Statistical Foundations

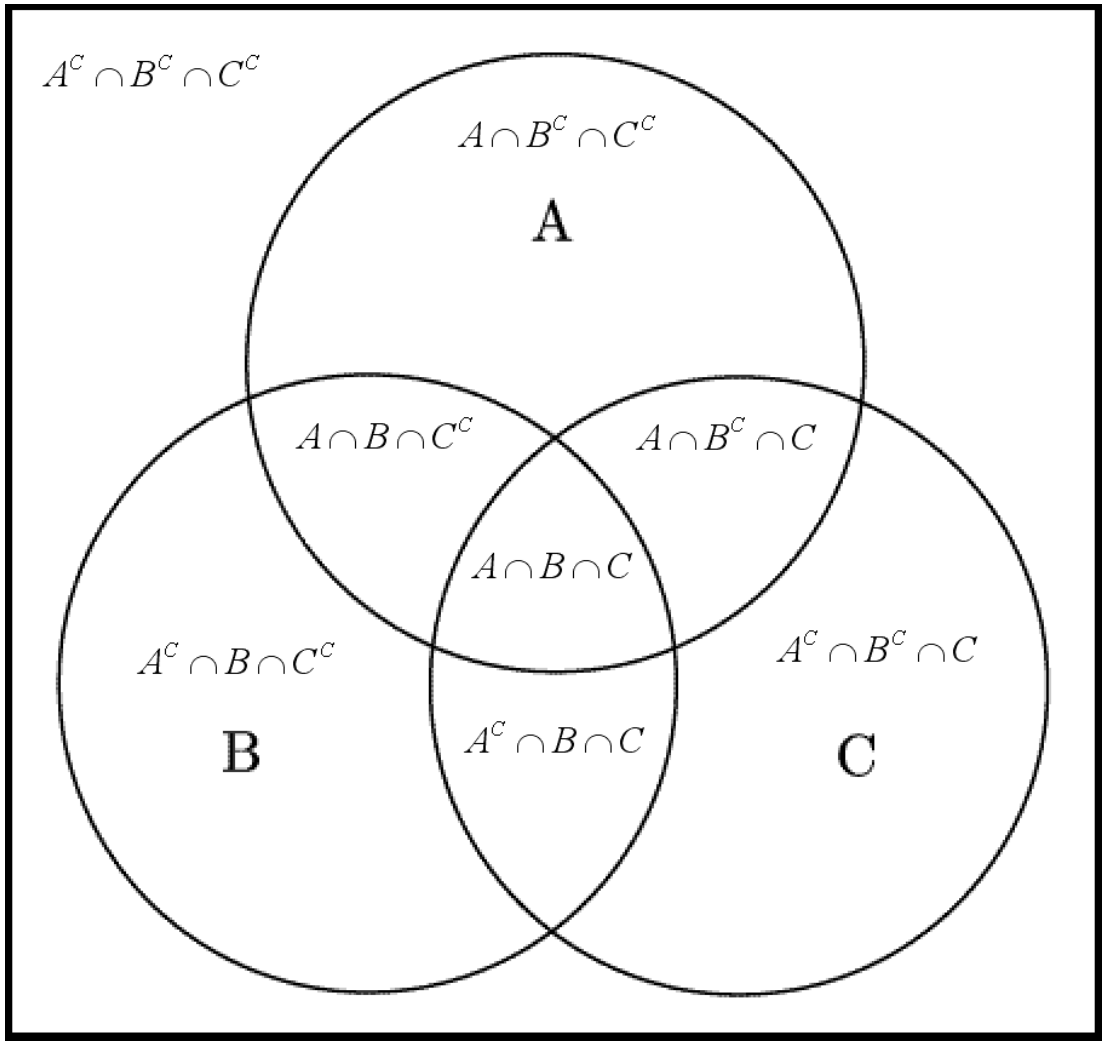
Arjun Mukherjee†

Course webpage:
http://www.cs.uh.edu/~arjun/courses/nlp

† Contains contents from [Casella and Berger, 2002] , and various other sources. Referenced in place.

Set and Probability Theory

- Sample space
- Event: Any collection of possible outcomes for an experiment.
- Events are subsets of sample space.
- Event occurrence ⇒ The outcome of the experiment lies in set A.



Set and Probability Theory

- Basic operation on events: Union($\cup$), Intersection ($\cap$), Complementation (^{C})

Theorem 1.1.4 *For any three events, A, B, and C, defined on a sample space S,*

a. Commutativity
$$A \cup B = B \cup A,$$
$$A \cap B = B \cap A;$$

b. Associativity
$$A \cup (B \cup C) = (A \cup B) \cup C,$$
$$A \cap (B \cap C) = (A \cap B) \cap C;$$

c. Distributive Laws
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C),$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C);$$

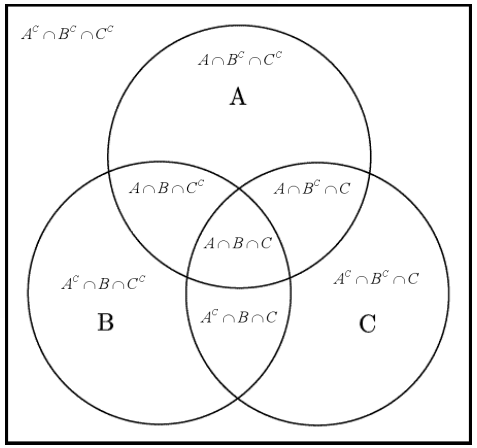
d. DeMorgan's Laws
$$(A \cup B)^c = A^c \cap B^c,$$
$$(A \cap B)^c = A^c \cup B^c.$$

Set and Probability Theory

- Probability is a set function whose domain is a sigma algebra, $\mathcal{B}$ (Borel field on S) and range is $[0, 1]$
- $\mathcal{B}$ can be thought of as set of all subsets of S

Definition 1.2.4 Given a sample space S and an associated sigma algebra $\mathcal{B}$, a *probability function* is a function P with domain $\mathcal{B}$ that satisfies

1. $P(A) \geq 0$ for all $A \in \mathcal{B}$.
2. $P(S) = 1$.
3. If $A_1, A_2, \ldots \in \mathcal{B}$ are pairwise disjoint, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.



Defining Probability

- Scoring on dart board

Example 1.2.7 (Defining probabilities–II) The game of darts is played by throwing a dart at a board and receiving a score corresponding to the number assigned to the region in which the dart lands. For a novice player, it seems reasonable to assume that the probability of the dart hitting a particular region is proportional to the area of the region. Thus, a bigger region has a higher probability of being hit.

Referring to Figure 1.2.1, we see that the dart board has radius r and the distance between rings is $r/5$. If we make the assumption that the board is always hit (see Exercise 1.7 for a variation on this), then we have

$$P(\text{scoring } i \text{ points}) = \frac{\text{Area of region } i}{\text{Area of dart board}}.$$

For example

$$P(\text{scoring 1 point}) = \frac{\pi r^2 - \pi(4r/5)^2}{\pi r^2} = 1 - \left(\frac{4}{5}\right)^2.$$

It is easy to derive the general formula, and we find that

$$P(\text{scoring } i \text{ points}) = \frac{(6-i)^2 - (5-i)^2}{5^2}, \quad i = 1, \ldots, 5,$$

independent of π and r. The sum of the areas of the disjoint regions equals the area of the dart board. Thus, the probabilities that have been assigned to the five outcomes sum to 1, and, by Theorem 1.2.6, this is a probability function (see Exercise 1.8). ||

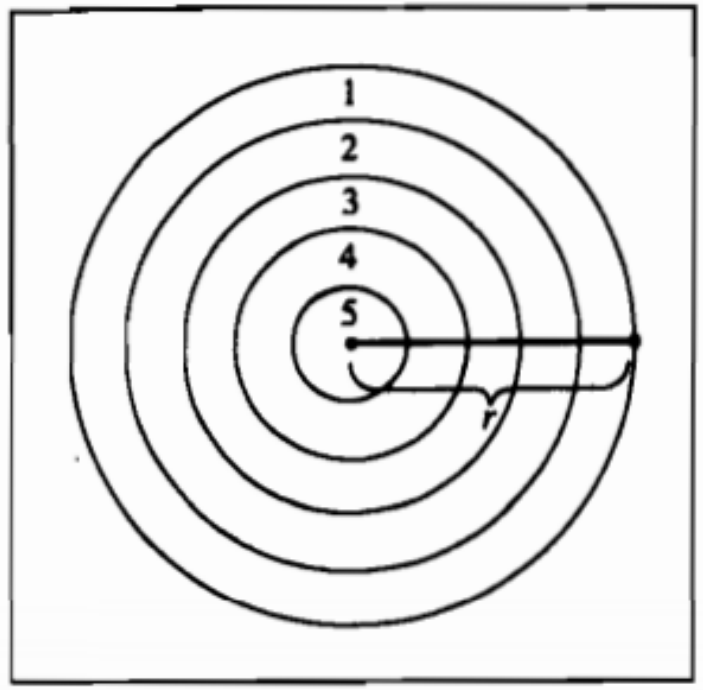


Figure 1.2.1. *Dart board for Example 1.2.7*

Probability Calculus

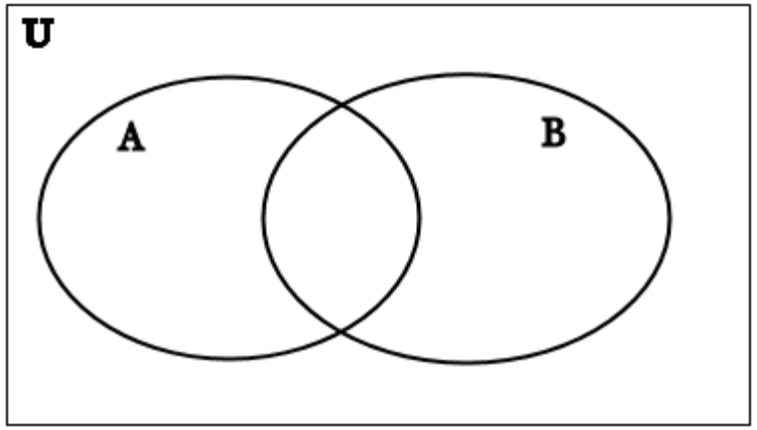
- Fundamental theorems

Theorem 1.2.8 *If P is a probability function and A is any set in $\mathcal{B}$, then*

a. $P(\emptyset) = 0$, *where $\emptyset$ is the empty set;*

b. $P(A) \leq 1$;

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



Theorem 1.2.9 *If P is a probability function and A and B are any sets in $\mathcal{B}$, then*

a. $P(B \cap A^c) = P(B) - P(A \cap B)$; **How?**

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

c. *If* $A \subset B$, *then* $P(A) \leq P(B)$.

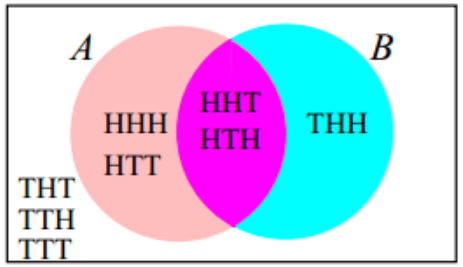
Conditional Probability and Independence

- Q: Flip a coin 3 times. If there are 2 heads, what's the probability that the first flip is heads?

Scenario: Flip a fair coin three times

A = "First flip is heads" $P(A) = \frac{4}{8}$
$= \{HHH, HHT, HTH, HTT\}$

B = "Two flips are heads" $P(B) = \frac{3}{8}$
$= \{HHT, HTH, THH\}$



Conditional probability

- Flip a coin 3 times. If there are 2 heads, what's the probability that the first flip is heads?
- *Rephrase:* Assuming B is true, what's the probability of A?
- Since B is true, the coin flips are one of HHT, HTH, or THH.
- Out of those, the outcomes where A is true are HHT and HTH (which is $A \cap B$). So 2 out of the 3 possible outcomes in B give A.
- The probability of A, given that B is true, is

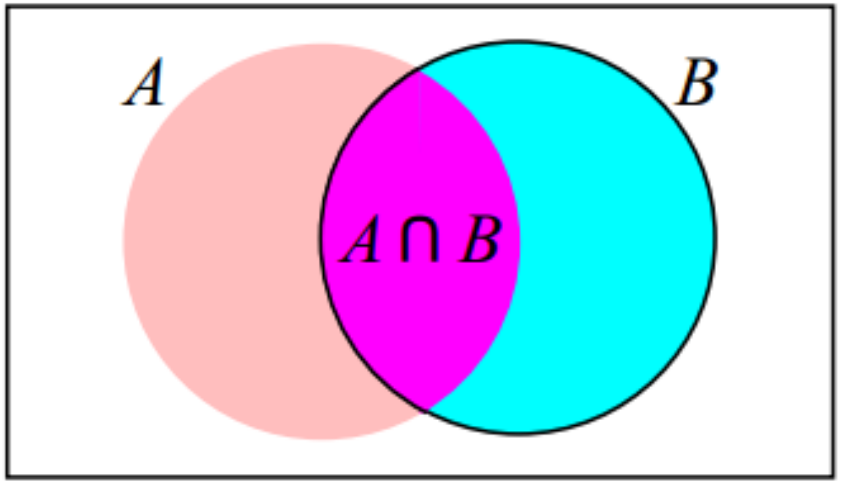
$$\frac{P(\{HHT, HTH\})}{P(\{HHT, HTH, THH\})} = \frac{2/8}{3/8} = \frac{2}{3} \qquad P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Prof. Tesler | Conditional Probability and Bayes' Theorem | Math 186 / January 8, 2014 | 3 / 31

† Slide courtesy of G.Tesler's Math 186

Conditional Probability and Independence

- Conditioning can be thought of as shrinkage of the effective sample space



$P(A)$ = probability of A
measures A as a fraction of the sample space.

$P(A \mid B)$ = probability of A, given B
measures $A \cap B$ as a fraction of B:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

† Venn diagram courtesy of G.Tesler Math 186

Conditional Probability and Independence

- Q: If a family has 2 children. Given that one of them is a girl what is the probability that both are girls?

  ½ (?) or more or less (?)

Conditional Probability and Independence

- Q: If a family has 2 children. Given that one of them is a girl what is the probability that both are girls?
- ½ (?) or more or less (?)
- Answer:

Sample space, S = {BB, BG, GB, GG}

$$P(\text{both girl} \mid \text{one of them girl}) = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}$$

Conditional Probability and Independence

- Three prisoners

Example 1.3.4 (Three prisoners) Three prisoners, A, B, and C, are on death row. The governor decides to pardon one of the three and chooses at random the prisoner to pardon. He informs the warden of his choice but requests that the name be kept secret for a few days.

The next day, A tries to get the warden to tell him who had been pardoned. The warden refuses. A then asks which of B or C will be executed. The warden thinks for a while, then tells A that B is to be executed.

Warden's reasoning: Each prisoner has a $\frac{1}{3}$ chance of being pardoned. Clearly, either B or C must be executed, so I have given A no information about whether A will be pardoned.

A's reasoning: Given that B will be executed, then either A or C will be pardoned. My chance of being pardoned has risen to $\frac{1}{2}$.

- Q: Whose reasoning (Warden's or Prisoner A's) is correct?
- Answer: Warden. Why? See example 1.3.4 in SI [Casella and Berger, 2002]

Conditional Probability and Independence

- Statistical independence

Definition 1.3.7 Two events, A and B, are *statistically independent* if

(1.3.8) $$P(A \cap B) = P(A)P(B).$$

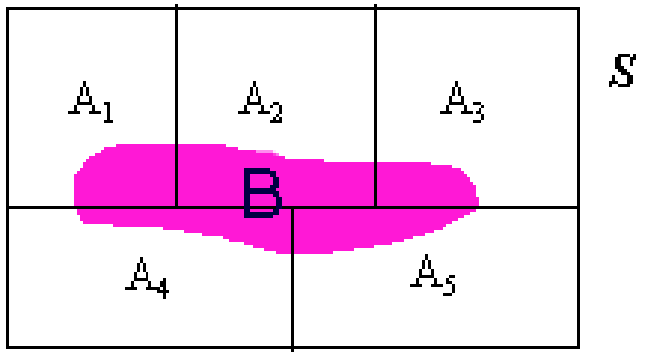
- Useful (algebraic) representations

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

- Bayes' rule

Theorem 1.3.5 (Bayes' Rule) *Let $A_1, A_2, \ldots$ be a partition of the sample space, and let B be any set. Then, for each $i = 1, 2, \ldots$,*

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{\infty} P(B|A_j)P(A_j)}.$$



Conditional Probability and Independence

- Example problem:

Two litters of a particular rodent species have been born, one with two brown-haired and one gray-haired (litter 1), and the other with three brown-haired and two gray-haired (litter 2). We select a litter at random and then select an offspring at random from the selected litter.

(a) What is the probability that the animal chosen is brown-haired?

(b) Given that a brown-haired offspring was selected, what is the probability that the sampling was from litter 1?

- Solution:

a.

$$\begin{aligned} & P(\text{Brown Hair}) \\ = \; & P(\text{Brown Hair}|\text{Litter 1})P(\text{Litter 1}) + P(\text{Brown Hair}|\text{Litter 2})P(\text{Litter 2}) \\ = \; & \left(\frac{2}{3}\right)\left(\frac{1}{2}\right) + \left(\frac{3}{5}\right)\left(\frac{1}{2}\right) = \frac{19}{30}. \end{aligned}$$

b. Use Bayes Theorem

$$P(\text{Litter 1}|\text{Brown Hair}) = \frac{P(BH|L1)P(L1)}{P(BH|L1)P(L1) + P(BH|L2)P(L2} = \frac{\left(\frac{2}{3}\right)\left(\frac{1}{2}\right)}{\frac{19}{30}} = \frac{10}{19}.$$

Random Variable

- A random variable is a mapping (function) from Sample space, S into real numbers, $\mathbb{R}$.

Example 1.4.3 (Three coin tosses–II) Consider again the experiment of tossing a fair coin three times from Example 1.3.13. Define the random variable X to be the number of heads obtained in the three tosses. A complete enumeration of the value of X for each point in the sample space is

s	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
$X(s)$	3	2	2	2	1	1	1	0

The range for the random variable X is $\mathcal{X} = \{0, 1, 2, 3\}$. Assuming that all eight points in S have probability $\frac{1}{8}$, by simply counting in the above display we see that the induced probability function on $\mathcal{X}$ is given by

x	0	1	2	3
$P_X(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

For example, $P_X(X = 1) = P(\{\text{HTT}, \text{THT}, \text{TTH}\}) = \frac{3}{8}$. ‖

Cumulative Distribution

- The cumulative distribution function (cdf) of a random variable is defined as $F_X(x) = P_X(X \leq x), \forall x$

Example 1.5.2 (Tossing three coins) Consider the experiment of tossing three fair coins, and let X = number of heads observed. The cdf of X is

$$F_X(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ \frac{1}{8} & \text{if } 0 \leq x < 1 \\ \frac{1}{2} & \text{if } 1 \leq x < 2 \\ \frac{7}{8} & \text{if } 2 \leq x < 3 \\ 1 & \text{if } 3 \leq x < \infty. \end{cases} \tag{1.5.1}$$

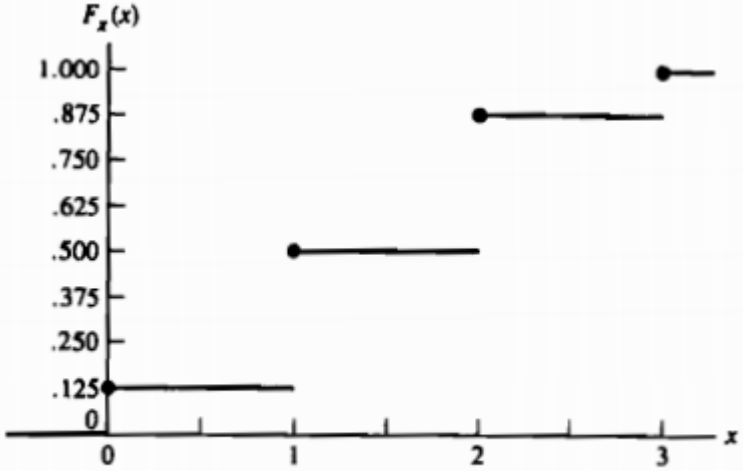


Figure 1.5.1. *Cdf of Example 1.5.2*

The step function $F_X(x)$ is graphed in Figure 1.5.1. There are several points to n from Figure 1.5.1. F_X is defined for all values of x, not just those in $\mathcal{X} = \{0, 1, 2, 3\}$. Thus, for example,

$$F_X(2.5) = P(X \leq 2.5) = P(X = 0, 1, \text{or } 2) = \frac{7}{8}.$$

Density and Mass Functions

- "Mass" and "density" function, $f_X(x)$ applies to *discrete* and *continuous* random variables respectively.

$$f_X(x) = P(X = x) \quad \text{for all } x. \qquad P(X \le x) = F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt.$$

Example 1.6.4 (Logistic probabilities) For the logistic distribution of Example 1.5.5 we have

$$F_X(x) = \frac{1}{1+e^{-x}}$$

and, hence,

$$f_X(x) = \frac{d}{dx}F_X(x) = \frac{e^{-x}}{(1+e^{-x})^2}.$$

The area under the curve $f_X(x)$ gives us interval probabilities (see Figure 1.6.1):

$$\begin{aligned} P(a < X < b) &= F_X(b) - F_X(a) \\ &= \int_{-\infty}^{b} f_X(x)\,dx - \int_{-\infty}^{a} f_X(x)\,dx \\ &= \int_{a}^{b} f_X(x)\,dx. \end{aligned}$$

There are really only two requirements for a pdf (or pmf), both of which are immediate consequences of the definition.

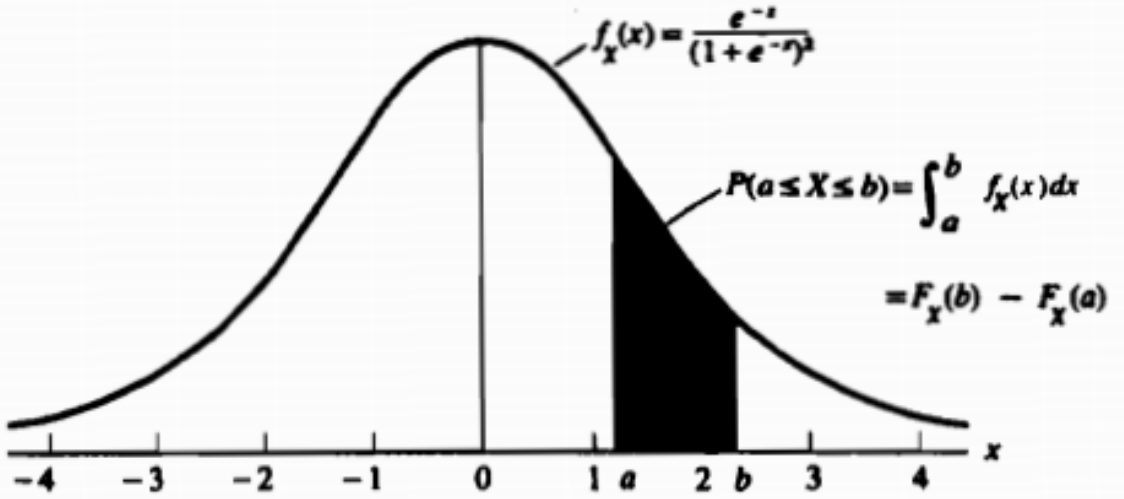


Figure 1.6.1. *Area under logistic curve*

Mean and Variance

- Mean/Expected value (1st moment) denoted by μ or $E[X]$

Definition 2.2.1 The *expected value* or *mean* of a random variable $g(X)$, denoted by $\mathrm{E}\,g(X)$, is

$$\mathrm{E}\,g(X) = \begin{cases} \int_{-\infty}^{\infty} g(x) f_X(x)\,dx & \text{if } X \text{ is continuous} \\ \sum_{x \in \mathcal{X}} g(x) f_X(x) = \sum_{x \in \mathcal{X}} g(x) P(X = x) & \text{if } X \text{ is discrete,} \end{cases}$$

- Variance (2nd moment) denoted by σ^2 or $Var[X]$

Definition 2.3.2 The *variance* of a random variable X is its second central moment, $\mathrm{Var}\,X = \mathrm{E}(X - \mathrm{E}\,X)^2$. The positive square root of $\mathrm{Var}\,X$ is the *standard deviation* of X.

$$\begin{aligned} \mathrm{Var}\,X = \mathrm{E}(X - \mathrm{E}\,X)^2 &= \mathrm{E}[X^2 - 2X\mathrm{E}\,X + (\mathrm{E}\,X)^2] \\ &= \mathrm{E}\,X^2 - 2(\mathrm{E}\,X)^2 + (\mathrm{E}\,X)^2 \\ &= \mathrm{E}\,X^2 - (\mathrm{E}\,X)^2, \end{aligned}$$

Discrete Uniform Distribution

- A random variable $X \sim Uni\ (N)$ when

(3.2.1) $$P(X = x|N) = \frac{1}{N}, \quad x = 1, 2, \ldots, N,$$

where N is a specified integer. This distribution puts equal mass on each of the outcomes $1, 2, \ldots, N$.

- Mean $E[X]$ and Variance $Var[X]$

We then have

$$\mathrm{E}X = \sum_{x=1}^{N} xP(X = x|N) = \sum_{x=1}^{N} x\frac{1}{N} = \frac{N+1}{2}$$

and

$$\mathrm{E}X^2 = \sum_{x=1}^{N} x^2\frac{1}{N} = \frac{(N+1)(2N+1)}{6},$$

and so

$$\begin{aligned}\mathrm{Var}\, X &= \mathrm{E}X^2 - (\mathrm{E}X)^2 \\ &= \frac{(N+1)(2N+1)}{6} - \left(\frac{N+1}{2}\right)^2 \\ &= \frac{(N+1)(N-1)}{12}.\end{aligned}$$

Bernoulli Distribution

- A *bernoulli trial* is an experiment which has exactly two outcomes: success/failure (e.g., tossing a coin yields S = {H, T}) with a parameter p (success probability)
- A random variable $X \sim Bern\ (p)$ when

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p, \end{cases} \qquad 0 \le p \le 1. \tag{3.2.3}$$

The value $X = 1$ is often termed a "success" and p is referred to as the success probability. The value $X = 0$ is termed a "failure." The mean and variance of a Bernoulli(p) random variable are easily seen to be

$$\mathrm{E}X = 1p + 0(1-p) = p,$$
$$\mathrm{Var}\,X = (1-p)^2 p + (0-p)^2(1-p) = p(1-p).$$

Binomial Distribution

- A *binomial distribution* models the total number of successes in a n identical and (independent) Bernoulli trails.

If n identical Bernoulli trials are performed, define the events

$$A_i = \{X = 1 \text{ on the } i\text{th trial}\}, \quad i = 1, 2, \ldots, n.$$

If we assume that the events $A_1, \ldots, A_n$ are a collection of independent events (as is the case in coin tossing), it is then easy to derive the distribution of the total number of successes in n trials. Define a random variable Y by

$$Y = \text{ total number of successes in } n \text{ trials.}$$

The event $\{Y = y\}$ will occur only if, out of the events $A_1, \ldots, A_n$, exactly y of them occur, and necessarily $n - y$ of them do not occur. One particular outcome (one particular ordering of occurrences and nonoccurrences) of the n Bernoulli trials might be $A_1 \cap A_2 \cap A_3^c \cap \cdots \cap A_{n-1} \cap A_n^c$. This has probability of occurrence

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3^c \cap \cdots \cap A_{n-1} \cap A_n^c) &= pp(1-p)\cdot\cdots\cdot p(1-p) \\ &= p^y(1-p)^{n-y}, \end{aligned}$$

Binomial Distribution

- However, note that there are $\binom{n}{y}$ ways in which one can obtain the event $\{Y = y\}$
- Since these are all independent trails, P($Y = y$) is given by:

$$P(Y = y|n, p) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0, 1, 2, \ldots, n,$$

- Which is also the pmf of the random variable $Y \sim Bin(n, p)$
- Mean $E[Y] = np$ and Variance $Var[Y] = np(1-p)$
- By the addition properties for independent random variables, the mean and variance of the binomial distribution are equal to the sum of the means and variances of the *n* independent Bernoulli variable

Binomial Distribution

Example 3.2.3 (Dice probabilities) Suppose we are interested in finding the probability of obtaining at least one 6 in four rolls of a fair die. This experiment can be modeled as a sequence of four Bernoulli trials with success probability $p = \frac{1}{6} = P(\text{die shows 6})$. Define the random variable X by

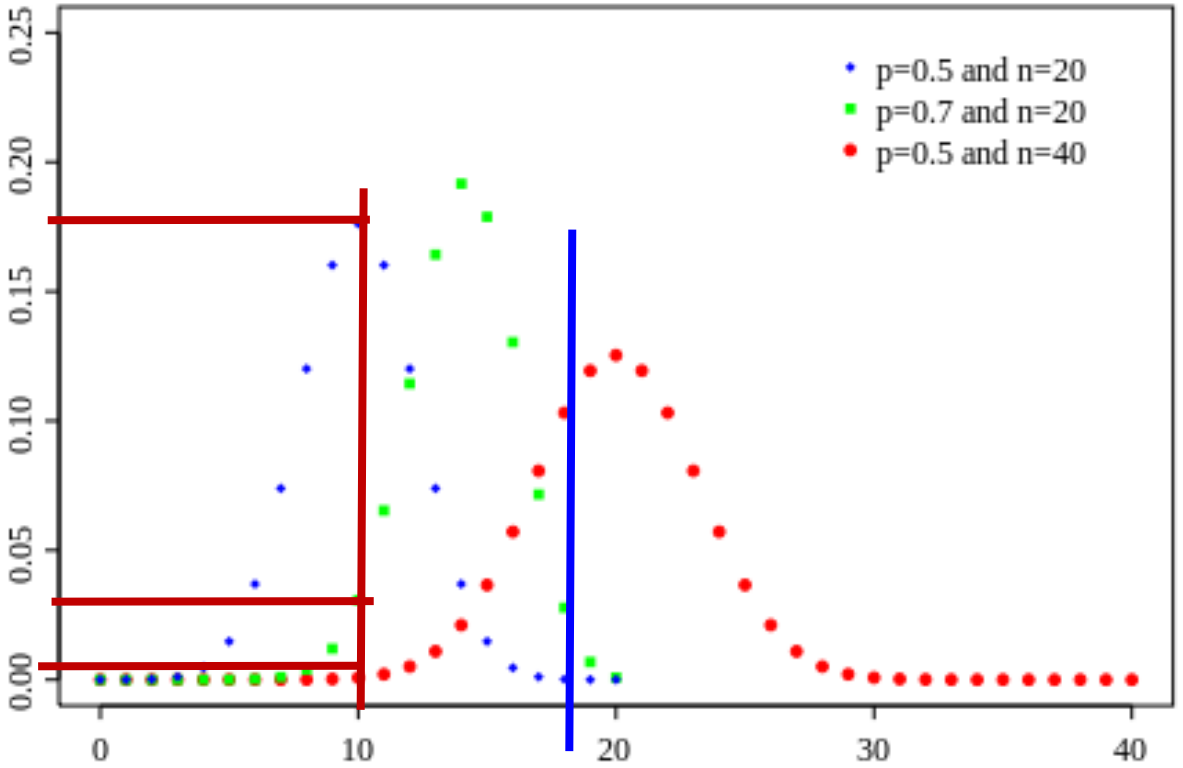
$$X = \text{total number of 6s in four rolls.}$$

Then $X \sim \text{binomial}\left(4, \frac{1}{6}\right)$ and

$$\begin{aligned} P(\text{at least one 6}) &= P(X > 0) = 1 - P(X = 0) \\ &= 1 - \binom{4}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^4 \\ &= 1 - \left(\frac{5}{6}\right)^4 \\ &= .518. \end{aligned}$$

Binomial Distribution

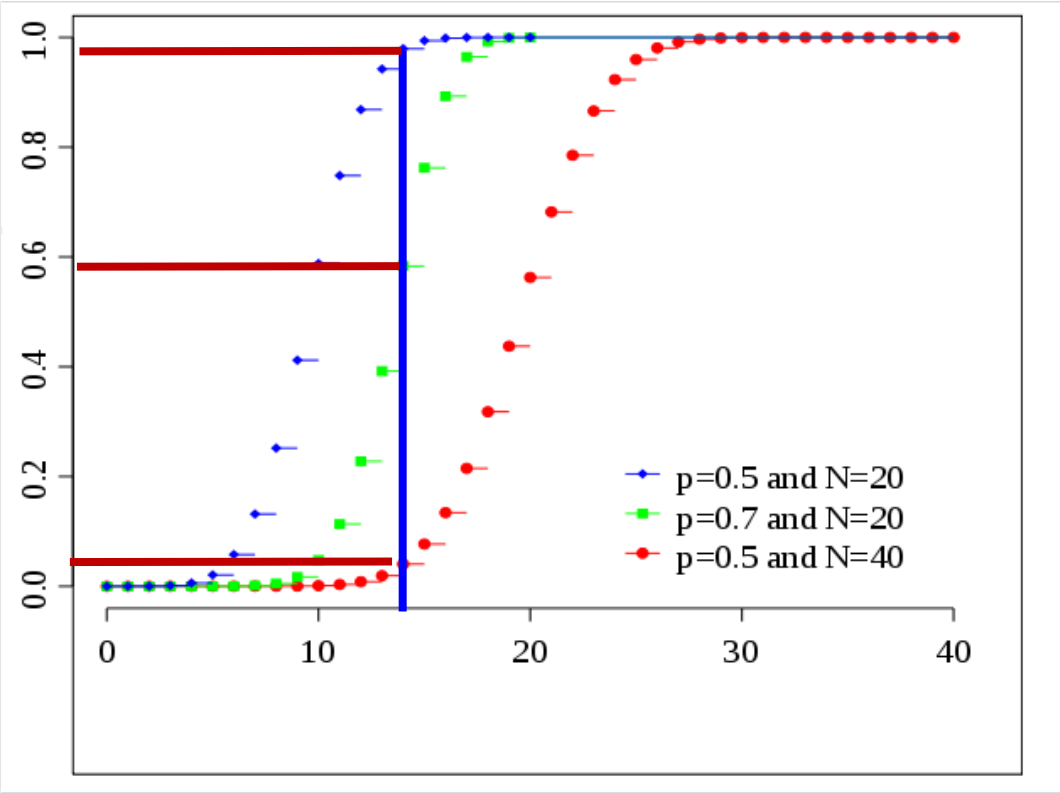
- Comparing $X_1 \sim Bin(n = 20, p = 0.5)$, $X_2 \sim Bin(n = 20, p = 0.7)$, $X_3 \sim Bin(n = 40, p = 0.5)$. The PMF plot is shown below.



- Note that $P(X_1 = 10) > P(X_2 = 10) > P(X_3 = 10)$. But $P(X_1 = 18) < P(X_2 = 18) < P(X_3 = 18)$What does this mean?

Binomial Distribution

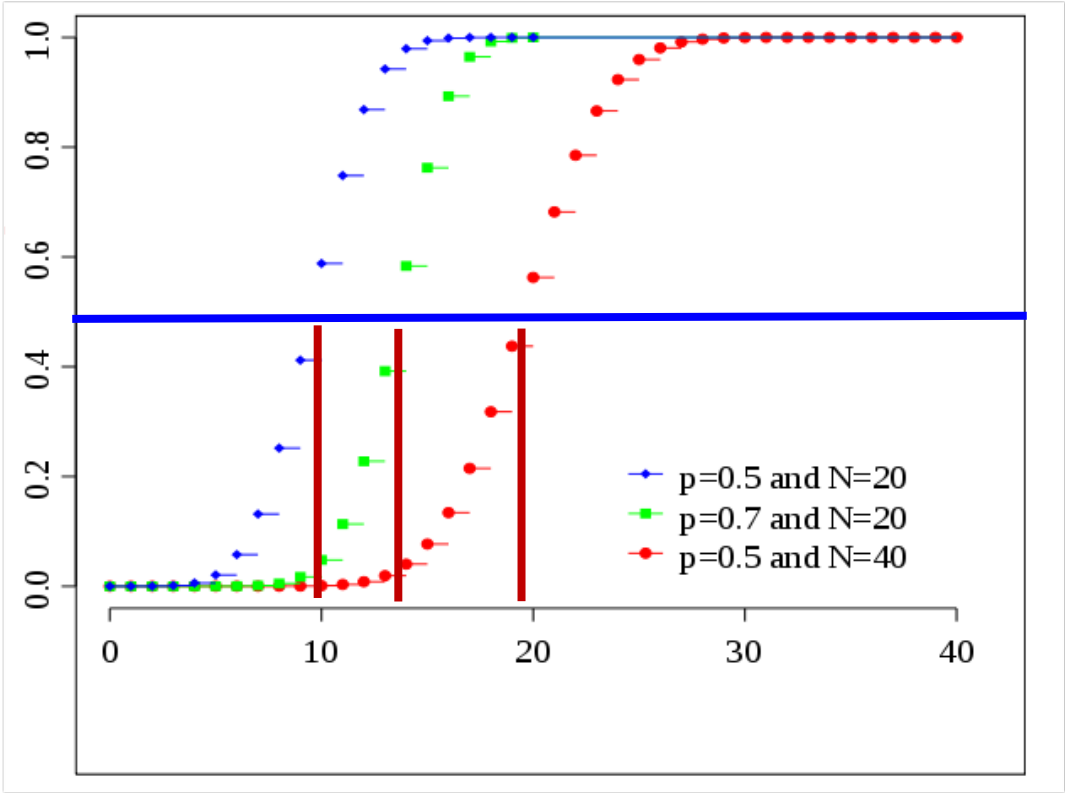
- Comparing $X_1 \sim Bin(n = 20, p = 0.5)$, $X_2 \sim Bin(n = 20, p = 0.7)$, $X_3 \sim Bin(n = 40, p = 0.5)$. The CDF plot is shown below.



- Note that $F_X(X_1 = 14) > F_X(X_2 = 14) > F_X(X_3 = 14)$. What does this mean?

Binomial Distribution

- Comparing $X_1 \sim Bin(n = 20, p = 0.5)$, $X_2 \sim Bin(n = 20, p = 0.7)$, $X_3 \sim Bin(n = 40, p = 0.5)$. The CDF plot is shown below.



- Note that 50% of the cdf space is bounded by $(X_1 < 10)$, $(X_2 < 14), (X_3 < 20)$. Where 10, 14 and 20 happens to be the means of X_1, X_2, X_3. What does this mean?
$X \sim InvCDF(U)$

Poisson Distribution

- Used for modeling events occurring in a fixed time interval with known average rate (λ). E.g, # of phone calls per hour with a avg. rate of 3 call/hr.
- A random variable $X \sim Poisson(\lambda)$ has the following PMF

$$P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, \ldots.$$

- Mean, $E[X] = \lambda$ (Obviously!)

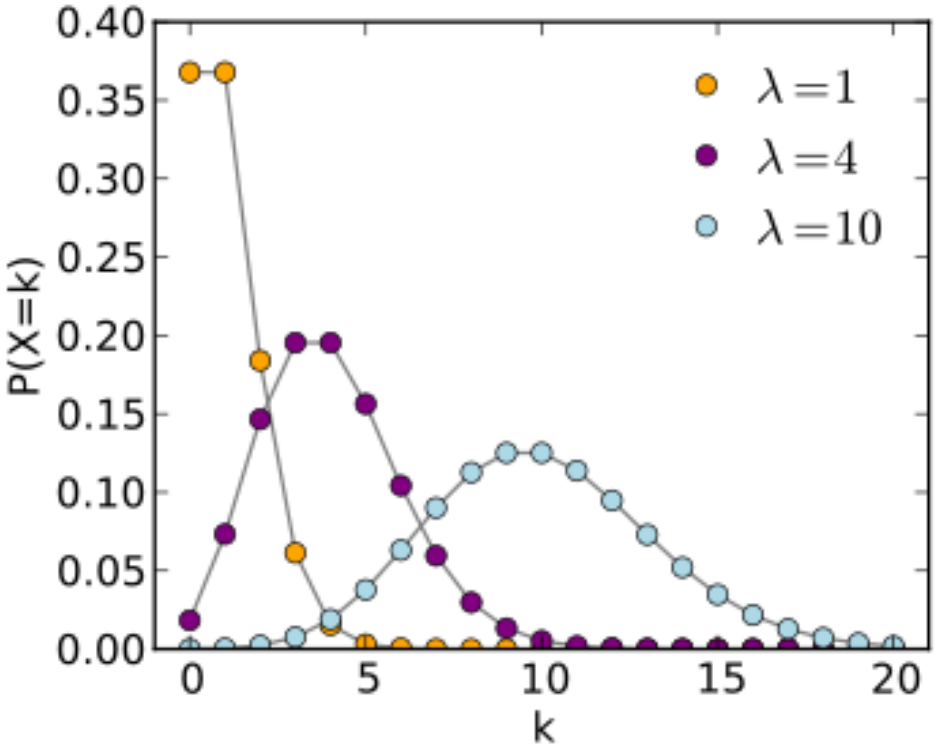
$$\begin{aligned}
EX &= \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!} \\
&= \sum_{x=1}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!} \\
&= \lambda e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!} \\
&= \lambda e^{-\lambda}\sum_{y=0}^{\infty}\frac{\lambda^y}{y!} \qquad \text{(substitute } y = x-1\text{)} \\
&= \lambda
\end{aligned}$$

Poisson Distribution

- Poisson distribution models the degree of spread around a known average rate of occurrence.
- $X_1 \sim Poisson(\lambda)$ models the # of occurrences in the next time interval.

- More precisely, given the average rate for temporal processes (e.g, mails per day, phone calls per hour, etc.), Poisson specifies the likelihood of counts (of mails, phone calls) during one/next period of observation.

- Comparing $X_1 \sim Poisson(\lambda = 1)$ vs. $X_2 \sim Poisson(\lambda = 4)$ vs. $X_3 \sim Poisson(\lambda = 10)$



Poisson Distribution

Example 3.2.4 (Waiting time) As an example of a waiting-for-occurrence application, consider a telephone operator who, on the average, handles five calls every 3 minutes. What is the probability that there will be no calls in the next minute? At least two calls?

If we let X = number of calls in a minute, then X has a Poisson distribution with $EX = \lambda = \frac{5}{3}$. So

$$\begin{aligned} P(\text{no calls in the next minute}) &= P(X = 0) \\ &= \frac{e^{-5/3}\left(\frac{5}{3}\right)^0}{0!} \\ &= e^{-5/3} = .189; \end{aligned}$$

$$\begin{aligned} P(\text{at least two calls in the next minute}) &= P(X \geq 2) \\ &= 1 - P(X = 0) - P(X = 1) \\ &= 1 - .189 - \frac{e^{-5/3}\left(\frac{5}{3}\right)^1}{1!} \\ &= .496. \end{aligned}$$

||

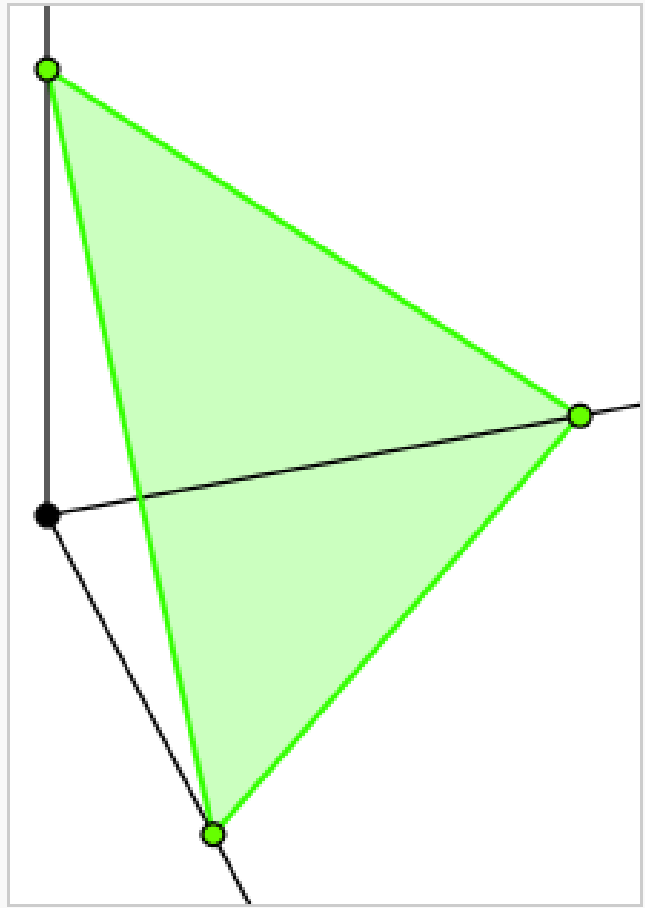
Categorical Distribution

- Generalization of a Bernoulli trail
- $X \sim Cat(K; \boldsymbol{p} =< p_1, p_2, \ldots . p_K >)$ is probability distribution that describes the result of a random event that can take exactly 1 out of K possible outcomes, with the probability of each outcome separately specified.
- It can be thought of as a K-way dice with K faces (or a roulette wheel!). $\mathrm{P}(i^{th}\ face) = p_i$. Clearly, $\sum_{i=1}^{K} p_i = 1$

- PMF is given by $P(X = i|\boldsymbol{p}) = p_i\ , 1 \leq i \leq K.$
- NOTE: This is important in various NLP/Text Mining models for sampling words out of a distribution over words and often (loosely) referred to as a Multinomial distribution

Categorical Distribution

- PMF formulation using inversion bracket $[X = i], 1 \leq i \leq K$
- $P(X|p) = \prod_{i=1}^{K} p_i^{[X=i]}$
- Mean $E[X = i] = p_i$ and Variance $Var[X = i] = p_i(1 - p_i)$. Where $[X = i]$ evaluates to 1 if $X = i$, otherwise 0.
- The possible probabilities form a standard K-1 dimensional simplex.
- Q: Given a random number generator, rand(0,1) how would you devise a sampling scheme to sample from a Categorical distribution?
- Hint: Try simulating a biased coin toss with Pr(H) = 0.7. Then extend to Cat(4, <0.1, 0.2, 0.4, 0.3>)



The possible probabilities for the categorical distribution with $k = 3$ are the 2-simplex $p_1 + p_2 + p_3 = 1$, embedded in 3-space.

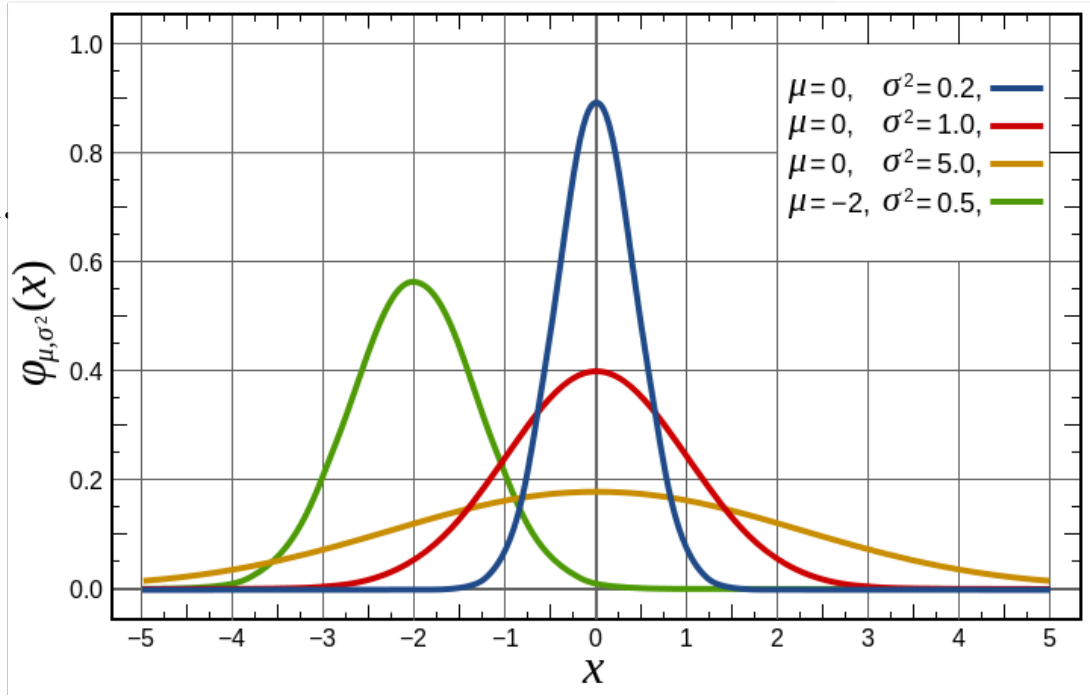
Gaussian Distribution

- Also called Normal distribution. Key distribution in all of Bayesian Statistics.

The normal distribution has two parameters, usually denoted by μ and σ^2, which are its mean and variance. The pdf of the *normal distribution* with mean μ and variance σ^2 (usually denoted by $\mathrm{n}(\mu, \sigma^2)$) is given by

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}, \quad -\infty < x < \infty. \tag{3.3.13}$$

- When $X \sim N(\mu = 0, \sigma^2 = 1)$,
  we say X follows standard normal distribution.



Gaussian Distribution

- For standard normal, $Z \sim N(\mu = 0, \sigma^2 = 1)$, actual probabilities are often looked up using a table.

If $X \sim \mathrm{n}(\mu, \sigma^2)$, then the random variable $Z = (X - \mu)/\sigma$ has a $\mathrm{n}(0, 1)$ distribution, also known as the *standard normal*. This is easily established by writing

$$\begin{aligned} P(Z \le z) &= P\left(\frac{X - \mu}{\sigma} \le z\right) \\ &= P(X \le z\sigma + \mu) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{z\sigma+\mu} e^{-(x-\mu)^2/(2\sigma^2)}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt, \qquad \left(\text{substitute } t = \frac{x - \mu}{\sigma}\right) \end{aligned}$$

showing that $P(Z \le z)$ is the standard normal cdf.

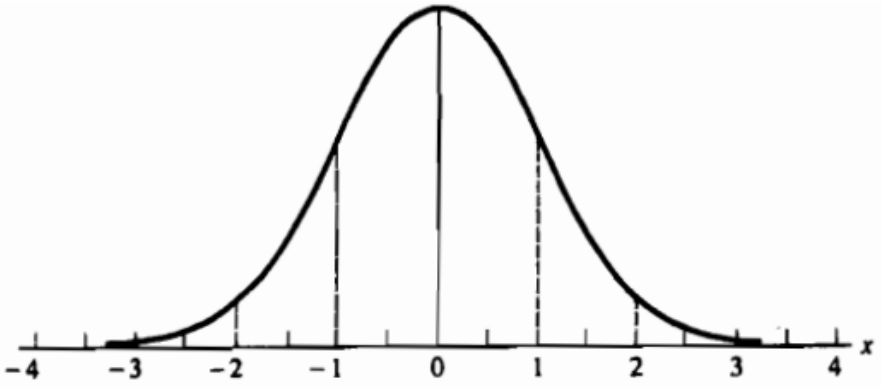


Figure 3.3.1. *Standard normal density*

Gaussian as Binomial Approximation

Example 3.3.2 (Normal approximation) Let $X \sim \text{binomial}(25, .6)$. We can approximate X with a normal random variable, Y, with mean $\mu = 25(.6) = 15$ and standard deviation $\sigma = ((25)(.6)(.4))^{1/2} = 2.45$. Thus

$$P(X \leq 13) \approx P(Y \leq 13) = P\left(Z \leq \frac{13-15}{2.45}\right) = P(Z \leq -.82) = .206,$$

How?

while the exact binomial calculation gives

$$P(X \leq 13) = \sum_{x=0}^{13} \binom{25}{x} (.6)^x (.4)^{25-x} = .267,$$

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.1	.46017	.45620	.45224	.44828	.44433	.44038	.43644	.43251	.42858	.42465
-0.0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414

Beta Distribution

- A random variable, X ~ $Beta(\alpha, \beta)$ is continuous distribution in [0, 1] with two shape parameters.

(3.3.16) $$f(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1, \quad \alpha > 0, \quad \beta > 0,$$

where $B(\alpha, \beta)$ denotes the beta function,

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx.$$

The beta function is related to the gamma function through the following identity:

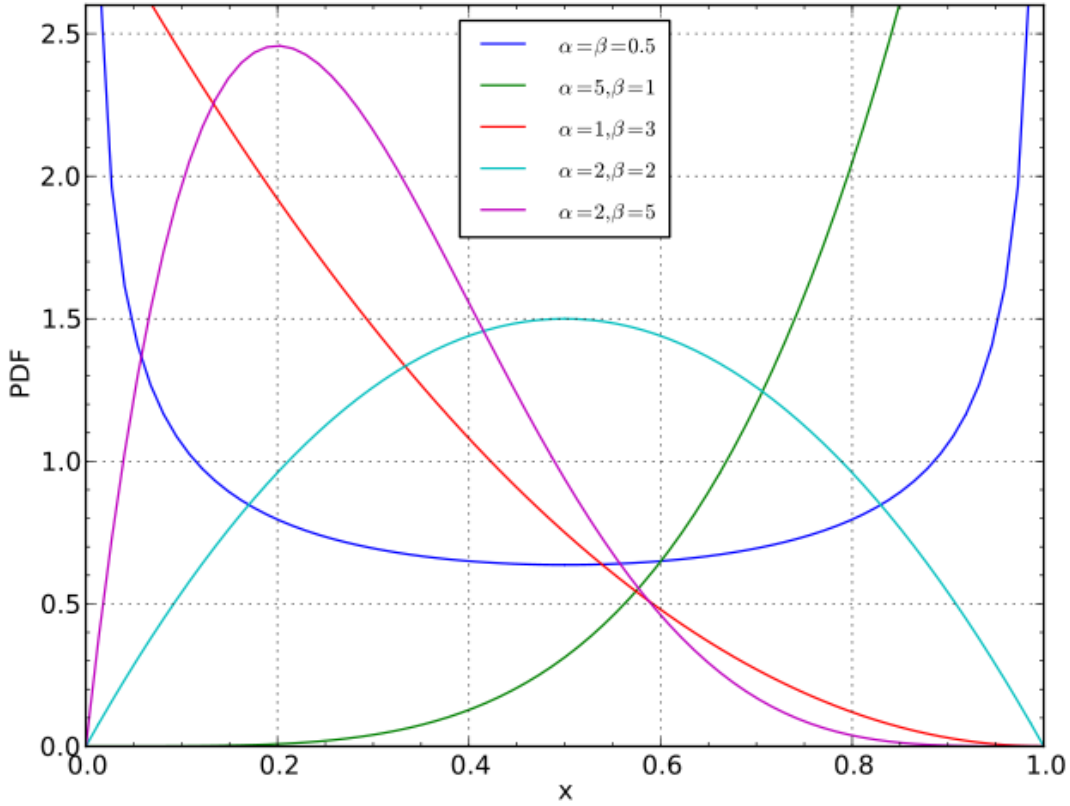
(3.3.17) $$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

- Beta distribution is often used to model proportions because they lie in [0, 1].
- Mean $E[X]$ and Variance $Var[X]$:

$$\mathrm{E}X = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \mathrm{Var}\,X = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Beta Distribution

- Comparing different Beta densities and shape parameters.
- Mean $E[X]$ and Variance $Var[X]$: $\mathrm{E}X = \frac{\alpha}{\alpha+\beta}$ and $\mathrm{Var}\,X = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
- **Q: What does the mean value tell us? How do we relate them to the curve?**



Random Vectors in Euclidean $\mathbb{R}^n$

- An n-dimensional random vector (which is also a random variable), $X =< X_1, X_2, \ldots X_n >$ is a function form sample space S to $\mathbb{R}^n$

- Bivariate case for two tosses of fair dice

Example 4.1.2 (Sample space for dice) Consider the experiment of tossing two fair dice. The sample space for this experiment has 36 equally likely points and was introduced in Example 1.3.10. For example, the sample point $(3, 3)$ denotes the outcome in which both dice show a 3; the sample point $(4, 1)$ denotes the outcome in which the first die shows a 4 and the second die a 1; etc. Now, with each of these 36 points associate two numbers, X and Y. Let

$$X = \text{ sum of the two dice} \quad \text{and} \quad Y = |\text{difference of the two dice}|.$$

For the sample point $(3, 3)$, $X = 3+3 = 6$ and $Y = |3 - 3| = 0$. For $(4, 1)$, $X = 5$ and $Y = 3$. These are also the values of X and Y for the sample point $(1, 4)$. For each of the 36 sample points we could compute the values of X and Y. In this way we have defined the bivariate random vector (X, Y).

Random Vectors in Euclidean $\mathbb{R}^n$

- Bivariate case for two tosses of fair dice

Having defined a random vector (X, Y), we can now discuss probabilities of events that are defined in terms of (X, Y). The probabilities of events defined in terms of X and Y are just defined in terms of the probabilities of the corresponding events in the sample space S. What is $P(X = 5 \text{ and } Y = 3)$? You can verify that the only two sample points that yield $X = 5$ and $Y = 3$ are $(4, 1)$ and $(1, 4)$. Thus the event "$X = 5$ and $Y = 3$" will occur if and only if the event $\{(4, 1), (1, 4)\}$ occurs. Since each of the 36 sample points in S is equally likely,

$$P(\{(4, 1), (1, 4)\}) = \frac{2}{36} = \frac{1}{18}.$$

Thus,

$$P(X = 5 \text{ and } Y = 3) = \frac{1}{18}.$$

- How does the joint distribution of (X, Y) look like?
- The joint PMF (or PDF) gives us the complete probability distribution of the random vector (X,Y) [i.e., the values it can take and the probabilities of those events (X,Y) attaining those values].

Joint and Marginal Distributions

- The joint distribution for (X, Y) in the previous example
- What happens if we care only about one random variable of our random vector? E.g., What is P(Y=0)?
- We need marginal distributions, i.e., P(Y=0, X is any allowable value).

		x										
		2	3	4	5	6	7	8	9	10	11	12
y	0	$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$
	1		$\frac{1}{18}$		$\frac{1}{18}$		$\frac{1}{18}$		$\frac{1}{18}$		$\frac{1}{18}$	
	2			$\frac{1}{18}$		$\frac{1}{18}$		$\frac{1}{18}$		$\frac{1}{18}$		
	3				$\frac{1}{18}$		$\frac{1}{18}$		$\frac{1}{18}$			
	4					$\frac{1}{18}$		$\frac{1}{18}$				
	5						$\frac{1}{18}$					

Table 4.1.1. *Values of the joint pmf* $f(x, y)$

Theorem 4.1.6 *Let* (X, Y) *be a discrete bivariate random vector with joint pmf* $f_{X,Y}(x, y)$. *Then the marginal pmfs of* X *and* Y, $f_X(x) = P(X = x)$ *and* $f_Y(y) = P(Y = y)$, *are given by*

$$f_X(x) = \sum_{y \in \Re} f_{X,Y}(x, y) \quad \text{and} \quad f_Y(y) = \sum_{x \in \Re} f_{X,Y}(x, y).$$

$$\begin{aligned} f_Y(0) &= f_{X,Y}(2,0) + f_{X,Y}(4,0) + f_{X,Y}(6,0) \\ &\quad + f_{X,Y}(8,0) + f_{X,Y}(10,0) + f_{X,Y}(12,0) \\ &= \frac{1}{6}. \end{aligned}$$

Joint and Marginal: Continuous Case

- Similarly, we can define joint and marginal for continuous random vectors

Definition 4.1.10 A function $f(x, y)$ from $\Re^2$ into $\Re$ is called a *joint probability density function* or *joint pdf of the continuous bivariate random vector* (X, Y) if, for every $A \subset \Re^2$,

$$P((X, Y) \in A) = \int_A \int f(x, y)\, dx\, dy.$$

The *marginal probability density functions* of X and Y are also defined as in the discrete case with integrals replacing sums. The marginal pdfs may be used to compute probabilities or expectations that involve only X or Y. Specifically, the marginal pdfs of X and Y are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy, \quad -\infty < x < \infty,$$

(4.1.3)

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx, \quad -\infty < y < \infty.$$

Any function $f(x, y)$ satisfying $f(x, y) \geq 0$ for all $(x, y) \in \Re^2$ and

$$1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy$$

Dirichlet Distribution

- Generalization of Beta. A multivariate K-dimensional random variable, $\mathbf{X} = < X_1, X_2, \ldots X_K > \sim Dir(K, \boldsymbol{\alpha} = < \alpha_1, \alpha_2, \alpha_3, \ldots \alpha_k >)$ on the K-1 simplex ($\mathbb{R}^{K-1}$) having the following density function.

$$f(x_1, \cdots, x_{K-1}; \alpha_1, \cdots, \alpha_K) = \frac{1}{\mathrm{B}(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}, \quad \mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}, \qquad \boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K).$$

- The density exists on the K-1 dimensional simplex defined by

$$x_1, \cdots, x_{K-1} > 0$$
$$x_1 + \cdots + x_{K-1} < 1$$
$$x_K = 1 - x_1 - \cdots - x_{K-1}$$

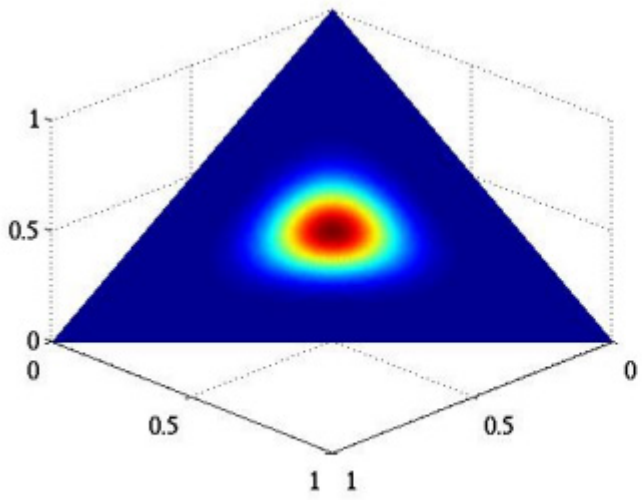
- Dirichlet distributions are often used (as priors) with multinomial/categorical distributions for modeling word emission.

- Mean $E[X_i] = \frac{\alpha_i}{\alpha_0}$ where $\alpha_0 = \sum \alpha_i$

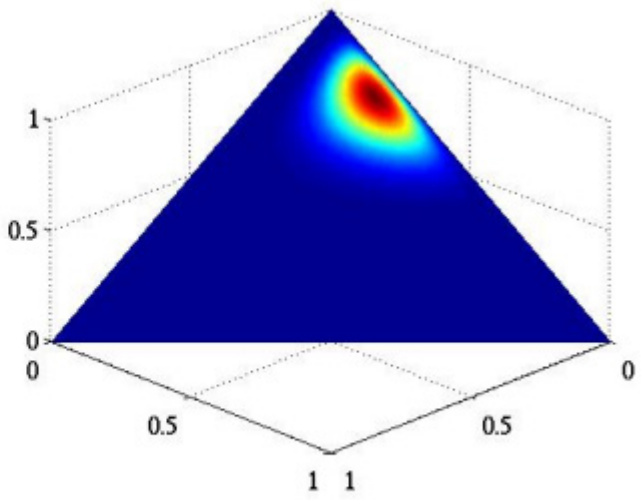
Dirichlet Distribution: Interpretation

- Consider the Dirichlet distribution of order K = 3.

- If we plot the samples of $\mathbf{X} = < X_1, X_2, X_3 > \sim Dir(K = 3, \boldsymbol{\alpha} =< \alpha_1, \alpha_2, \alpha_3 >)$ using a concentration heat map, we get



$\alpha = [10, 10, 10]$

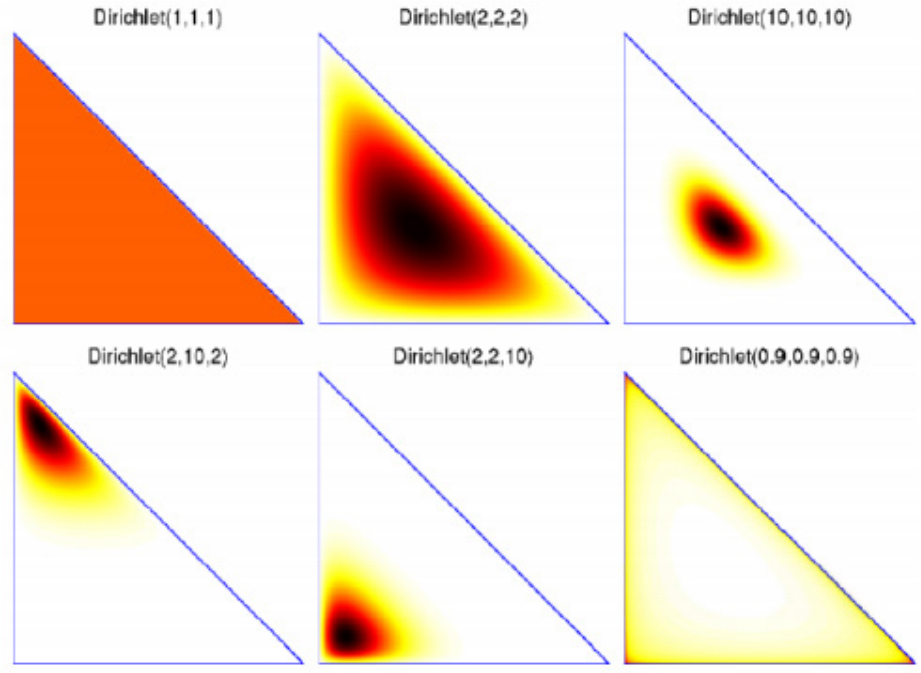


$\alpha = [2, 5, 15]$

- Base measure defines the mean distribution
- Concentration parameter (α) governs density. Values >/< 1 prefer dense/spare variates respectively (i.e., individual samples of a draw are close/far away from each other).

Dirichlet Distribution: Interpretation

- Another example of a Dirichlet distribution of order K = 3.

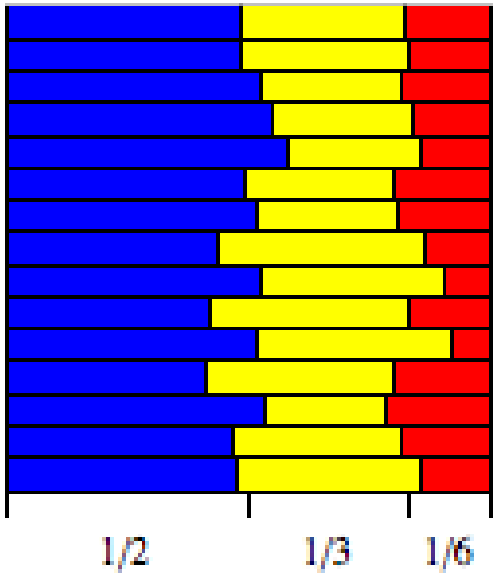


- Base measure defines the mean distribution
- Concentration parameter (α) governs density. Values >/< 1 prefer dense/spare variates respectively (i.e., individual samples of a draw are close/far away from each other).
- Also see D.Blei's tutorial (slides 32-39)

Dirichlet Distribution: String Cuts

- A more tangible example!

- Consider you want to cut a string of length = 1 unit to K=3 pieces of different lengths where each of the K=3 pieces had a designated average length ($\frac{1}{2}, \frac{1}{3}, \frac{1}{6}$ respectively) with some variance. How would different string cuts look like?



- The above cuts are nothing but samples of the Dirichlet distribution, $Dir(K = 3, \boldsymbol{\alpha} =< \frac{1}{2}, \frac{1}{3}, \frac{1}{6} >)$.