CSE383M and CS395T, Spring, 2013
Problems from Segments 1—18 (previously listed in the Course Wiki)

Segment 1

To Calculate

1. Prove that $P(ABC) = P(B)P(C|B)P(A|BC)$.

2. What is the probability that the sum of two dice is odd with neither being a 4?

To Think About

1. First-order logic is a type of propositional calculus with propositions a, b, c and quantifier symbols $\forall$ and $\exists$. This allows statements like "Socrates is a philosopher", "Socrates is a man", "There exists a philosopher who is not a man", etc. Can you use first-order logic as a calculus of inference? Is it the same as using the probability axioms? If not, then which of Cox's suppositions is violated?

2. You are an oracle that, when asked, says "yes" with probability P and "no" with probability $1 - P$. How do you do this using only a fair, two-sided coin?

3. For the trout/minnow problem, what if you want to know the probability that the Nth fish caught is a trout, for N=1,2,3,... What is an efficient way to set up this calculation? (Hint: If you ever learned the word "Markov", this might be a good time to remember it!)

Segment 2

To Calculate

1. If the knight had captured a Gnome instead of a Troll, what would his chances be of crossing safely?

2. Suppose that we have two identical boxes, A and B. A contains 5 red balls and 3 blue balls. B contains 2 red balls and 4 blue balls. A box is selected at random and exactly one ball is drawn from the box. What is the probability that it is blue? If it *is* blue, what is the probability that it came from box B?

To Think About

1. Do you think that the human brain's intuitive "inference engine" obeys the commutativity and associativity of evidence? For example, are we more likely to be swayed by recent, rather than older, evidence? How can evolution get this wrong if the mathematical formulation is correct?

2. How would you simulate the Knight/Troll/Gnome problem on a computer, so that you could run it 100,000 times and see if the Knights probability of crossing safely converges to 1/3?

3. Since different observers have different background information, isn't Bayesian inference useless for making social decisions (like what to do about climate change, for example)? How can there ever be any consensus on probabilities that are fundamentally subjective?

Segment 3

To Calculate

1. The slides used a symmetry argument ("relabeling") to simplify the calculation. Redo the calculation without any such relabeling. Assume that the doors have big numbers "1", "2", and "3" nailed onto them, and consider all possibilities. Do you still have to make an assumption about Monty's preferences (where the slide assumed 1/2)?

To Think About

1. Lawyers are supposed to be able to argue either side of a case. What is the best argument that you can make that switching doors can't possibly make any difference? In other words, how cleverly can you hide some wrong assumption?

2. We stated the problem as *requiring* the host to offer the contestant a chance to switch. But what if the host can offer that chance, or not, as he sees fit? Then, when offered the chance, should you still switch? (Spoiler alert: see this New York Times interview with Monte Hall.)

3. Mr. and Mrs. Smith tell you that they have two children, one of whom is a girl.
(a) What is the probability that the other child is a girl?
Mr. Smith then shows you a photo of his children on his iPhone. One is clearly a girl, but the other one's face is hidden behind the family dog, and you can't tell their gender.
(b) What is the probability that the hidden child is a girl?
(c) If your answers to (a) and (b) are different, explain why there is a difference.

Segment 4

To Calculate

1. Evaluate $\int_0^1 \delta(3x-2)dx$

2. Prove that $\delta(ax) = \frac{1}{a}\delta(x)$.

3. What is the numerical value of $P(A|S_B I)$ if the prior for $p(x)$ is a massed prior with half the mass at $x = 1/3$ and half the mass at $x = 2/3$?

To Think About

1. With respect to problem 3, above, since x is a probability, how can choosing x=1/3 half the time, and x=2/3 the other half of the time be different from choosing x=1/2 all the time?

2. Suppose A is some event that we view as stochastic with P(A), such as "will it rain today?". But the laws of physics (or meteorology) say that A actually depends on other weather variables X, Y, Z, etc., with conditional probabilities P(A|XYZ...). If we repeatedly sample just A, to naively measure P(A), are we correctly marginalizing over the other variables?

Segment 5

To Calculate

1. You throw a pair of fair dice 10 times and, each time, you record the total number of spots. When you are done, what is the probability that exactly 5 of the 10 recorded totals are prime?

2. If you flip a fair coin one billion times, what is the probability that the number of heads is between 500010000 and 500020000, inclusive? (Give answer to 4 significant figures.)

To Think About

1. Suppose that the assumption of independence (the first "i" in "i.i.d.") were violated. Specifically suppose that, after the first Bernoulli trial, every trial has a probability Q of simply reproducing the immediately previous outcome, and a probability (1-Q) of being an independent trial. How would you compute the probability of getting n events in N trials if the probability of each event (when it is independent) is p?

2. Try the Mathematica calculation on slide 5 without the magical "GenerateConditions -> False". Why is the output different?

Segment 6

To Calculate

1. Write down an explicit expression for what the slides denote as bin(n,N,r).

2. There is a small error on slide 7 that carries through to the first equation on slide 8 and the graph on slide 9. Find the error, fix it, and redo the graph of slide 9. Does it make a big difference? Why or why not?

To Think About

1. Suppose you knew the value of r (say, r = 0.0038). How would you simulate many instances of the Towne family data (e.g., the tables on slides 4 and 5?

2. How would you use your simulation to decide if the assumption of ignoring backmutations (the red note on slide 7) is justified?

3. How would you use your simulation to decide if our decision to trim T2, T11, and T13 from the estimation of r was justified? (This question anticipates several later discussions in the course, but thinking about it now will be a good start.)

Segment 7

To Calculate

1. Prove the result of slide 3 the "mechanical way" by setting the derivative of something equal to zero, and solving.

2. Give an example of a function $p(x)$, with a maximum at $x = 0$, whose third moment M_3 exists, but whose fourth moment M_4 doesn't exist.

3. List some good and bad things about using the median instead of the mean for summarizing a distribution's central value.

To Think About

1. This segment assumed that $p(x)$ is a known probability distribution. But what if you know $p(x)$ only experimentally. That is, you can draw random values of x from the distribution. How would you estimate its moments?

2. High moments (e.g., 4 or higher) are algebraically pretty, but they are rarely useful because they are very hard to measure accurately in experimental data. Why is this true?

3. Even knowing that it is useless, how would you find the formula for I_8, the eighth semi-invariant?

Segment 8

To Calculate

1. In Segment 6 (slide 8) we used the improper prior $1/r$. Show that this is just a limiting case of a (completely proper) Lognormal prior.

2. Prove that $\mathrm{Gamma}(\alpha, \beta)$ has a single mode at $(\alpha - 1)/\beta$ when $\alpha \geq 1$.

3. Show that the limiting case of the Student distribution as $\nu \to \infty$ is the Normal distribution.

To Think About

1. Suppose you have an algorithm that can compute a CDF, $P(x)$. How would you design an algorithm To Calculate its inverse (see slide 9) $x(P)$?

2. The lifetime t of a radioactive nucleus (say Uranium 238) is distributed as the Exponential distribution. Do you know why? (Hint: What is the distribution of an Exponential(β) random variable *conditioned on* its being greater than some given value?)

Segment 9

To Calculate

1. Use characteristic functions to show that the sum of two independent Gaussian random variables is itself a Gaussian random variable. What is its mean and variance?

2. Calculate (don't just look up) the characteristic function of the Exponential distribution.

To Think About

1. Learn enough about contour integration to be able to make sense of Saul's explanation at the bottom of slide 7. Then draw a picture of the contours, label the pole(s), and show how you calculate their residues.

2. Do you think that characteristic functions are ever useful computationally (that is, not just analytically to prove theorems)?

Segment 10

To Calculate

1. Take 12 random values, each uniform between 0 and 1. Add them up and subtract 6. Prove that the result is close to a random value drawn from the Normal distribution with mean zero and standard deviation 1.

2. Invent a family of functions, each different, that look like those in Slide 3: they all have value 1 at x = 0; they all have zero derivative at x = 0; and they generally (not necessarily monotonically) decrease to zero at large x. Now multiply 10 of them together and graph the result near the origin (i.e., reproduce what Slide 3 was sketching).

3. For what value(s) of ν does the Student distribution (Segment 8, Slide 4) have a convergent 1st and 2nd moment, but divergent 3rd and higher moments?

To Think About

1. A distribution with moments as in problem 3 above has a well-defined mean and variance. Does the CLT hold for the sum of RVs from such a distribution? If not, what goes wrong in the proof? Is the mean of the sum equal to the sum of the individual means? What about the variance of the sum? What, qualitatively, does the distribution of the sum of a bunch of them look like?

2. Give an explanation of Bessel's correction in the last expression on slide 5. If, as we see, the MAP calculation gives the factor 1/N, why would one ever want to use 1/(N-1) instead? (There are various wiki and stackoverflow pages on this. See if they make sense to you!)

Segment 11

To Calculate

1. For the Cauchy distribution (Segment 8, Slide 3), find the inverse function of the CDF.

2. In your favorite programming language, write a function that returns independent Cauchy deviates.

To Think About

1. Suppose you want a function that returns deviates for Student (ν). Could you use the Cauchy pdf (or some scaling of it) as a bounding function in a rejection method? How efficient is this (i.e., what fraction of the time does it reject)?

2. Explain the three inequality tests in the "while" statement in Leva's algorithm (slide 7) and why they are hooked together with logical operators in the way shown.

Segment 12

To Calculate

1. What is the critical region for a 5% two-sided test if, under the null hypothesis, the test statistic is distributed as $\text{Student}(0, \sigma, 4)$? That is, what values of the test statistic disprove the null hypothesis with $p < 0.05$? (OK to use Python, MATLAB, or Mathematica.)

2. For an exponentially distributed test statistic with mean μ (under the null hypothesis), when is the the null hypothesis disproved with $p < 0.01$ for a one-sided test? for a two-sided test?

To Think About

1. P-value tests require an initial choice of a test statistic. What goes wrong if you choose a poor test statistic? What would make it poor?

2. If the null hypothesis is that a coin is fair, and you record the results of N flips, what is a good test statistic? Are there any other possible test statistics?

3. Why is it so hard for a Bayesian to do something as simple as, given some data, disproving a null hypothesis? Can't she just compute a Bayes odds ratio, P(null hypothesis is true)/P(null hypothesis is false) and derive a probability that the null hypothesis is true?

Segment 13

To Calculate

1. With p=0.3, and various values of n, how big is the largest discrepancy between the Binomial probability pdf and the approximating Normal pdf? At what value of n does this value become smaller than 10^{-15}?

2. Show that if four random variables are (together) multinomially distributed, each separately is binomially distributed.

To Think About

1. The segment suggests that $A \neq T$ and $C \neq G$ comes about because genes are randomly distributed on one strand or the other. Could you use the observed discrepancies to estimate, even roughly, the number of genes in the yeast genome? If so, how? If not, why not?

2. Suppose that a Bayesian thinks that the prior probability of the hypothesis that "$P_A = P_T$" is 0.9, and that the set of all hypotheses that "$P_A \neq P_T$" have a total prior of 0.1. How might he calculate the odds ratio $\mathrm{Prob}(P_A = P_T)/\mathrm{Prob}(P_A \neq P_T)$? Hint: Are there nuisance variables to be marginalized over?

Segment 14

To Calculate

1. Suppose the stopping rule is "flip exactly 10 times" and the data is that 8 out of 10 flips are heads. With what p-value can you rule out the hypothesis that the coin is fair? Is this statistically significant?

2. Suppose that, as a Bayesian, you see 10 flips of which 8 are heads. Also suppose that your prior for the coin being fair is 0.75. What is the posterior probability that the coin is fair? (Make any other reasonable assumptions about your prior as necessary.)

3. For the experiment in the segment, what if the stopping rule was (perversely) "flip until I see five consecutive heads followed immediately by a tail, then count the total number of heads"? What would be the p-value?

To Think About

1. If biology journals require p<0.05 for results to be published, does this mean that one in twenty biology results are wrong (in the sense that the uninteresting null hypothesis is actually true rather than disproved)? Why might it be worse, or better, than this? (See also the provocative paper by Ioannidis, and this blog in Technology Review (whose main source is this article). Also this news story about ESP research. You can Google for other interesting references.)

Segment 16 (order is intentional)

To Calculate

1. Simulate the following: You have M=50 p-values, none actually causal, so that they are drawn from a uniform distribution. Not knowing this sad fact, you apply the Benjamini-Hochberg prescription with $\alpha = 0.05$ and possibly call some discoveries as true. By repeated simulation, estimate the probability of thus getting N wrongly-called discoveries, for N=0, 1, 2, and 3.

2. Does the distribution that you found in problem 1 depend on M? On α ? Derive its form analytically for the usual case of $\alpha \ll 1$?

To Think About

1. Suppose you have M independent trials of an experiment, each of which yields an independent p-value. Fisher proposed combining them by forming the statistic

$$S = -2 \sum_{i=0}^{i=M} \log(p_i)$$

Show that, under the null hypothesis, S is distributed as $\mathrm{Chisquare}(2M)$ and describe how you would obtain a combined p-value for this statistic.

2. Fisher is sometimes credited, on the basis of problem 1, with having invented "meta-analysis", whereby results from multiple investigations can be combined to get an overall more significant result. Can you see any pitfalls in this?

Segment 15 (order is intentional)

To Calculate

1. In slide 4, we used "posterior predictive p-value" to get the respective p-values 1.0e-13, .01, .12, and .0013. What if we had mistakenly just used the maximum likelihood estimate r=0.003, instead of integrating over r? What p-values would we have obtained?

To Think About

1. Can you think of a unified way to handle the Towne family problem (estimating r and deciding which family members are likely "non-paternal") without trimming the data? We'll show one such method in a later segment, but there is likely more than one possible good answer.

Segment 17

To Calculate

1. Calculate the Jacobian determinant of the transformation of variables defined by

$$y_1 = x_1/x_2, \qquad y_2 = x_2^2$$

2. Consider the 3-dimensional multivariate normal over (x_1, x_2, x_3) with $\mu = (-1, -1, -1)$ and

$$\Sigma^{-1} = \begin{pmatrix} 5 & -1 & 2 \\ -1 & 8 & 1 \\ 2 & 1 & 4 \end{pmatrix}$$. (Note the matrix inverse notation.)

What are 2-dimensional μ and Σ^{-1} for

(a) the distribution on the slice $x_3 = 0$?

(b) the marginalization over x_3 ?

Hint: The answers are all simple rationals, but I had to use Mathematica to work them out.

To Think About

1. Prove the assertions in slide 5. That is, implement the ideas in the blue text.

2. How would you plot an error ellipsoid in 3 dimensions? That is, what would be the 3-dimensional version of the code in slide 8? (You can assume the plotting capabilities of your favorite programming language.)

Segment 18

To Calculate

1. Random points i are chosen uniformly on a circle of radius 1, and their (x_i, y_i) coordinates in the plane are recorded. What is the 2x2 covariance matrix of the random variables X and Y? (Hint: Transform probabilities from θ to x . Second hint: Is there a symmetry argument that some components must be zero, or must be equal?)

2. Points are generated in 3 dimensions by this prescription: Choose λ uniformly random in $(0, 1)$. Then a point's (x, y, z) coordinates are $(\alpha\lambda, \beta\lambda, \gamma\lambda)$. What is the covariance matrix of the random variables (X, Y, Z) in terms of $\alpha, \beta, \text{ and } \gamma$? What is the linear correlation matrix of the same random variables?

To Think About

1. Suppose you want to get a feel for what a linear correlation $r = 0.3$ (say) looks like. How would you generate a bunch of points in the plane with this value of r ? Try it. Then try for different values of r . As r increases from zero, what is the smallest value where you would subjectively say "if I know one of the variables, I pretty much know the value of the other"?

2. Suppose that points in the (x, y) plane fall roughly on a 45-degree line between the points (0,0) and (10,10), but in a band of about width w (in these same units). What, roughly, is the linear correlation coefficient r ?