1 Probability Theory

1.1 Classical Terminology [DeGroot 1.3-1.4]

- Experiment E.g. toss a coin 10 times or sequence a genome
- Outcome A possible result of an experiment,
  - E.g HHTHTTHHHT or ACGCTTATC
- Sample space The set of all possible outcomes of some experiment
  - E.g. $\{H, T\}^{10}$ or $\{A, C, G, T\}^*$.
- Event Any subset of the sample space
  - E.g. ≥ 4 heads; DNA seqs w/no run of > 50 As.

1.2 Goals

- Compute probabilities of events given probability of each possible outcome.
- Revise probabilities when new info is obtained

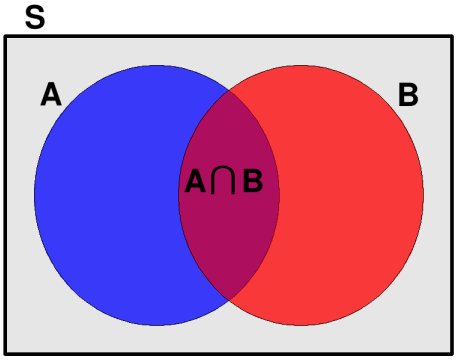
1.3 Definions, axioms, simple theorems [DeGroot 1.5]

- If S is a sample space and $A \subseteq S$ is an event, then $Pr(A)$ is a number representing the probability that the event will occur – i.e., the probability that the outcome of the experiment will be in the set of outcomes constituting the event.
- Axiom 1 For any event A, $Pr(A) \geq 0$
- Axiom 2 If S is a sample space, $Pr(S) = 1$.
- Events A, B are disjoint iff $A \cap B = \emptyset$. The set $\{A_1, A_2, ...\}$ is disjoint iff every pair is disjoint. Disjoint events are mutually exclusive.
- Axiom 3 For any finite or infinite collection of disjoint events $A_1, A_2, ...,$

$$Pr(\cup_i A_i) = \sum_i Pr(A_i)$$

- Theorem $Pr(\emptyset) = 0$.
- Theorem For any event A where A^c is the complement of A, $Pr(A^c) = 1 - Pr(A)$.

- Theorem For any event A, $0 \leq Pr(A) \leq 1$.
- Theorem If $A \subseteq B$, then $Pr(A) \leq Pr(B)$.
- Theorem $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$.



2 Finite Sample Spaces [DeGroot 1.6]

- When $S = \{s_1, ..., s_n\}$, s_i is called an elementary event. $p_i \equiv Pr(s_i)$.
- The axioms imply
  - $p_i \geq 0, i = 1$ to n
  - $\sum_{i=1}^{n} p_i = 1$
- If $p_i = 1/n$ for $i = 1, ..., n$, the events $\{s_i\}$ are called equiprobable.
  - Then if A is an event and $|A| = M$, $Pr(A) = M/n$.

3 Permutations - Sampling without replacement [DeGroot 1.7]

- E.g. picking 3 cards from a deck of n.
  - Outcomes are 3-place vectors, each place filled by a different card
  - # of possible outcomes is $n(n-1)(n-2)$
  - Each outcome is called a permutation. (vector would be more intuitive.)
- When k cards are selected, # of permutations (vectors) is
  - $P_{n,k} = n(n-1)...(n-k+1)$
  - "The number of permutations of n elements taken k at a time."

- Another way to write this is $P_{n,k} = \frac{n!}{(n-k)!}$
- When k = n, $P_{n,k} = n(n-1)...1 = n!$. ($0! = 1$ by convention)

4 Combinations [DeGroot 1.8]

- How many distinct subsets of k elements can be chosen from n?
- Each subset is a combination of k elts. (subset would be more intuitive.)
- Order is irrelevant. No two combos (subsets) have the same elements.
- How many combos (subsets) of k elements can be selected from a set of n ($C_{n,k}$) elements?
  - E.g. the combos (subsets) of 2 elts from $\{a, b, c, d\}$ are

$$\{a, b\}\{a, c\}\{a, d\}\{b, c\}\{b, d\}\{c, d\} : C_{4,2} = 6$$

- In general, $C_{n,k} = \frac{n!}{(n-k)!k!}$. Also written $\binom{n}{k}$.
- Derivation: For each combo of k elements from n, there are $k!$ permutations. So # of permutations of n elements= $k!$ # of combos, or $P_{n,k} = k!C_{n,k}$.

$$C_{n,k} = \frac{P_{n,k}}{k!} = \frac{n!}{(n-k)!k!}$$

- Example: Probability of exactly 3 heads on 10 tosses of a fair coin. 2^{10} equiprobable outcomes, so

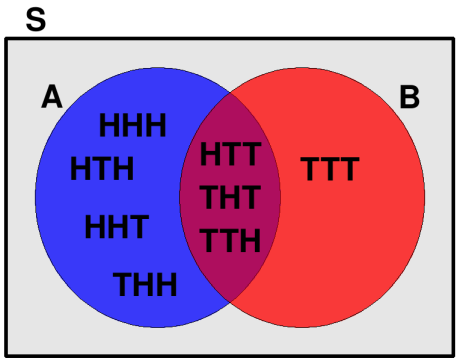
$$Pr(3H \text{ out of } 10 \text{ tosses}) = \frac{\binom{10}{3}}{2^{10}}$$

5 Conditional Probability [DeGroot 2.1]

- Suppose A and B are events in S and we know that the outcome of a given experiment was in B (i.e. B occurred).
- What is the probability that the outcome was also in A?
- Conditional probability of A given B, written $Pr(A|B)$.
- Definition: $Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$ (Memorize this!)

- Example: Given that the number of heads on tossing three fair coins is less than two, what is the probability that it is greater than 0?

$$A : \#H > 0, B : \#H < 2$$



$$Pr(A \cap B) = 3/8, Pr(B) = 4/8, Pr(A|B) = \frac{3/8}{4/8} = 3/4$$

Result is same as direct count of remaining outcomes.

6 Independence [DeGroot 2.2]

- If $Pr(A|B) = Pr(A)$, we say A is independent of B.
  - Then $\frac{Pr(A \cap B)}{Pr(B)} = Pr(A)$ so $Pr(A \cap B) = Pr(A)Pr(B)$.
- If A is independent of B, then B is independent of A.
- A and B are independent.

7 Random Variables [DeGroot 3.1]

- Definition: A random variable (r.v.) is a function from a sample space S into the real numbers ($\mathbb{R}$)
  - E.g. flip a coin 10 times. $S = \{H, T\}^{10}$. For any $s \in S$, let $X(s)$ be the # times H occurs in s. X is a r.v. w/range $\{0..10\}$.
- Equations and inequalities involving random variables are interpreted as events. If X is a r.v., S is its sample space, and $a \in \mathbb{R}$, then $X = a$ is equivalent to the event $\{s \in S$ such that $X(s) = a\}$).
  - Therefore, if $a \in \mathbb{R}, Pr(X = a) \equiv Pr(\{s : X(s) = a\})$.

  - Any statement about the values of an r.v. can be interpreted as an event. E.g. $X < 2.7$ is equivalent to $\{s$ such that $X(s) < 2.7\}$).
  - Any fact that is true of all events is also true of all statements about the values of r.v.'s.

- A r.v. whose range is a finite or countably infinite set $\{x_1, x_2, ...\}$ is called discrete. Otherwise, it is continuous.

- The probability function of a discrete r.v. X is the function that maps $a \in \mathbb{R}$ to $Pr(X = a)$.

- We denote the probability function of a r.v. X as $f(x) \equiv \Pr(X = x)$, where the lower case letter x ranges over the real numbers.

- Also write $\Pr(x)$ as shorthand for $\Pr(X = x)$. Probability fn of X is also called the *distribution of* X.
  - Example: $\Pr(x) = 1/2^x$ for $x \in \{1, 2, ...\}$ is short for $\Pr(X = x) = 1/2^x$

- If $H \subseteq S$ is a possible outcome of an experiment or a set of possible outcomes it is convenient to write $Pr(X = H)$, even though H is not a number. This is equivalent to $Pr(Y = 1)$ where Y is a r.v. defined by $Y(s) = 1$ if and only if $s \in H$.

7.1 Example: Bernoulli and binomial distributions

- Consider an experiment in which a single coin is flipped. Let X be a random variable that maps heads to 1 and tails to 0.

- The probability function for X has the form $\Pr(X = 1) = \Theta$ and $\Pr(X = 0) = (1 - \Theta)$, for some $\Theta \in [0, 1]$ (the interval between 0 and 1). *Bernoulli distribution* on X.

- If the coin is flipped N times, the probability of a specific sequence of outcomes $x_1, \ldots, x_n$ $(x_i \in 0, 1)$ is

$$\Theta^{\sum x_i}(1 - \Theta)^{N - \sum x_i}$$

- Let Y be an r.v. that maps outcomes of N samples from a Bernoulli distribution, $x_1, \ldots x_n$, to their sum. The probability function of Y is

$$Pr(Y = y) = \binom{N}{Y} \Theta^Y (1 - \Theta)^{N-Y}$$

for $y \in 0 \ldots N$ and 0 otherwise. I.e. the number of different binary sequences with Y 1's times the probability of any particular binary sequence with Y 1's.

7.2 Example: Hypergeometric distribution

- Genomics experiments frequently yield sets of genes. E.g., the set of genes whose expression levels are elevated in liver cancer cells.
- Most basic basic analysis is to ask whether certain types of genes are present in the result set more often than would be expected by chance. E.g. genes in a given signaling pathway.
- G: all genes in a particular pathway (or other gene set) of interest G.
- Genome contains n genes of which s are in G. The experiment yields a set of d genes of which x are in G.
- Given n, s, and d, what is the probability that exactly x of the d will be in G?
  - X is a discrete r.v. representing the number of genes resulting from the experiment that are in G. What probability function would make a good model for this question?
  - How many total outcomes are possible, given that d genes are selected from n? (Think combinations.)
  - How many ways are there in which to select x genes from among the s genes that belong to the pathway?
  - How many outcomes are there in which $d - x$ genes from among the $n - s$ genes that do not belong to the pathway?
  - What fraction of total outcomes have exactly x of the d genes selected overlapping the s genes that belong to the pathway?
  - Hypergeometric distribution with parameters n, s, and d. All must be integers of which $n > s, n > d$, and $d > x$.

7.3 Joint probability distributions

- Multiple r.v.'s can be defined on outcomes of the same experiment. E.g. sampling from the space of all people in the United States, X might represent the person's height (in centimeters) and Y his weight (in kilos).

Intuitively, X and Y are not independent – knowing a person's height is useful for guessing his weight, and vice-versa.

- Whenever we discuss two random variables, assume they are defined on outcomes of the same experiment.
- $Pr(x, y)$, the joint distribution of X and Y, is the probability that the outcome s of an experiment is such that $X(s) = x$ and $Y(s) = y$.
- For discrete r.v.'s X and Y,
  - $Pr(x|y) = \frac{Pr(x,y)}{Pr(y)}$ if $\Pr(y) > 0$; if $\Pr(y) = 0$ then $\Pr(x|y)$ is undefined.
  - X and Y are independent iff $Pr(x|y) = Pr(x)$ for all $x, y \in \mathbb{R}$

7.4 Continuous Random Variables

- For a continuous r.v. Θ, probability of any particular value is zero.
- Use a probability density function (p.d.f.) $f(\Theta)$ where

$$\Pr(y \leq \Theta \leq z) = \int_y^z f(\Theta)\, d\Theta$$

  (i.e. area under the pdf between y and z).
- Note: $Pr(-\infty \leq \Theta \leq +\infty) = 1$
- Example: Uniform p.d.f $f(\Theta) = 1$ for $0 \leq \Theta \leq 1$, $f(\Theta) = 0$ elsewhere.

$$\Pr(y \leq \Theta \leq z) = z - y$$

- Example: exponential p.d.f $f(\Theta) = ce^{-c\Theta}$ for $0 \leq \Theta$, $f(\Theta) = 0$ elsewhere. c is a positive constant affecting the shape of the exponential.

$$\Pr(y \leq \Theta \leq z) = -e^{-cz} + e^{-cy}$$

- For continuous r.v.'s X and Y with joint p.d.f. $f(x, y)$ and *marginal* p.d.f.s $f_1(x)$ and $f_2(y)$, we define the *conditional p.d.f.* $g(x|y)$ as

$$g(x|y) = \frac{f(x, y)}{f_2(y)}$$

- The p.d.f.s of continuous r.v.s can often be treated like the p.f.s of discrete r.v.s by replacing summations with integrals.

8 Four rules for manipulating probability expressions

8.1 Chain rule

Example:

$$\begin{aligned} Pr(x_1, x_2, x_3) &= \frac{Pr(x_1, x_2, x_3)}{Pr(x_2, x_3)} \frac{Pr(x_2, x_3)}{Pr(x_3)} Pr(x_3) \\ &= Pr(x_1|x_2, x_3) Pr(x_2|x_3) Pr(x_3) \end{aligned}$$

8.2 Bayes rule

Example:

$$\begin{aligned} Pr(x_1|x_2) Pr(x_2) &= Pr(x_1, x_2) \\ &= Pr(x_2|x_1) Pr(x_1) \\ Pr(x_1|x_2) &= \frac{Pr(x_2|x_1) Pr(x_1)}{Pr(x_2)} \end{aligned}$$

Similarly,

$$Pr(x_1|x_2, x_3) = \frac{Pr(x_2|x_1, x_3) Pr(x_1, x_3)}{Pr(x_2, x_3)}$$

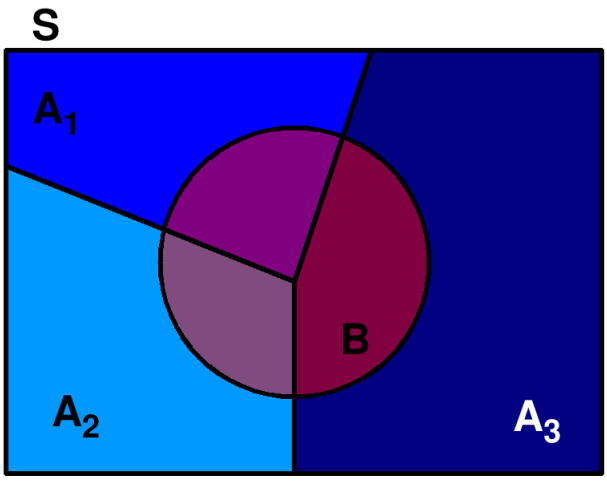
Hence,

$$Pr(x_1|x_2, x_3) = \frac{Pr(x_2|x_1, x_3) Pr(x_1|x_3)}{Pr(x_2|x_3)}$$

Extra variables behind the conditioning are carried along.

8.3 Summing out (Marginalizing)

- Suppose $A_1, A_2, ...$ are disjoint events w/union S, the sample space.
- Then $(B \cap A_1), (B \cap A_2), ...$ are disjoint, w/ union B.



$$\sum_i Pr(B \cap A_i) = Pr(\cup_i (B \cap A_i))$$
$$= Pr(B)$$

- Let X, Y be discrete r.v.'s and let $y_1, y_2, ...$ be all possible values of Y.
- Let $A_i = \{s|Y(s) = y_i\}$ and $B = \{s|X(s) = x\}$
- The A_i are disjoint, so $\sum_i Pr(B \cap A_i) = Pr(B)$
- Translating back into r.v. notation: $\sum_i Pr(x, y_i) = Pr(x)$

8.4 Exhaustive Conditionalization

$$\sum_y Pr(x|y)Pr(y) = \sum_y Pr(x, y)$$
$$= Pr(x)$$

9 Parameter estimation

- Suppose the probability function of a random variable belongs to a known family of distributions. Figure out which member of the family.
- Example: a biased coin turns up heads w/ uknown prob $\Theta, 0 \leq \Theta \leq 1$.
  - Determine Θ, the parameter.
  - (Sometimes Θ stands for a vector of parameters)

9.1 Maximum likelihood estimation

- Likelihood function: Given a set of observed outcomes O, can compute $Pr(O|\Theta)$ as a function of Θ.
  - Example: $Pr(x \text{ heads in } N \text{ flips } |\Theta) = \frac{N!}{x!(N-x)!}\Theta^x(1-\Theta)^{N-x}$
  - To emphasize that the likelihood is a function of the parameters once the observations are known, the notation $L(\Theta|N,x)$ is sometimes used.
- When there is little data, ML tends to underestimate probabilities of rare events, relative to our intuition.
- Example: Estimate probs for an unfair die from 10 rolls.
  - Suppose die has 6 nearly symmetrical faces.
  - Observations: 1,3,4,2,4,6,2,1,2,2
  - ML estimate for $Pr(5) = \frac{0}{10} = 0$. Implausible given die shape.
- Maximum likelihood estimate is the high point of the likelihood function.
  - Given observations O, $\hat{\Theta}_{ML} = \arg\max_\Theta Pr(O|\Theta)$
  - E.g.: coin flipping, x heads in N flips. $\hat{\Theta}_{ML} = \frac{x}{N}$.
  - To prove, find the maximum of the log of the likelihood by equating its derivative to zero and solving for Θ.

9.2 Maximum a postiori estimation [DeGroot 7.1, 7.2]

- The maximium a posterior (MAP) estimate is the most probable value of the parameter – i.e. the max of the posterior distribution.
- Posterior Distribution on Θ
  - If Θ has finitely many possible values we can treat it as a discrete r.v.:

$$Pr(\Theta|O) = \frac{Pr(O|\Theta)Pr(\Theta)}{Pr(O)}$$

    For a fixed set of observations,

$$Pr(\Theta|O) \propto Pr(O|\Theta)Pr(\Theta)$$

  - Posterior is proportional to likelihood times prior
  - Prior can be uninformed – don't know so use uniform prior

– Or used to encode actual knowledge about a problem

– If Θ is a continuous r.v. with prior p.d.f. $g(\Theta)$ then its posterior p.d.f. $f(\Theta|O)$ is given by:

$$f(\Theta|O) = \frac{Pr(O|\Theta)g(\Theta)}{Pr(O)}$$

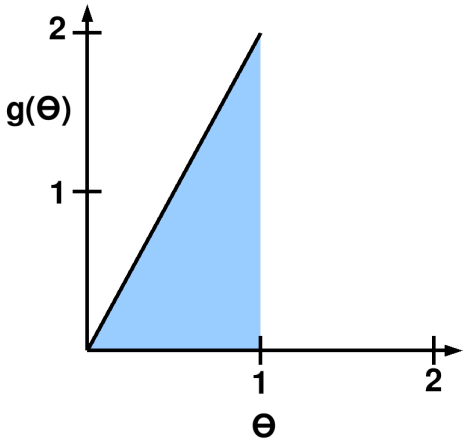
- Example: If Θ is the probability of flipping heads on a coin (i.e. the parameter of a Bernoulli distribution), possible priors $g(\Theta)$ include:

  – The uniform p.d.f: $g(\Theta) = 1$ for $0 \leq \Theta \leq 1$, $g(\Theta) = 0$ elsewhere. Then for any $y, z \in [0, 1]$:

  $$Pr(y \leq \Theta \leq z) = z - y$$

  – A prior representing an expectation of coins biased toward more heads:

  $$g(\Theta) = \begin{cases} 2\Theta & 0 \leq \Theta \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$



  – Using the latter prior, the posterior p.d.f. f is given by:

  $$f(\Theta|x \text{ heads in } N \text{ flips}) \propto \underbrace{\frac{N!}{x!(N-x)!}\Theta^x(1-\Theta)^{N-x}}_{\text{likelihood}} \underbrace{2\Theta}_{\text{prior}} \text{ for } 0 \leq \Theta \leq 1$$

  $$\propto \Theta^{x+1}(1-\Theta)^{N-x} \text{ (or zero if } \Theta < 0 \text{ or } \Theta > 1)$$

9.3 Conjugate Priors [DeGroot 5.8 and 7.3]

- The two example priors given above, the uniform prior on $[0, 1]$ and

$$g(\Theta) = \begin{cases} 2\Theta & 0 \leq \Theta \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

are both *conjugate priors* for the binomial likelihood function because the prior and the posterior have the same functional form. Distributions with this form are called *beta distributions*.

- In general, a beta distribution on Θ has the form

$$\text{beta}(\Theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\Theta^{a-1}(1-\Theta)^{b-1}$$

for $\Theta \in [0,1]$ and zero for $\Theta \notin [0,1]$. When n is a positive integer, $\Gamma(n) = (n-1)!$. Thus, when a and b are positive integers,

$$\text{beta}(\Theta|a,b) = \frac{(a+b-1)!}{(a-1)!(b-1)!}\Theta^{a-1}(1-\Theta)^{b-1}$$

- The p.d.f. that is uniform on $[0,1]$ and 0 elsewhere is $\text{beta}(\Theta|a=1,b=1)$.
- The p.d.f. that is 2Θ on $[0,1]$ and 0 elsewhere is $\text{beta}(\Theta|a=2,b=1)$.
- When a beta prior p.d.f. with integer parameters a, b, is multiplied by a binomial likelihood of x heads in N trials, the posterior p.d.f. on Θ is

$$f(\Theta|x,N,a,b) \propto \Theta^{x+a-1}(1-\Theta)^{N-x+b-1}$$

for $\Theta \in [0,1]$ and 0 elsewhere. The proportionality constant ensures that the integral of the posterior over all values of Θ is 1. It can be shown (DeGroot, Section 5.10) that this implies the normalized posterior p.d.f. is:

$$\frac{(N+a+b-2)!}{(x+a-1)!(N-x+b-1)!}\Theta^{x+a-1}(1-\Theta)^{N-x+b-1} \text{ (or zero if } \Theta < 0 \text{ or } \Theta > 1)$$

- Thus, when a prior p.d.f. is a beta distribution with parameters a, b and the likelihood is binomial, the posterior p.d.f. is a beta distribution with parameters $x+a$ and $N+a+b$. For this reason, the beta family of priors is said to be conjugate to binomial likelihood functions.
- The prior and posterior are always beta distributions – only the parameters change when data arrive.
- The beta distribution with parameters a and b takes on its maximum value when $\Theta = \frac{a-1}{a+b-2}$. (The derivation is just like the derivation of $\hat{\Theta}_{ML}$.)
  - Therefore, after observing x heads in N flips, the maximum a posteriori estimate is

$$\hat{\Theta}_{MAP} = \frac{x+a-1}{N+a+b-2}$$

  - Can continuously update the MAP estimate as new data come in.

- a and b can be thought of as pseudocounts or "imaginery data" for heads and tails, respectively.
- In the limit of large N, $\hat{\Theta}_{MAP}$ and $\hat{\Theta}_{ML}$ converge the same value.
- The larger a and b are, the more data it takes to overwhelm the prior. When they are large, the prior is said to be strong.

9.4 Posterior Predictive Distributions

- With a posterior pdf $g(\Theta|O)$ on parameters Θ given observations O, you can compute a distribution on future observations that does not depend on assuming any particular parameter values.

$$\Pr(X = x|O) = \int_{-\infty}^{\infty} \Pr(X = x|\Theta) g(\Theta|O) d\Theta$$

- This is exhaustive conditionalization on a continuous r.v. Θ.
- When the prior on Θ is beta(a, b), Y is a binary r.v. with $\Pr(Y = 1) = \Theta$, and N Bernoulli experiments have been observed in which $Y = 1$ occurred x times, the previous equation becomes:

$$\begin{aligned}
\Pr(Y = 1|O) &= \int_0^1 \Theta \frac{(N + a + b - 1)!}{(x + a - 1)!(N - x + b - 1)!} \Theta^{x+a-1} (1 - \Theta)^{N-x+b-1} d\Theta \\
&= \frac{(N + a + b - 1)!}{(x + a - 1)!(N - x + b - 1)!} \int_0^1 \Theta^{x+a} (1 - \Theta)^{N-x+b-1} d\Theta \\
&= \frac{(N + a + b - 1)!}{(x + a - 1)!(N - x + b - 1)!} \frac{(x + a)!(N - x + b - 1)!}{(N + a + b)!} \\
&= \frac{x + a}{N + a + b}
\end{aligned}$$

- The integral is derived in DeGroot section 5.10.
- Compare this posterior predictive to the MAP estimate of Θ.
  - MAP estimate of Θ is a point estimate while the posterior predictive considers all possible values of Θ.
  - As empirical data N and x grow large, the MAP estimate, posterior predictive, and ML estimate all converge on the same value.

9.5 Multinomial Distributions

- Formulas presented above for binary r.v.'s also apply to any variable with a finite number of possible values, with minor / obvious modifications.
- Binomial likelihood becomes multinomial.
- beta prior becomes *Dirichlet* prior.
- rv's with large numbers of possible values are common in bioinformatics.
- With many possible values, observations are often not sufficent to estimate parameters by ML.
- Dirichlet priors are common and important.

10 Expectation [DeGroot 4.1, 4.2]

- The expectation of a r.v. is its long-run expected average. For a discrete r.v.:
- $E[X] \equiv \sum_x Pr(x)x$
- If f is a real-valued function of X, $E[f(X)] = \sum_x Pr(x)f(X)$
- Example: If X is uniformly distributed over {1,2,3,4} and $f(X) = X^2$

$$E[f(X)] = \frac{1}{4} \cdot 1^2 + \frac{1}{4} \cdot 2^2 + \frac{1}{4} \cdot 3^2 + \frac{1}{4} \cdot 4^2 = \frac{30}{4}$$

- Expectation is linear. If $X_1, ...X_n$ are r.v.'s then $E(X_1 + ... + X_n) = E(X_1) + ...E(X_n)$.

11 Expectation Maximization

- Method for estimating parameters from experiments in which some key aspects of the outcomes are hidden.
- Example: A bag contains two types of six-sided dice. Each type has it's own probability of rolling the faces $1 \ldots 6$.
  - A series of experiments is carried out. In each one, a die is chosen at random from the bag, rolled m times, and put back in the bag.

- You are told the number that comes up on each roll, but not which type of die was used. (This is the hidden variable.)
- You must estimate the probability of selecting each a die of each type and the probability of rolling each number on a die of each type.

- Key idea 1 of EM approach
  - Let count(k,A) be the number of times face k was rolled on a die of type A.
  - If we knew the counts we could estimate Pr(face k|die type A) by M.L.:

$$\frac{\text{count}(k, A)}{\sum_{k=1}^{6} \text{count}(k, A)}$$

  - If we knew the counts we could also estimate Pr(die type A) by M.L.:

$$\frac{\sum_{k=1}^{6} \text{count}(k, A)}{\sum_{k=1}^{6} \text{count}(k, A) + \sum_{k=1}^{6} \text{count}(k, B)}$$

  - Can't get count(k, A) or count(k, B) from the data because die type is hidden.
  - Instead, use the expected counts with respect to the posterior distribution on the hidden variable (die type of each experiment).

- Formal derivation of expected counts for dice example.
  - Let $f_{i,j}$ be the face shown on the j^{th} roll of the i^{th} die selected.
  - Let D_i be the type of the i^{th} die chosen (A or B). Hidden variable.
  - Let

$$C_{i,j,k,A} = \begin{cases} 1 & \text{if } f_{i,j} = k \text{ and } D_i = A \\ 0 & \text{otherwise} \end{cases}$$

  - Then

$$\text{count}(k, A) = \sum_{i,j} C_{i,j,k,A}$$

  - So

$$
\begin{aligned}
E[\text{count}(k, A)] &= E\left[\sum_{i,j} C_{i,j,k,A}\right] \\
&= \sum_{i,j} E\left[C_{i,j,k,A}\right] \\
&= \sum_{i,j} \left(\Pr(C_{i,j,k,A} = 0) \cdot 0 + \Pr(C_{i,j,k,A} = 1) \cdot 1\right) \\
&= \sum_{i,j} \Pr(C_{i,j,k,A} = 1) \\
&= \sum_{i,j \text{ s.t. } f_{i,j}=k} \Pr(D_i = A)
\end{aligned}
$$

- We have some observations relevant to $\Pr(D_i = A)$, so we will use them by calculating the posterior probability $\Pr(D_i = A | f_{i,1}, f_{i,2}, \ldots)$.
- Key idea 2 of EM approach
  - Assume you know the probabilities. Start with random guesses.
  - Repeatedly improve probabilities by making max likelihood estimates from expected counts.
  - Use probabilities from the last round to compute expectations for the next.