a vanilla cookie. What's the probability it came from Bowl 1?

Remember the cookie problem?

- P(Bowl1 | Vanilla)? Not so obvious
- P(Vanilla | Bowl1)? Super easy

Bayes Theorem

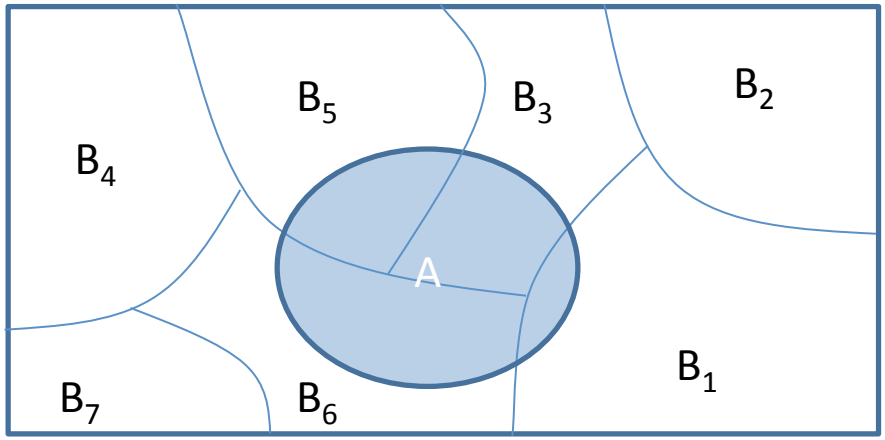


$$P(B|A) = \frac{P(A|B)*P(B)}{P(A)}$$

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)}$$

$$Recall: P(A \cap B) = P(B)*P(A|B) = P(A)*P(B|A)$$

Law of Total Probability



$$P(A) = \sum_i P(B_i \cap A) = \sum_i P(B_i) * P(A|B_i)$$

$$Recall: P(A \cap B) = P(B) * P(A|B) = P(A) * P(B|A)$$

Chain Rule

- Can write any joint distribution as incremental product of conditional distributions
- Particularly important for study of Bayesian Networks and Probability Graphical Models

$$P(X_1, X_2, X_3) = \underbrace{\underbrace{P(X_1) * P(X_2|X_1)}_{P(X_1, X_2)} * P(X_3|X_1, X_2)}_{P(X_1, X_2, X_3)}$$

- Not hard to see this extend for any number of variables…

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i|X_1, ..., X_{i-1})$$

Afternoon

Permutations & Combinations

- Permutations
  - # of **unique** ways to choose k out of n = n!/(n-k)!
  - Ex. 3 out of 10? 10!/7! = 10*9*8

- Combinations
  - # of ways to choose k out of n = n!/[(n-k)!*k!]
  - Ex. 3 out of 10? 10!/(7!*3!)
    - the 3! "removes" the duplicates

- 2 Uber and 3 Lyft cars are called at the same time. Arrival times are iid.
- What's the probability that the 2 Ubers arrive before the 3 Lyfts?

Q: 2 Uber and 3 Lyft cars are called at the same time. Arrival times are iid. What's the probability that the 2 Ubers arrive before the 3 Lyfts?

(i) $\underline{U}\ \underline{U}\ \_\ \_\ \_ = (2/5)*(1/4) = 1/10$

(ii) Permutations: $(2!*3!)/5! = 1/10$

- Think of each car as unique (U_1 U_2 L_1 L_2 L_3)
- Then there are 5! unique orderings
- Of those 5! unique orderings, 2!*3! result in 2 Uber before 3 Lyft

(iii) Combinations: $1/({}_5C_2) = 1/10$

- Don't think of cars as unique among types (U U L L L)
- Then there are 5 choose 2 total orderings, ${}_5C_2$
- Of those ${}_5C_2$ orderings, only 1 results in 2 Uber before 3 Lyft

Random Variables

A real valued function defined on a sample space, whose value is determined with probability

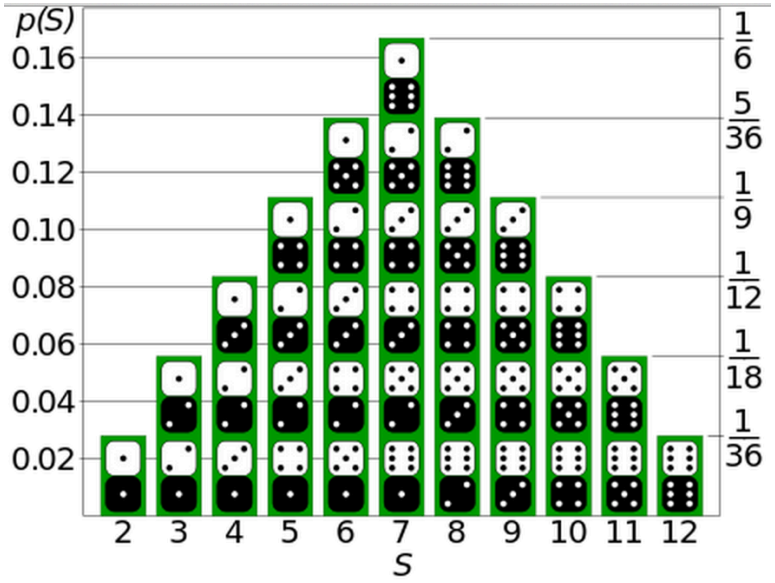
Random Variables

Discrete Case: P(X = i)

X = Sum of two rolled dice

P(X) => Probability Distribution of RV

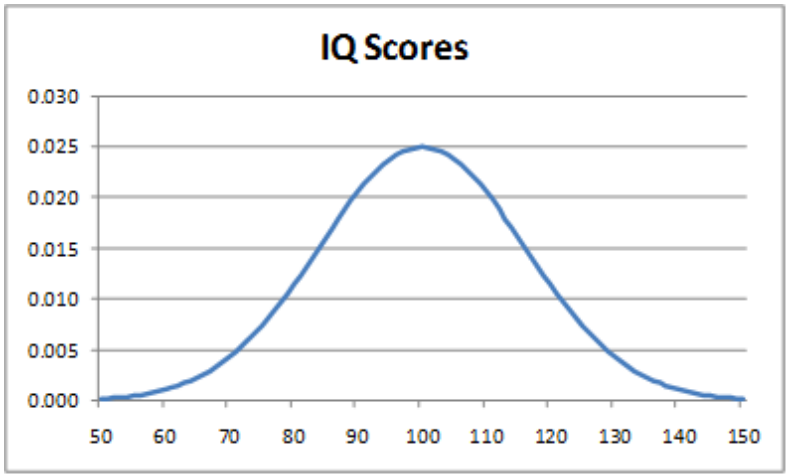
P(X = 12) = Probability sum of rolls is 12



Continuous Case: **f(x)** or p(x)

X = IQ Score of random individual

P(X < 120) = 0.909

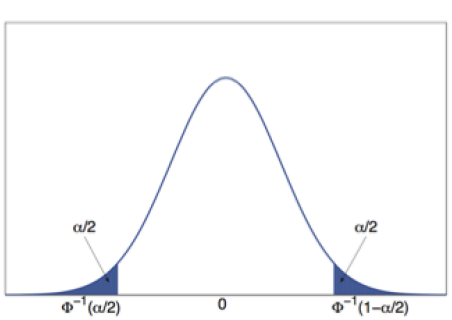


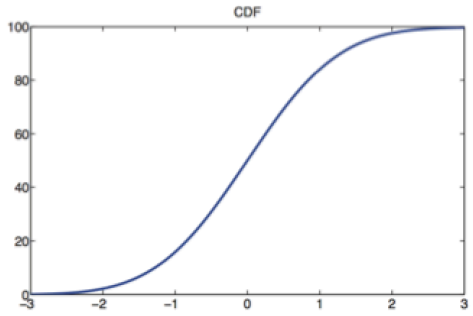
Random Variables - More on pdf

- The pdf, or the probability density function, often denoted f(x), describes the *relative* likelihood of a random variable taking on a given value.
  - What's probability some random man is 6 feet tall?

- The cdf, or cumulative distribution function, often denoted F(x), is the probability of that random variable **X** having a value <=x.

$$f(x) \qquad F(x) = P(X \leq x)$$



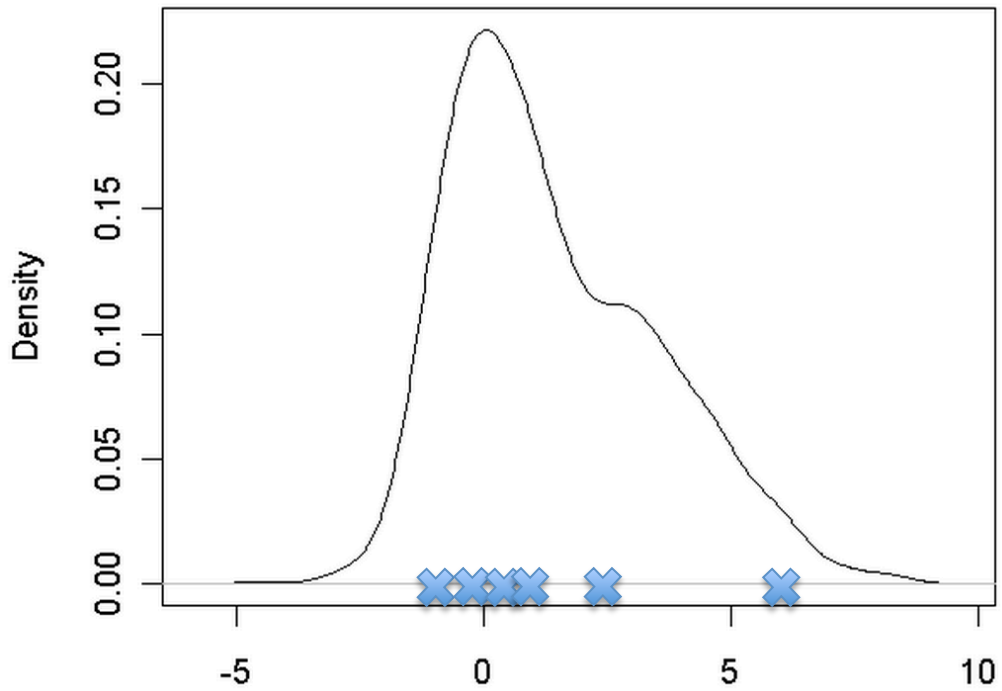


PDF ⇔ CDF

- pdf → cdf? Integrate $F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt.$
- cdf → pdf? Take derivative $F'(x) = \frac{d}{dx}F(x) = f(x)$

Expectation

- Intuition
  - Think of it as not as the average of possible values, but the average of infinite samples from the distribution



Expectation

- Discrete: Probability weighted average of all possible values

$$E(X) = x_1 * p_1 + x_2 * p_2 + \ldots + x_k * p_k$$

- Continuous: Same idea, except replace Σ with integral, and replace probabilities with probability densities

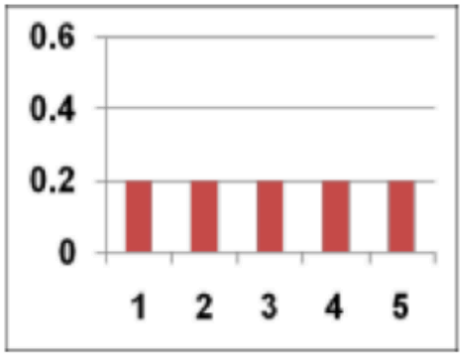
$$E(X) = \int_{-\infty}^{\infty} x * f(x) dx$$

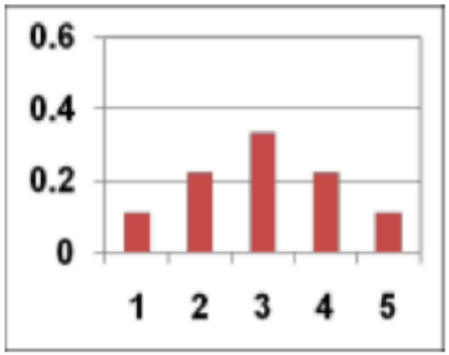
Variance

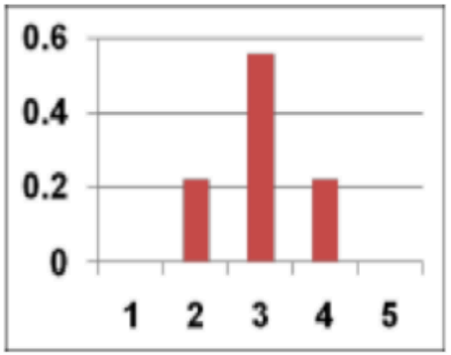
- Intuition
  - Measures how much "spread" there is in a set of numbers
  - The mean squared distance of random variable X from the mean, μ.

$$\operatorname{Var}(X) = \boxed{\operatorname{E}\left[(X - \mu)^2\right]}$$

$$\begin{aligned}
\operatorname{Var}(X) &= \operatorname{E}\left[(X - \operatorname{E}[X])^2\right] \\
&= \operatorname{E}\left[X^2 - 2X\operatorname{E}[X] + (\operatorname{E}[X])^2\right] \\
&= \operatorname{E}\left[X^2\right] - 2\operatorname{E}[X]\operatorname{E}[X] + (\operatorname{E}[X])^2 \\
&= \boxed{\operatorname{E}\left[X^2\right] - (\operatorname{E}[X])^2}
\end{aligned}$$







Variance

- Discrete: Probability weighted average of all possible deviations from mean

  Suppose discrete r.v. X can take on ***k*** distinct values.

$$E(X) = \mu$$

$$Var(X) = E[(X-\mu)^2] = \sum_{i=1}^{k} p_i \ * \ (x_i - \mu)^2$$

- Continuous: Same idea, except replace Σ with integral, and replace probabilities p_i with probability densities f(x)

$$Var(X) = \sigma^2 = \int (x-\mu)^2 f(x) dx$$

Covariance

- Covariance measures how much two random variables change together

$$Cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)]$$

$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

$$\hat{Cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

Relation to Var(X)?

Correlation

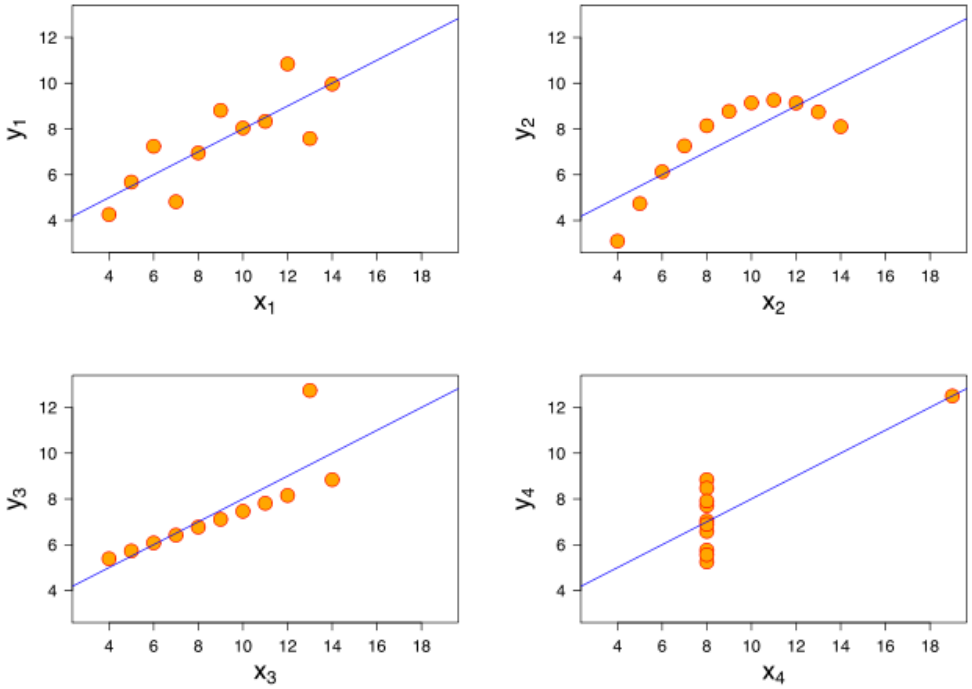
- Correlation is the normalized version of the correlation coefficient
  - Also measure of strength of linear relationship between two variables.

$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sigma_x \sigma_y}$$

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2)}}$$

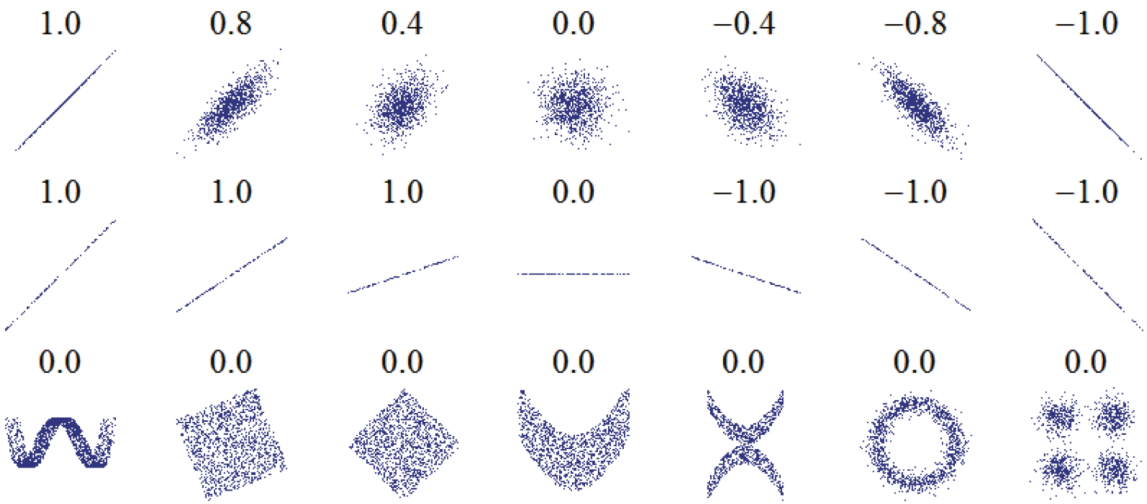
Correlation

- Correlation measures strength of linear relationship between two variables… but beware that one number can't capture what's really going on!



All of these have a correlation of 0.816

Correlation



Reflects **noisiness** and **direction**

But **not slope**

And **not other non-linearities**

Relationship between all my variables?

Let *X* be data matrix containing *m* features: $X = [X_1, X_2, X_3, \ldots, X_m]$

$$Var(X) = \begin{bmatrix} Var(X_1) & Cov(X_1,X_2) & \cdots & Cov(X_1,X_m) \\ Cov(X_1,X_2) & Var(X_2) & \cdots & Cov(X_2,X_m) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_1,X_m) & Cov(X_2,X_m) & \cdots & Var(X_m) \end{bmatrix}$$

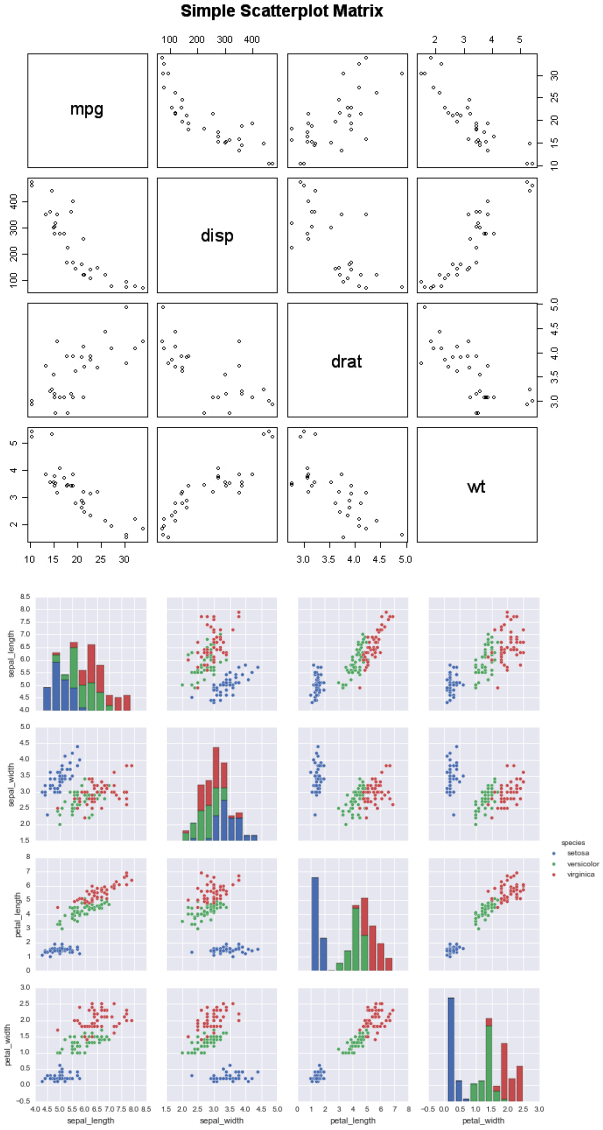
- Compact representation of all covariances
- Variances on diagonal

$$corr(X,Y) = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

$$corr(X) = \begin{bmatrix} 1 & corr(X_1,X_2) & \cdots & corr(X_1,X_m) \\ corr(X_1,X_2) & 1 & \cdots & corr(X_2,X_m) \\ \vdots & \vdots & \ddots & \vdots \\ corr(X_1,X_m) & corr(X_2,X_m) & \cdots & 1 \end{bmatrix}$$

Often use correlation matrices because it's normalized

	DJIA	S&P 500	Nasdaq	Canada	Mexico	Brazil	Stoxx 50	FTSE 100	CAC 40	DAX	IBEX	Italy	Netherlands	Sweden	Switzerland	Nikkei	Hang Seng	Australia
DJIA		0.97	0.85	0.57	0.56	0.52	0.52	0.48	0.51	0.56	0.49	0.50	0.50	0.42	0.42	0.09	0.11	0.07
S&P 500	0.97		0.91	0.62	0.58	0.55	0.50	0.47	0.50	0.55	0.48	0.50	0.49	0.41	0.41	0.09	0.11	0.05
Nasdaq	0.85	0.91		0.58	0.56	0.52	0.48	0.43	0.48	0.54	0.47	0.48	0.48	0.42	0.38	0.14	0.16	0.07
Canada	0.57	0.62	0.58		0.53	0.53	0.42	0.45	0.41	0.41	0.42	0.42	0.39	0.37	0.35	0.17	0.22	0.17
Mexico	0.56	0.58	0.56	0.53		0.56	0.42	0.42	0.44	0.43	0.43	0.44	0.39	0.38	0.38	0.17	0.25	0.17
Brazil	0.52	0.55	0.52	0.53	0.56		0.33	0.35	0.32	0.34	0.34	0.34	0.29	0.30	0.28	0.17	0.22	0.15
Stoxx 50	0.52	0.50	0.48	0.42	0.42	0.33		0.92	0.94	0.89	0.87	0.88	0.92	0.78	0.86	0.26	0.30	0.24
FTSE 100	0.48	0.47	0.43	0.45	0.42	0.35	0.92		0.86	0.80	0.80	0.82	0.84	0.73	0.78	0.26	0.30	0.26
CAC 40	0.51	0.50	0.48	0.41	0.44	0.32	0.94	0.86		0.89	0.88	0.89	0.92	0.78	0.84	0.28	0.32	0.25
DAX	0.56	0.55	0.54	0.41	0.43	0.34	0.89	0.80	0.89		0.83	0.84	0.86	0.75	0.77	0.26	0.29	0.21
IBEX	0.49	0.48	0.47	0.42	0.43	0.34	0.87	0.80	0.88	0.83		0.84	0.83	0.75	0.77	0.27	0.32	0.26
Italy	0.50	0.50	0.48	0.42	0.44	0.34	0.88	0.82	0.89	0.84	0.84		0.85	0.74	0.78	0.24	0.29	0.23
Netherlands	0.50	0.49	0.48	0.39	0.39	0.29	0.92	0.84	0.92	0.86	0.83	0.85		0.75	0.82	0.27	0.30	0.23
Sweden	0.42	0.41	0.42	0.37	0.38	0.30	0.78	0.73	0.78	0.75	0.75	0.74	0.75		0.75	0.29	0.33	0.27
Switzerland	0.42	0.41	0.38	0.35	0.38	0.28	0.86	0.78	0.84	0.77	0.77	0.78	0.82	0.75		0.29	0.32	0.29
Nikkei	0.09	0.09	0.14	0.17	0.17	0.17	0.26	0.26	0.28	0.26	0.27	0.24	0.27	0.29	0.29		0.52	0.49
Hang Seng	0.11	0.11	0.16	0.22	0.25	0.22	0.30	0.30	0.32	0.29	0.32	0.29	0.30	0.33	0.32	0.52		0.48
Australia	0.07	0.05	0.07	0.17	0.17	0.15	0.24	0.26	0.25	0.21	0.26	0.23	0.23	0.27	0.29	0.49	0.48	



Statistical Distributions

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$Binomial(n,p)$	$\binom{n}{k} p^k (1-p)^{n-k}$ for $0 \leq k \leq n$	np	npq
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$e^{-\lambda}\lambda^x/x!$ $for\ k = 0, 1, 2...$	λ	λ
$Uniform(a,b)$	$\frac{1}{b-a}$ $\forall x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x}$ $x \geq 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

- X ~ Bernoulli(p) = Single coin flip turns out to be Heads
- X ~ Binomial(100, p) = # of coin flips out of 100 that turn out to be Heads
- X ~ Geometric(p) = # of Trials until coin flip turns out to be Heads

- X ~ Poisson(λ=10) = # of taxis passing a street corner in a given hour (on avg 10/hr)
- X ~ Exponential(λ=10) = Time until taxi will pass street corner

- X ~ Uniform(0,360) = Degrees between hour hand and minute hand
- X ~ Gaussian(100, 10) = IQ Score

Bernoulli(p), Binomial(n,p), Geometric(p)

- Will the **1** coin flip turn out to be heads? 

$$\begin{aligned} P(X=1) &= 1 - P(X=0) \\ &= 1 - q \\ &= p \end{aligned}$$

- How many of the **n** coin flips will turn out to be heads?

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n$$

- How many trials until coin flip turns out to be heads?

$$\Pr(X=k) = (1-p)^{k-1} p \qquad k = 1, 2, 3, \ldots.$$

Link between Bernoulli and Binomial

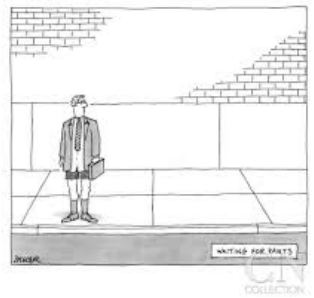
If $X_1, \ldots, X_n$ are independent Bernoulli(p) trials…

$$Y = \sum_{i=1}^{n} X_i \sim Binomial(n, p)$$

Poisson(λ)

- Number of events occurring in fixed interval of time or space.
  - Events occur at some average rate, λ, and independently of time since last event.

- Ex. Number of taxis passing a street corner in a given hour. On average there are 10/hour.



$$P(X = k) = \frac{\lambda^k * e^{-\lambda}}{k!} \qquad E(X) = \lambda$$

Link between Binomial and Poisson

- **Binomial**, fixed n trials, # of successes can be no greater than n.
- **Poisson**, in some sense, there are an infinite # of trials. Really, there could be any # of taxis that pass by.
- However they seem similar don't they? In fact, as $\lim_{n\to\infty} \lim_{p\to 0}$ and $\lambda = np$ stays constant, the Binomial distribution approximates Poisson.

Exponential(λ)

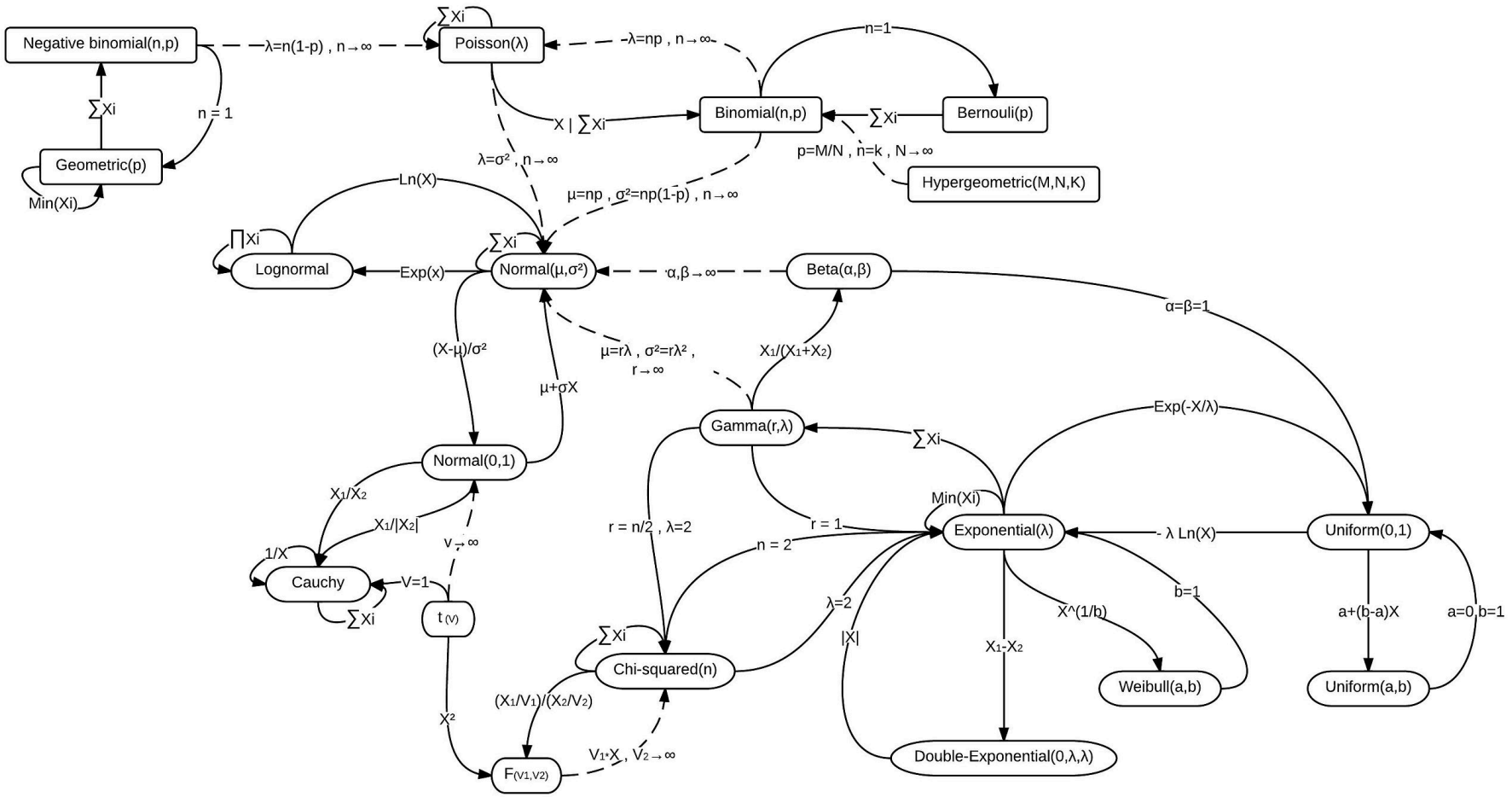
- Time between Poisson events
  - Note because Poisson events occur continuously and independently, Exponential is ”memoryless”

- Ex. Time until taxi arrives at street corner for pickup. On average there are 10/hour.

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0 \qquad E(X) = \frac{1}{\lambda}$$

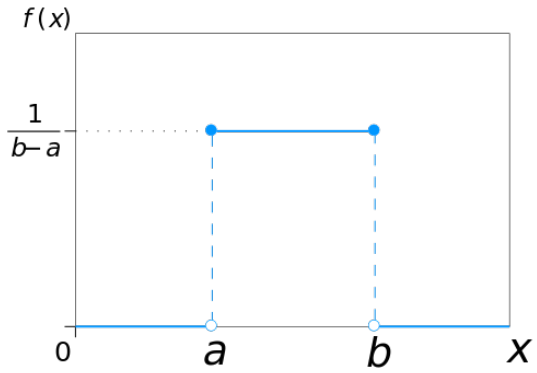
  - Why memoryless? Well even if a taxi just went by 2 minutes ago, assuming taxis are arriving continuously and independently, that doesn’t affect your expected wait time still.

A lot of distributions are related…



Uniform(a,b)

- Intervals of the same length on the distribution's support (a,b) are equally probably



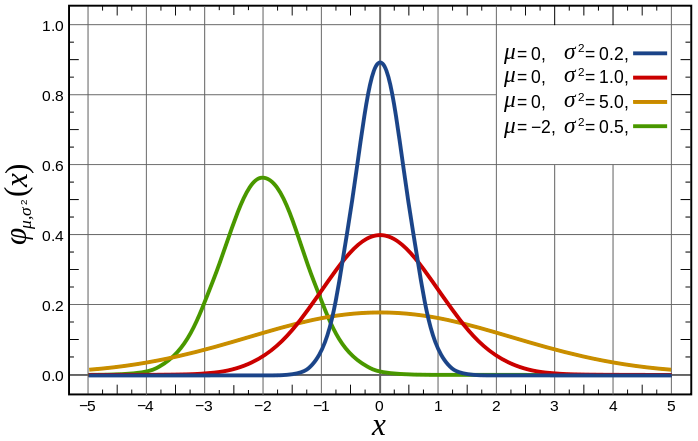
$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$

- Ex. What's the probability that the degrees between hour hand and minute hand is greater than 90?
  - X~Uniform(0,360). P(X>90) = (360-90)/360 = 0.75

Gaussian(μ,σ), or Normal(μ,σ)

- Very commonly occurring distribution shaped like a bell curve.

- Related to **Central Limit Theorem**, which states that under certain conditions, mean of r.v. independently drawn from same distribution is normally distributed!

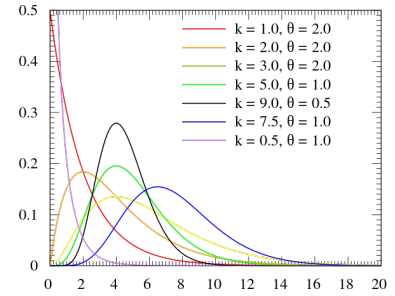


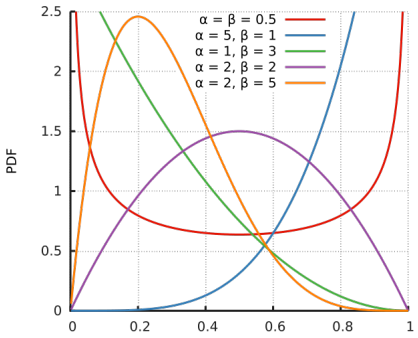
- Normal(μ,σ) is just a stretched and shifted Standard Normal(0,1).
  - If X~Normal(μ,σ) and Z~Normal(0,1),

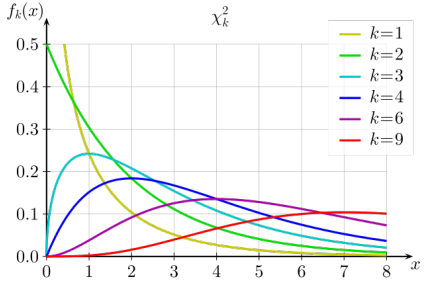
    $X = Z*\sigma + \mu$ and $Z = (X - \mu)/\sigma$

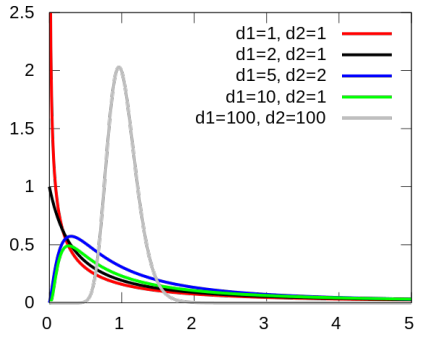
Briefly…

- Gamma(k, θ): time until n events in a process with no memory
  - Sum of k Exponentials
- Beta(α,β): useful in estimating success probabilities
  - Uniform is actually special case of Beta
  - Uniform(0,1) ⇔ Beta(1,1)
- Chi-Square: useful for goodness of fit tests
  - Sum of squares of *k* independent Gaussian r.v.s
- F-distribution: useful for some statistical tests
  - Ratio of two (normalized) chi-squared-distributed random variables









Joint Probability Distribution

P(X= x, Y= y) = P(x, y)

- How probable to observe these two attributes together?

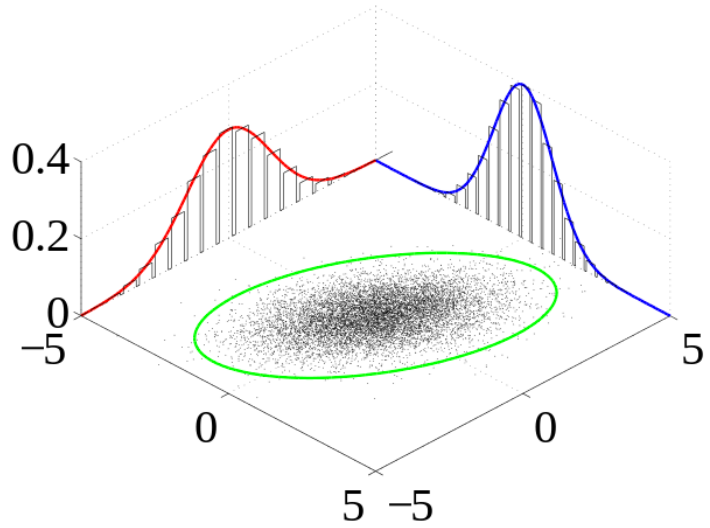
Discrete

roll some dice that results in A and B…

	1	2	3	4	5	6
A	1	1	1	1	0	1
B	0	1	1	0	1	0

$$P(A=1 \cap B=0) = \{1,4,6\} = 3/6$$

Continuous



Joint Probability Distribution

Discrete

$$P(X = x \cap Y = y) = P(Y = y \mid X = x) \cdot P(X = x) = P(X = x \mid Y = y) \cdot P(Y = y)$$

$$P(X = x \cap Y = y) = P(X = x) \cdot P(Y = y)$$

If X, Y independent

$$\sum_i \sum_j P(X = x_i \cap Y = y_i) = 1$$

Continuous

$$f_{X,Y}(x,y) = f_{Y|X}(y \mid x) f_X(x) = f_{X|Y}(x \mid y) f_Y(y)$$

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$$

If X, Y independent

$$\int_x \int_y f_{X,Y}(x,y)\, dy\, dx = 1.$$

(1) "Marginal Distributions"
(2) If independent, can split
(3) As usual, Σ or integrate over all possibilities = 1

Marginalization

We know P(X,Y), what is P(X=x)?

$\Longrightarrow$ Use Law of Total Probability

Discrete

$$P(X = x) = \sum_{y} P(X = x, Y = y)$$
$$= \sum_{y} P(X = x | Y = y) P(Y = y)$$

	x_1	x_2	x_3	x_4	$p_y(Y)\downarrow$
y_1	$4/32$	$2/32$	$1/32$	$1/32$	$8/32$
y_2	$2/32$	$4/32$	$1/32$	$1/32$	$8/32$
y_3	$2/32$	$2/32$	$2/32$	$2/32$	$8/32$
y_4	$8/32$	0	0	0	$8/32$
$p_x(X) \rightarrow$	$16/32$	$8/32$	$4/32$	$4/32$	$32/32$

Continuous

$$p_X(x) = \int_{y} p_{X,Y}(x, y)\, dy$$
$$= \int_{y} p_{X|Y}(x|y)\, p_Y(y)\, dy$$

Independence

	Always	Under independence
$P(Y\|X) =$		$P(Y)$
$P(X \cap Y) =$	$P(X) * P(Y\|X)$ $P(Y) * P(X\|Y)$	$P(X) * P(Y)$
$Var(X+Y) =$	$Var(X) + Var(Y)$ $+2 * Cov(X,Y)$	$Var(X) + Var(Y)$
$Cov(X,Y) =$	$E(XY) - E(X)E(Y)$	0

Questions

- Difference between permutations and combinations?
  - Be able to do Uber and Lyft problem both ways
- Bayes Rule (and when useful)?
- Law of Total Probability?
- In layman's terms, what's Variance? Covariance? Correlation?
  - How are they related?
  - What's wrong with simply using Correlation to describe X vs. Y?
- Describe distributions below. Give 1 clear example each.
  - Discrete: Bernoulli, Binomial, Geometric, Poisson
  - Continuous: Uniform, Gaussian, Exponential
- What is the Joint Distribution of X and Y?
  - How to marginalize?
  - What if they're independent?

Questions

- Difference between permutations and combinations?
  - Permutations: # of **unique** ways to choose k out of n = n!/(n-k)!
  - Combinations: # of ways to choose k out of n = n!/[(n-k)!*k!]
- Bayes Rule (and when useful)? $P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$ P(B|A) easy, but P(A|B) hard
- Law of Total Probability? $P(A) = \sum_i P(B_i \cap A) = \sum_i P(B_i) * P(A|B_i)$
- In layman's terms, what's Variance? Covariance? Correlation?
  - How are they related? Var(X) = Cov(X,X); Cor(X,Y) = Cov(X,Y)/(σ_X*σ_Y)
  - What's wrong with simply using Correlation to describe X vs. Y?

    Correlation captures linearity of relationship in a single number. Plotting is better.
- Describe distributions below. Give 1 clear example each.
  - Discrete: Bernoulli, Binomial, Geometric, Poisson See Slide 23
  - Continuous: Uniform, Gaussian, Exponential
- What is the Joint Distribution of X and Y?
  - How to marginalize? If have P(X,Y) and want P(X) just sum/integrate over all values of Y
  - What if they're independent? P(X,Y) = P(X)*P(Y) or f(x,y) = f(x) * f(y)

Appendix

Definition:

$$cov(X,Y) = E(X-\mu_X)(Y-\mu_Y)$$

This can be simplified as follows:

$$cov(X,Y) = E(X-\mu_X)(Y-\mu_Y) = E(XY) - \mu_Y E(X) - \mu_X E(Y) + \mu_X\mu_Y$$

Therefore,

$$cov(X,Y) = E(XY) - (EX)(EY)$$

Note: If X, Y are independent then $E(XY) = (EX)E(Y)$ Therefore $cov(X,Y) = 0$.

Let W, X, Y, Z random variables, and a, b, c, d constants:

- Find $cov(a+X, Y)$

  $$cov(a+X,Y) = E(a+X-\mu_{a+X})(Y-\mu_Y) = E(a+X-\mu_X-a)(Y-\mu_Y)$$

  Therefore, $cov(a+X,Y) = cov(X,Y)$.

- Find $cov(aX, bY)$

  $$cov(aX,bY) = E(aX-\mu_{aX})(bY-\mu_{bY}) = E(aX-a\mu_X)(bY-b\mu_Y)$$

  Therefore, $cov(aX,bY) = abE(X-\mu_X)(Y-\mu_Y) = ab\ cov(X,Y)$

- Find $cov(X, Y+Z)$

  $$cov(X,Y+Z) = E(X-\mu_X)(Y-\mu_{Y+Z}) = E(X-\mu_X)(Y+Z-\mu_Y-\mu_Z)$$

  Or

  $$cov(X,Y+Z) = E(X-\mu_X)(Y-\mu_Y+Z-\mu_Z) = E(X-\mu_X)(Y-\mu_Y) + E(X-\mu_X)(Z-\mu_Z)$$

  Therefore, $cov(X,Y+Z) = cov(X,Y) + cov(X,Z)$

- Using the results above we can find $cov(aW+bX, cY+dZ)$.

  $$cov(aW+bX,cY+dZ) = ab\ cov(W,Y) + ad\ cov(W,Z) + bc\ cov(X,Y) + bd\ cov(X,Z)$$

More about covariance between random variables

useful general formula,
Recall Var(X) = Cov(X,X)