

Probability and Statistics

Stat 238a/538a Fall, 2009

This is a full set of notes from a previous running of the class. Some of it may be a bit out of date (e.g. some parts having to do with computer software versions) and I may be producing updates and additions to these notes as necessary as we go along.

Contents

1	Introduction to probability, birthdays, R, simulation.....	3
1.1	Introduction.....	3
1.2	Monopoly.....	3
1.3	Some important terminology, and the mathematical framework.....	5
1.4	A special case: Equally likely outcomes.....	5
1.5	Introduction to R and simulation.....	8
1.6	The windshield example, probability measures, and more on the mathematical framework.....	11
1.7	Conditional probability and independence.....	14
1.8	Random variables and distributions.....	16
1.9	Binomial distribution.....	17
1.10	Law of total probability and Bayes' formula.....	19
1.11	An initial skirmish with likelihood and Bayesian statistics.....	21
1.12	Specifying distributions: Probability densities, cumulative distributions.....	26
1.13	More on the laser pole: Likelihood and inference.....	31
1.14	Joint distributions.....	34
1.15	Marginal and conditional distributions.....	36
1.16	Expectation.....	38
1.17	Variance.....	43
1.18	Time for a new distribution... The Geometric distribution.....	48
1.19	Notes and example on covariance.....	50
1.20	Law of Large Numbers: simulation, proof and pathology.....	52
1.20.1	Law of large numbers simulation.....	52
1.20.2	Law of Large Numbers ("LLN") statement.....	54
1.20.3	Pathology: Cauchy simulation and LLN "violation".....	54
1.20.4	Law of Large Numbers: Proof.....	55
1.21	Normal distributions.....	57
1.21.1	"Discovering" the normal distribution.....	57
1.21.2	Standard Normal distribution.....	60
1.21.3	General Normal distributions.....	62
1.21.4	The "68, 95, 99.7 rule".....	62
1.21.5	Using R and Normal tables.....	64
1.21.6	Another R Picture.....	65
1.22	A note on statistics and Statistics.....	65
1.23	Normal probability plots.....	66
1.24	Distributions of sums and the Central limit theorem.....	71
1.24.1	Distributions of sums of independent r.v.'s.....	71
1.24.2	Central limit theorem.....	72
1.24.3	Normal approximation to Binomial distributions.....	75
1.24.4	The "continuity correction".....	76
1.25	Conditional expectation.....	78
2	Markov chains and Markov chain Monte Carlo.....	82
2.1	Markov chains: introduction and examples.....	82
2.2	How the distribution of the state evolves over time, and how matrices come into the picture... ..	86

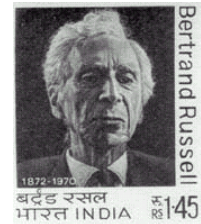
2.3	More on Markov chains: Stationary distributions.....	88
2.3.1	Stationary distributions	89
2.4	Limit Theorem for Markov Chains.....	93
2.5	Use of Markov chains for simulation.....	93
2.6	Designer Markov chains: The Metropolis-Hastings method	96
2.6.1	Example of the Metropolis method:.....	96
2.6.2	The general idea of Metropolis-Hastings (explained in the discrete case).....	96
2.6.3	Why does the Metropolis-Hastings recipe work?	97
3	Statistical models, estimation, mean squared error, maximum likelihood.....	100
3.1	Statistical models	100
3.2	Estimation	100
3.3	Estimation examples and relative efficiency: mean versus median, Normal and double exponential distributions	103
3.4	Maximum likelihood estimation	105
3.5	Bayesian estimation and conjugate priors.....	109
4	Bayesian data analysis and statistical inference using MCMC.....	117
4.1	Subliminal math improvement: an example “from scratch” using random walk Metropolis... 117	117
4.2	BUGS.....	122
4.3	Notes on BUGS and on running BUGS from R.	125
4.4	JAGS: installing and running it.....	127
4.4.1	Website and documentation	127
4.4.2	Installation.....	128
4.4.3	Running JAGS	128
4.5	Regression, crying and IQ.....	132
4.6	More on correlation and regression	134
4.7	Notes on algebra and geometry of regression	140
4.8	Multiple linear regression	142
4.9	Model selection and the Deviance Information Criterion.....	147
4.10	Gibbs sampler.	149
4.11	Logistic Regression.....	151
4.12	Hidden Markov models.....	156
4.13	Epilogue	164

1 Introduction to probability, birthdays, R, simulation

Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means.

The whole problem with the world is that fools and fanatics are always so certain of themselves, but wiser people so full of doubts.

A habit of basing convictions upon evidence, and of giving to them only that degree or certainty which the evidence warrants, would, if it became general, cure most of the ills from which the world suffers.



1.1 Introduction

The theory of probability is an intellectual achievement representing a remarkable combination of abstract theoretical depth with immense practical utility. This theory is the fundamental tool for dealing with uncertainty, and its applications pervade an amazing variety of fields.

Why probability? What does probability have to do with statistics? What is Statistics anyway? The *Yale College Programs of Study* says "Statistics is the science and art of prediction and explanation." Sounds interesting but not all that specific. (Actually I think I wrote that some years ago myself...) Another concise way of looking at it is that Statistics provides a framework and tools for addressing questions involving uncertainty. In fact, most difficult, interesting questions involve uncertainty; if they did not, then they would not be so difficult or interesting. Certain mathematical questions lack uncertainty; for example, in Euclidean plane geometry, the sum of the angles in a triangle is 180° . This question is settled; nobody asks "What is the latest thinking on the sum of the angles in a triangle in Euclidean plane geometry?" This question could be answered by a person sitting in a dark room with his eyes closed, ears covered, nose plugged, etc., simply by pure thought. Other questions require interaction with the world to answer; the investigator must perform experiments or collect data. Does smoking cause cancer? Does listening to Mozart make people smarter? Is there a gene that causes cystic fibrosis? If so, where is it in the genome? Should Bush really have won the 2000 election? Having collected relevant data, we use probability theory to assess the strength of evidence that the data provides for the question. The question is never fully, 100% settled; it is only settled to a degree of certainty, and part of our job is to quantify how certain we are.

Statistics also plays a key role in developing ways for computers to do things, such as recognize patterns and sift through mounds of data and draw conclusions. In pattern recognition, for example, we might use all available "training data" to construct probability models that help distinguish one class of patterns from another. Then we would fit those models to a new example. We could guess that the new example belongs to the class whose probability model fits the best (this is the idea of "maximum likelihood"). We would use the degree of fit of all the models to help us quantify our degree of certainty about the classification we have guessed.

1.2 Monopoly

The theory of probability was first conceived by thinking about games of chance and gambling, and this remains a good place to start. My 6 year old son Corey loves to play Monopoly (well, he was 6 and loved Monopoly when I first wrote this).



Suppose his piece is currently on the corner square labeled "GO". He is about to roll a pair of dice, and move his piece that many squares in the direction of the arrow. You see those 3 hotels that he will visit if he rolls a 6 or an 8 or a 9? Those are my hotels, and he will have to pay me a lot of money and go bankrupt if he visits one of them. Corey wants to calculate the probability that he will go bankrupt on the next move. [He actually knows how to do this, and he insists on calculating a probability like this whenever this sort of situation arises.] OK, if you don't know about Monopoly don't worry – our problem is simply this: When a pair of dice is rolled,

what is the probability that the sum of the two dice gives 6, 8, or 9?

In order to analyze this we need to think of all the possible outcomes that could take place, and how likely each of them is. To avoid confusion, let's suppose Corey does not drop the dice perfectly simultaneously, but rather one of the dice falls before the other. If we write what happens to the two dice in order, we obtain the following 36 outcomes:

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

You see why there are 36 outcomes, right? For each of the 6 possible rolls for the first die, there are 6 possible rolls for the second, which leads to a total of $6 \times 6 = 36$ outcomes. Let us assume that each of these 36 outcomes is equally likely. Tiny imperfections or asymmetries in the dice, or some sort of magical rolling skill on Corey's part, might perturb the probabilities slightly so that some outcomes are ever so slightly more likely than others, but for all practical purposes, and because we do not know any better information, we adopt the assumption of equally likely outcomes here. So the question becomes: in

how many of the 36 outcomes does Corey go bankrupt? For example, if there were only 1 way out of 36 that Corey would go bankrupt, the probability would be $1/36$. If there were 27, the probability would be $27/36 = 3/4$. And so on. In this case, the cases in which Corey goes bankrupt are those in which the total is 6 or 8 or 9. These are shown in red below

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

and we can see that there are 14 such outcomes. The probability of going bankrupt in Corey's next move is $14/36 = 0.389$.

1.3 Some important terminology, and the mathematical framework

If we want to put Probability Theory on a firm mathematical foundation, then we should define our terms as mathematical objects. This makes Probability Theory part of Mathematics, and makes it possible for various propositions and theorems about probabilities to be mathematically *proved*. For example, great results in the theory of probability include the Law of Large Numbers and the Central Limit Theorem, both of which will be discussed later.

The *sample space*, which we will typically denote by S , is the set of all possible outcomes. For example, in modeling the experiment of rolling a pair of dice as above, the set S would consist of the 36 listed possible outcomes.

An *event* is a subset of the sample space. For example, the event that Corey goes bankrupt after rolling his pair of dice is the subset of S consisting of the 14 outcomes shown in red above.

A *probability measure* is a function that assigns a number between 0 and 1 to each event; this number is called the probability of the event. (The function must satisfy additional properties that we will discuss shortly.) We write $P(A)$ for the probability of the event A .

So that's a good start. In the mathematical theory, fundamental terms like "sample space," "event," and "probability measure," instead of being vague intuitive concepts, are simply examples of the simple mathematical objects of *set*, *subset*, and *function*. We'll continue some more with the mathematical framework below.

Many probabilities have a *relative frequency* interpretation. (I say many and not all because some probabilities are better interpreted as a reflection of subjective degrees of belief. More on this later.) For example, our finding that the probability of getting a 6 or 8 or 9 in one roll of a pair of dice is $14/36$ may be interpreted as saying that if we repeat this experiment many times, in the long run the fraction of trials on which a 6 or 8 or 9 appears will converge to $14/36$.

1.4 A special case: Equally likely outcomes

Many other problems rely on the same principle used in analyzing the Monopoly problem above. If we *assume* that all of the possible outcomes in S are equally likely, then the probability of an event A is obtained by dividing the number of elements in A by the total number of elements in S . That is, in symbols,

$$(1.1) \quad P(A) = \frac{\#(A)}{\#(S)}$$

To evaluate probabilities of the form (1.1) we need to solve some counting problems. A simple but useful tool here is the following.

Fundamental counting principle: Suppose we are about to perform a sequence of k actions, one after another, and

- the first action can be performed in n_1 ways,
- for each of the ways of performing the first action, the second action can be performed in n_2 ways,
- for each of the ways of performing the first two actions, the third action can be performed in n_3 ways, and so on, up to
- for each of the ways of performing the first $k - 1$ actions, the last action can be performed in n_k ways.

Then the total number of ways that the full sequence of k actions can be performed is the product $n_1 n_2 n_3 \cdots n_k$. Since the answer is obtained by multiplying, this is sometimes called the *multiplication principle*.

Example: The university's Statistics Club has 12 members. How many ways are there to choose a president, vice president, and secretary?

Answer: We will think of this as performing a sequence of 3 actions: choosing the president, then vice president, then the secretary. The president can be chosen in 12 ways. Having done that, for each possible choice of president, there are 11 ways to choose the vice president. And finally, having chosen a president and vice president, there are 10 ways to choose a secretary. Multiplying gives an answer of $12 \times 11 \times 10 = 1320$ ways to choose those core statisticians of power.

Here is a classic example in which the same simple idea becomes much more interesting.

The Birthday Problem. Suppose there are k people in a room. What is the probability that there is at least one match among the k birthdays?

For example, suppose the room contains $k = 23$ people. There are 365 possible birthdays (we will ignore leap years in this problem). Let us assume that each of those 365 days is equally likely as a birthday. (We adopt this assumption as an approximation that is probably not too far from the truth, and because we have no specific information that allows us to improve the approximation at this point.) Just think about this intuitively and take a guess. With 23 people, is there a good chance of a match among their birthdays? Most people would say the chances are slim because there are 365 possible birthdays but only 23 people. In fact, the probability is slightly larger than 0.5. So if you were given the chance to bet on a birthday match among 23 people, or getting a "heads" on a toss of a coin, you would do better to bet on the birthday match. If there were more people in the room, the probabilities rise quite quickly. For example, with $k = 40$, then the probability is 89%, and with $k = 50$, the probability is 97%!

How did I know these amazing things? You already know the secret – these probabilities are calculated by applying the fundamental counting principle we have discussed. This principle applies directly to the "complementary" event of *no* birthday match. So let A denote the event that there is *no* match among the k birthdays, that is, the k birthdays are all distinct. We will calculate the probability $P(A)$. The sample space consists of all possible ways of assigning birthdays to k people. We are *assuming* that each of these assignments is equally likely, so the probability of A will be a quotient of counts, as (1.1) prescribes. For the denominator, we need to know $\#(S)$, that is, how many ways there are to assign birthdays to k people. By the Counting Principle, $\#(S) = 365^k$, because for each person, there are 365 ways to assign

that person's birthday. To count the numerator $\#(A)$, which is the number of ways of assigning k birthdays in such a way that no two birthdays match, we imagine sequentially assigning birthdays to person 1, and then person 2, and so on. There are 365 ways to assign a birthday to person 1. Having made this assignment, there are only 364 ways to assign a birthday to person 2. Having made the first two assignments, there are 363 ways to assign a birthday to person 3. And so on. For the last person, person k , having previously assigned $k - 1$ birthdays, there are $365 - (k - 1)$ ways to assign a birthday to person k . So by multiplying we obtain $\#(A) = (365)(364)(363)\cdots(365 - k + 1)$. So here is our answer for the probability of no birthday match:

$$P(A) = \frac{(365)(364)(363)\cdots(365 - k + 1)}{365^k}.$$

A more insightful way to write this might be

$$P(A) = \left(\frac{365}{365}\right)\left(\frac{364}{365}\right)\left(\frac{363}{365}\right)\cdots\left(\frac{365 - k + 1}{365}\right),$$

a product of k fractions, the first of which is 1, and with each factor successively decreasing by $1/365$. This can be easily calculated on a computer. For example, if $k = 40$, the probability of no match is

$$P(A) = \left(\frac{365}{365}\right)\left(\frac{364}{365}\right)\left(\frac{363}{365}\right)\cdots\left(\frac{327}{365}\right)\left(\frac{326}{365}\right) = 0.1088.$$

To answer our original question, which is the probability of a match (the opposite of "no match"), we subtract from 1, getting $P(\text{match}) = 1 - P(\text{no match}) = 1 - 0.1088 = .8912$ in the case of $k = 40$.

Q: Why did we subtract from 1? Justify this by using (1.1).

Most people find the answers to this type of birthday question surprising. This is a good reason to learn to calculate probabilities – we cannot always trust our intuitive estimates of what is likely and what is unlikely!

Example: (Back to the university Statistics club.) The statisticians are getting ready to party, but statisticians are nothing if not tidy. How many ways can they choose a cleanup committee of 3 members?

Solution: Didn't we already do this one? I mean, we figured out that there are $12 \times 11 \times 10 = 1320$ ways to choose a president, vice-president, and secretary from the same club. The difference here is that we want to consider members of the cleanup committee indistinguishable in a sense. That is, there are no ranks within the committee. If there were a head cleanerupper, associate cleanerupper, and assistant cleanerupper, then we would know the answer: there are $12 \times 11 \times 10$ ways to choose such a "ranked committee." We want to know the number of ways of choosing an "unranked committee" having 3 members. For now, let us call that unknown number C . Observe that for each unranked committee, there are $3 \times 2 \times 1 = 6$ ranked committees, because by the Counting Principle, there are $3 \times 2 \times 1$ ways to assign ranks to 3 people. Therefore, since we may choose a ranked committee by first choosing an unranked committee and then assigning ranks within that committee, the Counting Principle again gives

$$C \times (3 \times 2 \times 1) = (12 \times 11 \times 10), \text{ or } C = \frac{12 \times 11 \times 10}{3 \times 2 \times 1}.$$

We will use the notation $\binom{12}{3}$ for the number of ways of choosing an unranked committee of size 3 from a club of 12 members, or, in other words the number of subsets of size 3 from a set of size 12. Such unranked committees, or simply *sets* where there is no idea of ordering within the set, are called *combinations*. The same reasoning that we just used in the example shows that the number of combinations of n items taken k at a time is

$$\binom{n}{k} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k(k-1)(k-2)\cdots 1}.$$

Defining the factorial function $r! = r(r-1)(r-2)\cdots(2)(1)$ and noting that

$n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{k!(n-k)!}$, we can write the above formula in a more concise (and certainly more exciting looking – 3 exclamation points!!!) way

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

The notation $\binom{n}{k}$ is typically read, by people odd enough to find themselves talking about these things, as "*n choose k*".

1.5 Introduction to R and simulation

There is some wonderful statistical computing software called "R" that is even more wonderful because it is free and it's available for your computer. So go download it! To do this, go to <http://cran.us.r-project.org/>. Under "Precompiled Binary Distributions" you will see Windows, Linux, and Macintosh versions. Click on the appropriate one. I'll talk about the Windows version here. After you select the Windows version, click on "base" – you want the "base distribution". There you will see a README file with installation and other instructions. The file you really want to download is R-2.3.1-win32.exe, which is the setup program that you use to install R (at least that's the name it had when I wrote this; by now it could be R-2.4.0-win32.exe or whatever). You will see the file name there, and you can download it from there: click on it, and say you want to "Save to Disk" (for example, you could save it to your Desktop. After saving R-2.3.1-win32.exe to your computer somewhere, double-click it to start the installation. After installing R, and after you start it up, you will see a Help menu, which contains a lot of useful documentation. For example, you could choose "Html help" and then choose "An introduction to R". Also in the Help menu for R, if you select Manuals, you will see 5 manuals. These are in Adobe Acrobat form, and also include that "An introduction to R" document as well as other more detailed manuals.

You can start by using R as a calculator. At the R prompt, ">", type an expression to calculate it and R will return the answer.

```
> 2*3
[1] 6
> 2^10
[1] 1024
> sin(pi/4)
[1] 0.7071068
```

That's handy. You can use "=" to assign a variable a value:

```
> x = 5
> x^2
[1] 25
```

Use "c" as follows to create a vector:

```
> y = c(3, 7, 5, 1, 2, 3, 2, 5, 5)
```

We can extract elements of the vector by using square brackets as follows:

```
> y[2]
[1] 7
> 2:4
```



```
[1] 2 3 4
> y[2:4]
[1] 7 5 1
```

We can use various built in functions on vectors. For example, "length" returns the number of components in the vector, and "table" gives a list of the elements in the vector together with their frequencies:

```
> length(y)
[1] 9
> table(y)
y
1 2 3 5 7
1 2 2 3 1
```

Here are 12 random numbers drawn from a “uniform distribution” over the interval between 0 and 1:

```
> z = runif(12)
> z
[1] 0.79763720 0.25793181 0.04354754 0.30732140 0.22922168
[6] 0.15864316 0.33307461 0.52337870 0.58317311 0.45802095
[11] 0.63019991 0.46671983
```

We can see which of these is less than 0.5 with the expression "z < 0.5"

```
> z < 0.5
[1] FALSE TRUE TRUE TRUE TRUE TRUE TRUE FALSE FALSE
[10] TRUE FALSE TRUE
```

Here is how you can do random sampling, with and without replacement:

```
> sample(10,5)
[1] 3 9 1 2 8
> x = sample(10, 10, replace = F)
> x
[1] 10 6 3 8 2 5 9 4 1 7
> y = sample(10, 10, replace = T)
> y
[1] 3 3 5 6 4 5 3 2 7 3
> table(x)
x
 1  2  3  4  5  6  7  8  9 10
1  1  1  1  1  1  1  1  1  1
> table(y)
y
2 3 4 5 6 7
1 4 1 2 1 1
```

You know, I don't know if I really believe that the probability of a birthday match with just $k = 40$ people is nearly 90%. I'd like to get 40 people, ask them their birthdays, and see if I really get a match. We can simulate this process on the computer.

To get 40 birthdays, we'll sample from the 365 days, with replacement.

```
> k = 40
> bdays = sample(365, k, replace = T)
> bdays
[1] 347 35 349 185 194 275 354 235 321 155 114 182 71 287
```

```
[15] 8 108 201 216 183 96 303 199 285 88 285 237 111 75
[29] 288 103 55 140 134 42 287 46 176 191 224 206
```

Well, we did it. Now, is there a match somewhere within that 40-long vector `bdays`? Scanning by eyeball is a bit tedious. But we've already met an R function that can help: the "table" function. Let's try it out:

```
> tally = table(bdays)
> tally
bdays
 8 35 42 46 55 71 75 88 96 103 108 111 114 134 140
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
155 176 182 183 185 191 194 199 201 206 216 224 235 237 275
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
285 287 288 303 321 347 349 354
2 2 1 1 1 1 1 1
```

Now we can see that in fact there are two different birthdays – days 285 and 287 – that are both found twice among the 40 birthdays. If we don't care about the details of which days matched and so on, we could test for a birthday match even more simply by seeing whether there are any frequencies higher than 1 in the tally, as follows:

```
> max(tally)
[1] 2
> max(tally) > 1
[1] TRUE
```

So, indeed, we have simulated the process of checking for a birthday match in one room containing 40 people, and we did find a match, as the calculation indicates is likely, with probability nearly 90%.

Wouldn't you like repeat this many times to see if our success on the first trial was just a fluke? In fact, if we repeated this procedure many times, we could check the theoretical prediction that about 90% of these trials should have birthday matches. But the beauty of computing is that, if the computer did this once, it can do it many times. We can do this by putting a loop around what we just did. For now, let's repeat the experiment 100 times. Here's the program:

```
k = 40
ntrials = 100
results = c()
for(i in 1:ntrials){
  bdays = sample(365, k, replace = T)
  tally = table(bdays)
  results = c(results, (max(tally) > 1))
}
```

and here are the 100 indicators of whether there was a birthday match on each trial:

```
> results
[1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE
[10] TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE FALSE TRUE
[19] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE
[28] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[37] TRUE TRUE TRUE FALSE TRUE TRUE TRUE FALSE TRUE
[46] TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE FALSE
[55] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE
[64] TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE
[73] TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE FALSE
[82] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[91] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE
```

```
[100] TRUE
```

You can see that most of these trials contained birthday matches. How many? The function "sum" applied to a numerical vector returns the sum of the elements in the vector; applied to a logical vector, the "sum" function considers a TRUE to be "1" and FALSE to be "0", so that the sum is the number of "TRUE" elements in the vector.

```
> sum(results)
[1] 87
```

So for 40 people in the room, 87 out of 100 birthday trials resulted in a birthday match. This is very much consistent with the answer 0.8912 we obtained earlier for the probability of a match.

1.6 *The windshield example, probability measures, and more on the mathematical framework*

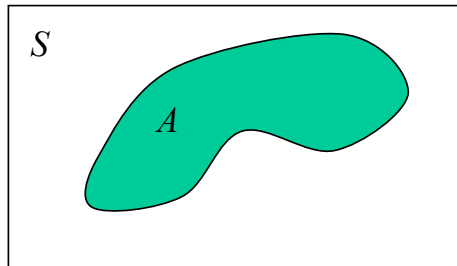
Here is another simple but very useful example to keep in mind as we go along:

The windshield example. You are driving your car. It is about to start raining. Where will the first drop hit your windshield?

Here the sample space S is the set of all of the points of the windshield.

An event here is a subset of the windshield.

Can draw familiar Venn diagrams now to picture events.



Assume there are no "preferred" locations or regions of the windshield -- the probability is spread uniformly over the windshield.

$$P(A) = \frac{\text{area of } A}{\text{area of } S}.$$

← So remember: In the windshield example, probabilities are ratios of areas

For convenience, suppose the area of the windshield S happens to be exactly 1. In that case,

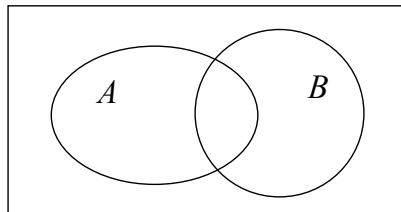
$$P(A) = \text{area of } A$$

E.g. • $P(S) = 1$.

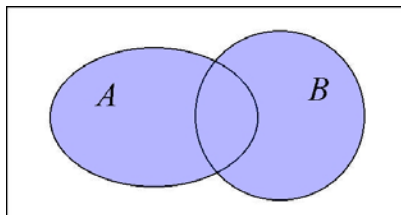
• If an event A takes up $5/8$ the area of the windshield, then $P(A) = 5/8$. And so on.

→ This explains why people often reason about the properties of probabilities by drawing Venn diagrams and thinking about areas – the windshield example shows that area is in fact an example of a probability measure.

Familiar set-theoretic concepts. Given two events A and B as follows.



The subset shaded in blue below is a new event called the *union* of A and B , and denoted $A \cup B$.

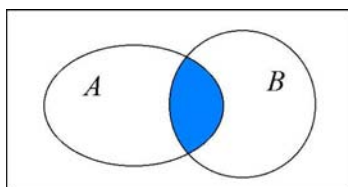


In order to apply these concepts to “word problems,” models, situations of real life, and so on, we should know how these operations correspond to constructions in ordinary English. So here is a question: Which is the appropriate way to refer to the blue event? Is it “A and B” or “A or B”?

Now, it would be better if you would stop reading until you try to answer this question for yourself. So I’ll stall for a sentence to give your eyes a chance to come to a screeching halt... right here.

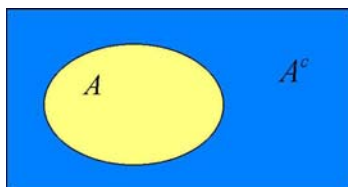
Now, some people are tempted to say “A *and* B,” perhaps because they are thinking that both A *and* B are shaded. But the right answer is “A or B.” A good way to think about this is to say to yourself: a point is shaded blue in the diagram if it is in the event A *or* it is in the event B .

The *intersection* of two sets A and B consists of the points of S that are in both A and B .



Accordingly, this set can be referred to as “A and B.” We will denote this intersection by $A \cap B$ or simply by AB . We will see later that there is good reason to think of the intersection as a product, so the notation AB makes sense.

The *complement* of a set A , denoted A^c , consists of the outcomes *not* in A .



This event can be referred to as “not A”: The event A^c is the event that A *does not occur*.

Define \emptyset to be the empty set.

Two events A and B are *disjoint* if $A \cap B = \emptyset$.

The windshield example suggests: if A and B are disjoint then $P(A \cup B) = P(A) + P(B)$.

More generally: define a sequence of events A_1, A_2, \dots to be disjoint if $A_i \cap A_j = \emptyset$ for each pair $i \neq j$.

Definition: A *probability measure* is a function that assigns a number to each event. A probability measure P is required to satisfy the following axioms:

☺ $P(A) \geq 0$ for each event A .

☺ $P(S) = 1$.

☺ If events A_1, A_2, \dots are disjoint, then $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$. That is, the probability of the union is the sum of the probabilities.

All general properties of probabilities can be derived from these axioms.

Example: The complement rule. $P(A^c) = 1 - P(A)$.

To derive this, note that A and A^c are disjoint, and $A \cup A^c = S$, so that, by properties (i) and (iii),
 $1 = P(S) = P(A \cup A^c) = P(A) + P(A^c)$.

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

Proof:

By logic, $A \cup B = (A \cap B^c) \cup (A^c \cap B) \cup (A \cap B)$.

And $(A \cap B^c)$, $(A^c \cap B)$, and $(A \cap B)$ are disjoint.

So $P(A \cup B) = P(A \cap B^c) + P(A^c \cap B) + P(A \cap B)$.

But $A = (A \cap B) \cup (A \cap B^c)$ [union of 2 disjoint sets]

implies $P(A) = P(A \cap B) + P(A \cap B^c)$, or $P(A \cap B^c) = P(A) - P(A \cap B)$.

Similarly, $P(A^c \cap B) = P(B) - P(A \cap B)$.

$$\begin{aligned} P(A \cup B) &= P(A) - P(A \cap B) + P(B) - P(A \cap B) + P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B). \end{aligned}$$



More general "inclusion-exclusion principle": e.g.,

$$\begin{aligned}
P(A \cup B \cup C \cup D) &= P(A) + P(B) + P(C) + P(D) \\
&\quad - P(AB) - P(AC) - P(AD) - P(BC) - P(BD) - P(CD) \\
&\quad + P(ABC) + P(ABD) + P(ACD) + P(BCD) \\
&\quad - P(ABCD)
\end{aligned}$$

(note: using product notation here -- " AB " means " $A \cap B$ ") and in general,

$$\begin{aligned}
P(A_1 \cup A_2 \cup \dots \cup A_n) &= \sum_i P(A_i) - \sum_{i_1 < i_2} P(A_{i_1} A_{i_2}) + \sum_{i_1 < i_2 < i_3} P(A_{i_1} A_{i_2} A_{i_3}) \\
&\quad - \dots + (-1)^{n-1} P(A_1 A_2 \dots A_n).
\end{aligned}$$

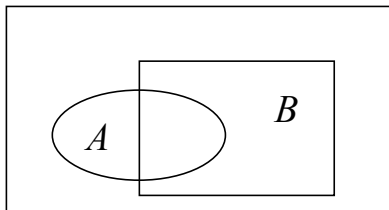
If you want a proof now, you can take a look at Theorem 3.8 on p.112 of the Grinstead and Snell book.

We'll be able to derive this formula in a slicker and simpler way after introducing expectation of random variables.

1.7 Conditional probability and independence

Statistics is about learning from data. We can think of collecting data as observing events, like "The outcome of this experiment was such-and-such." So a critical part of probability and statistics is to work out how observing the occurrence of one event affects the probability of something else.

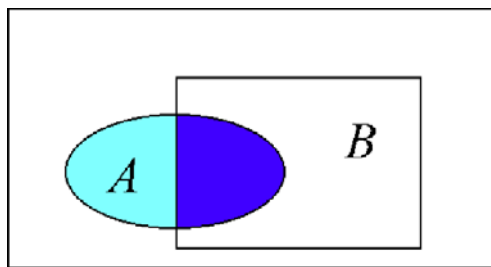
Idea of $P(B|A)$: Given that A occurs, what is the probability that B also occurs?



Q: By eyeball, what is $P(B|A)$?

[I'm not writing the answer down – that would spoil this opportunity to learn! First guess...]

Given that the raindrop fell in A , we restrict our attention to the set A . The drop is equally likely to fall anywhere within A .



Given A , the event B also occurs when the drop falls in the dark blue region, i.e., the event $(A \text{ and } B)$.

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

Independence

E.g. two tosses of a fair coin.

$A = \{\text{heads on first toss}\}$, $B = \{\text{heads on second toss}\}$.

E.g. *not* independent: choose a random person

$A = \{\text{person's height more than 75 inches}\}$

$B = \{\text{person's father's height more than 75 inches}\}$

“ B is independent of A ” means “being told that A occurred does not affect your probability that B occurs.”

That is, $P(B | A) = P(B)$, i.e., $\frac{P(A \cap B)}{P(A)} = P(B)$.

Definition: Two events A and B are *independent* if $P(A \cap B) = P(A)P(B)$.

[Note the concept is symmetric in A and B]

Note: Definition of conditional probability often used in form: $P(AB) = P(A)P(B | A)$.

This gives $P(ABC) = P(AB)P(C | AB) = P(A)P(B | A)P(C | AB)$

and in general

$$(1.2) \quad P\left(\bigcap_{i=1}^k A_i\right) = P(A_1)P(A_2 | A_1)P(A_3 | A_1 A_2) \cdots P(A_k | A_1 \cdots A_{k-1}) \leftarrow \text{“Multiplication rule”}$$

Example: birthday problem prob of no match among k people:

$$P(A) = \left(\frac{365}{365}\right)\left(\frac{364}{365}\right)\left(\frac{363}{365}\right) \cdots \left(\frac{365 - k + 1}{365}\right)$$

Think of this as (1.2), where $A_i = \{\text{no birthday match among the first } i \text{ people}\}$.

Another way of thinking about conditional probabilities: Sometimes it is nice to think of conditional probs in terms of "contingency tables"

	Hospital A	Hospital B
Died	300	50
Survived	3000	1000

This table represents two variables for 4350 patients.

Suppose we choose a random patient from these 4350. Then, for example:

$$P\{\text{Hospital B}\} = \frac{1000 + 50}{4350} = .241$$

$$P\{\text{Died}\} = \frac{300 + 50}{4350} = .080$$

$$P\{\text{Died} \cap \text{Hospital B}\} = \frac{50}{4350} = .011$$

$$P\{\text{Died} | \text{Hospital B}\} = \frac{P\{\text{Died} \cap \text{Hospital B}\}}{P\{\text{Hospital B}\}} = \frac{50}{1000 + 50} = .048$$

$$P\{\text{Hospital A} | \text{Survived}\} ?$$

1.8 Random variables and distributions

The intuitive meaning of the term "random variable" might be expressed as "a numerical feature or description of a random outcome." It is a random number, that is, a numerical outcome whose value depends on chance occurrences. For example, you might say, "Let X_1, X_2, X_3, X_4 , and X_5 denote the heights in millimeters of the next 5 people that pass by me on the sidewalk." You would think of X_1, X_2, X_3, X_4 , and X_5 as five random variables. Or if you are about to toss a coin 3 times, you could define a random variable X to be the number of Heads obtained.

The concept of random variable can be formulated in the mathematical framework we have been developing, as follows.

DEFINITION. A *random variable* is a function defined on the sample space. That is, it is a function that assigns a number to each outcome in the sample space.

Example. Toss a coin 3 times. Define X = number of heads in the 3 tosses.

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

Many events are defined in terms of random variables. Remember an event is a subset of the sample space, S .

E.g., $\{X = 2\} = \{s : X(s) = 2\} = \{\text{THH}, \text{HTH}, \text{HHT}\}.$

Let's assume the 8 outcomes in this sample space are equally likely, so that each outcome has probability $1/8$.

So, for example, since the event $\{X = 2\}$ contains 3 outcomes, we have $P\{X = 2\} = 3/8$. Similarly, by looking at the other possible values for X we get $P\{X = 0\} = P\{TTT\} = 1/8$, $P\{X = 1\} = P\{TTH, THT, HTT\} = 3/8$, and $P\{X = 3\} = P\{HHH\} = 1/8$.

Collecting all of the possible values for the random variable X together with the probabilities, we arrive at

$$X = \begin{cases} 0 & \text{with probability } 1/8 \\ 1 & \text{with probability } 3/8 \\ 2 & \text{with probability } 3/8 \\ 3 & \text{with probability } 1/8 \end{cases}$$

→ This is called the *distribution* of X .

Definition. A random variable is called *discrete* if it can take on only finitely many or countably infinitely many values.

[A random variable is called *discreet* if it can keep a secret well and does not go around insulting all the other random variables.]

Random variables that take on values in an uncountably infinite set (like an interval of real numbers, for example) are sometimes called *continuous*.

DEFINITION. The *distribution* of a discrete random variable is given by a list of the possible values of the random variable together with the probabilities that the random variable takes on those values. This is also called a *probability mass function*. The probability mass function f_X of a random variable X , given by $f_X(x) = P\{X = x\}$, tells how much probability "mass" is assigned to each of the possible values of X .

1.9 Binomial distribution

Now that we know what a distribution is, let's introduce a very common and useful family of distributions: the Binomial distributions. We'll be introducing new distributions gradually over the coming sections. Binomial distributions are used as models for distributions of *counts*.

Generic setup:

- Performing n independent "trials" of an "experiment."
- Each trial could be a "success" or a "failure."
- Let p = probability of success on each trial.
- Let X = number of successes among the n trials.

Then: We say the r.v. X has a **Binomial distribution with parameters n and p** .

Notation: $X \sim B(n, p)$

Probability mass function:

$$P\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, \dots, n.$$

Where does this formula come from? To see how it is derived, let's do an example. Suppose we roll a die 5 times, and let X denote the number of 6's obtained. Then $X \sim B(5, \frac{1}{6})$, and we want to calculate distribution (the pmf) of X . For example, let's start with the question: What is $P\{X = 2\}$?

One way to get $X = 2$: SSFFF

Probability of this way: $\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{5}{6}\right) = \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3$

Another way to get $X = 2$: FSFSF

Prob: $\left(\frac{5}{6}\right)\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)\left(\frac{1}{6}\right)\left(\frac{5}{6}\right) = \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3$ again!

All possible ways to get $X = 2$:

SSFFF, SFSFF, SFFSF, SFFFS, FSSFF,
FSFSF, FSFFS, FFSSF, FFSFS, FFFSS

There are $\binom{5}{2} = 10$ of these. Each has prob. $\left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3$.

So $P\{X = 2\} = \binom{5}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3$.

[[In general $P\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k}$]]

Example (noisy communication channel): [Rice p. 37]

For a single bit (0 or 1) of information, have probability $p = 0.1$ of erroneous transmission.

Suppose we transmit the bit 5 times and use "majority decoder."
Also assume 5 transmissions have independent errors.

X = number of errors $\sim B(5, 0.1)$.

Bit decoded incorrectly if $X \geq 3$.

So error prob is decreased from 0.1 to

$$\begin{aligned} & P\{X = 3\} + P\{X = 4\} + P\{X = 5\} \\ &= \binom{5}{3} (.1)^3 (.9)^2 + \binom{5}{4} (.1)^4 (.9)^1 + \binom{5}{5} (.1)^5 (.9)^0 \\ &= (10)(.001)(.81) + (5)(.0001)(.9) + (.00001) = 0.00856 \end{aligned}$$

Using R: For $X \sim B(n, p)$ built-in functions are

$$\text{dbinom}(x, n, p) = P\{X = x\}$$

$$\text{pbinom}(x, n, p) = P\{X \leq x\}$$

$\text{rbinom}(s, n, p)$ returns a vector of s observations from the $B(n, p)$ distribution

[[Think “d” for “density,” “p” for “(cumulative) probability,” “r” for “random”]]

So, e.g., can get $P\{X = 3\} = .0081$ by

```
> dbinom(3, 5, .1)
[1] 0.0081
```

and get $P\{X \geq 3\} = 1 - P\{X \leq 2\}$ by

```
> 1 - pbinom(2, 5, .1)
[1] 0.00856
```

Can see a picture of the $B(50, 0.6)$ prob mass function with

```
plot(0:50, dbinom(0:50, 50, .6))
```

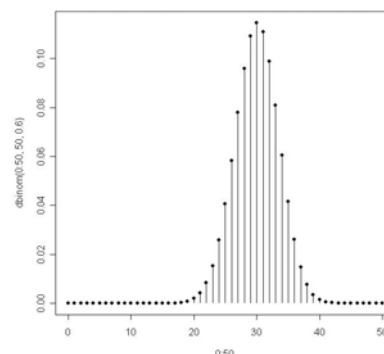
or if you prefer,

```
plot(0:50, dbinom(0:50, 50, .6), type="h")
```

or even

```
plot(0:50, dbinom(0:50, 50, .6), type="h")
points(0:50, dbinom(0:50, 50, .6), pch=19)
```

[Incidentally, easiest way to see choices for plotting character (pch) is to do `example(points)`].



1.10 Law of total probability and Bayes' formula

Imagine a "consequence" that can have a number of possible "causes."

Law of total probability expresses the probability of a consequence as a sum over the possible causes.

Bayes' formula finds conditional probabilities of causes, given an observed consequence.

Example: A blood test screening for a virus is given to people randomly chosen from a population. Under the assumptions to be listed below,

- For each person, what is the probability of a positive test result?
- If a given person has a positive test result, what is the conditional probability that the person indeed has the virus?

Let $A = \{\text{person infected with virus}\}$

$B = \{\text{blood test positive}\}$

Assume that

1% of the population has the virus, i.e. $P(A) = .01$.

Test has a false positive rate of 1.5%, i.e. $P(B | A^c) = .015$.

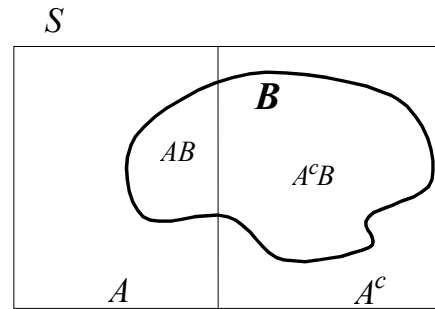
Test has a false negative rate of 0.3%, i.e. $P(B^c | A) = .003$. [So $P(B | A) = .997$]

We want: $P(B)$ and $P(A | B)$.

$$B = AB \cup A^c B$$

$$P(B) = P(AB) + P(A^c B)$$

$$P(AB) = P(A)P(B | A), \quad P(A^c B) = P(A^c)P(B | A^c)$$



$$P(B) = P(A)P(B | A) + P(A^c)P(B | A^c) \quad \text{"law of total probability"}$$

Here $P(B) = (.01)(.997) + (.99)(.015) = .02482$

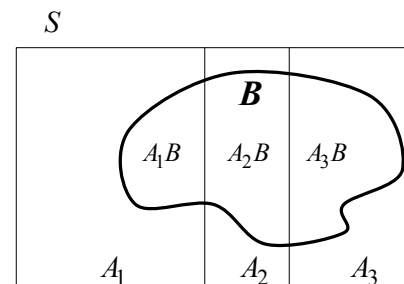
$$P(A | B) = \frac{P(AB)}{P(B)} = \frac{P(A)P(B | A)}{P(B)} = \frac{(.01)(.997)}{.02482} = \frac{.00997}{.02482} = .4017.$$

(Surprised?)

General Law of Total Probability and Bayes' formula:

Consider a *partition* of S into events A_1, A_2, \dots, A_k .

["Partition" means "disjoint and union is S ."]



Law of total probability:
$$P(B) = \sum_{j=1}^k P(A_j B) = \sum_{j=1}^k P(A_j)P(B | A_j)$$

Bayes' rule:
$$P(A_i | B) = \frac{P(A_i B)}{P(B)} = \frac{P(A_i)P(B | A_i)}{\sum_{j=1}^k P(A_j)P(B | A_j)}$$

[Don't memorize – understand and know how to derive!]

A windshield picture of Bayes' formula: Remember probabilities can be pictured as areas.

Example: $P(A_1) = 0.5$, $P(A_2) = 0.2$, $P(A_3) = 0.3$

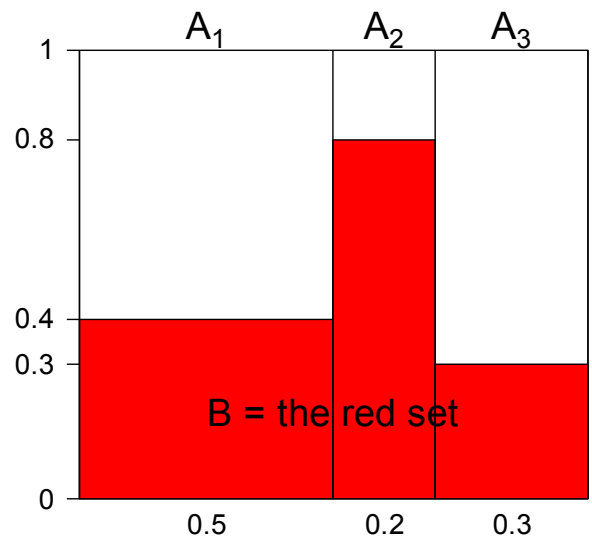
$P(B|A_1) = 0.4$, $P(B|A_2) = 0.8$, $P(B|A_3) = 0.3$

$$P(B) = (0.5)(0.4) + (0.2)(0.8) + (0.3)(0.3) = 0.45$$

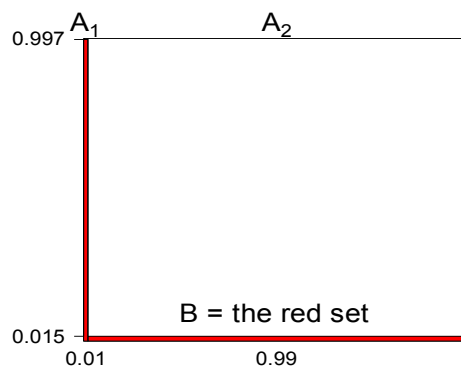
$$P(A_1|B) = \frac{(0.5)(0.4)}{0.45} = \frac{0.20}{0.45}, \quad P(A_2|B) = \frac{(0.2)(0.8)}{0.45} = \frac{0.16}{0.45},$$

$$P(A_3|B) = \frac{(0.3)(0.3)}{0.45} = \frac{0.09}{0.45}.$$

Makes Bayes' formula pretty obvious, right?



← Just for fun: $A_1 = \{\text{virus}\}$, $B = \{\text{positive blood test}\}$



1.11 An initial skirmish with likelihood and Bayesian statistics.

We have now learned enough to be able to take a look at the ideas we will use in doing Statistics and analyzing data with probabilistic models. The humble Bayes' rule, which is really little more than the definition of conditional probability, will be the key tool. And I was just joshing about the "skirmish"; I trust that likelihood and Bayesian statistics will become our close friends.

Example: Clinical trial [Prototype of Bayesian Statistical Inference].

21 patients are recruited into a trial. Each patient takes the drug on one occasion and is given a placebo on the other, with the order randomized.

The drug worked better than the placebo in 18 of the 21 pairs.

In the sample, the fraction where drug works better is 18/21.

We are interested in the unknown fraction in the population.

Let's call this unknown fraction Θ .

Θ is also interpretable as a probability. If we imagine choosing a random patient from the population, then Θ would be the probability that our randomly chosen patient would do better on the drug than on the placebo.

Model: Take 21 patients from the population.

Each has probability Θ of "success" (drug working better than placebo).

Number of successes: $X \sim B(21, \Theta)$.

[More terminology: Θ is called a parameter. A parameter is a number, typically unknown, that describes a probability distribution.]

In a nutshell:

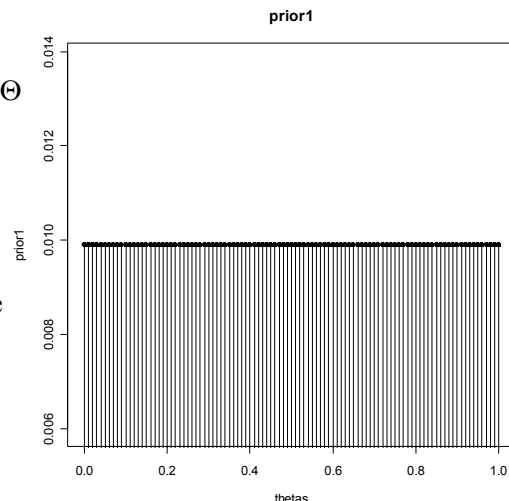
- $X \sim B(21, \Theta)$. We observed $X = 18$.
- Now what do we think about Θ ?

Θ is the unknown "cause".
 $X = 18$ is the "consequence".

In the Bayesian approach, the quantities that we do not know (like the unknown probability Θ) are thought of mathematically as random variables. Random variables have probability distributions. So, the final ingredient we need to specify as part of our model is a probability distribution for Θ . This is known as the *prior distribution* for the unknown parameter. This is typically chosen to reflect whatever beliefs and uncertainties about the unknown parameter value before we perform our experiment and observe the data. That is the sense of the word "prior" here; it is our distribution before seeing the data. If we don't have strong beliefs about plausible and implausible values for the parameter, we might just spread our prior probability over a wide range of possible parameter values.

Since we have talked only about discrete probability distributions so far, let us choose a discrete distribution for Θ now. For example, to model windshield-like beliefs that Θ could equally well lie anywhere in the interval $[0,1]$ we could imagine a fine grid of possible values for Θ , such as $\{0, .01, .02, .03, \dots, 1\}$, say, and postulate that each of these values has equal prior probability. Going along with that idea for now, we take our prior distribution for Θ to be uniform on $\{0, .01, .02, .03, \dots, 1\}$. That is,

$$P\{\Theta = \theta\} = \frac{1}{101} \text{ for } \theta \in \{0, .01, .02, .03, \dots, 1\}.$$



[[Why do I say "for now"? We are free to try other prior distributions later. We might want to do this in order to see how different choices of prior distribution affect our conclusions.]]

We'll use Bayes' rule to find the *posterior distribution* for Θ , given $X = 18$.

[Hey, it's just conditional probability. But it's statistics too...]

$$\begin{aligned} P\{\Theta = \theta | X = 18\} &= \frac{P(\{\Theta = \theta\} \cap \{X = 18\})}{P\{X = 18\}} \\ &= \frac{P\{\Theta = \theta\} P\{X = 18 | \Theta = \theta\}}{\sum_{\theta} P\{\Theta = \theta\} P\{X = 18 | \Theta = \theta\}} \end{aligned}$$

Note that the denominator, $P\{X = 18\}$, is just some number.
 It's nice to write Bayes' rule without this clutter as:

$$P\{\Theta = \theta \mid X = 18\} \propto P\{\Theta = \theta\} P\{X = 18 \mid \Theta = \theta\}.$$

We haven't lost any information here – we can always recover the proportionality constant at the end because the posterior distrib must add to 1 over the different possible θ values.

The first factor is the prior distribution, $P\{\Theta = \theta\} = \frac{1}{101}$, as specified above.

The other factor, $P\{X = 18 \mid \Theta = \theta\}$, is called the *likelihood*.

Bayes' rule: $\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$

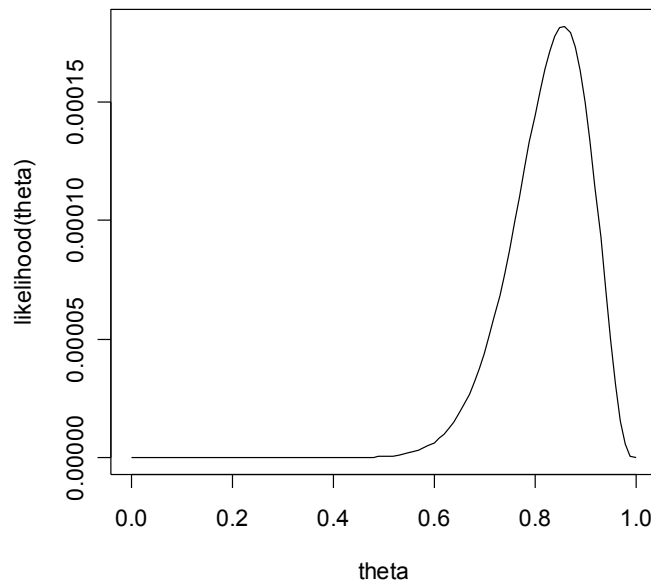
In abbreviated notation: $p(\theta \mid x) \propto p(\theta)p(x \mid \theta)$

Definition: The *likelihood* is the probability of the observed data, as a function of the unknown parameter.

Our model says that if $\Theta = \theta$, then $X \sim B(21, \theta)$. So the likelihood is

$$p(X = 18 \mid \theta) = \binom{21}{18} \theta^{18} (1 - \theta)^3 \propto \theta^{18} (1 - \theta)^3.$$

Here is a graph of the *likelihood function*:



ASIDE:

We can already understand what a *maximum likelihood estimator* is. Where is the maximum of the likelihood function? That is, what value of Θ maximizes the likelihood? From the graph, by eye we can see it's somewhere between 0.8 and 0.9, and using R we can get it as accurate numerically as we'd like, just by plotting over finer regions. In fact, some calculus shows that the answer is $18/21$ – a very intuitive guess for Θ having observed 18 successes out of 21 trials! Here $\hat{\Theta}_{MLE} = 18/21 = 0.857$.

BACK TO BAYESIAN FRAMEWORK:

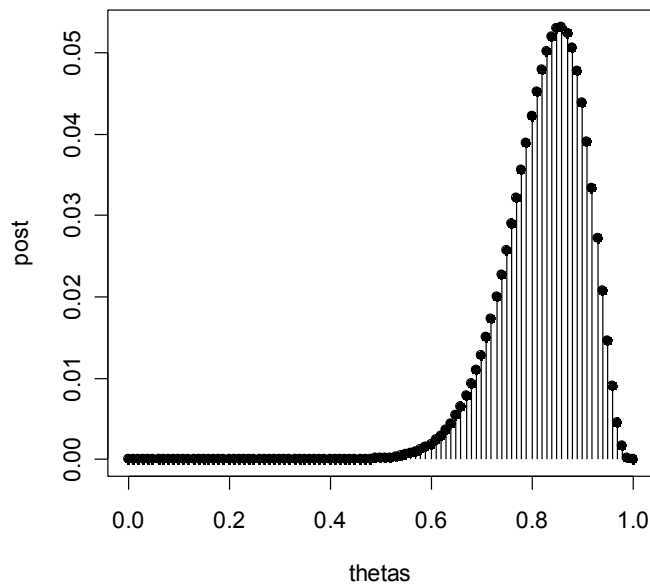
The prior is simply another constant here. In general we would have to multiply by the prior explicitly, but here the posterior is simply

$$p(\theta | X=18) \propto \theta^{18}(1-\theta)^3.$$

We can use R to compute and graph this:

```
thetas = seq(0,1,.01)
post = thetas^18*(1-thetas)^3
post = post/sum(post)
plot(thetas,post,type="h")
points(thetas,post,pch=19)
```

[[The line `post = post/sum(post)` resolves the proportionality “ \propto ”, making the sum equal to 1]]



E.g.,

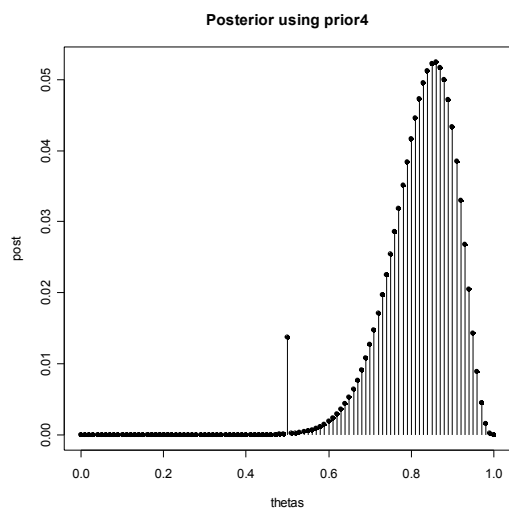
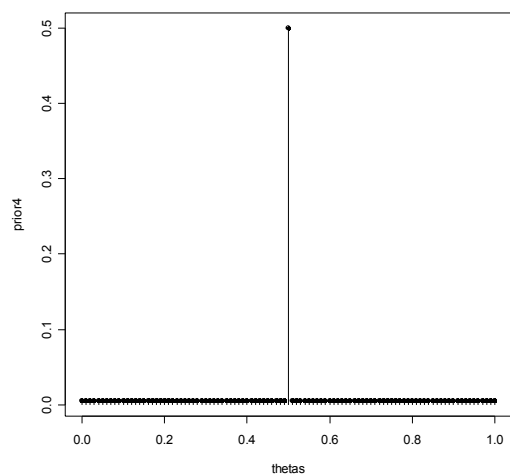
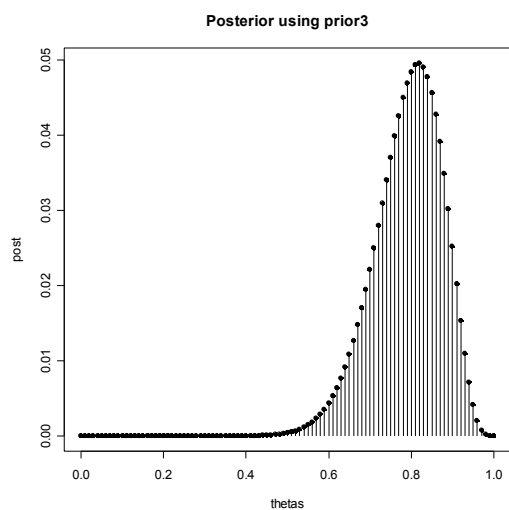
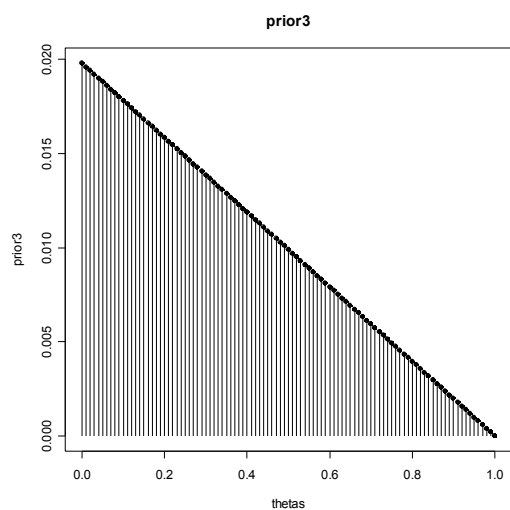
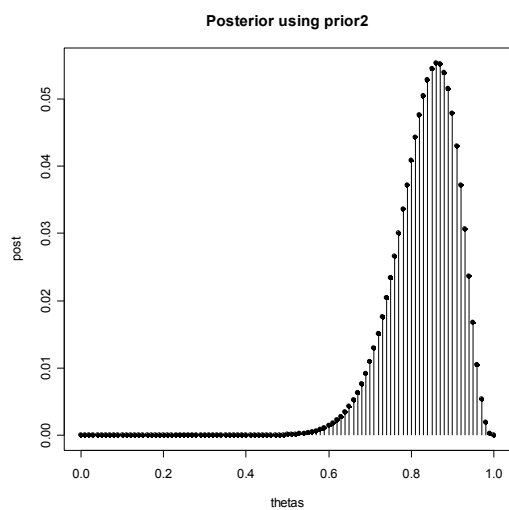
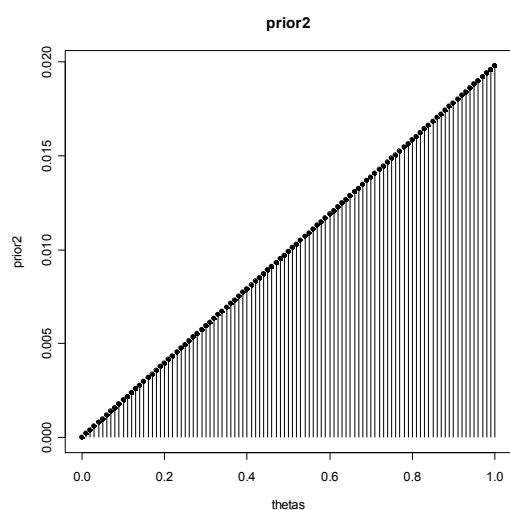
```
> sum(post[thetas>0.5])
[1] 0.999499
```

That is, our posterior probability that $\Theta > 0.5$ is 0.999499.

That's an example of how conclusions are drawn in Bayesian statistics:

- Call the data X , and call the parameter of interest Θ .
- We assume a distribution for Θ [called the *prior* “we’re writing it as $p(\theta)$ ”]
- Also assume a *model*, which is a collection of conditional distributions for X given the various possible values for Θ . [$p(x | \theta)$]
- Calculate the conditional distribution of Θ given the observed data $X = x$. This is called the *posterior distribution* [$p(\theta | x)$]
- Use your posterior distribution to give the probabilities of any statements you are interested in about Θ .

More examples of priors and the resulting posteriors:



```
> post4[51]
```

[1] 0.01376213

In prior4, the prior prob of $\Theta = 0.5$ was 0.5. Posterior prob of $\Theta = 0.5$ is 0.0138.

1.12 Specifying distributions: Probability densities, cumulative distributions

THE
NORMAL
LAW OF ERROR
STANDS OUT IN THE
EXPERIENCE OF MANKIND
AS ONE OF THE BROADEST
GENERALIZATIONS OF NATURAL
PHILOSOPHY. IT SERVES AS THE
GUIDING INSTRUMENT IN RESEARCHES
IN THE PHYSICAL AND SOCIAL SCIENCES AND
IN MEDICINE, AGRICULTURE, AND ENGINEERING.
IT IS AN INDISPENSABLE TOOL FOR THE ANALYSIS AND THE
INTERPRETATION OF THE BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT.

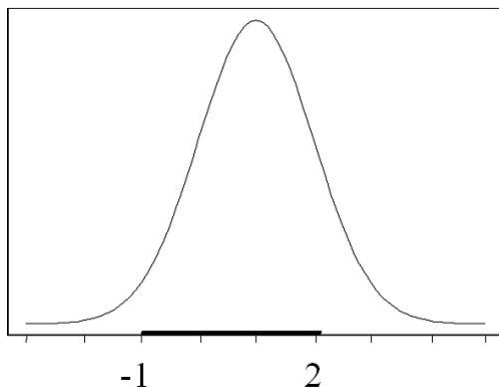
-- William Youden

Where are we? We have seen some of the framework of probability theory. We know what a probability distribution is for a discrete random variable; such distributions are described by probability *mass* functions. We've met an example of a family of such distributions: Binomial. We also have seen some of important fundamental ideas of Probability and Statistics, including Bayes' formula and likelihood functions. We'll get back to these shortly. Today we will extend our scope to "continuous" random variables, whose distributions can be described by probability *density* functions. Another description that works for both discrete and continuous distributions is the cumulative distribution function. We'll define these and also look at more examples of probability distributions. We'll also discuss joint, marginal, and conditional distributions, which come up in studying the joint behavior of more than one random variable---how random variables depend on each other.

For continuous random variables, there is no way to give a list of all possible values and their probabilities. In fact, for continuous random variables, as we will see, each individual value has probability 0.

Probability distributions of continuous random variables are described by densities.

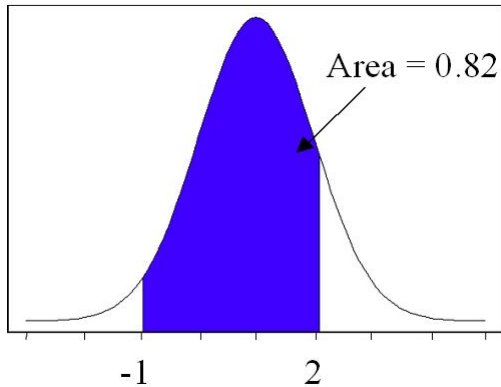
A probability density function is a nonnegative function $f : \mathbb{R} \rightarrow [0, \infty)$.



The probability density function of a random variable X is denoted f_X .

The job of a probability density is to determine probabilities, as follows:

$P\{a \leq X \leq b\} = \int_a^b f_X(x)dx$. That is, probabilities are given by areas under the probability density function. For the density shown above, the probability that the random variable lies between -1 and 2 is 0.82 :



A probability density function f_X satisfies the following conditions:

- $f_X(x) \geq 0$ for all x
- $\int_{-\infty}^{\infty} f_X(x)dx = 1$

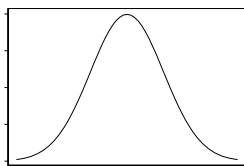
Note. Continuous distributions give probability 0 to individual values:

$$P\{X = a\} = \int_a^a f_X(x)dx = 0.$$

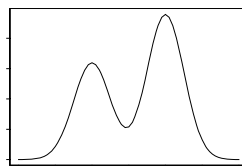
So, for example, with continuous random variables we don't need to be fastidious about whether or not we are including the endpoints of an interval, since

$$P\{a \leq X \leq b\} = P\{a < X < b\} = P\{a \leq X < b\} = P\{a < X \leq b\}.$$

Some words that describe distributions

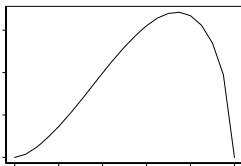


Symmetric,
unimodal

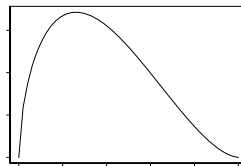


Bimodal

Skewness: Which is skewed to the left? Skewed to the right?



Skewed to the left



Skewed to the right

[[Most people want to give the opposite answers!]]

Definition: The *cumulative distribution function* F_X [“cdf”] is defined by

$$F_X(x) = P\{X \leq x\} = \int_{-\infty}^x f_X(t)dt.$$

Note: The Fundamental Theorem of Calculus says: $f_X(x) = F'_X(x)$.

That is: differentiating the cdf F gives the pdf f .

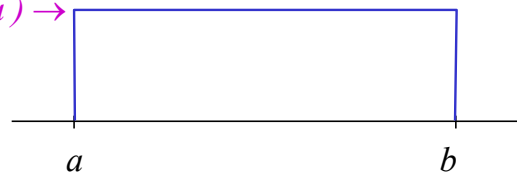
This is often useful!

Note: The cumulative distribution is defined for all random variables, discrete or continuous or whatever, in the same way: $F_X(x) = P\{X \leq x\}$.

Example: Uniform density. "Uniform" means "constant."

$X \sim U(a, b)$ has $f_X(x) = \frac{1}{b-a}$ for $a < x < b$.

height $1/(b-a) \rightarrow$



$$F_X(x) = \begin{cases} 0 & \text{for } x \leq a \\ \frac{x-a}{b-a} & \text{for } a < x < b \\ 1 & \text{for } x \geq b \end{cases}$$

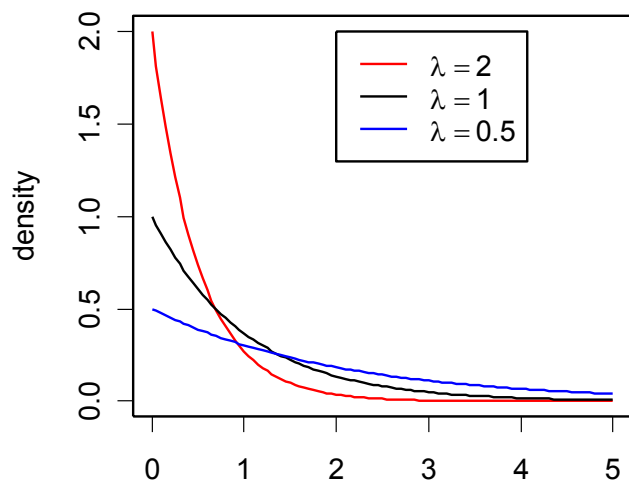
Example: Exponential distribution

The exponential distribution with "rate" parameter $\lambda > 0$ has density $f(t) = \lambda e^{-\lambda t}$ for $t > 0$.

If T has this density we write $T \sim \text{Exp}(\lambda)$.

This distribution is often used in modeling waiting times, times between events, etc. For example, the famous and useful "Poisson process" has exponentially distributed times between events.

Exponential densities



Cdf: $F_T(t) \triangleq P\{T \leq t\} = \int_{-\infty}^t f(s)ds = \int_0^t \lambda e^{-\lambda s} ds = -e^{-\lambda s} \Big|_0^t = (-e^{-\lambda t}) - (-1) = 1 - e^{-\lambda t}$ for $t \geq 0$.
(And of course $F_T(t) = 0$ for $t < 0$).

A nice, clean summary: $P\{T > t\} = e^{-\lambda t}$ for $t \geq 0$.

Example. The “memoryless property” of the exponential distribution:

$$\begin{aligned} P\{T > c + t \mid T > c\} &= \frac{P(\{T > c + t\} \cap \{T > c\})}{P\{T > c\}} \\ &= \frac{P\{T > c + t\}}{P\{T > c\}} = \frac{e^{-\lambda(c+t)}}{e^{-\lambda c}} = e^{-\lambda t} = P\{T > t\}. \end{aligned}$$

Thinking through the meaning of the last equation suggests an interpretation of the interesting “memoryless” name. Imagine you are waiting for something to happen, and T is the time you need to wait. Suppose you have already waited for c minutes, so that your information is that $T > c$. What is the probability $P\{T > c + t \mid T > c\}$ that you will need to wait more than t additional minutes? The memoryless property would say that this probability does not depend on how long you have already waited (that is, c) and in fact it is the same as if you were starting over again. Many processes are well modeled by this distribution; for example you might be waiting for the next pop when cooking popcorn in your microwave oven. How would a waiting time violate this property? For example, suppose you are waiting to pay at a checkout line in a store, and there is one customer in front of you. From your experience at this store, you have found typical customers may take just around 30 seconds to pay and be out of your way. But you wait and wait for this customer, and find that it has already been 5 minutes and counting. At that point, do you still expect this customer will be finished in about 30 additional seconds? Probably not – instead, you would probably figure that there is some sort of problem and it is quite likely that this service time will drag on much longer.

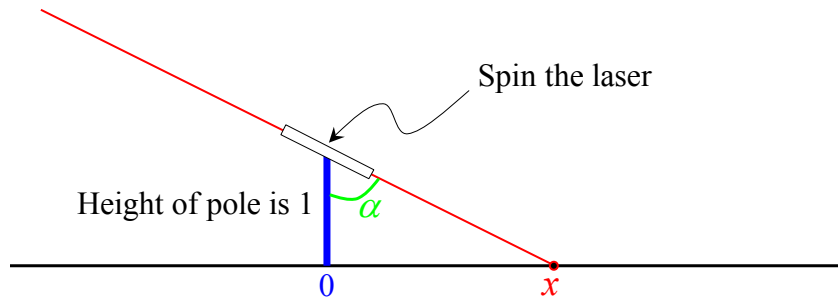
Example. Suppose X has pdf f_X and a and b are numbers, with $a \neq 0$. Define $Y = aX + b$. Then Y has pdf $f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$. [Derivation: Exercise.] (Remember graphing trig functions?)

[Example of the example: If $T \sim \text{Exp}(1)$, then $\frac{T}{\lambda} \sim \text{Exp}(\lambda)$.

The exponential distributions form a "scale family."]

Example. THE LASER POLE

Spin the laser to a random angle α . Here we'll say "random" means "uniformly distributed."



We'll be making further use of this story as we go along; it provides interesting examples in probability and statistical inference.

Trigonometry: $x = \tan(\alpha)$.

Say the angle $A \sim U\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ and $X = \tan(A)$.

What is the pdf of X ?

To find the pdf, f , first find the cdf, F , then differentiate. [Common trick!]

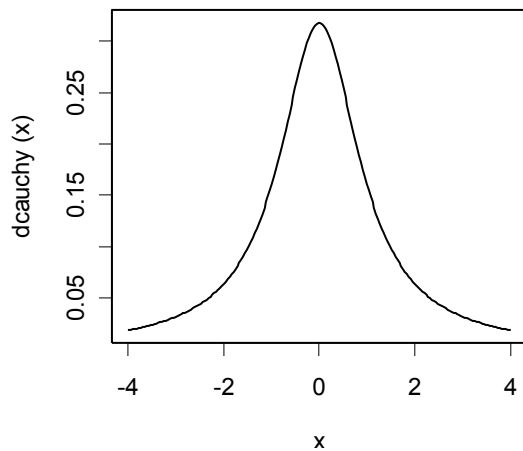
$$\begin{aligned} F(x) &= P\{X \leq x\} = P\{\tan(A) \leq x\} = P\left\{A \leq \tan^{-1} x\right\} \\ &= \frac{\tan^{-1} x - (-\pi/2)}{(\pi/2) - (-\pi/2)} = \frac{1}{\pi} \left(\tan^{-1} x + \frac{\pi}{2} \right) \end{aligned}$$

Of course you remember your useful calculus facts, like $\frac{d}{dx} \tan^{-1}(x) = \frac{1}{1+x^2}$ (haven't we all used this a million times?). So the density is

$$f(x) = F'(x) = \frac{1}{\pi(1+x^2)}.$$

This is called the *Cauchy distribution*. We have just found the pdf of the Cauchy distribution. It's also the "*t distribution with one degree of freedom*"

```
> plot(dcauchy, -4, 4)
```



From the calculation

```
> 1-2*pcauchy(-4)
[1] 0.8440417
```

we can see that the above plot actually shows only about 84% of the Cauchy distrib's mass, with the remaining 16% lying beyond -4 or 4 .

A simultaneous plot of the Normal and Cauchy densities shows that the Cauchy has “fatter tails” than the Normal.

Here, try this:

```
win.graph(); par(cex=1.3,lwd = 2)
plot(dnorm,-5,5,ylab="density",xlab="")
plot(dcauchy,-5,5,add=T,col="red")
legend(1,.4,c("Normal","Cauchy"),col=c("black","red"),lty=c(1,1))
```

It is much easier to get extreme outliers from a Cauchy distribution than from a Normal; you can see this by doing

```
?rcauchy
x=rcauchy(1000); hist(x)
x=rnorm(1000); hist(x)
```

Summary: We now have 3 ways to describe – actually, to specify – a probability distribution:

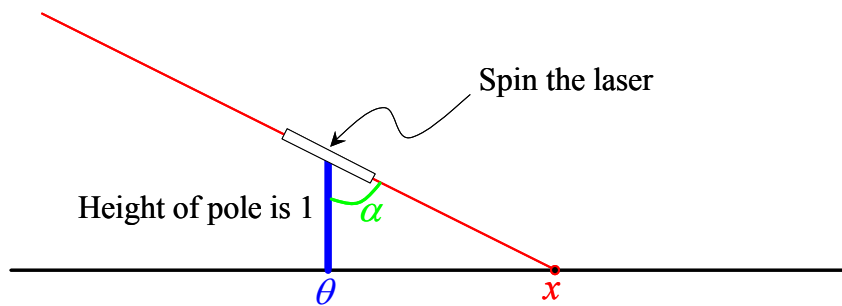
- **probability mass function:** for discrete distributions. Doesn't work for continuous distributions.
- **probability density function:** for continuous distributions. Doesn't work for discrete distributions.
- **cumulative distribution function:** works for any distribution: discrete, continuous, or "mixed".

1.13 More on the laser pole: Likelihood and inference.

Having looked at the clinical trial example, I figured that before forgetting about the ideas developed there, I'd show you a second example of using our basic paradigm for doing statistics in order to solidify the ideas. Hey, if it's a paradigm, it must be important!

This is also a nice example that illustrates the value of the probability infrastructure we have been developing. Here we will develop a statistical inference from first principles for a situation that is not one of the standard ones for which “canned” statistical procedures may be found among the menu items in a

standard statistical package. Starting from a physical description of the problem, we use the probability we just worked out in formulating a model, and apply our Bayesian paradigm.



Suppose we get to observe 4 values of X .
E.g. we see $X_1 = 7.8$, $X_2 = 1.0$, $X_3 = 2.0$, $X_4 = 9.2$.

We do not know θ ; you could imagine that the pole and the laser are invisible. We get to observe only the locations of the 4 points of light.

Our task is to estimate θ .

Real Statisticians use the likelihood function.

The likelihood of a given θ value is the probability [mass or density] of the observed data (the x 's), if that θ were the true value.

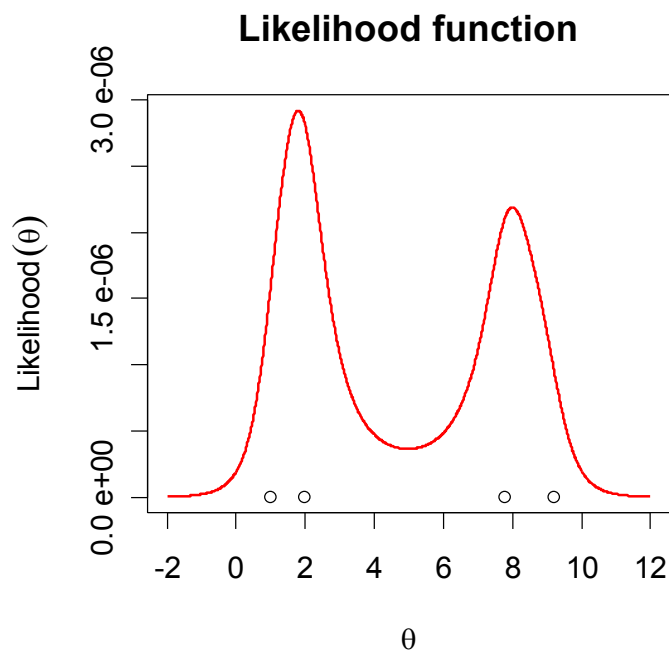
If the true value were θ , then the probability density of the X_i 's would be $f(x|\theta) = \text{dcauchy}(x - \theta)$, where dcauchy is the Cauchy density function $\frac{1}{\pi(1+x^2)}$ that we calculated in detail last time.

That is, $f(x|\theta) = \text{dcauchy}(x - \theta) = \frac{1}{\pi(1+(x-\theta)^2)}$.

This is a payoff for knowing how to work with probabilities, densities, etc. Now we can model probabilistic phenomena and, given data from such a random phenomenon, do inference about unknown features of that phenomenon.

For our particular data [$X_1 = 7.8$, $X_2 = 1.0$, $X_3 = 2.0$, $X_4 = 9.2$], the likelihood function is $L(\theta) = f(7.8|\theta)f(1.0|\theta)f(2.0|\theta)f(9.2|\theta)$.

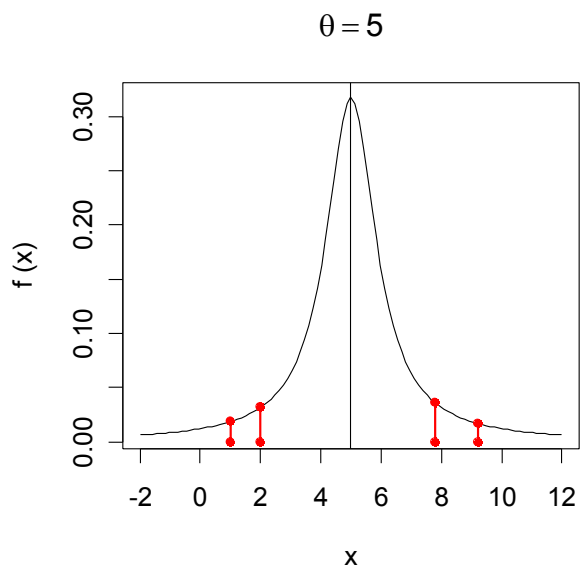
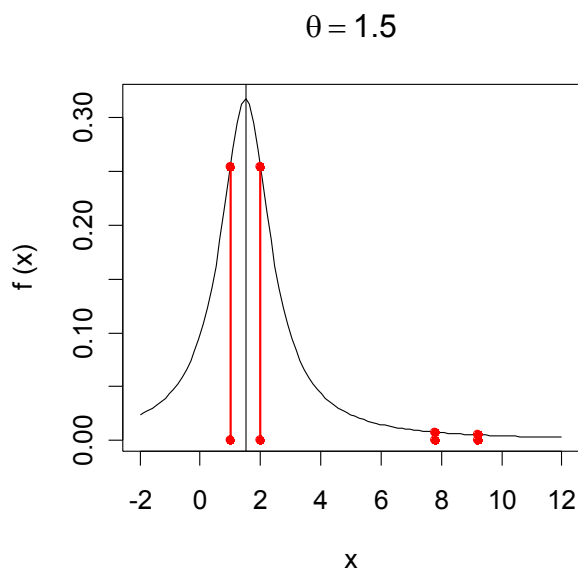
We can calculate this for a fine grid of θ values and plot it.

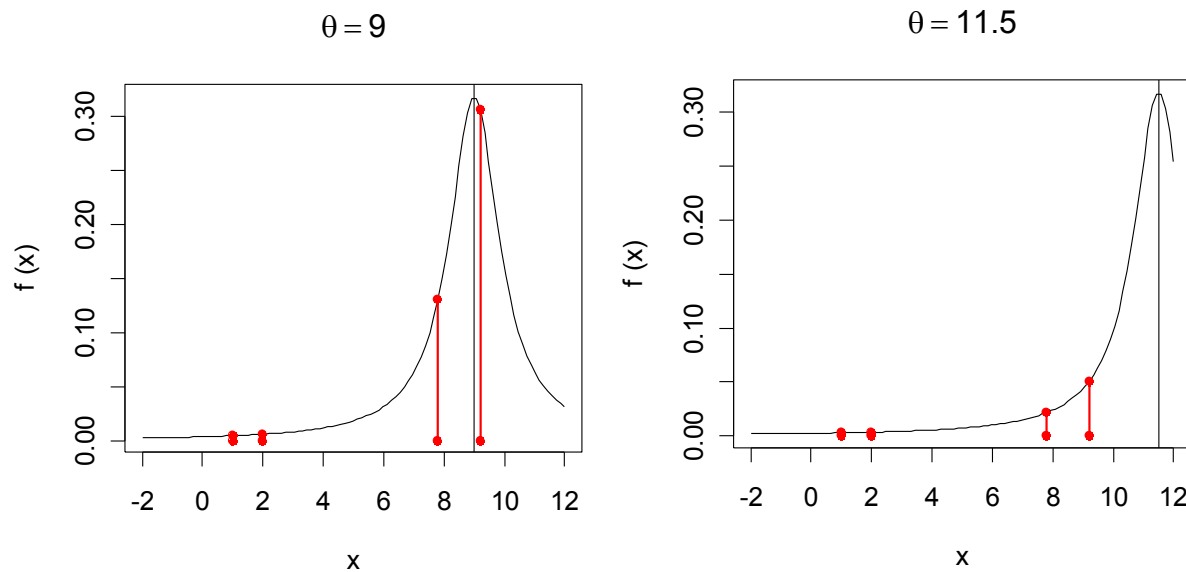


The likelihood function is the main object underlying all of our statistical inferences.

I'll show you pictures illustrating [[conceptually]] how the likelihood is calculated at four theta values: $\theta = 1.5, 5, 9, \text{ and } 11.5$.

In each case the likelihood for that theta value is the product of the heights of the 4 red lines.





To complete a Bayesian analysis of this data, we would want a posterior distribution for θ , not just the likelihood function.

By Bayes' rule, we know $\text{posterior} \propto \text{prior} \times \text{likelihood}$

So we need to choose a prior and multiply it by the likelihood function we just calculated to get the posterior.

E.g. if we choose a uniform prior, then the posterior looks just like the likelihood (except that it is scaled by whichever proportionality constant converts the likelihood function into a probability density function).

An example of such a posterior probability calculation using R is given in the file 050914.R.

Here are the commands to do it, given more concisely:

```
x = c(1, 2, 7.8, 9.2)    # observed data

thetas = seq(-2, 12, .01) # again we'll consider a fine grid of thetas

m = length(thetas)
lik = thetas
for(i in 1:m){lik[i] <- Lcauchy(thetas[i],x)}    # lik is now the likelihoods of the thetas

# If we take a uniform prior, then posterior is proportional to likelihood:
post = lik/sum(lik)

# E.g. let's find the posterior prob that 0<theta<4:
indices = (thetas > 0)&(thetas<4)
sum(post[indices])
```

1.14 Joint distributions

We've discussed the distribution of a random variable. Most statistical questions involve consideration of more than one variable, how they depend on each other, how using one can help to predict another, and so

on. Here we'll set up the framework and terminology of distributions for more than a single random variable at a time.

For example, if I tell you I have two random variables X and Y each of which has the $\text{Bern}(1/2)$

distribution $\begin{cases} 0 \text{ w.p. } 1/2 \\ 1 \text{ w.p. } 1/2 \end{cases}$, that says nothing about how X and Y are related. E.g. they could be

independent coin tosses. Or they could be the same coin toss recorded twice. Or

In language we are about to introduce, in this simple example, we would say that knowing X and Y both have the "marginal" distribution $\text{Bern}(1/2)$ does not determine their "joint" distribution.

A **random vector** is a vector of random variables – i.e. something like (X, Y) or (X_1, \dots, X_n) .

The probability distribution of a random vector is often called a **joint distribution**. This emphasizes that we are considering the probability distribution of more than one random variable simultaneously.

For a discrete random vector, the joint distribution, like any distribution, tells the possible values of the random vector and their probabilities. It is a probability mass function.

$$\begin{aligned} f_{X_1, \dots, X_n}(x_1, \dots, x_n) &= P\{(X_1, \dots, X_n) = (x_1, \dots, x_n)\} \\ &= P\{X_1 = x_1, \dots, X_n = x_n\} \end{aligned}$$

So in a sense the word "joint" is unnecessary. The joint distribution of a collection of random variables is just a distribution (same idea as we've been discussing), applied to a vector.

To avoid subscripts and dot-dot-dots for today we'll mostly just look at 2 variables X and Y .

$$f_{X,Y}(x, y) = P\{(X, Y) = (x, y)\} = P\{X = x, Y = y\}.$$

Of course, $f_{X,Y}(x, y) \geq 0$ for all x and y , and $\sum_{x,y} f_{X,Y}(x, y) = 1$.

For a continuously distributed random vector, the probability distribution may be given by a "joint density." Again, it's just the same idea as we have seen, applied to more than one dimension. Probabilities are determined by integrals.

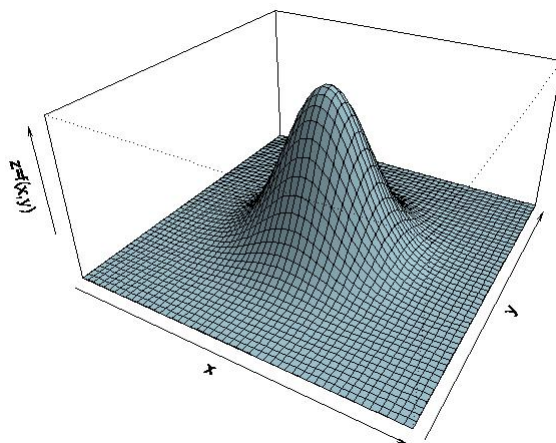
E.g.: " (X, Y) has joint density f " means that, for sets A in \mathbb{R}^2 ,

$$P\{(X, Y) \in A\} = \iint_{(x,y) \in A} f(x, y) \, dx \, dy.$$

E.g. of the e.g.:

$$P\{X \leq a, b \leq Y \leq c\} = \int_{y=b}^c \int_{x=-\infty}^a f(x, y) \, dx \, dy.$$

Of course, $f(x, y) \geq 0$ for all x and y ,



and $\iint f(x, y) \, dx \, dy = 1$.

Here is an example of a "bivariate Normal" joint density function:

1.15 Marginal and conditional distributions

The joint distribution of X and Y contains full probabilistic information about X and Y .

In particular, from it we can derive the distribution of X considered separately as just a single random variable. Such a distribution is called a **marginal distribution**.

And, of course, the same with Y – it has a marginal distribution too.

We get marginal distributions as follows:

$$f_X(x) = \sum_y f_{X,Y}(x, y), \quad f_Y(y) = \sum_x f_{X,Y}(x, y) \quad \text{in the discrete case.}$$

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dx \quad \text{in the continuous case.}$$

Why? E.g., for the discrete case,

$$f_X(x) = P\{X = x\} = \sum_y P\{X = x, Y = y\} = \sum_y f_{X,Y}(x, y) \quad \text{[because } Y \text{ has to take some value } y \text{]}$$

Example: We toss a coin 3 times, and define

X = number of heads on first toss (0 or 1)

Y = total number of heads in all 3 tosses

The joint distribution of X and Y is given by this table:

	Y			
X	0	1	2	3
0	1/8	1/4	1/8	0
1	0	1/8	1/4	1/8

We get the marginal distributions of X and Y by summing, because

$$P\{X = x\} = \sum_y P\{X = x, Y = y\} \quad \text{and} \quad P\{Y = y\} = \sum_x P\{X = x, Y = y\}.$$

	Y			
X	0	1	2	3
0	1/8	1/4	1/8	0
1	0	1/8	1/4	1/8
	1/8	3/8	3/8	1/8
				1/2

(Now I think you can see the origin of the name "marginal.")

That is, $X = \begin{cases} 0 & \text{with prob } 1/2 \\ 1 & \text{with prob } 1/2 \end{cases}$, $Y = \begin{cases} 0 & \text{with prob } 1/8 \\ 1 & \text{with prob } 3/8 \\ 2 & \text{with prob } 3/8 \\ 3 & \text{with prob } 1/8 \end{cases}$.

Similarly, given a joint distribution, we can find **conditional distributions** by using the definition of conditional probability.

Example: (Continuation of previous example)

What is the conditional distribution of X given that $Y = 2$?

E.g., $P\{X = 0 \mid Y = 2\} = \frac{P\{X = 0, Y = 2\}}{P\{Y = 2\}} = \frac{f_{X,Y}(0,2)}{f_Y(2)} = \frac{1/8}{3/8} = \frac{1}{3}$.

The conditional distrib of X given that $Y = 2$ is $\begin{cases} 0 & \text{with prob } 1/3 \\ 1 & \text{with prob } 2/3 \end{cases}$.

Note: We get this conditional distrib by "normalizing" the " $Y = 2$ " column

1/8
1/4

 of the joint distribution table, getting

1/3
2/3

.

Similarly, to get the conditional distrib of Y given $X = 0$, we would normalize the " $X = 0$ " row

0	1	2	3
1/8	1/4	1/8	0

of the table, getting the distrib

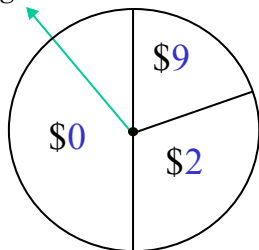
0	1	2	3
1/4	1/2	1/4	0

General formula: $f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$. [And of course $f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$]

NOTE: These formulas hold for both discrete and continuous variables, interpreting the f 's as prob mass functions and prob density functions, respectively.

1.16 Expectation

Example -- The spinner game:



$$\text{Winnings: } X = \begin{cases} 0 & \text{with prob } 0.5 \\ 2 & \text{with prob } 0.3 \\ 9 & \text{with prob } 0.2 \end{cases}$$

For this random variable we define the **expectation** (or **mean**) of X as $E(X) = (0)(.5) + (2)(.3) + (9)(.2) = 2.4$ dollars .

Motivation: WHY IS THE MEAN DEFINED THIS WAY?

This definition makes the "Law of Large Numbers" (LLN) work.

LLN says: The mean of a random variable is the limiting long-run average that would result from taking repeated independent copies of the random variable (having the same probability distribution).

[The LLN is a generalization of the relationship between probabilities and long-run frequencies. We'll prove a version of this soon!]

Let X_1, X_2, \dots be independent having the same distrib as X .

Define the running average $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$.

We want to define $E(X)$ so that $\bar{X}_n \rightarrow E(X)$ as $n \rightarrow \infty$.

$$X_1 + X_2 + \dots + X_n = 0(\# \text{ of } 0\text{'s}) + 2(\# \text{ of } 2\text{'s}) + 9(\# \text{ of } 9\text{'s})$$

$$\bar{X}_n = 0\left(\frac{\# \text{ of } 0\text{'s}}{n}\right) + 2\left(\frac{\# \text{ of } 2\text{'s}}{n}\right) + 9\left(\frac{\# \text{ of } 9\text{'s}}{n}\right)$$

$$\bar{X}_n = 0\left(\underbrace{\frac{\# \text{ of } 0\text{'s}}{n}}_{0.5}\right) + 2\left(\underbrace{\frac{\# \text{ of } 2\text{'s}}{n}}_{0.3}\right) + 9\left(\underbrace{\frac{\# \text{ of } 9\text{'s}}{n}}_{0.2}\right)$$

That is, $\bar{X}_n \rightarrow E(X)$, where $E(X)$ is as defined above.

- The pattern is: $E(X)$ is a sum of values times probabilities.

Definition: For a discrete random variable X having probability mass function $f_X(x) = P\{X = x\}$, the mean [or expectation] $E(X)$ is defined to be $E(X) = \sum_x x f_X(x)$. For a continuous random variable having pdf f_X we define $E(X) = \int x f_X(x) dx$.

A physical interpretation: The expectation of a distribution is its center of mass, or "balance point."

Example: Let $X \sim \text{Unif}(a, b)$. Find $E(X)$.

First, should be able to guess the answer... what is it? *Recommended mental habit: Always take a guess first!*

OK, having guessed what the answer should be, we'll let ourselves calculate.

X has pdf $f(x) = \frac{1}{b-a}$ on the interval $a < x < b$. So

$$E(X) = \int_a^b x \left(\frac{1}{b-a} \right) dx = \frac{1}{b-a} \frac{x^2}{2} \Big|_a^b = \frac{1}{b-a} \left(\frac{b^2}{2} - \frac{a^2}{2} \right) = \frac{b+a}{2}$$

"Law of the unconscious statistician"

Suppose X has density f_X . Let g be a function and let $Y = g(X)$.

Then the *Law of the Unconscious Statistician* says: $E(Y) = E(g(X)) = \int g(x) f_X(x) dx$.

[[It's a really silly name, but I think it's nice to use names whenever possible...]]

That is, we do not have to calculate, f_Y , the density of Y , and calculate $E(Y)$ as $\int y f_Y(y) dy$.

You can still think of $\int g(x) f_X(x) dx$ as of the form $\int (\text{values})(\text{probabilities})$. So it seems a pretty intuitive way to calculate the expectation – so much so that it is often treated as obvious and used without being aware that it is not in fact just the definition of expectation.

Discrete case: $E(g(X)) = \sum_x g(x) P\{X = x\}$.

Example: Define $Y = g(U) = U^2$ where $U \sim U(0,1)$. Find $E(Y)$.

Direct-from-the-definition method: We could do this by finding the density of Y as follows.

$$F_Y(y) = P\{U^2 \leq y\} = P\{U \leq y^{1/2}\} = y^{1/2} \text{ for } 0 < y < 1.$$

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{1}{2} y^{-1/2}.$$

$$E(Y) = \int_0^1 y f_Y(y) dy = \int_0^1 y \frac{1}{2} y^{-1/2} dy = \frac{1}{2} \int_0^1 y^{1/2} dy = \left(\frac{1}{2} \right) \left(\frac{2}{3} \right) = \frac{1}{3}.$$

The Law of the Unconscious Statistician tells us we can also do it this way:

$$E(Y) = \int u^2 f(u) du = \int_0^1 u^2 du = \frac{1}{3}, \text{ which is considerably simpler.} \quad \blacktriangleright$$

The LOUS also works for functions of more than one variable.

E.g. suppose $Z = g(X, Y)$ and we want $E(Z)$. Here's how:

$$E(Z) = \sum_{x,y} g(x,y) P\{X=x, Y=y\} \text{ in discrete case,}$$

$$E(Z) = \iint_{\text{all } x,y} g(x,y) f_{X,Y}(x,y) dx dy \text{ in continuous case.}$$

Example: Linearity of expectation Let $Z = g(X, Y) = X + Y$.

$$\begin{aligned} E(Z) &= \sum_{x,y} (x+y) P\{X=x, Y=y\} \\ &= \sum_x x \sum_y P\{X=x, Y=y\} + \sum_y y \sum_x P\{X=x, Y=y\} \\ &= \sum_x x P\{X=x\} + \sum_y y P\{Y=y\} = E(X) + E(Y) \end{aligned}$$

Hey, $E(X+Y) = E(X) + E(Y)$!

[The "hey" was ironical; this was "expected", right?] ▶

Example: Expectations of products and independence. Let $Z = g(X, Y) = XY$.

Suppose X and Y are independent; some people write this as $X \perp\!\!\!\perp Y$.

$$\begin{aligned} E(Z) &= \sum_{x,y} (xy) \underbrace{P\{X=x, Y=y\}}_{P\{X=x\}P\{Y=y\} \text{ if } X,Y \text{ indep}} \\ &= \sum_x x P\{X=x\} \sum_y y P\{Y=y\} \\ &= E(X) E(Y) \end{aligned}$$

☺ That is, if $X \perp\!\!\!\perp Y$ then $E(XY) = E(X)E(Y)$.

Indicator variables. An indicator variable is a random variable that takes only two values: 0 and 1. We speak of the "indicator of an event". For an event A , the indicator $I(A)$ is the random variable that takes value 1 when A occurs, and 0 when A does not occur.

Expectation of an indicator variable: Since $I(A)$ has distrib $\begin{cases} 1 \text{ with prob } P(A) \\ 0 \text{ with prob } 1-P(A) \end{cases}$,

$$E(I(A)) = P(A).$$

That is, the expectation of an indicator variable is the probability of the event it's indicating. So "expectation" generalizes "probability," and "random variable" generalizes "event".

Expectation of the binomial distribution (an example of use of indicator variables):

Let $X \sim B(n, p)$, so that $P\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, \dots, n$.

First, guess: What is the mean of X ??

$$(1.3) \quad E(X) = \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k} = \dots \text{heroic virtuoso calculations} \dots$$

"THE INDICATOR TRICK" is to write X as a sum of indicators.

Define I_1, I_2, \dots, I_n by $I_k = I\{k^{\text{th}} \text{ trial is a success}\}$, so that $X = I_1 + I_2 + \dots + I_n$ counts the number of successes in n trials.

So $E(X) = E(I_1 + \dots + I_n) = E(I_1) + \dots + E(I_n) = p + \dots + p = np$.

This is much easier than using (1.3), and also more intuitive.

Inclusion-exclusion principle (another illustration of indicator variables):

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

And how about more general statements of the same kind, like

$$(1.4) \quad \begin{aligned} P(A \cup B \cup C \cup D) = & P(A) + P(B) + P(C) + P(D) \\ & - P(AB) - P(AC) - P(AD) - P(BC) - P(BD) - P(CD) \\ & + P(ABC) + P(ABD) + P(ACD) + P(BCD) \\ & - P(ABCD) \end{aligned}$$

How can these be derived?

One way is to work very hard to keep track of all the regions in a Venn Diagram and make sure we count each region the appropriate number of times [as in Degroot and Schervish pages 41-42, or Grinstead and Snell pages 104-105].

Here is the cool method using indicators and simple mechanical algebra. First note that, in general,

$$(1.5) \quad I(A^c) = 1 - I(A),$$

$$(1.6) \quad I(AB) = I(A)I(B).$$

In (1.6), the " AB " in " $I(AB)$ " means the intersection, $AB \triangleq A \cap B$.

We want $P(A \cup B \cup C \cup D) = E[I(A \cup B \cup C \cup D)]$.

DeMorgan's Law, a simple piece of logic, says $(A \cup B \cup C \cup D)^c = A^c B^c C^c D^c$.

$$\begin{aligned}
 1 - I(A \cup B \cup C \cup D) &= (I(A^c))(I(B^c))(I(C^c))(I(D^c)) \\
 &= (1 - I(A))(1 - I(B))(1 - I(C))(1 - I(D)) \\
 &= 1 - I(A) - I(B) - I(C) - I(D) \\
 &\quad + I(A)I(B) + I(A)I(C) + I(A)I(D) + I(B)I(C) + I(B)I(D) + I(C)I(D) \\
 &\quad - I(A)I(B)I(C) - I(A)I(B)I(D) - I(A)I(C)I(D) - I(B)I(C)I(D) \\
 &\quad + I(A)I(B)I(C)I(D) \\
 &= 1 - I(A) - I(B) - I(C) - I(D) \\
 &\quad + I(AB) + I(AC) + I(AD) + I(BC) + I(BD) + I(CD) \\
 &\quad - I(ABC) - I(ABD) - I(ACD) - I(BCD) \\
 &\quad + I(ABCD)
 \end{aligned}$$

Take expected values of the red stuff:

$$\begin{aligned}
 1 - P(A \cup B \cup C \cup D) \\
 &= 1 - P(A) - P(B) - P(C) - P(D) \\
 &\quad + P(AB) + P(AC) + P(AD) + P(BC) + P(BD) + P(CD) \\
 &\quad - P(ABC) - P(ABD) - P(ACD) - P(BCD) \\
 &\quad + P(ABCD)
 \end{aligned}$$

This proves (1.4)!

Expectation as the "best prediction" for a value drawn from a distribution:

Here is a game: I am about to draw a random variable X from a probability distribution. You are to make a guess, hoping to be close to the actual realized random value of X . In fact, if you guess c , then according to this game, you are penalized $(X - c)^2$ dollars. That is, the penalty is your "squared error." You want to choose a guess that minimizes your expected squared error.

The answer is that the best choice for this game is to choose $c = E(X)$. To derive this fact, we simply think of the expected squared error $E((X - c)^2)$ as a function of c by defining $g(c) = E((X - c)^2)$, and we find the value of c that minimizes the function g . We can write g as

$$g(c) = E((X - c)^2) = E(X^2 - 2cX + c^2) = c^2 - 2\mu c + E(X^2), \text{ where } \mu \text{ denotes } E(X). \text{ So } g \text{ is}$$

simply a quadratic function of c . Since the coefficient of the squared term c^2 is positive (it is 1), the graph of g as a function of c is a parabola that opens upward, and so to find the c minimizing g we can set the derivative $g'(c)$ equal to 0, as follows: $g'(c) = 2c - 2\mu = 0$, which is solved by the choice $c = \mu$. This is what we wanted to show: taking $c = \mu = E(X)$ minimizes the expected squared loss in our game.

Note: If our game were altered so that your penalty were $|X - c|$, your best choice of c would not be the mean, but rather the median.

Notes about the mean and median

Both indicate some sort of "center" of the distribution of values.

Robustness: This means "insensitive to a few extreme observations." E.g. imagine typo of adding a few extra zeros to a number when entering the sample data.

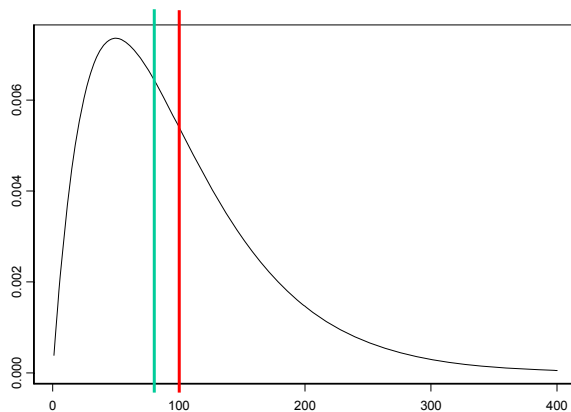
Mean is not robust, median is.

E.g.: compare

57, 72, 76, 93, 94
to

57, 72, 7

E.g.: for the following density, which value is the mean and which is the median?



1.17 Variance.

Definition: Let X be a random variable having mean $E(X) = \mu$. The *variance* of X , denoted $\text{var}(X)$, is defined to be $\text{var}(X) = E[(X - \mu)^2] = \int (x - \mu)^2 f_X(x) dx$. [or $\sum_x (x - \mu)^2 f_X(x)$]

The random variable $(X - \mu)^2$ measures [well, in a squared way] how far away X is from its mean.

The variance is just the expected value of $(X - \mu)^2$. So the variance is small for rv's that are always close to their mean, and large for random variables having a distribution that is spread out over a wide range of values.

Definition: The *standard deviation* of X , denoted $\text{SD}(X)$, is defined to be the square root of the variance: $\text{SD}(X) = \sqrt{\text{var}(X)}$.

Example: The spinner.

Here $X = \begin{cases} 0 & \text{with prob } 0.5 \\ 2 & \text{with prob } 0.3 \\ 9 & \text{with prob } 0.2 \end{cases}$. We already calculated its mean, $\mu = 2.4$.

$$\text{So: } X - \mu = \begin{cases} 0 - 2.4 = -2.4 & \text{with prob } 0.5 \\ 2 - 2.4 = -0.4 & \text{with prob } 0.3 \\ 9 - 2.4 = 5.6 & \text{with prob } 0.2 \end{cases} \quad (X - \mu)^2 = \begin{cases} (-2.4)^2 & \text{with prob } 0.5 \\ (-0.4)^2 & \text{with prob } 0.3 \\ (5.6)^2 & \text{with prob } 0.2 \end{cases}$$

$$\text{var}(X) = (0 - 2.4)^2 \cdot 0.5 + (2 - 2.4)^2 \cdot 0.3 + (9 - 2.4)^2 \cdot 0.2 = 9.2. \quad \leftarrow \text{Units? Squared dollars!}$$

$$\text{SD}(X) = \sqrt{9.2} = 3.03. \quad \leftarrow \text{Dollars.} \quad \blacktriangleright$$

Comments:

- The variance is the mean squared deviation of the random variable from its mean. The standard deviation is a “root mean square.”
- Remember that gambling interpretation of the mean that we discussed last time? The mean μ is the best guess “ c ” to minimize your expected squared error when guessing the value of a random variable; that is, $E[(X - c)^2]$ is minimized by taking $c = \mu$. Well, the variance $\text{var}(X)$ is your expected loss playing this game as well as you can.
- The standard deviation is more naturally interpretable as a measure of the “spread” of a distribution. Its units are the same as the units of X , whereas the variance is in squared units.
- The variance has nicer mathematical properties.

Those “nicer mathematical properties” also explain why the variance has dominated in practice over other measures of “spread” that might seem more intuitive, such as $E(|X - \mu|)$ where μ is the mean, or perhaps median, of the distribution of X .

Another related thought for comfort: The “squaring” operation would please Pythagorus, and summing squares is in a sense more natural than summing absolute values.

(↑) What am I talking about here?

Ask yourself: “In the plane, how far apart are the points (2,1) and (6,4)?”

Obvious note: the mean and variance of a random variable depend only on the distribution of the random variable. It's the distribution that's important. So it's convenient to write equations like $E[B(n, p)] = np$, and we'll do this sort of thing.

Two simple properties: Let X be a random variable, and let c be a number.

$$\text{var}(X + c) = \text{var}(X)$$

$$\text{var}(cX) = c^2 \text{var}(X)$$

An alternative formula that is often convenient: $\text{var}(X) = E(X^2) - (E(X))^2$.

Why? Because:

$$\begin{aligned}
\text{var}(X) &= E[(X - \mu)^2] \\
&= E[X^2 - 2\mu X + \mu^2] \\
&= E(X^2) - 2\mu E(X) + \mu^2 \\
&= E(X^2) - 2\mu^2 + \mu^2 \\
&= E(X^2) - \mu^2
\end{aligned}$$

A common mistake seen among the uninitiated is to believe that $E(X^2)$ should be the same as $(E(X))^2$. Now in fact we know that

- $E(X^2)$ is always at least as large as $(E(X))^2$
- in fact $E(X^2)$ is always larger than $(E(X))^2$ unless $\text{var}(X) = 0$
- the amount by which $E(X^2)$ is larger than $(E(X))^2$ is the variance $\text{var}(X)$.

Example: Uniform distribution

If $V \sim U(a, b)$, we already know that $E(V) = \frac{a+b}{2}$.

Claim: $\text{var}(V) = \frac{(b-a)^2}{12}$.

Why? Note

- “ $U(a, b) \sim a + U(0, b-a)$ ” \leftarrow explain what this type of loose but convenient notation means
- $\text{var}(U(a, b)) = \text{var}(U(0, b-a))$
- $U(0, b-a) \sim (b-a)U(0, 1)$
- $\text{var}(U(0, b-a)) = (b-a)^2 \text{var}(U(0, 1))$

So it's sufficient to find $\text{var}(U(0, 1))$. Let $U \sim U(0, 1)$.

$$\begin{aligned}
E(U^2) &= \int_0^1 u^2 du = \frac{1}{3}. \quad E(U) = \frac{1}{2}. \\
\text{var}(U) &= E(U^2) - (EU)^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}.
\end{aligned}$$



Example: Indicator variable

Let $I \sim \text{Bern}(p)$, that is, I has (or is that “have”?) distribution

$$I = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases}.$$

$$\text{var}(I) = E(\underbrace{I^2}_I) - (EI)^2 = (EI) - (EI)^2 = p - p^2 = p(1-p).$$



Next, how about $\text{var}(X + Y)$ for two random variables X and Y ? E.g., is it true that

$$\text{var}(X + Y) \stackrel{??}{=} \text{var}(X) + \text{var}(Y) ?$$

No. Clear counterexamples include:

$$\text{var}(X + X) = \text{var}(2X) = 4 \text{var}(X) \neq \text{var}(X) + \text{var}(X)$$

$$\text{var}(X + (-X)) = \text{var}(0) = 0 \neq 2 \text{var}(X) = \text{var}(X) + \text{var}(-X)$$

However, if X and Y are **independent**, then we **do** have
 $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$

[This last property is the main “nice mathematical property” for the variance.]

Proof: The general result is derived as follows.

Given X with mean μ_x and Y with mean μ_y .

We want the variance of $X + Y$, which has mean $\mu_x + \mu_y$.

$$\begin{aligned} \text{var}(X + Y) &= E\left[\left((X + Y) - (\mu_x + \mu_y)\right)^2\right] \\ &= E\left[\left((X - \mu_x) + (Y - \mu_y)\right)^2\right] \\ &= E\left[(X - \mu_x)^2 + (Y - \mu_y)^2 + 2(X - \mu_x)(Y - \mu_y)\right] \\ &= \text{var}(X) + \text{var}(Y) + 2 \underbrace{E\left[(X - \mu_x)(Y - \mu_y)\right]}_{\text{"cov}(X, Y)\text{"}} \end{aligned}$$

But when X and Y are independent,

$$\text{cov}(X, Y) = E\left[(X - \mu_x)(Y - \mu_y)\right] = \underbrace{E(X - \mu_x)}_0 \underbrace{E(Y - \mu_y)}_0 = 0.$$



Summary:

- In general, $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2 \text{cov}(X, Y)$. [[More about covariance later]]
- When X and Y are independent, $\text{cov}(X, Y) = 0$, so that $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$.

Definition: If $\text{cov}(X, Y) = 0$, then we say that X and Y are *uncorrelated*.

[[We will define the correlation below; the term uncorrelated comes from the correlation being 0.]]

From the above we see that in order to have $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$ we do not really need X and Y to be independent, but rather it is enough for X and Y to be uncorrelated.

As an aside, note again the Pythagorean metaphor: A variance is like a kind of squared length.

- In geometry, if two vectors are orthogonal, then the squared length of the sum is the sum of the squared lengths.
- In probability, if two random variables are uncorrelated, then the variance of the sum is the sum of the variances.

Example: variance of the Binomial distribution

Let $X \sim B(n, p)$. We know $E(X) = np$, so to get $\text{var}(X)$ it is enough to calculate

$$E(X^2) = \sum_{k=0}^n k^2 \binom{n}{k} p^k (1-p)^{n-k} = \dots ???, \text{ which again is unappetizing.}$$

But again the "indicator trick" makes this easy. Think of X as a sum of indicators, $X = I_1 + I_2 + \dots + I_n$.

Using the independence of the I_k 's, we can say

$$\text{var}(X) = \underbrace{\text{var}(I_1)}_{p(1-p)} + \dots + \underbrace{\text{var}(I_n)}_{p(1-p)} = np(1-p) ! \leftarrow \text{punctuation, not factorial}$$

Moral: it's often better to think of a count X as a sum of indicators than as something like

$$P\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k}.$$



Important example: Sample means

Suppose X_1, X_2, \dots, X_n are iid – “**independent and identically distributed**” – with mean μ and variance σ^2 (and so SD = σ). We think of X_1, X_2, \dots, X_n as a random sample from a population having mean μ and variance σ^2 .

Define $S = X_1 + \dots + X_n$ and $\bar{X} = \frac{S}{n}$. So the random variable \bar{X} is called the “sample mean.” Let's find the mean and variance of the random variable \bar{X} .

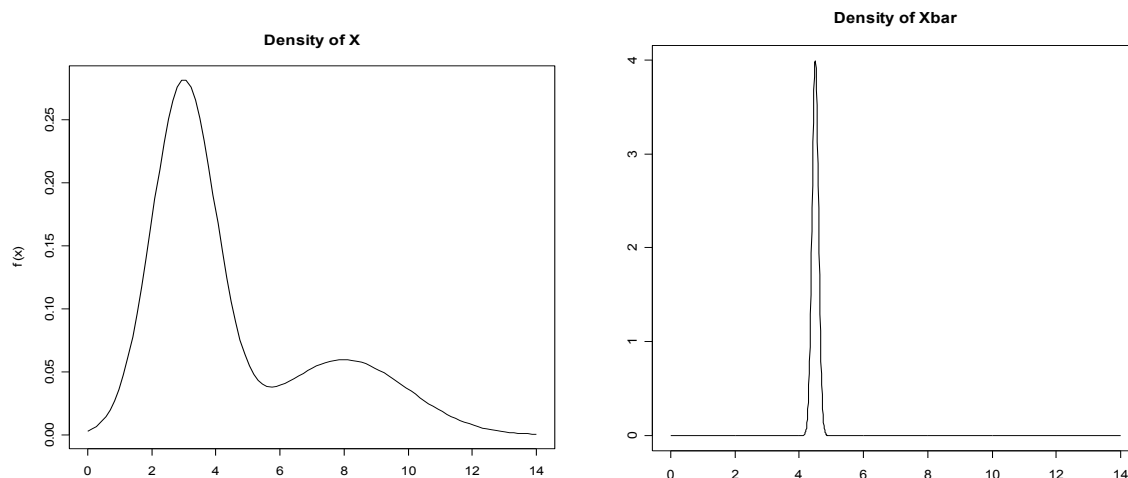
$E(S) = E(X_1) + \dots + E(X_n) = n\mu$, so $E(\bar{X}) = \mu$. This can be summarized (or obfuscated) by the statement: The mean of the sample mean is the population mean. Next let's find the variance and the standard deviation of the sample mean.

$$\text{var}(S) = \text{var}(X_1) + \dots + \text{var}(X_n) = n\sigma^2.$$

$$\text{var}(\bar{X}) = \text{var}\left(\frac{S}{n}\right) = \frac{\text{var}(S)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

$$\text{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

The \sqrt{n} in the denominator shows that the sample mean is less variable for a large sample than it is for a smaller sample.



1.18 Time for a new distribution The Geometric distribution.

Imagine repeated independent trials with success probability p .

You repeat until the first success, and stop.

Let T denote the total number of trials you perform.

$$P\{T = k\} = (1 - p)^{k-1} p \quad \text{for } k = 1, 2, 3, \dots$$

We say T has a *Geometric distribution with success probability p* , and write this as $T \sim \text{Geom}(p)$.

E.g. if T is the number of time needed to roll a die until getting the first “2”, then $T \sim \text{Geom}(1/6)$.

Mean of Geometric distribution:

$$E(T) = \sum_{k=1}^{\infty} k P\{T = k\} = \sum_{k=1}^{\infty} k (1 - p)^{k-1} p.$$

One could figure out the sum starting with a calculus trick:

$$\sum_{k=1}^{\infty} x^k = \frac{1}{1-x} - 1 \quad \text{for } 0 < x < 1 \quad \text{suggests}$$

$$\sum_{k=1}^{\infty} k x^{k-1} = \frac{d}{dx} \left(\frac{1}{1-x} - 1 \right) = \frac{1}{(1-x)^2}.$$

Taking $x = 1 - p$ gives $\sum_{k=1}^{\infty} k (1 - p)^{k-1} = \frac{1}{(1 - (1 - p))^2} = \frac{1}{p^2}$, so that $E(T) = p \frac{1}{p^2} = \frac{1}{p}$.

$$pE(T) = p + (1 - p) = 1$$

→ There is a more insightful way to explain the answer using the idea of the Law of Large Numbers.

☺ Later we'll also see how to do it as an example of the “Law of Total Expectation.”

The Law of Large Numbers, which we will prove a version of shortly, if I may randomly permute my prepositions and appositive phrases, says that the average of many independent random variables coming from a certain distribution will converge to the expected value of that distribution.



Suppose n is huge. Think of the n trials as consisting of a succession of some [[random]] number K of “cycles,” with each cycle being the segment of trials between successive successes, so to speak.

Question: about how big is K ? [[Hint: Note K is simply the number of successes in n trials!]]

Answer: $K \approx np$, i.e., $\frac{K}{n} \rightarrow p$. [E.g. $\frac{K}{n} \rightarrow \frac{1}{6}$ for success = die rolling “2”]

Note $\sum_{i=1}^K T_i \approx n$ (just a few less than n actually).

$$\underbrace{\bar{T}_K = \frac{1}{K} \sum_{i=1}^K T_i}_{E(T)} \approx \frac{n}{K} = \frac{1}{(K/n)} \rightarrow \frac{1}{p}.$$

So $E(T)$ must be $\frac{1}{p}$!

Example of a modeling application: This type of modeling is done quite a lot in population genetics; for example it is the basis of the famous “coalescent” model. We imagine comparing Y chromosomes from two randomly sampled men, and want to model how they are different. We think about how the two men are descended from a common ancestor some number of generations back. How many generations?

We consider a simplistic model, with:

- discrete generations,
- a fixed population size n [[number of men]] in each generation
- “random parenting,” so that the father of each man is equally likely to be any of the n men in the previous generation, and all “choices” of fathers are independent.

Say T = number of generations back to the MRCA [“most recent common ancestor”] of the two men.

Then: $T \sim \text{Geom}(1/n)$.

So $E(T) = n$ generations.

If we were looking at 3 men, the expected time until at least one pair “coalesces” is about $n/3$ generations; for 4 men, it’s about $n/6$; for k men it’s about $n/\left(\frac{k}{2}\right)$. Can you see that this coalescence process is like a succession of birthday problems? An interesting fact about this model is that the expected time to get to the MRCA of all n men in the current generation is $\sim 2n$ generations ago!

1.19 Notes and example on covariance

Recall: We defined $\text{cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$.

Intuition: $\text{cov}(X, Y)$ tends to be larger if $X - \mu_x$ and $Y - \mu_y$ tend to behave similarly, that is, have the same sign and tend to be large together.

This arose as the extra term that came up in the formula

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2 \text{cov}(X, Y).$$

Note $\text{var}(X) = \text{cov}(X, X)$. This is sometimes a useful way to find a variance.

→ Properties of covariance:

An alternative formula: $\text{cov}(X, Y) = E(XY) - E(X)E(Y)$.

$\text{cov}(X, Y) = 0$ if X and Y are independent.

$$\text{cov}(X, Y + c) = \text{cov}(X, Y)$$

$$\text{cov}(X, cY) = c \text{cov}(X, Y)$$

$$\text{cov}(X, Y) = \text{cov}(Y, X). \quad \leftarrow [\text{wow}]$$

$$\text{cov}(X, Y + Z) = \text{cov}(X, Y) + \text{cov}(X, Z).$$

From (5) and (6) get more general linear behavior of cov, such as

$$\text{cov}(X + W, Y + Z) = \text{cov}(X, Y) + \text{cov}(X, Z) + \text{cov}(W, Y) + \text{cov}(W, Z),$$

$$\text{and in general: } \text{cov}\left(\sum_i X_i, \sum_j Y_j\right) = \sum_i \sum_j \text{cov}(X_i, Y_j)$$

→ Examples:

$$\odot \text{var}(cX) = \text{cov}(cX, cX) = c^2 \text{cov}(X, X) = c^2 \text{var}(X)$$

$$\begin{aligned} \odot \text{var}(X + Y) &= \text{cov}(X + Y, X + Y) \\ &= \text{cov}(X, X) + \text{cov}(X, Y) + \text{cov}(Y, X) + \text{cov}(Y, Y) \\ &= \text{var}(X) + 2 \text{cov}(X, Y) + \text{var}(Y) \end{aligned}$$

Example: Consider a simple model of a DNA sequence, as a sequence of n iid draws from the uniform distribution on $\{1, 2, 3, 4\}$.

[E.g., 1 means "a", 2 means "c", 3 means "g", 4 means "t".

"Uniform means "prob $\frac{1}{4}$ each". Of course we could generalize this.]

Suppose we have a sequence that is $n = 10000$ characters long.
Let N_1 and N_2 denote the number of 1's and 2's in the sequence.

Consider the excess of 1's over 2's, that is, $N_1 - N_2$. Of course, $E(N_1 - N_2) = 0$.

How variable is $N_1 - N_2$? What is its standard deviation?

Hey, why don't you make your own private guesses... e.g.,

- About how much variability would you expect to see in $N_1 - N_2$?
- [Just for fun] Which is more variable: $N_1 - N_2$ or $N_1 + N_2$?

We can investigate this distribution by simulation, or by using Math and Probability Theory. We'll do both.

The simulation in R can be done as in the R command file for today, and shown in class.

For the paper and pencil method...

Problem: Find the variance of $N_1 - N_2$.

Answer: $\text{var}(N_1 - N_2) = \text{var}(N_1) + \text{var}(N_2) - 2\text{cov}(N_1, N_2)$... *...to blackboard in class...*

Since $N_1 \sim B(10000, 0.25)$, $\text{var}(N_1) = 10000 \times (1/4) \times (3/4) = 1875$ and so of course $\text{var}(N_2) = 1875$.

Write $N_1 = \sum_{i=1}^{10000} I\{X_i = 1\}$ and $N_2 = \sum_{i=1}^{10000} I\{X_i = 2\}$. So

$$\text{cov}(N_1, N_2) = \text{cov}\left(\sum_{i=1}^{10000} I\{X_i = 1\}, \sum_{j=1}^{10000} I\{X_j = 2\}\right) = \sum_{i=1}^{10000} \sum_{j=1}^{10000} \text{cov}(I\{X_i = 1\}, I\{X_j = 2\}).$$

But for $i \neq j$, $I\{X_i = 1\}$ and $I\{X_j = 1\}$ are independent, so that $\text{cov}(I\{X_i = 1\}, I\{X_j = 2\}) = 0$.

$$\text{So } \text{cov}(N_1, N_2) = \sum_{i=1}^{10000} \text{cov}(I\{X_i = 1\}, I\{X_i = 2\}).$$

$$\text{But } \text{cov}(I\{X_i = 1\}, I\{X_i = 2\}) = E(I\{X_i = 1\}I\{X_i = 2\}) - \underbrace{E(I\{X_i = 1\})}_{1/4} \underbrace{E(I\{X_i = 2\})}_{1/4},$$

and $E(I\{X_i = 1\}I\{X_i = 2\}) = 0$, because the product $I\{X_i = 1\}I\{X_i = 2\}$ is always 0!

So $\text{cov}(I\{X_i = 1\}, I\{X_i = 2\}) = -1/16$, and so

$$\text{cov}(N_1, N_2) = \sum_{i=1}^{10000} \text{cov}(I\{X_i = 1\}, I\{X_i = 2\}) = 10000 \times (-1/16) = -625, \text{ and so}$$

$$\text{var}(N_1 - N_2) = 1875 + 1875 - 2(-625) = 5000. \quad \text{SD}(N_1 - N_2) = \sqrt{5000} = 70.71.$$

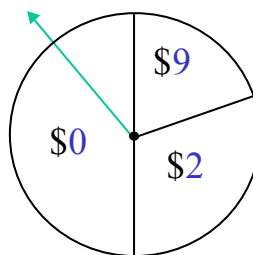


1.20 Law of Large Numbers: simulation, proof and pathology.

1.20.1 Law of large numbers simulation

Recall our spinner game:

$$\text{Winnings: } X = \begin{cases} 0 & \text{with prob } 0.5 \\ 2 & \text{with prob } 0.3 \\ 9 & \text{with prob } 0.2 \end{cases}$$



$$E(X) = (0)(.5) + (2)(.3) + (9)(.2) = 2.4 \text{ dollars.}$$

Let's sample from this distribution and look at the sample means.

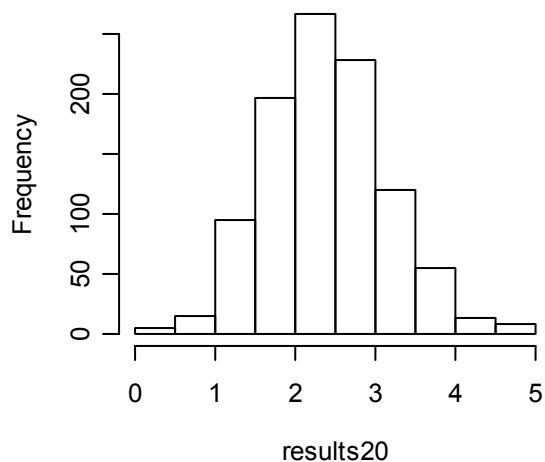
E.g. to investigate average winnings \bar{X}_n for $n = 20$:

```
n = 20
nit = 1000
results20 = numeric(nit)
for(i in 1:nit){
  spins = sample(c(0,2,9),n,prob=c(.5,.3,.2),replace=T)
  results20[i] = mean(spins)
}
```

Note: Here we are repeating the experiment "average the winnings from 20 spins of the wheel" 1000 times.

```
hist(results20)
```

Histogram of results20

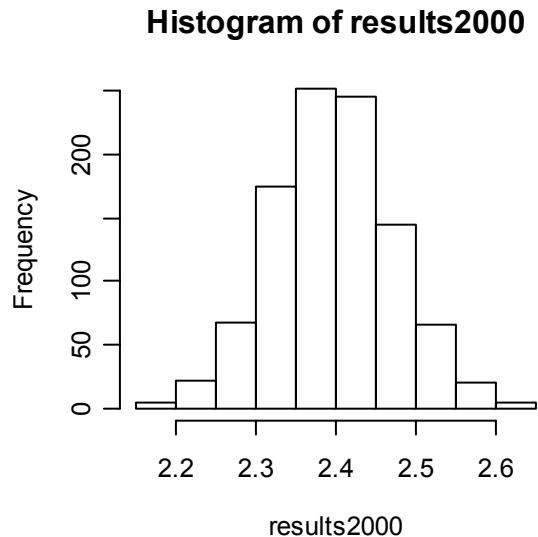


```
> quantile(results20,c(.05,.95))
5% 95%
1.30 3.65
```

```

n = 2000
nit = 1000
results2000 = numeric(nit)
for(i in 1:nit){
  spins = sample(c(0,2,9),n,prob=c(.5,.3,.2),replace=T)
  results2000[i] = mean(spins)
}
hist(results2000)

```



```

> quantile(results2000,c(.05,.95))
      5%      95%
2.271475 2.521500

```

Look what happened to the width of the distribution, as measured, e.g., by the separation between the 5th and 95th percentiles:

For $n = 20$ the separation was $3.65 - 1.30 = 2.35$
 For $n = 2000$ the separation is $2.5215 - 2.2715 = 0.250$.

The width of the distribution went down by a factor of about 10. This agrees with the theory we have developed about standard deviations, since 10 is the square root of the ratio of the sample sizes.

Evidently the sample mean has a tighter distribution around the true mean $[\mu = E(X) = 2.4]$ when $n = 2000$ than when $n = 20$ --- larger sample sizes lead to more accurate estimation.

For example, we could compare the fraction of times our simulated sample means fell within 0.1 of μ [that is, $|\bar{X} - 2.4| < 0.1$, or, in other words, $2.3 < \bar{X} < 2.5$] when $n = 2000$ and when $n = 20$:

```

> sum(results20 > 2.3 & results20 < 2.5)/nit
[1] 0.058      #← Few of these sample means were accurate to within 0.1

```

```

> sum(results2000 > 2.3 & results2000 < 2.5)/nit
[1] 0.817      #← Most of these sample means were accurate to within 0.1

```

1.20.2 Law of Large Numbers (“LLN”) statement

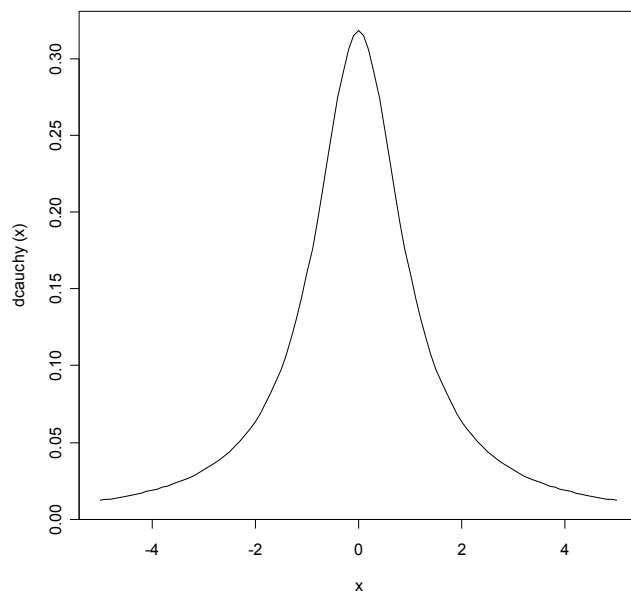
THEOREM: Let X_1, X_2, \dots be iid with mean μ and variance σ^2 , and define $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. Then $\bar{X}_n \rightarrow \mu$ “in probability” as $n \rightarrow \infty$. That is, for each $\varepsilon > 0$, $P\{|\bar{X}_n - \mu| < \varepsilon\} \rightarrow 1$ as $n \rightarrow \infty$.

For example, the simulation of the spinner game we just did indicated that for $\varepsilon = 0.1$, the probability $P\{|\bar{X}_{2000} - 2.4| < 0.1\}$ is much bigger than $P\{|\bar{X}_{20} - 2.4| < 0.1\}$, and the Theorem says that in fact as $n \rightarrow \infty$, the probability $P\{|\bar{X}_n - 2.4| < 0.1\}$ continues to increase, converging to 1 in the limit.

1.20.3 Pathology: Cauchy simulation and LLN “violation”

Recall the Cauchy distribution has density $f(x) = \frac{1}{\pi(1+x^2)}$.

```
plot(dcauchy, -5, 5)
```



```
n = 20
nit = 1000
results = numeric(nit)
for(i in 1:nit){
  spins = rcauchy(n)
  results[i] = mean(spins)
}
hist(results, main="Cauchy with n = 20")
quantile(results, c(.05, .95))
hist(results[abs(results)<10], breaks=100, col="red")
sum(abs(results)<10)/length(results)

> quantile(results, c(.05, .95))
      5%      95%
-6.814480  5.730381
```

Try the same experiment with $n = 2000$

```
> quantile(results, c(.05, .95))  
      5%      95%  
-7.409742  7.267072
```

What's going on? Have we discovered a violation of the LLN??

Check the assumptions stated in the theorem.

Let X_1, X_2, \dots be *iid* with mean μ and variance σ^2 , and define $\bar{X}_n = \sum_{i=1}^n X_i$. Then $\bar{X}_n \rightarrow \mu$ "in probability" as $n \rightarrow \infty$.

Q: What is the mean of the Cauchy distribution?

Obviously 0, right? After all, the density is symmetric about $x = 0$.

⊗ Actually, No: the mean of the Cauchy distribution does not exist!

If X has a Cauchy distribution centered at 0 (the *median* is indeed 0), and we try to find the mean $E(X)$, we are led to the calculation

$$\int_{-\infty}^{\infty} x \left(\frac{1}{\pi(1+x^2)} \right) dx = \int_0^{\infty} x \left(\frac{1}{\pi(1+x^2)} \right) dx + \int_{-\infty}^0 x \left(\frac{1}{\pi(1+x^2)} \right) dx$$

But since $\frac{x}{\pi(1+x^2)} \sim \frac{1}{\pi x}$ as $x \rightarrow \infty$ and $\int_c^{\infty} \frac{1}{\pi x} dx = \infty$, we see that $\int_0^{\infty} x \left(\frac{1}{\pi(1+x^2)} \right) dx = \infty$,

and our purported calculation of $E(X)$ gives the prototypical undefined quantity, " $\infty - \infty$ ".

No mean, no convergence to a mean, no LLN. No joy in Meanville.

In fact, the average of a sample of n Cauchy random variables has a Cauchy distribution – the average of n observations has the same distribution as a single observation!

1.20.4 Law of Large Numbers: Proof

Finally we will get to prove a law of large numbers.

I said "*a* law..." rather than "*the* law..." because there are other stronger results (perhaps saying something stronger, or assuming different conditions). Our version is called a "weak law of large numbers," in fact.

We'll start with some tools that are also of more general use.

Markov inequality: Let Z be a nonnegative random variable and let $c > 0$. Then $P\{Z \geq c\} \leq \frac{E(Z)}{c}$.

Proof: Note $c I\{Z \geq c\} = \begin{cases} c & \text{if } Z \geq c \\ 0 & \text{otherwise} \end{cases}$, so $c I\{Z \geq c\} \leq Z$.

Taking expected values gives $cP\{Z \geq c\} \leq E(Z)$, which is Markov's inequality. ►

Chebyshev inequality. Let Y be a random variable with mean μ_y and SD σ_y , and let $c > 0$. Then

$$P\{|Y - \mu_y| \geq c \sigma_y\} \leq \frac{1}{c^2}.$$

For example, the probability that Y differs from its mean by more than 10 standard deviations can be no more than $\frac{1}{10^2} = \frac{1}{100}$.

Proof: Define $Z = \frac{(Y - \mu_y)^2}{\sigma_y^2}$, a nonnegative r.v. Note $E(Z) = 1$, which is convenient.

Observe that

$$P\{|Y - \mu_y| \geq c \sigma_y\} = P\left\{\frac{(Y - \mu_y)^2}{\sigma_y^2} \geq c^2\right\} = P\{Z \geq c^2\},$$

and by the Markov inequality, $P\{Z \geq c^2\} \leq \frac{E(Z)}{c^2} = \frac{1}{c^2}$. ►

Proof of law of large numbers.

We want to show: $P\{|\bar{X}_n - \mu| < \varepsilon\} \rightarrow 1$ as $n \rightarrow \infty$.

[[This is also written as “ $\bar{X}_n \xrightarrow{P} \mu$ as $n \rightarrow \infty$,” and we say “ \bar{X}_n converges in probability to μ as $n \rightarrow \infty$.”]]

We apply the Chebyshev inequality to the rv $Y = \bar{X}_n$, which has mean $\mu_y = \mu$ and SD $\sigma_y = \frac{\sigma}{\sqrt{n}}$:

$$P\left\{|\bar{X}_n - \mu| \geq c \frac{\sigma}{\sqrt{n}}\right\} \leq \frac{1}{c^2}.$$

Choosing $c = \frac{\varepsilon\sqrt{n}}{\sigma}$, so that $c \frac{\sigma}{\sqrt{n}} = \varepsilon$ and $\frac{1}{c^2} = \frac{\sigma^2}{\varepsilon^2 n}$, we get

$$P\{|\bar{X}_n - \mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2 n} \rightarrow 0 \text{ as } n \rightarrow \infty,$$

that is,

$$\lim_{n \rightarrow \infty} P\{|\bar{X}_n - \mu| \geq \varepsilon\} = 0.$$

Taking complements gives

$$\lim_{n \rightarrow \infty} P \left\{ \left| \bar{X}_n - \mu \right| < \varepsilon \right\} = 1 .$$

That is, $\bar{X}_n \xrightarrow{P} \mu$ as $n \rightarrow \infty$.



1.21 Normal distributions

THE
NORMAL
LAW OF ERROR
STANDS OUT IN THE
EXPERIENCE OF MANKIND
AS ONE OF THE BROADEST
GENERALIZATIONS OF NATURAL
PHILOSOPHY. IT SERVES AS THE
GUIDING INSTRUMENT IN RESEARCHES
IN THE PHYSICAL AND SOCIAL SCIENCES AND
IN MEDICINE, AGRICULTURE, AND ENGINEERING.
IT IS AN INDISPENSABLE TOOL FOR THE ANALYSIS AND THE
INTERPRETATION OF THE BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT.

-- William Youden

1.21.1 "Discovering" the normal distribution

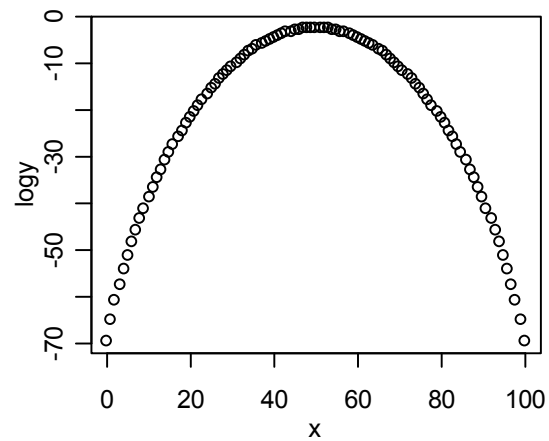
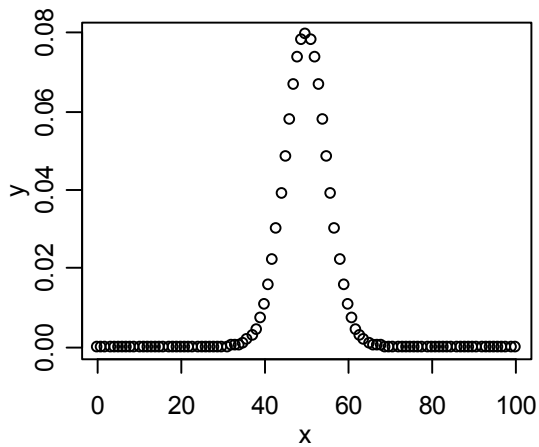
The Normal distribution is the famous "Bell curve," also known as the Gaussian distribution, after Carl Friedrich Gauss, possibly the greatest mathematician of all time.

Gauss, Schmauss. Could we have discovered this distribution ourselves?

The Binomial is one of our favorite distributions, the first one we met. Can we find a simple approximation to the Binomial? We'll use R to calculate, visualize, and compensate for our mathematical abilities that may be meager compared with those of Gauss. Take, for example, the number of Heads in 100 tosses of a coin – the $\text{Bin}(100, \frac{1}{2})$ distribution.

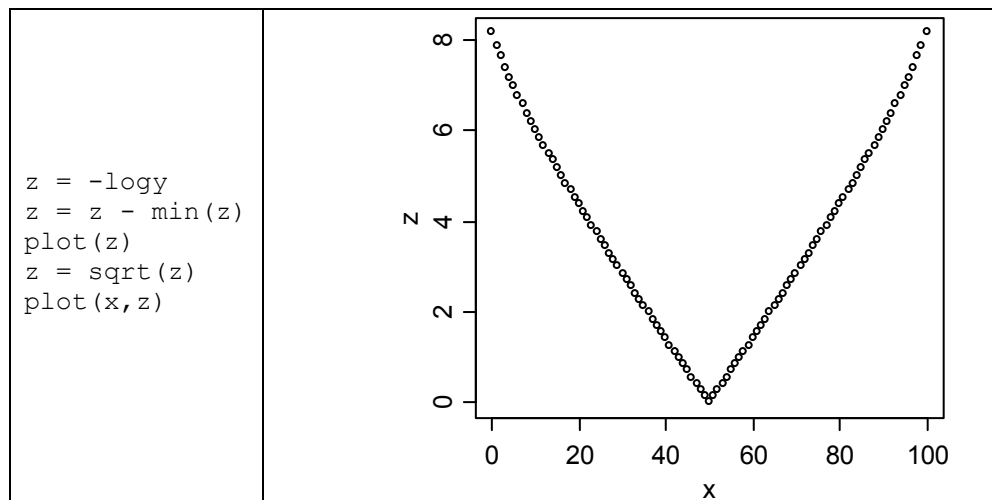
```
x = 0:100
p = 0.5 # can also use other p's, e.g., 0.6 or 0.7
y = dbinom(x,100,p)
plot(x,y)

logy = log(y)
plot(x,logy)
```



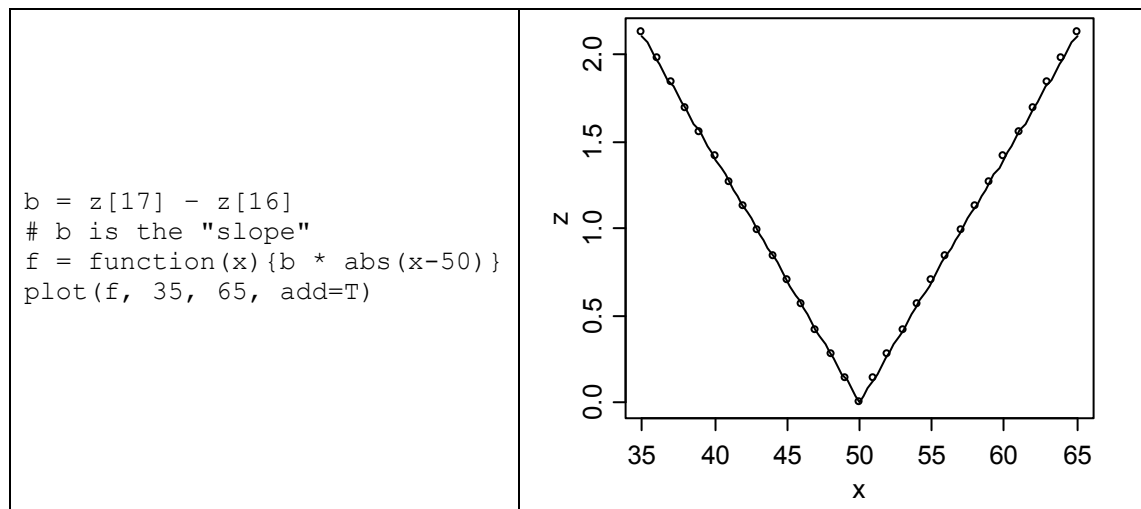
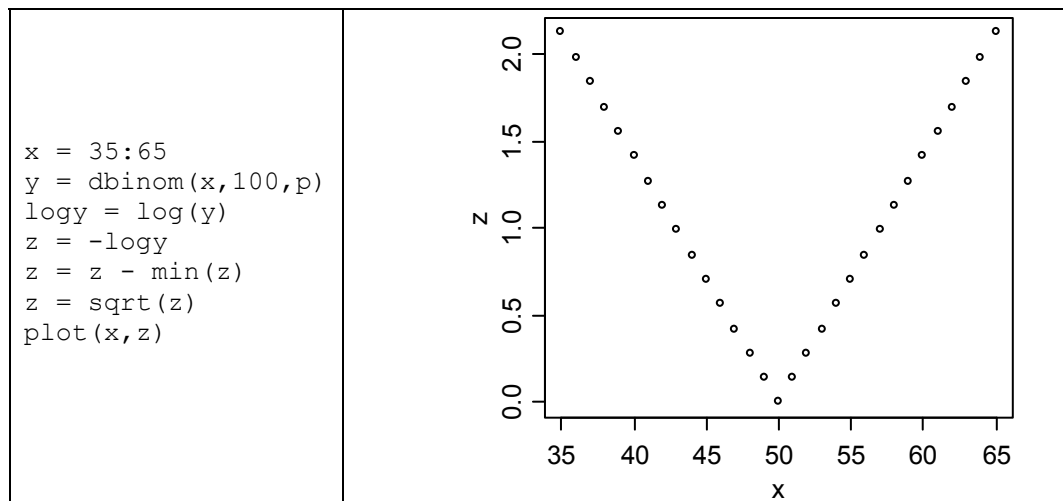
Hmm, the log probabilities look kind of parabolic...

How close is this to being a parabola [something of the form $c_1 - c_2(x - 50)^2$]?



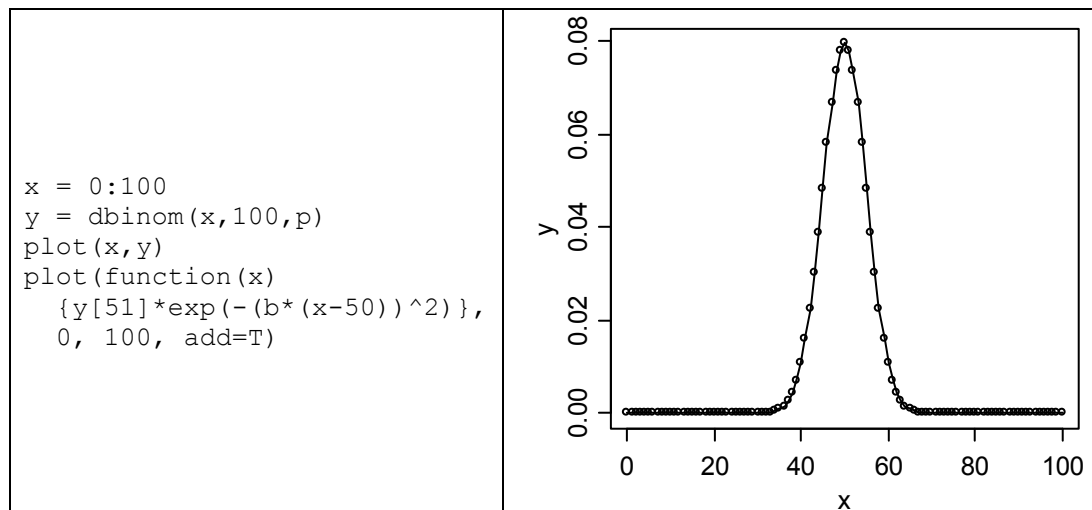
```
# > sum(dbinom(35:65,100,p))
# [1] 0.99821
```

Nearly all the probability lies between 35 and 65 [3 SD's of the binomial!]. So let's repeat the picture just for those x's:



The fit deteriorates near the ends of the distribution (which have negligible probability), but in the **center** (like from 35 to 65 or so) the fit is great. Later we will study (not prove) the “**Central** Limit Theorem,” which says that this sort of phenomenon occurs more generally.

Finally, we can go back and check the fit of the exponential of the parabola to the probabilities in the original picture.



You can try this same sort of thing, with the appropriate modifications, for other choices of n and p . From this we can see that when n becomes at all large (like 100 is plenty) and p is not too close to 0 or 1, binomial density functions (i.e. probability mass functions) are very close to having the form

$$p(x) = \exp[-\text{quadratic function of } x].$$

This is the general form of the pdf for Normal distributions.

Definition: A Normal distribution has a probability density function of the form $f(x) = \exp[-\text{quadratic function of } x]$ for all $-\infty < x < \infty$.

1.21.2 Standard Normal distribution

Definition: The *standard Normal distribution* has pdf $f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right)$.

What's so "standard" about that?

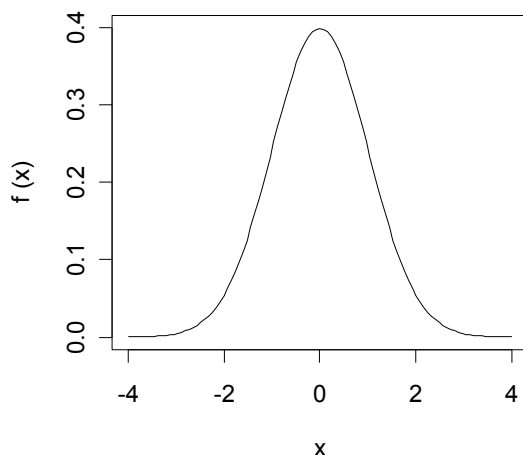
- Why the funny $\sqrt{2\pi}$?
- Why divide the x^2 by 2? Wouldn't it look nicer and more standard with just plain x^2 ?

The $\sqrt{2\pi}$ is a normalizing constant – it is the right number to divide by to make the integral of f equal to

1. So if we have our hearts set on $f(x) \propto \exp\left(\frac{-x^2}{2}\right)$, the $\frac{1}{\sqrt{2\pi}}$ is forced on us and can be derived by calculus.

2. If we had $f(x) \propto \exp(-x^2)$, the resulting distrib would have variance $\frac{1}{2}$. *The standard Normal distribution has mean 0 and variance 1.*

```
> f = function(x) {1/sqrt(2*pi)*exp(-x^2/2)}
> plot(f,-4,4)
```



We could have done this with `plot(dnorm, -4, 4)` but I didn't want you to suspect cheating.

Shall we check some of this calculus?

The normalizing constant $\sqrt{2\pi}$. [Note: This is more for curiosity, amusement, and the feeling of comfort that comes from having seen it once. It's *not* important to be able to reproduce these calculations.]

We want to evaluate $J = \int_{-\infty}^{\infty} e^{-x^2/2} dx$. Here is a trick:

$$\begin{aligned} J^2 &= \left(\int_{-\infty}^{\infty} e^{-x^2/2} dx \right) \left(\int_{-\infty}^{\infty} e^{-y^2/2} dy \right) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (e^{-x^2/2})(e^{-y^2/2}) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)/2} dx dy \end{aligned}$$

As my high school teacher would say, "This is crying out for polar coordinates." You remember...

$$x^2 + y^2 = r^2, \quad dx dy = r dr d\theta \dots$$

$$J^2 = \int_0^{2\pi} \int_0^{\infty} e^{-r^2/2} r dr d\theta$$

But $\int_0^{\infty} e^{-r^2/2} r dr = \int_0^{\infty} e^{-u} du = 1$, so $J^2 = \int_0^{2\pi} 1 d\theta = 2\pi$, and so $J = \sqrt{2\pi}$.

Lord Kelvin said, "A mathematician is one to whom *that* is as obvious as twice two makes four is to you. Liouville was a mathematician." [Quoted from Spivak's *Calculus on manifolds*.]

Mean: $\int_{-\infty}^{\infty} x f(x) dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-x^2/2} dx = 0$. [The integrand is an odd function and it decays to zero fast enough so the integral is not of the undefined form $\infty - \infty$.]

Variance: Integrate by parts to get

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} dx &= \frac{-1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x \left(-x e^{-x^2/2} \right) dx = \frac{-1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x d \left(e^{-x^2/2} \right) = \\ &= \frac{-1}{\sqrt{2\pi}} x e^{-x^2/2} \Big|_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(e^{-x^2/2} \right) dx = 0 + 1 = 1 \end{aligned}$$

Indeed: The standard Normal distribution has mean 0 and variance 1.

Inflection points: [Nice to know for drawing pictures]

If $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, then $f'(x) = -xf(x)$, and

$f''(x) = -f(x) - xf'(x) = -f(x) - x(-xf(x)) = f(x)(x^2 - 1)$, so $f''(x) = 0$ at $x = \pm 1$.

1.21.3 General Normal distributions

Having gained some intimacy with the standard Normal density, we move on to general Normal distributions.

Notation: We write $N(\mu, \sigma^2)$ for the Normal distribution with mean μ and variance σ^2 .

So the standard Normal distrib is written $N(0,1)$.

We typically use Z to denote a r.v. with a $N(0,1)$ distrib.

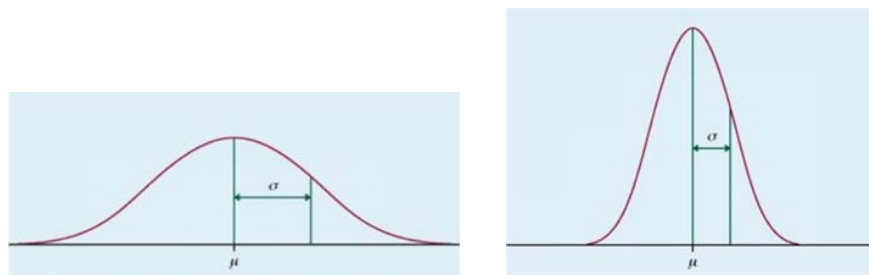
Simple linear relationship: If $Z \sim N(0,1)$, then $\mu + \sigma Z \sim N(\mu, \sigma^2)$. That is, we say that

$X \sim N(\mu, \sigma^2)$ if $\frac{X - \mu}{\sigma} \sim N(0,1)$.

The $N(\mu, \sigma^2)$ density has the same bell shape as $N(0,1)$, except it is stretched to be fatter by a factor of σ , and shifted over by μ units to be centered at μ .

[And since it's stretched to be fatter by σ , its height must be compressed by a factor of σ to keep the integral under the density equal to 1.]

$N(\mu, \sigma^2)$ density: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right]$



Note the inflection points of the $N(\mu, \sigma^2)$ density are at $\mu \pm \sigma$. This gives a way of visualizing σ graphically.

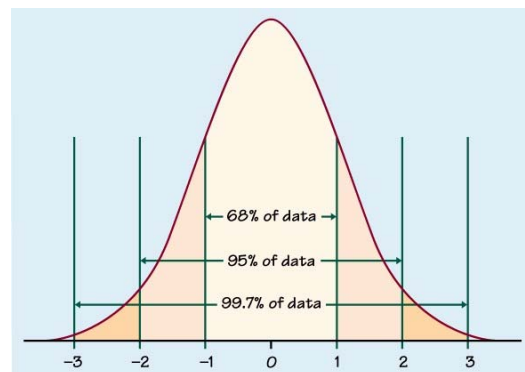
1.21.4 The "68, 95, 99.7 rule"

This "rule" is actually just 3 handy numbers to memorize.

In any Normal distribution:

- 68% of the distribution is within 1 SD of the mean (i.e. between $\mu - \sigma$ and $\mu + \sigma$)
- 95% of the distribution is within 2 SD's of the mean (i.e.

Stat 238 notes, 9/2/09



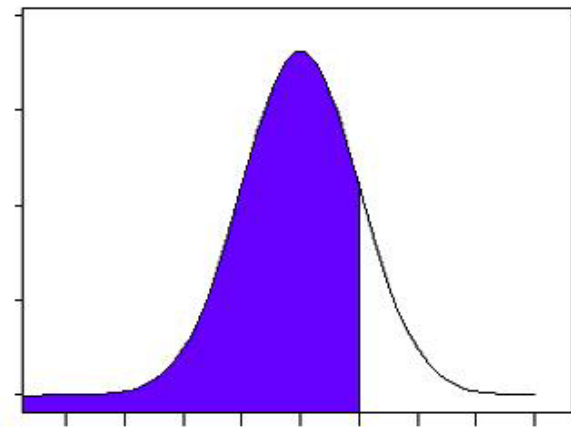
- between $\mu - 2\sigma$ and $\mu + 2\sigma$)
- 99.7% of the distribution is within 3 SD's of the mean (i.e. between $\mu - 3\sigma$ and $\mu + 3\sigma$)

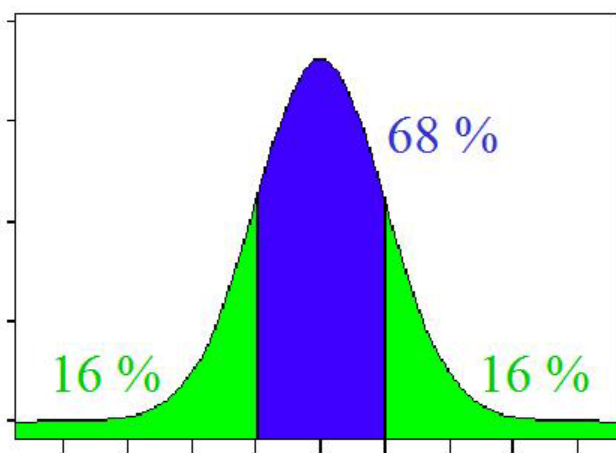
The rule is very handy, and can sometimes help you impress your friends.

Example: Suppose verbal SAT scores have a distribution that is approximately $N(\mu, \sigma^2)$ with $\mu = 505$ and $\sigma = 110$. What is the percentile of the score 615?

You can start by drawing a rough picture of the density, with the horizontal scale marked with μ , $\mu \pm \sigma$, $\mu \pm 2\sigma$.

The area we want is this:





Answer: the middle blue region, plus the left green region, that is, $68\% + 16\% = 84\%$. The score 615 is approximately the 84th percentile of the distribution. ►

1.21.5 Using R and Normal tables

Using R:

```
> pnorm(615, 505, 110)
[1] 0.8413447
```

As usual, for a quick check on the syntax of how to use the pnorm function, in R you can type "?pnorm". This is useful when you remember the name of the function but not exactly how to use it.

Normal tables: these are useful for exams, and for desert islands.

On the web, you will find a table that gives lots of values of the cumulative distribution function of the standard Normal distribution.

Notation: This cdf (of the standard Normal distrib) is denoted by Φ .

The table gives the values of $\Phi(x)$ for $x = 0.00, 0.01, 0.02, \dots, 3.09$. It is arranged in the way that has become typical for these tables, which might take a moment's explanation in class...

Simple sample questions:

- What is $\Phi(0)$?
- You want to know $\Phi(-1.2)$. How would you use the table? [Ans: 0.1151]

Standardization

You are on a desert island with your Normal distribution table, but no computer and no R.

Example: Continue to assume verbal SAT scores have a distribution that is approximately $N(\mu, \sigma^2)$ with $\mu = 505$ and $\sigma = 110$. What is the percentile of the score 700?

Ask: how many standard deviations is the score 700 above the mean score?

Answer: $\frac{700 - 505}{110} = 1.77$.

[You remember your long division]

So we look in our table to find the answer $\Phi(1.77) = 0.9616$.

The score 700 is about the 96th percentile.



In general, if we have $X \sim N(\mu, \sigma^2)$, we can *standardize* X by doing the linear transformation

$$Z = \frac{X - \mu}{\sigma}.$$

The r.v. Z thus formed will have a standard Normal distribution.

It is interpreted as "the number of SD's X by which X differs from its mean."

1.21.6 Another R Picture

An approach to fitting a quadratic function to the log binomial probs:

```
x = 0:100
y = dbinom(x,100,.5)
plot(x,y)
logy = log(y)
plot(x,logy)
fit = lm(logy ~ x + I(x^2))
summary(fit)
predicted = model.matrix(fit) %*% coef(fit)
lines(x, predicted)
```

Let's see how a quadratic fits to just the center part of the distrib having nearly all of the mass i.e. don't worry about fitting the tails of the distrib having extremely tiny mass.

So we try the same thing with $x = 35:65$ replacing $x = 0:100$.

```
x = 35:65
y = dbinom(x,100,.5)
logy = log(y)
plot(x,logy)
fit = lm(logy ~ x + I(x^2))
summary(fit)
predicted = model.matrix(fit) %*% coef(fit)
lines(x, predicted)
```

1.22 A note on statistics and Statistics

A definition of Statistics [quoted from *Introduction to the Practice of Statistics* by Moore and McCabe]:

Statistics is the science of collecting, organizing, and interpreting numerical facts, which we call *data*.

We usually regard data as the realized values of some random variables.

Typical example: random variables X_1, X_2, \dots, X_n are a *sample* from a probability distrib we will call the *population distribution*.

A *statistic* is a computable function of the data.

By "computable" here we mean the computation involves no unknown quantities, such as unknown parameters.

Examples:

- Sample mean: $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$,
- Sample variance: $s^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$.
- *Not* a statistic: $\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$ (unless we're assuming μ is known).

Since data are random variables and statistics are functions of the data, statistics are random variables too, and they have probability distributions. Such a distribution is called a *sampling distribution*. We imagine repeating the following:

- Take a sample of size n from the population
- Calculate our statistic on the sample

The probability distribution over such repeated samples is the sampling distribution.

E.g. we might hear, or even find ourselves saying, utterances such as

- "The sample mean is approximately Normally distributed."
- "The sampling distribution of the sample mean is approximately Normal."
- "The mean of the sampling distribution of the sample mean is the population mean."

These are all statements about the sampling distribution of a statistic – here, the statistic is the sample mean.

1.23 Normal probability plots

OK, we know about these wonderful Normal distributions. Many statistical procedures are based on assuming we are looking at samples from a Normal population distribution. How do we check this assumption?

A nice visual tool is the Normal probability plot. It relies on the fact that, although it is hard for us to judge whether a histogram is shaped like the Normal density rather than some other roughly bell shaped curve or whatever, it is relatively easy for us to judge whether or not a line is straight.

☺ A nice way to get data into R:

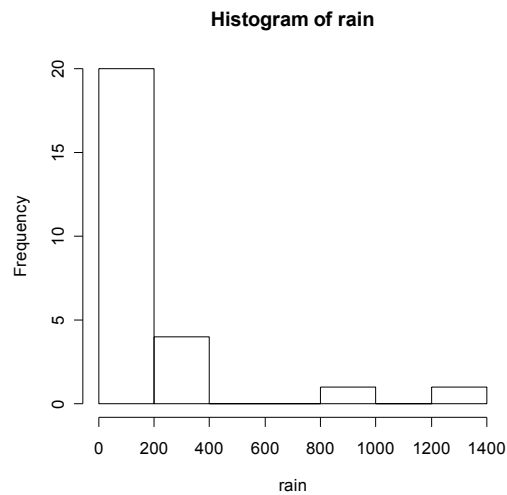
For an example let's get some data: Here is how to get data from, e.g., an Excel "comma separated values" file. [If you are given an ordinary Excel file, say, blah.xls, then open blah.xls in Excel, and click File → Save As... and in the little "Save as type..." drop-down list choose CSV.]

```
dat = read.csv("c:/_jtc/238/data/RainAndSeedingClouds.csv")
```

If you want alternative methods, do `?read.table` for more on related functions for reading data.

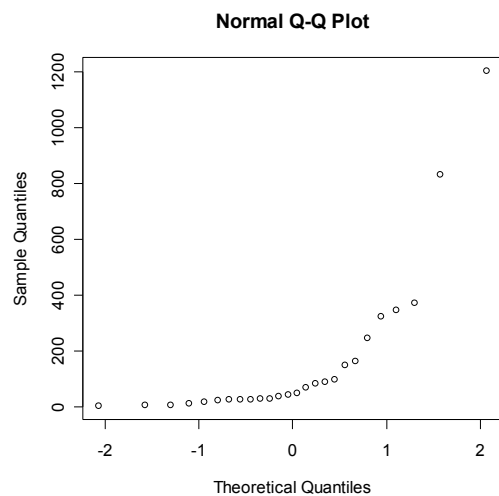
Suppose we want to do a Normal probability plot of rainfall from the unseeded clouds. But first a histogram

```
rain = dat[1:26,1]  
hist(rain)
```



Doesn't look very Normal at all. Here is a Normal probability plot: use the function `qqnorm`

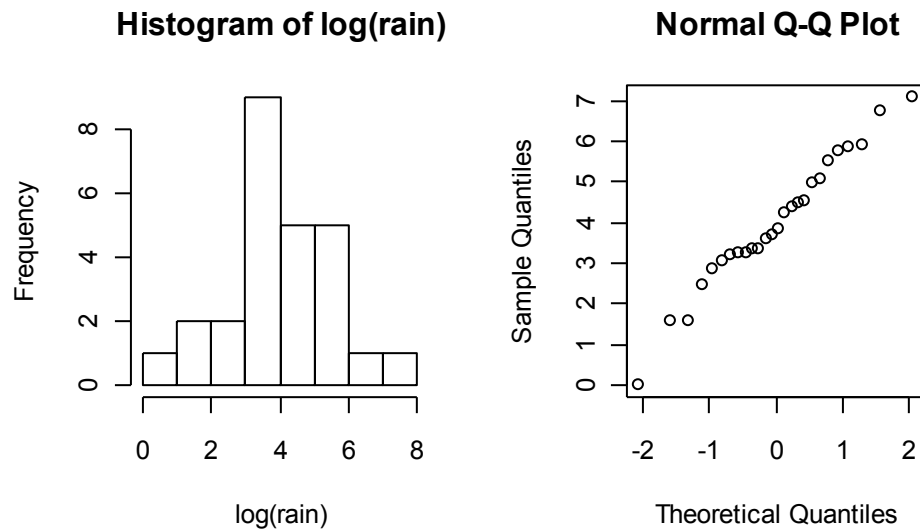
```
qqnorm(rain)
```



The non-Normality is reflected in the curvature of the plot.

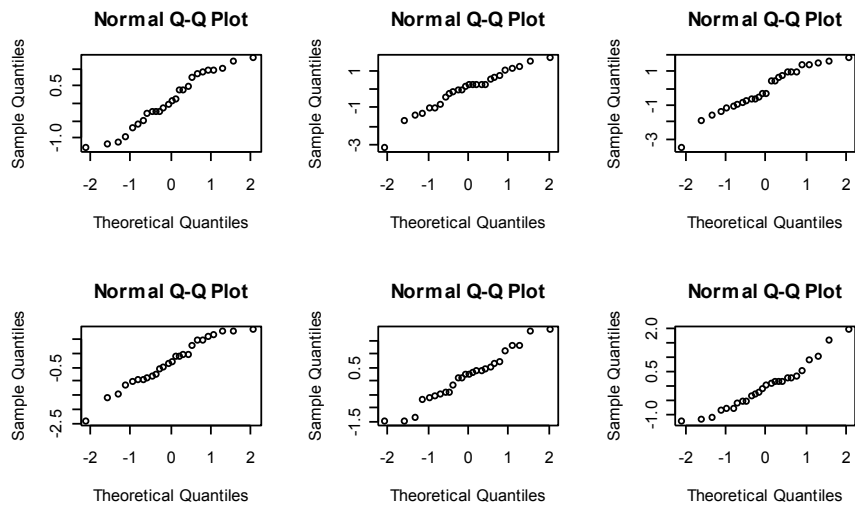
Sometimes data that contains different orders of magnitude can be transformed to have much more nearly a Normal distribution by taking logs.

```
lograin = log(rain)
hist(lograin)
qqnorm(lograin)
```



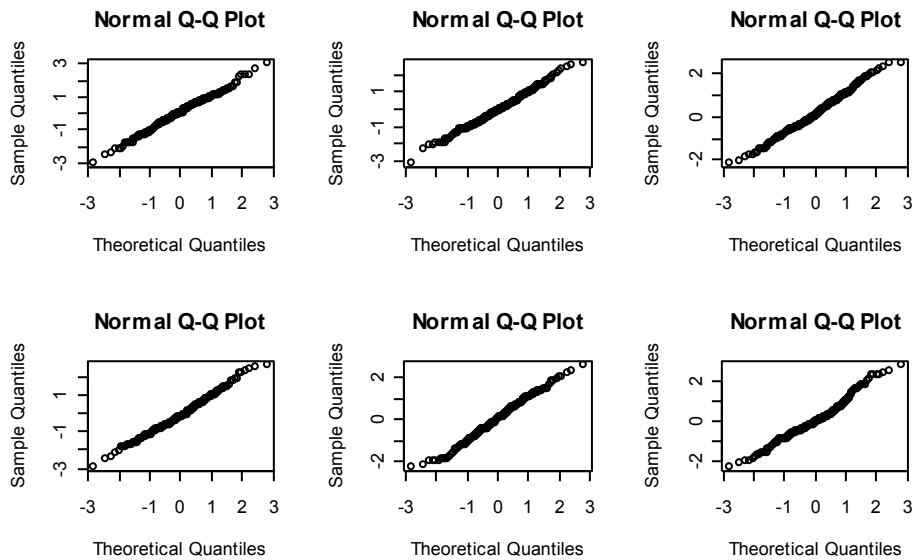
For comparison, let's do Normal probability plots for 6 artificially generated data sets that are really supposed to be Normally distributed:

```
par(mfrow=c(2,3))
for(i in 1:6) qqnorm(rnorm(26))
```



These suggest that our rainfall data set is not distinguishable from an actual sample of size 26 from a Normal distribution.

Here are some plots for Normally distributed samples of size 200.



They look quite straight though the middle, but we shouldn't get too excited if we see a bit of bending or wavering near the ends.

How are Normal probability plots constructed?

[[Note details will vary with different pieces of software; e.g. some may flip the horizontal and vertical axes from the description I'm about to give.]]

Vertical axis = "sample quantiles" = sorted *data* ← easy

Horizontal axis = "theoretical quantiles" ← what are these?

The idea of "theoretical quantiles" is values we would "expect" from a sample of the same size from a $N(0,1)$ distribution. E.g. for our rain data, with sample size $n = 26$, these are 26 quantiles of the standard Normal distribution.

Which quantiles?

R uses (I think) $\Phi^{-1}\left(\frac{k-0.5}{n}\right)$.

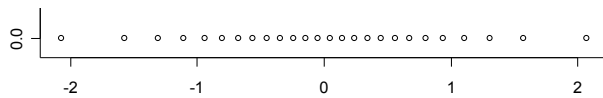
E.g., if n were 5, R would plot the values

```
> qnorm(c(0.1, 0.3, 0.5, 0.7, 0.9))
[1] -1.282 -0.524  0.000  0.524  1.282
```

on the horizontal axis, versus the sorted data on the vertical axis.

For $n = 26$, the quantiles look like this:

```
x = ((1:26)-0.5)/26
plot(qnorm(x), rep(0,26), ylim=c(0,2))
```

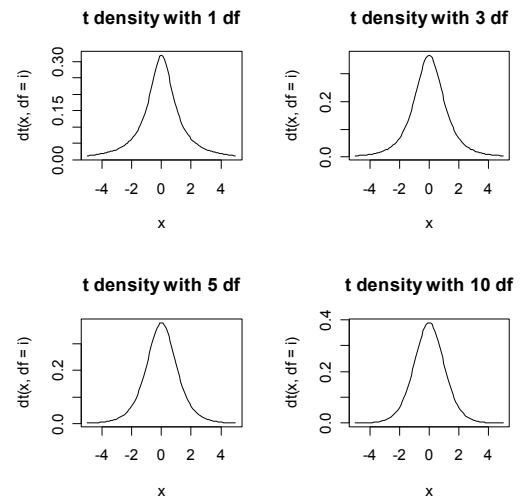


These are the "theoretical quantiles" plotted by `qqnorm` for the rain example – in both the plots of the original data and the logs. Transformations change the data, but not the "theoretical quantiles"!

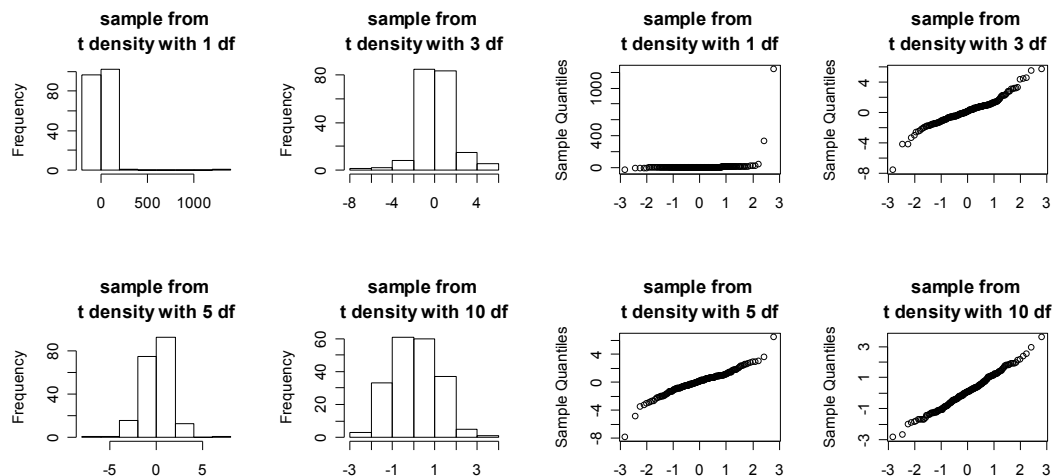
Another Example:

The "*t* distributions" also have bell-shaped densities, but fatter tails than a Normal distribution. They are characterized by a parameter called the "degrees of freedom". In fact, the *t* distribution with 1 degree of freedom is the same as the Cauchy (laser pole) density we have studied. As the degrees of freedom parameter increases, the densities get closer and closer to the standard Normal distribution.

This is what some *t* densities look like:



And here are some histograms and Normal probability plots from simulated samples from these distributions.



1.24 Distributions of sums and the Central limit theorem

1.24.1 Distributions of sums of independent r.v.'s

Suppose X and Y are independent with probability mass functions (or densities, if in the continuous case) f_X and f_Y . Let $Z = X + Y$. What is the pmf (or pdf) f_Z ?

We have to add up (or integrate) all of the possible ways of getting a particular value for Z .

$$\text{Discrete case: } f_Z(z) = P\{X + Y = z\} = \sum_x P\{X = x, Y = z - x\} = \sum_x f_X(x) f_Y(z - x).$$

[Note we've used the assumed independence in the last step.]

$$\text{That is, } f_Z(z) = \sum_x f_X(x) f_Y(z - x).$$

Continuous case:

$$f_Z(z) = \int f_X(x) f_Y(z - x) dx.$$

This operation is called *convolution* – in both cases we say that f_Z is the convolution of f_X and f_Y .

Example: Our old spinner game.

$$\text{Winnings from one play of game: } X = \begin{cases} 0 & \text{with prob } 0.5 \\ 2 & \text{with prob } 0.3 \\ 9 & \text{with prob } 0.2 \end{cases}$$

If we do it twice independently, getting X_1 and X_2 and sum $S_2 = X_1 + X_2$, our possible values for S_2 are 0, 2, 4, 9, 11, and 18. These values have probabilities

$$P\{S_2 = 0\} = f_{X_1}(0)f_{X_2}(0) = (0.5)(0.5) = 0.25$$

$$P\{S_2 = 2\} = f_{X_1}(0)f_{X_2}(2) + f_{X_1}(2)f_{X_2}(0) = (0.5)(0.3) + (0.3)(0.5) = 0.3$$

$$P\{S_2 = 4\} = f_{X_1}(2)f_{X_2}(2) = (0.3)(0.3) = 0.09$$

...

$$P\{S_2 = 18\} = f_{X_1}(9)f_{X_2}(9) = (0.2)(0.2) = 0.04.$$

And if we played the game 4 times, it would be like adding two independent r.v.'s from the previously calculated pmf of S_2 . S_4 would take values 0, 2, 4, 6, 8, 9, 11, ..., 36. And, for example,

$$\begin{aligned} f_{S_4}(4) &= f_{S_2}(0)f_{S_2}(4) + f_{S_2}(2)f_{S_2}(2) + f_{S_2}(4)f_{S_2}(0) \\ &= (.25)(.09) + (.3)(.3) + (.09)(.25) = .135 \end{aligned}$$



Example: Sum of two independent Normally distributed r.v.'s.

[This is an example of the convolution technique, but also, more importantly, an important result to know for its own sake.]

We might as well assume our r.v.'s have mean 0, but let's allow them to have arbitrary variances.

$X \sim N(0, \sigma^2)$, $Y \sim N(0, \tau^2)$, X and Y assumed independent.

$Z = X + Y$. What is the distrib of Z ?

Answer: $Z \sim N(0, \sigma^2 + \tau^2)$.

☺ **The sum of independent Normally distributed r.v.'s is another Normally distributed r.v.**

Note the content of this statement: It is obvious that Z has mean 0 and variance $\sigma^2 + \tau^2$. The part that requires proof is that the density of Z has the Normal shape, that is, $f(z)$ is of the form $f(z) \propto \exp[\text{quadratic function of } z]$. Here goes:

$$\begin{aligned} f_Z(z) &= \int f_X(x) f_Y(z-x) dx \\ &\propto \int \exp\left(\frac{-x^2}{2\sigma^2}\right) \exp\left(\frac{-(z-x)^2}{2\tau^2}\right) dx \\ &= \exp\left(\frac{-z^2}{2\tau^2}\right) \int \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)x^2 + \frac{zx}{\tau^2}\right\} dx \end{aligned}$$

Complete the square in the exponent:

$$\begin{aligned} -\frac{1}{2}\left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)x^2 + \frac{zx}{\tau^2} &= -\frac{1}{2}\left(\frac{\sigma^2 + \tau^2}{\sigma^2\tau^2}\right)\left[x^2 - 2\left(\frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right)\frac{zx}{\tau^2}\right] \\ &= -\frac{1}{2}\left(\frac{\sigma^2 + \tau^2}{\sigma^2\tau^2}\right)\left[\left(x - \left(\frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right)\frac{z}{\tau^2}\right)^2 - \left(\frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right)^2\frac{z^2}{\tau^4}\right] \\ &= -\frac{1}{2}\left(\frac{\sigma^2 + \tau^2}{\sigma^2\tau^2}\right)\left[\left(x - \left(\frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right)\frac{z}{\tau^2}\right)^2\right] + \frac{1}{2}\left(\frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right)\frac{z^2}{\tau^4} \\ f_Z(z) &\propto \exp\left(\frac{-z^2}{2\tau^2} + \frac{1}{2}\left(\frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right)\frac{z^2}{\tau^4}\right) \underbrace{\int \exp\left\{-\frac{1}{2}\left(\frac{\sigma^2 + \tau^2}{\sigma^2\tau^2}\right)\left[x - \left(\frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right)\frac{z}{\tau^2}\right]^2\right\} dx}_{\text{Does not depend on } z!} \\ &\propto \exp\left(\frac{-z^2}{2\tau^2} + \frac{1}{2}\left(\frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right)\frac{z^2}{\tau^4}\right) \leftarrow \begin{pmatrix} \text{Can already see } Z \text{ is Normally distributed} \\ \text{We could really stop here} \end{pmatrix} \\ &= \exp\left(\frac{-z^2}{2\tau^2}\left[1 - \frac{\sigma^2}{\sigma^2 + \tau^2}\right]\right) \\ &= \exp\left(\frac{-z^2}{2\tau^2}\left[\frac{\tau^2}{\sigma^2 + \tau^2}\right]\right) = \exp\left(\frac{-z^2}{2(\sigma^2 + \tau^2)}\right) \end{aligned}$$

This shows f_Z must be the $N(0, \sigma^2 + \tau^2)$ density! ►

1.24.2 Central limit theorem

The name is kind of punny. The theorem is of central importance in Probability and Statistics, and it also is most accurate in describing the “central” region of certain distributions (as opposed to the extreme “tails” of the distribution).

The theorem says that the probability distribution of the sum of many *iid* (independent and identically distributed) r.v.'s is nearly Normally distributed. It's a type of limit theorem; note the *many* qualifier – *many iid r.v.'s*.

So, suppose X_1, X_2, \dots are *iid* from a distribution having mean μ and variance σ^2 . Define $S_n = X_1 + \dots + X_n$.

We know the mean and variance of S_n : $E(S_n) = n\mu$, $\text{var}(S_n) = n\sigma^2$

The CLT says that S_n approximately has the Normal distribution $N(n\mu, n\sigma^2)$.

How to formulate this precisely as a limit statement?

E.g. “The distribution of S_n converges to $N(n\mu, n\sigma^2)$ ” is not a precise way to say it, particularly since the purported limit, $N(n\mu, n\sigma^2)$, is a moving target.

[Analogy: for many purposes, it's fine to approximate $n(n-1)/2$ by $n^2/2$ for large n , but we wouldn't say that “ $n(n-1)/2 \rightarrow n^2/2$ as $n \rightarrow \infty$ ” – in fact the difference between the left side and the right side is $n/2$, which goes to infinity.]

To have a limit that is not a moving target we can *standardize* the r.v.'s S_n .

The standardized version of S_n is $\frac{S_n - E(S_n)}{\text{SD}(S_n)} = \frac{S_n - n\mu}{\sqrt{n\sigma^2}}$, which has mean 0 and variance 1 for each n .

The CLT says that $\frac{S_n - n\mu}{\sqrt{n\sigma^2}}$ converges in distribution to $N(0,1)$ as $n \rightarrow \infty$.

This means that for each z , $\lim_{n \rightarrow \infty} P\left\{\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \leq z\right\} = P\{N(0,1) \leq z\}$.

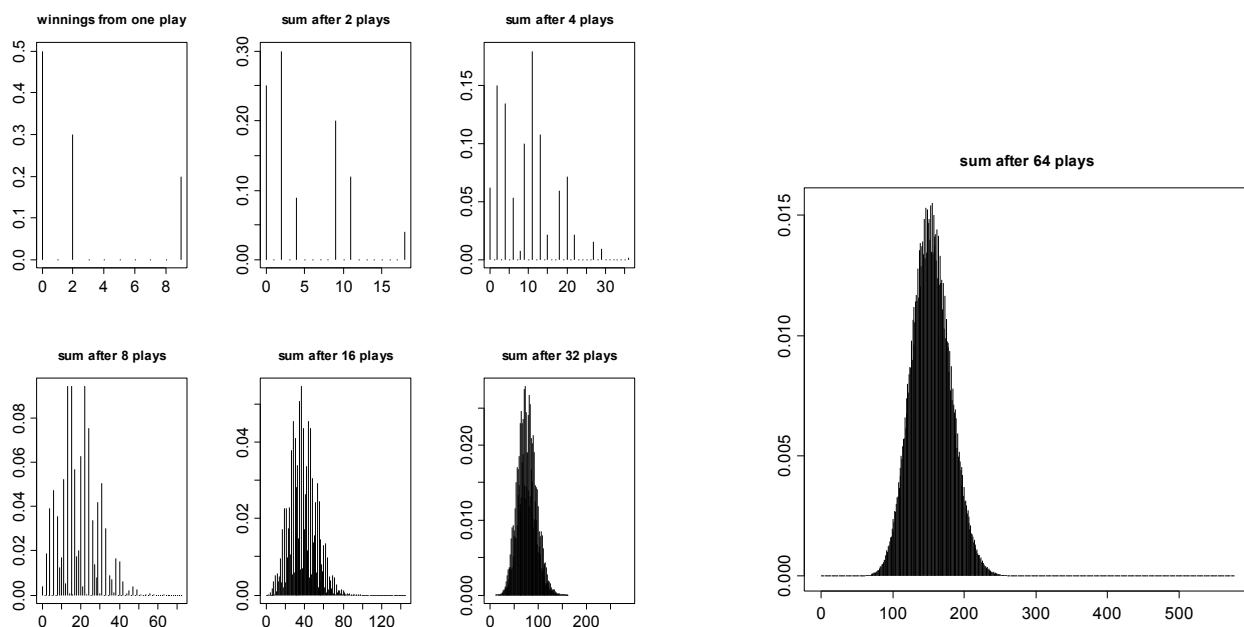
We write $\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{d} N(0,1)$.

Having learned to write the result precisely, it's now probably safe to allow ourselves to write Normal approximations in the less formal way $S_n \dot{\sim} N(n\mu, n\sigma^2)$, where “ $\dot{\sim}$ ” means “is approximately distributed as.”

Example: The old spinner.

I did the convolutions in R – see the file of commands for today on the web.

The pmf's of S_n for $n = 1, 2, 4, 8, 16, 32$, and 64 look like this:



Hmm... looks pretty normal. We could have predicted this with our Central Limit Theorem.

We know for large n , S_n will be nearly Normally distributed, with mean and variance $n\mu$ and $n\sigma^2$.

$$\text{Here } X = \begin{cases} 0 & \text{with prob } 0.5 \\ 2 & \text{with prob } 0.3 \\ 9 & \text{with prob } 0.2 \end{cases} \quad \text{so } \mu = (0)(.5) + (2)(.3) + (9)(.2) = 2.4, \text{ and}$$

$$\sigma^2 = E(X^2) - (EX)^2 = (0)(.5) + (4)(.3) + (81)(.2) - (2.4)^2 = 11.64 \quad [\text{and so } \sigma = \sqrt{11.64} = 3.412].$$

So, e.g., S_{64} is nearly Normal with mean $64(2.4) = 153.6$, and variance $64 \cdot 11.64 = 744.96$.

[Standard deviation $\sqrt{744.96} = 27.29$.]

E.g., this would lead to approximate statements like

$$P\{S_{64} \geq 185\} \approx P\{N(153.6, 744.96) \geq 185\} = P\left\{N(0,1) \geq \underbrace{\frac{185 - 153.6}{\sqrt{744.96}}}_{1.15}\right\} = 1 - \Phi(1.15) = .125.$$

Using R, I calculate

```
> sum(dist64[186:577])
[1] 0.1305863
```

Our Normal approximation of 0.125 was in the right ballpark.

[[These Normal approximations can be made noticeably better (although not exact of course; it's just a simple approximation) by a minor modification called the "continuity correction"... more later]]



1.24.3 Normal approximation to Binomial distributions

If $X \sim B(n, p)$, we already know the mean and variance of X :

$$E(X) = np, \text{ var}(X) = np(1 - p).$$

If $X \sim B(n, p)$ and n is large, can think of X as the sum of many *iid* r.v.'s.

[Remember X is a sum of indicators $X = I_1 + \dots + I_n$]

So the Central Limit Theorem says the distrib of X is approximately Normal.

So: for large n , if $X \sim B(n, p)$, then $X \dot{\sim} N(np, np(1 - p))$.

If we're thinking about estimating p by $\hat{p} = \frac{X}{n}$, the form that is useful is $\hat{p} \dot{\sim} N\left(p, \frac{p(1 - p)}{n}\right)$.

[[Note: X is a sum of indicators, and \hat{p} is a mean of indicators! So \hat{p} has the usual $\frac{\text{const}}{\sqrt{n}}$ sort of SD.]]

Example: Margin of error in a poll.

Accuracy of a poll depends on sample size.

E.g. suppose we take a poll of $n = 1600$ people. There is some true p that we don't know. We estimate p by $\hat{p} = \frac{X}{n}$, where $X \sim B(n, p)$.

So $\hat{p} \dot{\sim} N\left(p, \frac{p(1 - p)}{n}\right)$.

That is, our **error** $\hat{p} - p$ has approximately a Normal distrib with mean 0 and SD $\sqrt{p(1 - p)/n}$.

So the SD depends on both n and the unknown p .

E.g. imagine the unknown truth is that $p = 0.5$. Then the error has SD $= \sqrt{\frac{p(1 - p)}{n}} = \frac{1}{2\sqrt{n}}$.

For our example with $n = 1600$, this would be SD $= 1/80 = 1.25\%$.

The CLT Normal approximation (68, 95, 99.7 rule!) then tells us that the probability that our error is less than 2 SD's (or 2.5%) is 95%.

So, in this situation, with "confidence level" 95%, we can say that our estimator is within the range: true $p \pm (2.5\%)$.

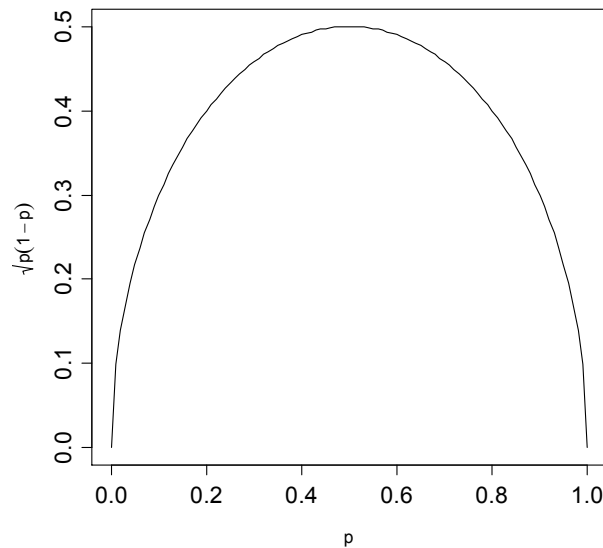
[A "95% confidence interval" would turn this statement around and say we have "95% confidence" that the true p is in the range $\hat{p} \pm (2.5\%)$.]

Note that the margin of error does not depend so severely on the true p , especially when p is somewhere near 0.5 (i.e. not too close to 0 or 1).

E.g.: If the true p were 0.6 (or 0.4), the margin of error [again corresponding to 95% confidence] would change from $\pm 2.5\%$ to $\pm 2.45\%$.

If the true p were 0.7 (or 0.3), the margin of error would be $\pm 2.3\%$.

This is because the function $\sqrt{p(1-p)}$ is quite flat near $p = 1/2$.



Example: (Continuation of previous)

If we want our 95%-confidence margin of error to be $\pm 1\%$ rather than $\pm 2.5\%$, how large a sample size would we need?

Solution: again we'll imagine that $p = 0.5$, noting that our answer does not depend so strongly on p near $p = 0.5$, and in fact the case $p = 0.5$ is the most “conservative,” in the sense that the margin of error is the largest there.

So we want $2SD(\hat{p}) = .01$, that is, $2\left(\frac{1}{2\sqrt{n}}\right) = .01$, that is, $\frac{1}{\sqrt{n}} = .01$, that is, $n = 10,000$. That's a relatively large poll.



1.24.4 The “continuity correction”

Example: Suppose we toss a coin 100 times.
What is the probability that we get 50 or fewer heads?

Let $X \sim B(100, 1/2)$ be the number of heads.

Our Normal approx says $X \sim N(50, 25)$, that is, X is approximately Normal with mean 50 and SD 5.

So we could answer our question this way:

$$P\{X \leq 50\} \approx P\{N(50, 5^2) \leq 50\} = \frac{1}{2}, \text{ which is clearly about right.}$$

Exactly? No; clearly the true answer for $P\{X \leq 50\}$ should be greater than $\frac{1}{2}$.

To get an idea of what is a bit cockeyed here, note that we could have rephrased the desired probability that we get “50 or fewer heads” to be the probability that we get “less than 51 heads.”

Doing a mechanical translation to a Normal approx in the same way would have led to the answer

$$P\{X \leq 50\} = P\{X < 51\} \approx P\{N(50, 5^2) < 51\} = P\left\{N(0, 1) < \frac{51 - 50}{5}\right\} = \Phi(.2) = .579.$$

Idea: Instead of approximating by $P\{N(50, 5^2) \leq 50\}$ or $P\{N(50, 5^2) < 51\}$ [which in fact is the same as $P\{N(50, 5^2) \leq 51\}$ because the Normal distrib is continuous!] the *continuity correction* proposes using the approximation $P\{X \leq 50\} \approx P\{N(50, 5^2) < 50.5\}$.

In our case this gives the approximation

$$P\{X \leq 50\} \approx P\{N(50, 5^2) < 50.5\} = P\left\{N(0, 1) < \frac{50.5 - 50}{5}\right\} = \Phi(.1) = .5389$$

R check:

```
> pbinom(50, 100, .5)
[1] 0.5397946
```

Our approx of .5389 was not bad!

Example: The continuity correction is *really* needed to get a nontrivial approximation for a probability like this. Suppose we toss a coin 100 times. What is the probability of *exactly* 50 heads?

Again, we would not want to do

$$P\{X = 50\} \approx P\{N(50, 5^2) = 50\} = 0 \quad \leftarrow \text{again, continuous distrib!}$$

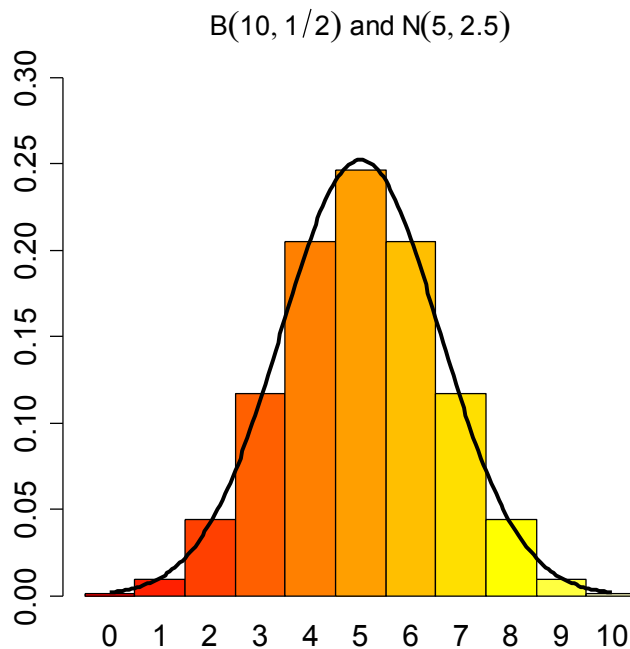
or $P\{49 < X < 51\}$. So we do

$$\begin{aligned} P\{X = 50\} &\approx P\{49.5 < N(50, 5^2) < 50.5\} \\ &= P\left\{\frac{49.5 - 50}{5} < N(0, 1) < \frac{50.5 - 50}{5}\right\} \\ &= P\{-.1 < N(0, 1) < .1\} = .0797 \end{aligned}$$

Again we'll let R be the judge:

```
> dbinom(50, 100, .5)
[1] 0.07958924
```

A picture of what's going on with the continuity correction:



It says, for example, $P\{X \leq 5\} \approx P\{\text{Normal} < 5.5\}$,
 $P\{X = 6\} \approx P\{5.5 < \text{Normal} < 6.5\}$.

Looks sensible in the picture: e.g., the sum of the areas of the bars, up to the bar over “5”, is approximated by the area under the Normal density up to 5.5.

1.25 Conditional expectation

We know, for example, if Y is a discrete r.v. then $E(Y) = \sum y P\{Y = y\}$. Suppose we know some information about another r.v., X . Suppose we know that X took the particular value x . Now our probabilities $P\{Y = y\}$ are no longer appropriate and should be replaced by conditional probabilities $P\{Y = y \mid X = x\}$. The expectation of Y , conditional on knowing the information $X = x$, becomes a **conditional expectation**, denoted $E(Y \mid X = x)$, and defined by

$$E(Y \mid X = x) = \sum_y y P\{Y = y \mid X = x\}.$$

More generally, given any piece of information that can be expressed as saying that an event A has occurred, we define the conditional expectation of Y given the event A by $E(Y \mid A) = \sum_y y P\{Y = y \mid A\}$.

Conditional expectations behave like ordinary expectations; they are just based on conditional probabilities. So, for example, we have the same kind of LOUS as before:

$$E(g(Y) \mid X = x) = \sum_y g(y) P\{Y = y \mid X = x\}.$$

In the continuous case, say we have two continuous r.v.’s X and Y having a joint density $f_{X,Y}(x, y)$.

The ordinary expectation of Y is, of course, $E(Y) = \int y f_Y(y) dy$.

If we are told that $X = x$ occurred and want the conditional expectation of Y , we replace the pdf $f_Y(y)$ by the conditional density of Y given that $X = x$, which is defined to be

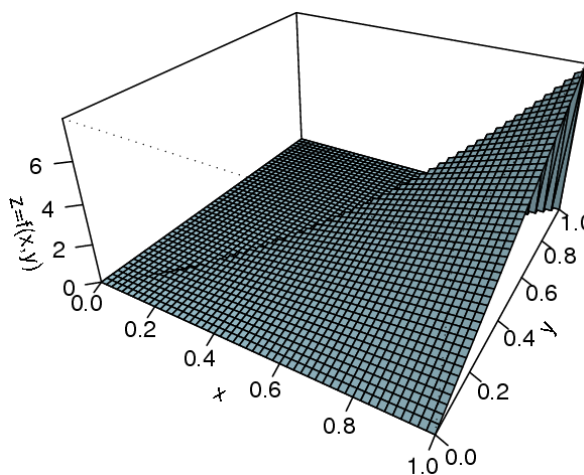
$$f_Y(y | X = x) = \frac{f_{X,Y}(x, y)}{f_X(x)}.$$

That is, $E(Y | X = x) = \int y f_Y(y | X = x) dy$.

Example: Suppose X and Y have joint density $f_{X,Y}(x, y) = \begin{cases} 8xy & \text{if } 0 < y < x < 1 \\ 0 & \text{otherwise} \end{cases}$.

Find $E(Y | X = x)$ for $0 < x < 1$.

Solution: Well, we don't really need a picture but I thought this looks kind of nice...



$$E(Y | X = x) = \int y f_Y(y | X = x) dy$$

We need the conditional density

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

Here we don't want to ignore the denominator or write $f_Y(y | X = x) \propto f_{X,Y}(x, y)$ because we are

trying to determine the answer as a function of x , so we can't drop any functions of x along the way.

So we patiently work out the denominator:

$$f_X(x) = \int f_{X,Y}(x, y) dy = \int_0^x 8xy dy = 8x \int_0^x y dy = 4x^3.$$

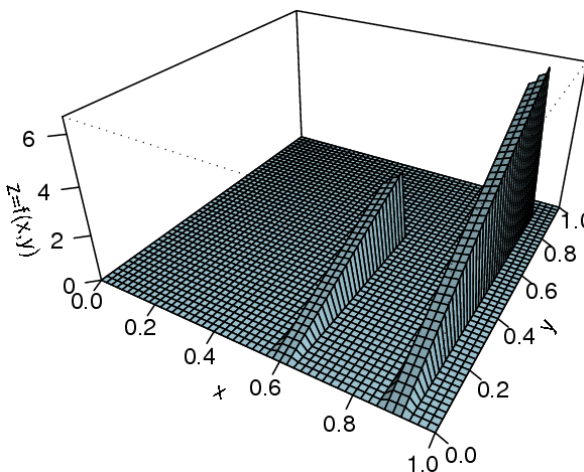
$$f_Y(y | X = x) = \frac{8xy}{4x^3} = \frac{2y}{x^2} \text{ for } 0 < y < x, \text{ and}$$

$$f_Y(y | X = x) = 0 \text{ otherwise.}$$

$$\begin{aligned} E(Y | X = x) &= \int y f_Y(y | X = x) dy \\ &= \int_0^x y \frac{2y}{x^2} dy = \frac{2}{x^2} \int_0^x y^2 dy = \frac{2}{x^2} \frac{x^3}{3} = \frac{2x}{3} \end{aligned}$$

We are simply finding the y coordinate of the center of mass of the thin slice of the density at x . E.g. in the picture we see two slices of the density, at $x = 0.6$ and $x = 0.9$. Our formula,

$E(Y | X = x) = (2/3)x$ tells us that the center of mass of the $x = 0.6$ slice is at $y = 0.4$, and that for $x = 0.9$ is at $y = 0.6$.



Law of Total Expectation

This is an important fact that generalizes the law of total probability:

$$E(Y) = \sum_x E(Y | X = x) P\{X = x\},$$

with the analogous continuous version being

$$E(Y) = \int E(Y | X = x) f_X(x) dx.$$

This is saying something intuitive. For example, in the discrete case, suppose X takes 4 values and Y takes 3 values. Think of Y as your winnings from a spinner game where the spinner has 3 amounts of money, with different probabilities. Except that here, we have 4 different spinners. The spinners have the same 3 amounts of money written on them, but the probabilities of those amounts of money could be different. X represents which of the 4 spinners is chosen. We choose X from its marginal distribution, with spinner x being chosen with probability $P\{X = x\}$. Then given the choice of x , we spin spinner x getting one of the 3 possible amounts of money as our Y . If we were just playing spinner x repeatedly our expected winnings per play of the game would be $E(Y | X = x)$. In this two-stage setting of a mixture of 4 spinners, our expected winnings per play of the game is a weighted average of the values $E(Y | X = x)$ for the individual spinners, with the weights being the probability $P\{X = x\}$ that the corresponding spinner is chosen.

Derivation (discrete case):

$$\begin{aligned} \sum_x E(Y | X = x) P\{X = x\} &= \sum_x \sum_y y P\{Y = y | X = x\} P\{X = x\} \\ &= \sum_x \sum_y y P\{Y = y, X = x\} \\ &= \sum_x y \sum_y P\{Y = y, X = x\} \\ &= \sum_y y P\{Y = y\} = E(Y) \end{aligned}$$

Derivation (continuous case):

$$\begin{aligned} \int E(Y | X = x) f_X(x) dx &= \int \int y f_Y(y | X = x) dy f_X(x) dx \\ &= \int \int y \frac{f_{X,Y}(x, y)}{f_X(x)} dy f_X(x) dx \\ &= \int \int y f_{X,Y}(x, y) dy dx \\ &= \int y \underbrace{\int f_{X,Y}(x, y) dx}_{f_Y(y)} dy \\ &= \int y f_Y(y) dy = E(Y) \end{aligned}$$

Example: On the television show “Fear factor,” the contestant is in a pitch black maze filled with horrible and smelly items.

He starts in a room, “the hub,” with 3 doors. He is whirled around and hit on the head until he is completely disoriented. Then he chooses a door. Behind each door is a tunnel.

Tunnel #1 leads him out of the maze in 3 minutes.

Tunnel #2 leads him back to the hub after 5 minutes.

Tunnel #3 leads him back to the hub after 7 minutes.

The doors are one-way; once a door is chosen, the contestant must follow that tunnel to its end.

At each return to the hub, the contestant is disoriented again and so again chooses each door with probability $1/3$.

What is the expected time the contestant spends in the maze?

Answer: Let Y denote the time spent in the maze.

☺ We'll use a great trick, often used in probability: condition on what happens first. In this case, let X denote the first door chosen. X takes possible values 1, 2, and 3.

The Law of Total Expectation says:

$$\begin{aligned} E(Y) &= E(Y \mid X = 1)P\{X = 1\} + E(Y \mid X = 2)P\{X = 2\} + E(Y \mid X = 3)P\{X = 3\} \\ &= \frac{1}{3}(E(Y \mid X = 1) + E(Y \mid X = 2) + E(Y \mid X = 3)) \end{aligned}$$

Obviously $E(Y \mid X = 1) = 3$.

What if we know that $X = 2$? Conditional on $X = 2$, we can think of $Y = 5 + \tilde{Y}$, where \tilde{Y} has the same distribution as a new, fresh Y .

So $E(Y \mid X = 2) = 5 + E(Y)$. Similarly, $E(Y \mid X = 3) = 7 + E(Y)$

So $E(Y) = \frac{1}{3}[3 + (5 + E(Y)) + (7 + E(Y))]$. We have found a little equation for $E(Y)$ in terms of itself. Solving this linear equation gives $E(Y) = 15$. ►

Example: Mean of $T \sim \text{Geom}(p)$.

We'll use that same trick again: condition on the first trial.

Let $I_1 = 1$ if the first trial is a success, $I_1 = 0$ otherwise.

$$\begin{aligned} E(T) &= E(T \mid I_1 = 1)P\{I_1 = 1\} + E(T \mid I_1 = 0)P\{I_1 = 0\} \\ &= \underbrace{E(T \mid I_1 = 1)}_1 p + \underbrace{E(T \mid I_1 = 0)}_? (1 - p) \end{aligned}$$

The conditional distribution of T , given that $I_1 = 0$, is the same as the unconditional distribution of $1 + T$. [If the first trial is a failure, you've already used one trial (that's the " $1 +$ "), and then the remaining time is as if you were starting over from scratch.]

So $E(T \mid I_1 = 0) = E(1 + T) = 1 + E(T)$.

So $E(T) = p + (1 + E(T))(1 - p)$, or $pE(T) = p + (1 - p) = 1$, or $E(T) = 1/p$. ►

2 Markov chains and Markov chain Monte Carlo

Main Entry: sto·chas·tic Pronunciation: st&-'kas-tik, stO- Function: adjective

Etymology: Greek stochastikos skillful in aiming, from stochazesthai to aim at, guess at, from stochos target, aim, guess

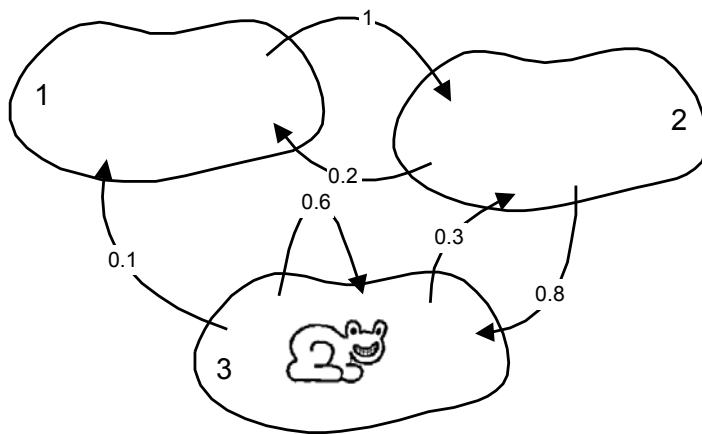
1: involving a random variable <a stochastic process>

2: involving chance or probability <a stochastic model of radiation-induced mutation>

Both to the world and to us in this course, Markov chains are important and useful in two ways: as models of random phenomena (“stochastic processes”) and, more recently, as computational tools. We’ll start by thinking of MC’s as stochastic models.

2.1 Markov chains: introduction and examples

Imagine a frog hopping around among lily pads. He hops according to the probabilities shown on the arrows. These jumping probabilities depend only on where he currently is, and not on the history of how he got there.



A Markov chain is a sequence of random variables X_0, X_1, \dots .

X_t represents the state of the system at time t .

A Markov chain is specified by:

- A *state space* $\mathbb{S} = \{\text{possible states}\}$. For the frog, $\mathbb{S} = \{1, 2, 3\}$.
- An *initial distribution*, π_0 .
 $\pi_0(i) = P\{X_0 = i\}$.
For example if the frog makes a random start in either state 1 or state 3, choosing by tossing a coin, then $\pi_0 = \begin{pmatrix} .5 & 0 & .5 \end{pmatrix}$.
- A *probability transition rule* (or *probability transition matrix*), P .
 $P(i, j) = P\{X_{t+1} = j \mid X_t = i\}$. [“Time homogeneity”: transition probabilities do not depend on t .]
For the frog, we have

$$P = \begin{matrix} & \begin{matrix} 1 & 2 & 3 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} & \begin{pmatrix} 0 & 1 & 0 \\ .2 & 0 & .8 \\ .1 & .3 & .6 \end{pmatrix} \end{matrix}$$

```

P = matrix(c(0,1,0,.2,0,.8,.1,.3,.6),nrow=3,byrow=T)
n = 10000
x = numeric(n)
pi0 = c(.5, 0, .5)
state = sample(1:3, 1, prob=pi0)
x[1] = state
for(i in 2:n){
  state = sample(1:3, 1, prob = P[state,])
  x[i] = state
}

```

After running this, x is a vector of length 10,000 containing a simulated path $X_0, X_1, \dots, X_{9999}$ of the frog.

Example: Random walks: symmetric and otherwise.

For random walks, state space is set of all integers.

Have probability p of going up, $1 - p$ of going down. Symmetric random walk case is $p = 1/2$.

Here the probability transition "matrix" $[[\text{its dimensions are } \infty \text{ by } \infty]]$ has $P(i, i+1) = p$,

$P(i, i-1) = 1 - p$ for all integers i .

Example: "Gambler's ruin" chain.

Same type of transitions as in a random walk, but now the state space is $\{0, 1, \dots, n\}$ and states 0 and n are "absorbing." Once the process hits an absorbing state it stays there forever.

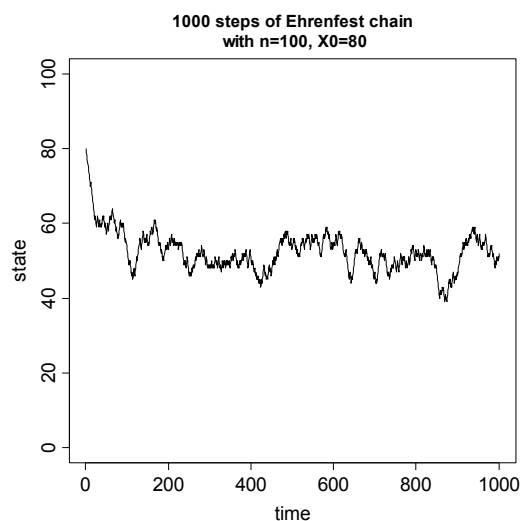
Example: Ehrenfest chain (diffusion, or dogs and fleas)

This is a model that arose in physics as a simple conceptual model of "mixing", for example, two volumes of gas connected by a small hole. One can use it to think about questions like, "Why is it that we don't hear more often about people dying because all the molecules of air in the room happened to go over into one corner, or maybe up near the ceiling?"

n balls in two urns; x balls in urn 1, $n - x$ balls in urn 2.

At each time, choose one of the n balls at random (each ball equally likely), and move that ball to the other urn.

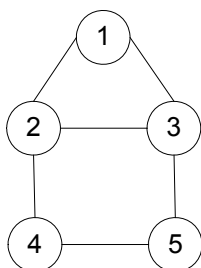
The state X_t is the number of balls in urn 1 at time t .



This is a prototypical model of a probabilistic system with a stable equilibrium; the state hovers around $n/2$, and if taken away from the equilibrium there is a probabilistic “restoring force” back toward equilibrium.

[[A discrete version of “diffusion in a potential well” – this might be discussed further in a course on stochastic processes.]]

Example: Random walk on a graph

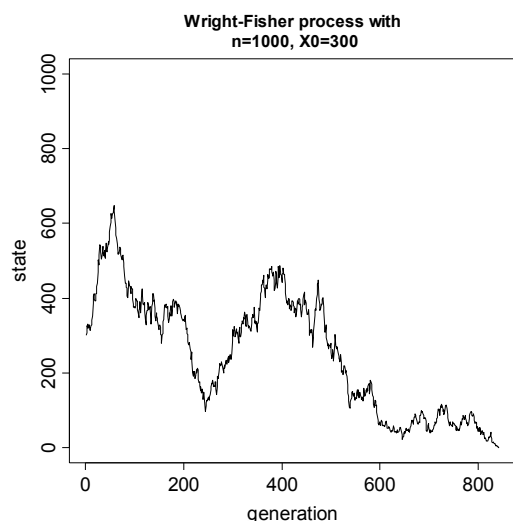


From each node, go to a neighbor chosen at random (with each neighbor being equally likely).

Example: Wright-Fisher chain from genetics.

n balls in an urn. At time t (or “generation t ”) we have X_t white balls, $n - X_t$ black balls. Imagine these as n individuals of two types. None of these individuals will survive to the next generation; that next generation will be filled with offspring from these individuals. An offspring of an individual is simply a copy of that individual. The next generation (generation $t + 1$) will be formed by n offspring, formed by randomly choosing an individual from generation t to reproduce, then choosing an individual again (independently, with replacement, e.g. so that the same individual could reproduce twice), and so on, until we have n offspring. That is, generation $t + 1$ is formed by choosing n times with replacement from generation t .

So here is how the transitions work: Given $X_t = x$, the conditional distribution of X_{t+1} is $\text{Bin}(n, x)$.



The Markov property:

The idea: Suppose I told you that I simulated a realization of the frog chain and came up with $X_0 = 3$, $X_1 = 1$, and $X_2 = 2$. I ask you for the conditional distribution of X_3 .

The answer is $\begin{pmatrix} .2 & 0 & .8 \end{pmatrix}$.

This answer depends only on $X_2 = 2$, that is, the most recent piece of historical information, and not on the earlier values of the process.

$$P\{X_3 = j \mid X_2 = 2, X_1 = 1, X_0 = 3\} = P\{X_3 = j \mid X_2 = 2\} \text{ for all } j.$$

In general, the Markov property says:

$$P\{X_{t+1} = x_{t+1} \mid X_t = x_t, X_{t-1} = x_{t-1}, \dots, X_0 = x_0\} = P\{X_{t+1} = x_{t+1} \mid X_t = x_t\}$$

Example: The probability of a path. For example, we can do calculations like this:

$$\begin{aligned} & P\{X_0 = 2, X_1 = 3, X_2 = 3, X_3 = 1\} \\ & \stackrel{(a)}{=} P\{X_0 = 2\}P\{X_1 = 3 \mid X_0 = 2\}P\{X_2 = 3 \mid X_0 = 2, X_1 = 3\}P\{X_3 = 1 \mid X_0 = 2, X_1 = 3, X_2 = 3\} \\ & \stackrel{(b)}{=} P\{X_0 = 2\}P\{X_1 = 3 \mid X_0 = 2\}P\{X_2 = 3 \mid X_1 = 3\}P\{X_3 = 1 \mid X_2 = 3\} \\ & = \pi_0(2)P(2,3)P(3,3)P(3,1) \end{aligned}$$

Here equality (a) uses the product rule for probabilities (i.e. the definition of conditional probability), which says, for example, that $P(ABCD) = P(A)P(B \mid A)P(C \mid AB)P(D \mid ABC)$, and equality (b) uses the Markov property to eliminate the irrelevant conditioning on previous states. So, in general, the probability of any finite path is simply the probability of the initial state multiplied by a succession of the probabilities of the transitions in the path. ►

Question: How about describing an example of a process that is *not* Markov?

2.2 How the distribution of the state evolves over time, and how matrices come into the picture

Let π_t denote the distribution at time t : that is, $\pi_t(j) = P\{X_t = j\}$.

The relationship between π_{t+1} and π_t is given by the Law of Total Probability:

$$\begin{aligned}\pi_{t+1}(j) &= P\{X_{t+1} = j\} \\ &= \sum_i P\{X_t = i\}P\{X_{t+1} = j \mid X_t = i\} \\ &= \sum_i \pi_t(i)P(i, j)\end{aligned}$$

This is really saying something simple. We are interested in the probability of being in state j at time $t + 1$. In order for this to happen, the chain must be in some state i at time t and then move to state j , and the probability of this happening is $\pi_t(i)P(i, j)$. We get the total probability of being in state j at time $t + 1$ by adding these probabilities over all the possible previous states i .

In the language of matrices

$$\begin{pmatrix} \pi_{t+1}(1) & \pi_{t+1}(2) & \pi_{t+1}(3) \end{pmatrix} = \begin{pmatrix} \pi_t(1) & \pi_t(2) & \pi_t(3) \end{pmatrix} \begin{pmatrix} P(1,1) & P(1,2) & P(1,3) \\ P(2,1) & P(2,2) & P(2,3) \\ P(3,1) & P(3,2) & P(3,3) \end{pmatrix}.$$

That is, $\pi_{t+1} = \pi_t P$.

Remember we think of the π_t 's as row vectors to make this work out.

Example: Suppose we start our frog out at state 2, deterministically (i.e. with probability 1). That is, $\pi_0 = (0 \ 1 \ 0)$.

$$\text{Then } \pi_1 = (0 \ 1 \ 0) \begin{pmatrix} 0 & 1 & 0 \\ .2 & 0 & .8 \\ .1 & .3 & .6 \end{pmatrix} = (.2 \ 0 \ .8)$$

$$\pi_2 = \pi_1 P = (.2 \ 0 \ .8) \begin{pmatrix} 0 & 1 & 0 \\ .2 & 0 & .8 \\ .1 & .3 & .6 \end{pmatrix} = (.08 \ .44 \ .48)$$

And here is the next step in more detail:

$$\begin{aligned}\pi_3 &= \pi_2 P = (.08 \ .44 \ .48) \begin{pmatrix} 0 & 1 & 0 \\ .2 & 0 & .8 \\ .1 & .3 & .6 \end{pmatrix} \\ &= \left((.44)(.2) + (.48)(.1) \quad (.08)(1) + (.48)(.3) \quad (.44)(.8) + (.48)(.6) \right) \\ &= (.136 \ .224 \ .64)\end{aligned}$$

→ Relationship between π_t and π_0 : $\pi_t = \pi_0 P^t$

```

matpow = function(M,n) {
# finds matrix power M^n
# use this for n = integer greater than 1
  ans = M
  for(i in 1:(n-1)) {
    ans = ans %*% M
  }
  ans
}

> c(0,1,0) %*% matpow(P,2)
      [,1] [,2] [,3]
[1,] 0.08 0.44 0.48
> c(0,1,0) %*% matpow(P,3)
      [,1] [,2] [,3]
[1,] 0.136 0.224 0.64

```

Check: Agrees with what we got above.

Let's find the distributions of the frog at times 10, 20, 50, 100, and 101:

```

> c(0,1,0) %*% matpow(P,10)
      [,1] [,2] [,3]
[1,] 0.1175349 0.2945469 0.5879181

> c(0,1,0) %*% matpow(P,20)
      [,1] [,2] [,3]
[1,] 0.1176470 0.2941179 0.5882351

> c(0,1,0) %*% matpow(P,50)
      [,1] [,2] [,3]
[1,] 0.1176471 0.2941176 0.5882353

> c(0,1,0) %*% matpow(P,101)
      [,1] [,2] [,3]
[1,] 0.1176471 0.2941176 0.5882353

```

Another way to see what's going on is to look at the matrix powers P^t .

Interpretation of (i,j) th entry in the matrix P^t :

$$P^t(i,j) = P\{X_t = j \mid X_0 = i\}$$

```

> matpow(P,3)
      [,1] [,2] [,3]
[1,] 0.080 0.440 0.480
[2,] 0.136 0.224 0.640
[3,] 0.116 0.300 0.584

```

We calculated the red row as our π_3 starting from $\pi_0 = (0 \ 1 \ 0)$.

```

> matpow(P,10)
      [,1] [,2] [,3]
[1,] 0.1178793 0.2932285 0.5888922
[2,] 0.1175349 0.2945469 0.5879181
[3,] 0.1176567 0.2940808 0.5882625

```

```
> matpow(P, 20)
      [,1]      [,2]      [,3]
[1,] 0.1176472 0.2941170 0.5882357
[2,] 0.1176470 0.2941179 0.5882351
[3,] 0.1176471 0.2941176 0.5882353

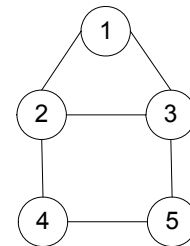
> matpow(P, 50)
      [,1]      [,2]      [,3]
[1,] 0.1176471 0.2941176 0.5882353
[2,] 0.1176471 0.2941176 0.5882353
[3,] 0.1176471 0.2941176 0.5882353
```

This says that, at large times (e.g. 50 is large enough), the frog has probabilities (0.1176471 0.2941176 0.5882353) of being in the various states, *no matter which state the frog started from at time 0*.

These probabilities (0.1176471 0.2941176 0.5882353) are called “steady state probabilities” or “limiting probabilities” for this Markov chain.

Another example of limiting probabilities: Random walk on the house graph

```
> P
      [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.0000000 0.5000000 0.5000000 0.0000000 0.0000000
[2,] 0.3333333 0.0000000 0.3333333 0.3333333 0.0000000
[3,] 0.3333333 0.3333333 0.0000000 0.0000000 0.3333333
[4,] 0.0000000 0.5000000 0.0000000 0.0000000 0.5000000
[5,] 0.0000000 0.0000000 0.5000000 0.5000000 0.0000000
```



Probabilities after 100 steps

```
> matpow(P, 100)
      [,1] [,2] [,3] [,4] [,5]
[1,] 0.1666667 0.25 0.25 0.1666667 0.1666667
[2,] 0.1666667 0.25 0.25 0.1666667 0.1666667
[3,] 0.1666667 0.25 0.25 0.1666667 0.1666667
[4,] 0.1666667 0.25 0.25 0.1666667 0.1666667
[5,] 0.1666667 0.25 0.25 0.1666667 0.1666667
```

2.3 More on Markov chains: Stationary distributions

Markov property: $P\{X_{t+1} = x_{t+1} \mid X_t = x_t, X_{t-1} = x_{t-1}, \dots, X_0 = x_0\} = P\{X_{t+1} = x_{t+1} \mid X_t = x_t\}$

Let π_t denote the distrib of X_t , that is, $\pi_t(i) = P\{X_t = i\}$.

We found $\pi_{t+1}(j) = \sum_i \pi_t(i)P(i, j)$.

That is, $\pi_{t+1} = \pi_t P$. [Think of the π_t 's as row vectors.]

So $\pi_t = \pi_0 P^t$ where P^t is the matrix power (P multiplied by itself t times)

$P^t(i, j) = P\{X_t = j \mid X_0 = i\}$.

E.g. for our frog example,

$$P = \begin{matrix} & \begin{matrix} 1 & 2 & 3 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} & \begin{pmatrix} 0 & 1 & 0 \\ .2 & 0 & .8 \\ .1 & .3 & .6 \end{pmatrix} \end{matrix}$$

```
> matpow(P, 3)
      [,1] [,2] [,3]
[1,] 0.080 0.440 0.480  ← this row is distrib at time 3, starting from  $X_0 = 1$ 
[2,] 0.136 0.224 0.640  ← this row is distrib at time 3, starting from  $X_0 = 2$ 
[3,] 0.116 0.300 0.584  ← this row is distrib at time 3, starting from  $X_0 = 3$ 

> matpow(P, 10)
      [,1] [,2] [,3]
[1,] 0.1178793 0.2932285 0.5888922
[2,] 0.1175349 0.2945469 0.5879181
[3,] 0.1176567 0.2940808 0.5882625

> matpow(P, 50)
      [,1] [,2] [,3]
[1,] 0.1176471 0.2941176 0.5882353
[2,] 0.1176471 0.2941176 0.5882353
[3,] 0.1176471 0.2941176 0.5882353
```

This says that, at large times (e.g. 50 is large enough), the frog has probabilities (0.1176471 0.2941176 0.5882353) of being in the various states, *no matter which state the frog started from at time 0*.

These probabilities (0.1176471 0.2941176 0.5882353) are called “**steady state probabilities**” or “**limiting probabilities**” for this Markov chain.

2.3.1 Stationary distributions

How do we characterize those limiting probabilities mathematically?

Suppose we have $\pi_t \rightarrow \pi$ as $t \rightarrow \infty$. [i.e. $\pi_t(i) \rightarrow \pi(i)$ as $t \rightarrow \infty$, for all states i]

But we know $\pi_{t+1}(j) = \sum_i \pi_t(i)P(i, j)$

So the limiting probabilities must satisfy: $\pi(j) = \sum_i \pi(i)P(i, j)$

As vectors: $\pi_{t+1} = \pi_t P$ for all t and $\pi_t \rightarrow \pi$ as $t \rightarrow \infty$ implies $\pi = \pi P$.

A probability mass function π satisfying the equation $\pi = \pi P$ is called a **stationary distribution** for the probability transition matrix P .

We have shown that a limiting distribution must be stationary.

Another interpretation: Suppose we use a stationary distribution π as our initial distribution: $\pi_0 = \pi$. Then $\pi_1 = \pi_0 P = \pi P = \pi$, and in fact $\pi_t = \pi$ for all t . The distribution does not change – that’s why it’s called “stationary.”

Example: Stationary distribution of Ehrenfest chain is Binomial($n, \frac{1}{2}$).

Intuitive explanation: Imagine a string of n bits and think of a step of the Ehrenfest chain as choosing a random bit and flipping it.

If we start with $X_0 \sim B(n, 1/2)$ then we also have $X_1 \sim B(n, 1/2)$. ►

In the previous example we could guess by having sufficient insight into the process.

We can calculate stationary distributions systematically by solving linear equations.

Example: The frog.

For $P = \begin{pmatrix} 0 & 1 & 0 \\ .2 & 0 & .8 \\ .1 & .3 & .6 \end{pmatrix}$ the equations $\pi = \pi P$ become

$$\pi_1 = .2\pi_2 + .1\pi_3$$

$$\pi_2 = \pi_1 + .3\pi_3$$

$$\pi_3 = .8\pi_2 + .6\pi_3$$

There is more than one way to go about solving these equations. Here is one that is not particularly carefully chosen.

Substituting the first equation in the second gives

$$\pi_2 = (.2\pi_2 + .1\pi_3) + .3\pi_3 = .2\pi_2 + .4\pi_3, \text{ or } .8\pi_2 = .4\pi_3, \text{ or } \pi_2 = (1/2)\pi_3.$$

Putting this relationship back into the first equation gives

$$\pi_1 = .2[(1/2)\pi_3] + .1\pi_3 = .2\pi_3 = (1/5)\pi_3$$

So we have found $\pi_1 = (1/5)\pi_3$ and $\pi_2 = (1/2)\pi_3$.

There are no more relationships to be found from the 3 equations above. [[In fact the 3rd equation, which we haven't used yet, just says $.4\pi_3 = .8\pi_2$, or $\pi_2 = (1/2)\pi_3$, which we already knew.]]

Using $\sum_i \pi_i = 1$ we get $(1/5)\pi_3 + (1/2)\pi_3 + \pi_3 = 1$, or $\pi_3 = \frac{10}{17}$.

$$\pi_1 = \frac{2}{17} = 0.1176471, \quad \pi_2 = \frac{5}{17} = 0.2941176, \quad \pi_3 = \frac{10}{17} = 0.5882353.$$

You know, for a problem like this one, it's less effort to get the computer to take the matrix P to a large power to see the answer rather than doing the algebra...

Remember?

```
> matpow(P, 50)
      [,1]      [,2]      [,3]
[1,] 0.1176471 0.2941176 0.5882353
[2,] 0.1176471 0.2941176 0.5882353
```

[3,] 0.1176471 0.2941176 0.5882353

...but it's good to know how to do the algebra.

Example: If a prob transition matrix is symmetric, then the uniform distrib is stationary.

Why? Note symmetry implies the column sums of the transition matrix are all 1's... ►

A simple generalization of this symmetry idea turns out to be very important. In fact, later on we'll mainly be talking about and using MC's that have a more general kind of symmetry property called "**time reversibility**."

Fact: If a probability mass function μ on the state space \mathbb{S} of a MC satisfies $\mu(i)P(i, j) = \mu(j)P(j, i)$ for all states i and j , then in fact μ is a stationary distribution for P .

In this case we say the chain P is *time reversible*. We'll explain the interesting name in a moment.

Proof of the fact: We are given $\mu(i)P(i, j) = \mu(j)P(j, i)$.

Sum it over i to get $\sum_i \mu(i)P(i, j) = \mu(j) \underbrace{\sum_i P(j, i)}_1 = \mu(j)$.

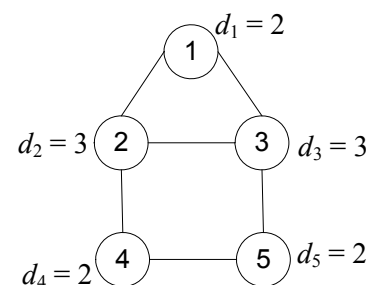
That is, $\mu(j) = \sum_i \mu(i)P(i, j)$, which says that μ is stationary. ►

Why is it called "time reversibility"? It's not necessarily important to know this, but it's kind of interesting. Suppose we are given that $\pi(i)P(i, j) = \pi(j)P(j, i)$ for all i, j , so that we know that π is a stationary distribution for the transition matrix P . Suppose X is a Markov chain having transition matrix P and initial distribution $X_0 \sim \pi$. Then we claim that the chain X_0, X_1, \dots is "time reversible" in the sense that if I showed you a movie of the chain running backward in time, you would not be able to detect that the movie is running backwards by any statistical tests. To see this, we will show that every possible path of the chain has the same probability as the time-reversed version of the path. For example, let's check this on paths of length 4; we'll show that the two paths $i \rightarrow j \rightarrow k \rightarrow l$ and $l \rightarrow k \rightarrow j \rightarrow i$ have the same probability:

$$\begin{aligned} P\{X_0 = i, X_1 = j, X_2 = k, X_3 = l\} \\ &= \pi(i)P(i, j)P(j, k)P(k, l) \\ &= \pi(j)P(j, i)P(j, k)P(k, l) \\ &= \pi(k)P(j, i)P(k, j)P(k, l) \\ &= \pi(l)P(j, i)P(k, j)P(l, k) \\ &= \pi(l)P(l, k)P(k, j)P(j, i) \\ &= P\{X_0 = l, X_1 = k, X_2 = j, X_3 = i\} \end{aligned}$$

Example: Random walks on graphs

Let $d(i)$ = the *degree* of node i , that is, the number of edges touching node i .



Here $P(i, j) = \frac{1}{d(i)}$ for each neighbor j of i , and $P(i, j) = 0$ otherwise.

Can check that the distribution $\mu(i) = \frac{d(i)}{\sum_j d(j)}$ satisfies the time-reversibility condition,

and therefore is stationary. ▶

Example: Random walk of a King on a 4 by 4 chessboard.

all possible moves up&down, right&left, diagonal equally likely

Degrees:

3	5	5	3
5	8	8	5
5	8	8	5
3	5	5	3

Here sum of degrees is 84, and so $\pi(i) = d(i) / 84$. ▶

Existence and uniqueness of stationary distributions

→ It is possible for a transition matrix to have more than one stationary distribution. For example, if

$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, the identity matrix, then of course *every* distribution is stationary: $\pi P = \pi$.

→ It is also possible to have no stationary distribution. For example, if we have a random walk on the integers with probability $p = 1/2$ of moving up and probability $1/2$ of moving down, the equations for a stationary distribution become $\pi(j) = \frac{1}{2}(\pi(j-1) + \pi(j+1))$

A bit of thought then shows these equations are inconsistent with the existence of a stationary distribution.

However...

Fact: If a MC has finitely many states, then it must have at least one stationary distribution.

Fact: If a MC is *irreducible*, that is, from each state it is possible to get to every other state in finite time, the chain has at most one stationary distribution.

2.4 Limit Theorem for Markov Chains

For the frog chain, we found that as $t \rightarrow \infty$, the distribution of X_t converges to the same limiting distribution $\pi = (2/17 \ 5/17 \ 10/17)$, no matter where we start the frog at time 0.

And fractions of time spent in states 1, 2, and 3 converge to these same limits.

Which chains have this convergence property (that is, convergence to a single limit distribution that does not depend on where the chain is started)? There are some counterexamples to rule out.

For example, the chain $P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

Here every distribution is stationary and the chain does not "forget" its past.

The problem here is that the states do not "communicate."

Definition: Given a Markov chain transition matrix P , we say that state j is **accessible** from state i if $P^t(i, j) > 0$ for some t . States i and j **communicate** if j is accessible from i and i is accessible from j .

Definition: We say that a Markov chain transition matrix P is **irreducible** if all pairs of states in the state space communicate.

Example: The frog chain is irreducible.

$$P = \begin{matrix} & \begin{matrix} 1 & 2 & 3 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} & \begin{pmatrix} 0 & 1 & 0 \\ .2 & 0 & .8 \\ .1 & .3 & .6 \end{pmatrix} \end{matrix}$$

Ergodic Theorem for Markov chains:

Let P be an irreducible transition matrix having a stationary distribution π . We can choose any arbitrary starting state for the chain: $X_0 = i$, say. Then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n I\{X_t = j\} = \pi(j) \text{ for all states } j.$$

[More generally, $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n f(X_t) = E_\pi(f(X))$ where $E_\pi(f(X)) = \sum_j f(j) \pi(j)$]

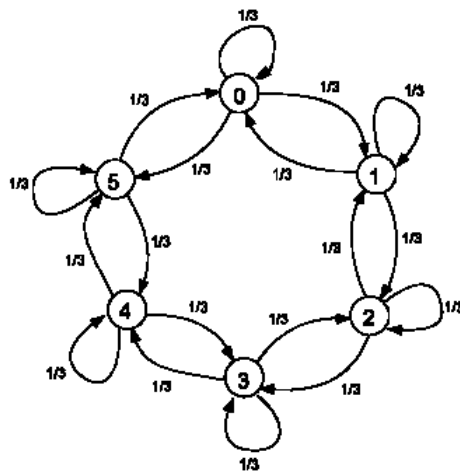
"Ergodic theorems" have conclusions of the form: as time goes to infinity, a time average converges to an expectation (a probabilistic average).

2.5 Use of Markov chains for simulation

The idea here is to use the Limit Theorem backward, in a sense, to generate random objects having a desired distribution.

A silly example: random walk on a clock. Suppose we want to simulate a random draw from the set $\mathbb{S} = \{0, 1, 2, 3, 4, 5\}$.

The obvious way is to generate a uniform random



number $U \sim \text{Unif}(0,1)$ and then take $X=0$ if U is between 0 and $1/6$, $X=1$ if U is between $1/6$ and $2/6$, ..., and $X=5$ if U is between $5/6$ and 1.

Another way is to perform this random walk for a long time.

A Not-Silly Example: Generating a random (uniformly distributed) 4 by 4 table of nonnegative integers having specified row and column sums. The interesting thing is that this example is using the same idea as the silly "random walk on a clock" example, but here we are doing a problem that we'd have a very difficult time finding another good way to do!

68	119	26	7	220
20	84	17	94	215
15	54	14	10	93
5	29	14	16	64
108	286	71	127	

Let \mathbb{S} denote the set of all 4 by 4 tables of nonnegative integers (counts) with the same row and column sums as this table.

Say we want to produce a sample of tables from \mathbb{S} that are uniformly distributed. This means Prob $\frac{1}{\#(\mathbb{S})}$ for each table in \mathbb{S} .

We don't even know how many tables there are in \mathbb{S} !

There is no straightforward way to do this, as there was for the "clock". But we can still use Markov chains.

E.g. one way: choose two rows and two columns at random. The intersections give a 2 by 2 sub-table.

Now toss a coin. If it comes up heads, add

1	-1
-1	1

 to that sub-table, otherwise add

-1	1
1	-1

, if you can do so without violating the constraints. Otherwise, just stay where you are.

Note the MC probability transition matrix is symmetric, so the uniform distrib is stationary.

This chain can be shown to be irreducible, so the Ergodic Thm says this works! ►

Illustrating the idea of the MC on tables with fixed row and col sums

```
x0 = matrix(c(8,1,4,2,7,1,3,1,1),nrow = 3)
```

```
#> x0
```

```
#      [,1] [,2] [,3]
# [1,]    8    2    3
# [2,]    1    7    1
# [3,]    4    1    1
```

```
move = function(x, print=T) {
  m = dim(x)[1]
  n = dim(x)[2]
  temp = sample(m,2)
  i1 = min(temp)
  i2 = max(temp)
  temp = sample(n,2)
  j1 = min(temp)
```

```

j2 = max(temp)
delta1 = matrix(c(1,-1,-1,1),nrow=2)
delta2 = matrix(c(-1,1,1,-1),nrow=2)
change = F
if(runif(1) < 0.5){
  if(min(x[i1,j2],x[i2,j1]) > 0){
    change = T
    x[c(i1,i2),c(j1,j2)]=x[c(i1,i2),c(j1,j2)]+delta1
  }
}
else{
  if(min(x[i1,j1],x[i2,j2]) > 0){
    change = T
    x[c(i1,i2),c(j1,j2)]=x[c(i1,i2),c(j1,j2)]+delta2
  }
}
if(print){
  cat(paste("\n(i1,i2)=(",i1," ",i2," ")
(j1,j2)=( ",j1," ",j2," )",sep=""))
  if(!change)cat("  no change")
  cat("\n")
  print(x)
}
x
}

x = x0
x = move(x) # repeat this several times to see how it works

x = x0
results = list(x)
for(i in 2:10){ # change to 1000
  x = move(x,print=F)
  results[[i]] = x
}

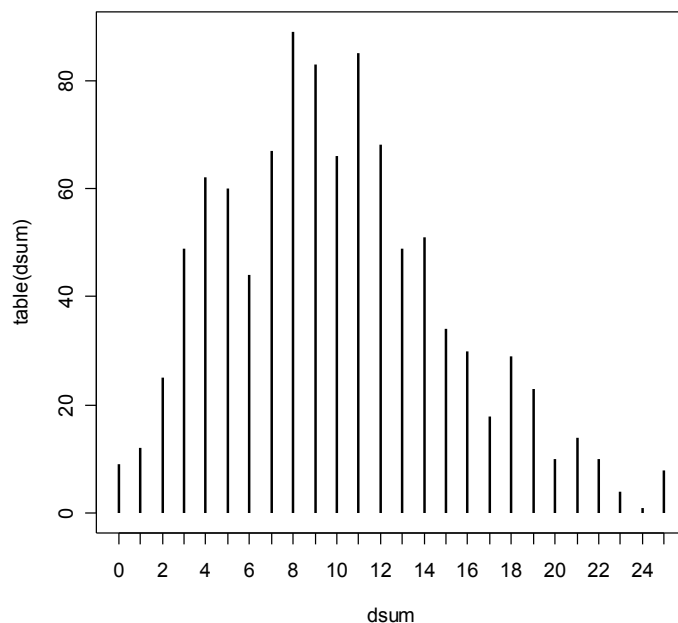
f = function(x){x[1,1]} # change this to collect different elements
corner = lapply(results,f)
corner = unlist(corner)

plot(corner)
table(corner)
plot(table(corner))

g = function(x){sum(diag(x))}

dsum = lapply(results,g)
dsum = unlist(dsum)
plot(dsum)
table(dsum)
plot(table(dsum))

```



There is a also a more efficient way included in a file on the web; see the function `moveMore`.

2.6 Designer Markov chains: The Metropolis-Hastings method

E.g. One can get a uniformly distributed point on the floor by walking around randomly for a long time. If we want a different density, we can bias our moves toward regions of higher density by sometimes rejecting moves if they would lead us to a state of lower density than the current state.

Suppose we want to simulate a representative random sample from a distribution π on a set \mathbb{S} .

By the ergodic theorem, we can do this by running a Markov chain on \mathbb{S} for a sufficiently long time. All we need is to set up a Markov chain that is irreducible and has π as its stationary distribution!

2.6.1 Example of the Metropolis method:

Suppose we want to have a MC on the real line that has a density f as its stationary distribution.

Suppose we are at the current state $X_t = x$. We want to decide where to move next. This is done in two stages.

Stage 1:

Choose a *candidate* state, y , from a particular density centered on x . For example, we could take y from the uniform density $U(x-1, x+1)$.

Stage 2:

Now we make a probabilistic decision about whether to *accept* the candidate.

- If $f(y) \geq f(x)$, accept the candidate; that is, take $X_{t+1} = y$.
- If $f(y) < f(x)$, accept the candidate with probability $\frac{f(y)}{f(x)}$, that is, take

$$X_{t+1} = \begin{cases} y & \text{with prob } f(y) / f(x) \\ x & \text{with prob } 1 - (f(y) / f(x)) \end{cases}$$

The Metropolis MC consists of a long sequence of repetitions of this step.

Let's look at some simple examples: using Metropolis to generate Normal and Exponentially distributed random variables. These will be done to illustrate the method and see how it works in some simple cases; of course we don't *need* Metropolis to do these simulation problems – e.g. we could just have R do `rnorm(1000)` or `rexp(1000)` and be done with it.

We'll try out some R commands (see below) and see how they work in class. Isn't it wonderful how dumb and mindless this is? Kind of like life... (e.g. evolution)

2.6.2 The general idea of Metropolis-Hastings (explained in the discrete case)

Problem: Given f , a probability mass function on a set \mathbb{S} , find a probability transition matrix P such that $fP = f$, i.e., $\sum_{x \in \mathbb{S}} f(x)P(x, y) = f(y)$.

...Note sometimes it is not easy or convenient to use symmetric proposals...

Let Q be *any* probability transition matrix on \mathbb{S} . We use Q as a way of generating candidates.

One Metropolis-Hastings step consists of 2 stages.

Stage 1:

Suppose we are currently at state x . Propose a candidate according to the distribution specified by row x of Q . That is, let the candidate be y with probability $Q(x, y)$.

Stage 2:

Now let y denote the candidate actually proposed.

Look at the ratio $r = \frac{f(y) Q(y, x)}{f(x) Q(x, y)}$.

- If $r \geq 1$, accept the candidate [take $X_{t+1} = y$].
- If $r < 1$: accept the candidate [take $X_{t+1} = y$] with probability r , and reject the candidate [take $X_{t+1} = x$] with probability $1 - r$.

Note: if the proposal mechanism Q is symmetric [$Q(x, y) = Q(y, x)$], then the acceptance ratio r is simply $\frac{f(y)}{f(x)}$, as in our initial example above.

A concise way to describe the Metropolis chain's probability transition matrix:

$$P(x, y) = Q(x, y) \min \left\{ 1, \frac{f(y) Q(y, x)}{f(x) Q(x, y)} \right\} \quad \text{for } y \neq x.$$

$$[\text{And } P(x, x) = 1 - \sum_{y \neq x} P(x, y) \text{ of course}]$$

[This is the same as what we said before because the probability of accepting a candidate is

$$\begin{cases} 1 & \text{if } r \geq 1 \\ r & \text{if } r < 1 \end{cases} = \min\{1, r\} = \min \left\{ 1, \frac{f(y) Q(y, x)}{f(x) Q(x, y)} \right\}, \text{ and because in order to move to } y, \text{ we need to}$$

propose y as a candidate and then accept the candidate.]

2.6.3 Why does the Metropolis-Hastings recipe work?

A simple trick!

For $y \neq x$,

$$f(x)P(x, y) = f(x)Q(x, y) \min \left\{ 1, \frac{f(y) Q(y, x)}{f(x) Q(x, y)} \right\} = \min \{ f(x) Q(x, y), f(y) Q(y, x) \}$$

Note this is symmetric in x and y . So $f(x)P(x,y) = f(y)P(y,x)$.

We know this is a sufficient condition for f to be stationary for the matrix P .

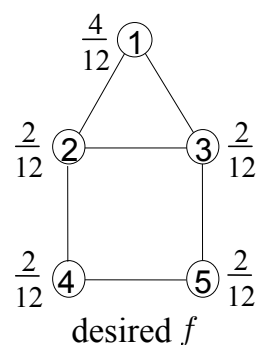
[[This is an example of the “time reversibility” fact from before:

Fact: If a probability mass function μ on the state space \mathbb{S} of a MC satisfies $\mu(i)P(i,j) = \mu(j)P(j,i)$ for all states i and j , then in fact μ is a stationary distribution for P .]]

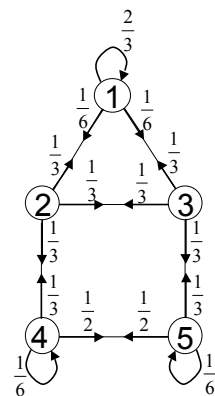
An Important Good Thing: In order to run the method, we do not need to be able to calculate individual values $f(x)$, but rather *ratios* of the form $f(y)/f(x)$. There are many problems in which nobody knows how to calculate the individual values, but ratios are easy!

[E.g. the four by four table problem is a simple example.]

Example: a desired distribution for the small house graph



Here let's take our Q to correspond to an ordinary random walk on the graph. The Metropolis-Hastings chain making the desired f stationary is



E.g. the $P(1,3) = 1/6$ comes from the calculation

$$P(1,3) = Q(1,3) \min \left\{ 1, \frac{f(3)Q(3,1)}{f(1)Q(1,3)} \right\} = \frac{1}{2} \min \left\{ 1, \frac{(2/12)(1/3)}{(4/12)(1/2)} \right\} = \frac{1}{2} \left\{ \frac{1}{3} \right\} = \frac{1}{6}.$$



Try out simple examples of the Metropolis method in R:

Standard Normal distribution example

Stat 238 notes, 9/2/09

```

f = function(x) {exp(-x^2/2)}

# We can write f(x) as exp(-x^2/2), without
# the 1/sqrt(2*pi), since normalization constants are not needed
# in the metropolis method; it uses only *ratios* of probabilities.

nit = 10000
results = numeric(nit)
scale = 1
initial = 3
state = initial
results[1] = initial

for(i in 2:nit){
  candidate = runif(1,state-scale,state+scale)
  ratio = f(candidate)/f(state)
  if(runif(1) < ratio) state = candidate
  results[i] = state
}

plot(results)
hist(results)
qqnorm(results)

ppnorm = function(x){
  L = length(x)
  x.sort = sort(x)
  plot(c(0, (1:L - 0.5)/L, 1), c(0, pnorm(x.sort), 1),
       ylab = "theoretical prob", xlab = "prob in data", type="b")
}

ppnorm(results)

# Next repeat with this weird one:
f = function(th){
  if (th>0 & th < 5) return((sin(th))^2*abs(th-2)^.3)
  return(0)
}

# Oops, is this a density? Don't worry...

par(mfrow=c(2,1))
myhist(results)
thetas=seq(0.01, 4.95, length=100)
y = numeric(100)
for(i in 1:100) y[i]=f(thetas[i])
plot(thetas, y, type="l")

```

<i>Guts of Metropolis</i>

3 Statistical models, estimation, mean squared error, maximum likelihood

3.1 Statistical models

Probability versus statistics:

- Probability problems start with a given probability distribution and ask about the probability of observing various values for random variables coming from that distribution.
- Statistics problems start with some observed values of random variables and ask about which probability distribution the random variables came from.

A *model* involves (is?) a collection of probability measures $\{P_\theta : \theta \in \Theta\}$.

Θ is the *parameter space*. Could be multi-dimensional, discrete, partly discrete and partly continuous...

Data: X . E.g. could be a vector $X = (X_1, \dots, X_n)$. [Note to myself: Below is slightly vague on this point and on notation.]

Example: $P_\theta = N(\theta, 1)$, a probability measure on \mathbb{R} . Here Θ is also \mathbb{R} . Suppose X_1, X_2, \dots, X_n are iid according to P_θ . We use P_θ and E_θ to talk about probability and expectation. For example, $P_\theta\{X_1 \leq \theta\} = 1/2$, $E_\theta(X_1) = \theta$, $\text{var}_\theta(\bar{X}) = 1/n$, etc.

Example: In our original “laser pole” problem, P_θ is the Cauchy density centered at θ , and Θ is again \mathbb{R} .

Example: If we imagine X_1, X_2, \dots, X_n are iid according to $N(\mu, \sigma^2)$, and we know neither μ nor σ^2 , then we would think of $\theta = (\mu, \sigma^2)$. So the parameter space is now 2-dimensional:

$$\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, 0 \leq \sigma^2 < \infty\} = \mathbb{R} \times \mathbb{R}_+.$$

Example: [simple regression] We might imagine x_1, x_2, \dots, x_n are nonrandom numbers and, for $i = 1, \dots, n$, Y_i is normally distributed with mean $\alpha + \beta x_i$ and variance σ^2 . In this case, $\Theta = \{(\alpha, \beta, \sigma^2) : -\infty < \alpha < \infty, -\infty < \beta < \infty, 0 \leq \sigma^2 < \infty\} = \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+$

All statisticians agree that a statistical model involves a family of probability distributions in this way.

This part of the model gives rise to the *likelihood function*, which is the probability of the data, as a function of θ . [“Probability” here could be either discrete or density.] Pretty much all statisticians agree that likelihood functions are meaningful and important.

Bayesians also view θ as a random variable, and so Bayesian models include a probability distrib for θ , the *prior distribution*. The model is viewed as a collection of conditional distributions for the data X , given the parameter θ .

For nonBayesian statisticians, θ is not viewed as a random variable.

3.2 Estimation

A *statistic* is a function of the data. It must be computable, in the sense that it should not involve any unknown quantities, like unknown parameters.

An *estimator* is a statistic that we use as a guess for the value of an unknown parameter (maybe parameter vector). Since it's a function of the data, X , we write an estimator δ as $\delta = \delta(X)$.

What is a *good* estimator?

For example, why is the median better than the mean in some problems? Why might a weighted average of the sample values be a better estimator of a mean in some problems than a straight sample average? What do we mean by "better"?

The *sampling distribution* of an estimator $\delta(X)$ is the probability distribution of $\delta(X)$.

The sampling distribution will depend on the unknown θ .

Example: Let X_1, \dots, X_n be iid $\sim \text{Exp}(\theta)$. We want to estimate the mean, $1/\theta$, using $\delta(X) = \bar{X}$. $E(X_i) = 1/\theta$, $\text{var}(X_i) = \theta^2 \Rightarrow E(\bar{X}) = 1/\theta$, $\text{var}(\bar{X}) = 1/(n\theta^2)$. By the Central Limit Theorem, the sampling distribution of $\delta(X)$ is approximately $N(1/\theta, 1/(n\theta^2))$. ►

One common measure of the typical error committed by an estimator:

The **mean squared error** of an estimator $\delta(X)$ is $\text{MSE}_\theta(\delta) = E_\theta[(\delta(X) - \theta)^2]$.

This is a function of θ .

[[The smaller the better, of course. But the comparison of two *functions* does not in general give a clear winner – one estimator might have the smaller MSE for some values of θ and the other estimator might have the smaller MSE for other θ 's.]]

To introduce other related quantities: We would like an estimator such that

- the mean of the sampling distribution is as close as possible to θ
- the variance of the sampling distribution is as small as possible.

⊗ The *bias* of the estimator $\delta(X)$ at θ measures the discrepancy between the mean of the sampling distribution and the true θ :

$$\text{bias}(\delta(X)) = E_\theta(\delta(X)) - \theta$$

⊗ The variance of the sampling distribution at θ is $\text{var}_\theta(\delta(X))$.

→ Relationship: $\text{MSE} = \text{bias}^2 + \text{var}$.

That is, $E_\theta[(\delta(X) - \theta)^2] = (E_\theta(\delta(X)) - \theta)^2 + \text{var}_\theta(\delta(X))$.

☺ Small MSE is good. Small bias is good. Small variance is good.

But there is sometimes (often) a tradeoff between bias and variance – estimators that have smaller variance have larger bias. We'll see more of this when we look at regression and model selection.

For now, another well known example.

[[People are probably more concerned about this than they should be, if their concern is practical. But it's useful as an example that helps give conceptual clarity about the issues.]]

Example: Biased and unbiased estimates of variance.

Suppose X_1, \dots, X_n are a sample (that is, iid) from some Normal distribution, but we know neither the mean nor the variance. So our model is the family $\{N(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 \geq 0\}$. We are interested in estimating σ^2 .

Let's compare two estimators: $\delta_0(X) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ and $\delta_1(X) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$,

where of course \bar{X} is the sample mean $(1/n) \sum_{i=1}^n X_i$.

A preliminary calculation:

$$\begin{aligned} \sum (X_i - \mu)^2 &= \sum ((X_i - \bar{X}) + (\bar{X} - \mu))^2 \\ &= \sum (X_i - \bar{X})^2 + 2 \underbrace{\sum (X_i - \bar{X})(\bar{X} - \mu)}_{(\bar{X} - \mu) \sum (X_i - \bar{X}) = 0} + \sum (\bar{X} - \mu)^2 \\ &= \left(\sum (X_i - \bar{X})^2 \right) + n(\bar{X} - \mu)^2, \end{aligned}$$

$$\text{or } \sum (X_i - \bar{X})^2 = \sum (X_i - \mu)^2 - n(\bar{X} - \mu)^2.$$

So

$$\begin{aligned} E \left[\sum (X_i - \bar{X})^2 \right] &= E \left[\sum (X_i - \mu)^2 \right] - nE \left[(\bar{X} - \mu)^2 \right] \\ &= n\sigma^2 - n \underbrace{\text{var}(\bar{X})}_{\sigma^2/n} = (n-1)\sigma^2. \end{aligned}$$

From this we can see that $E(\delta_1(X)) = E \left[\frac{1}{n-1} \sum (X_i - \bar{X})^2 \right] = \sigma^2$, that is, $\delta_1(X)$ is an unbiased estimator of σ^2 .

The estimator $\delta_0(X)$ is biased: $E(\delta_0(X)) = \frac{n-1}{n} \sigma^2$, so that

$$\text{bias}_\theta(\delta_0(X)) = E_\theta(\delta_0(X)) - \sigma^2 = \frac{-\sigma^2}{n}.$$

Now that we have compared the biases of δ_0 and δ_1 , how about the variances?

A more detailed analysis shows that $\text{var} \left(\sum_{i=1}^n (X_i - \bar{X})^2 \right) = 2(n-1)\sigma^4$.

[Note: this is the first place in this derivation that we have used the assumption that we are working with a Normal distribution.]

$$\text{So } \text{var}(\delta_0(X)) = \frac{2(n-1)\sigma^4}{n^2} < \frac{2(n-1)\sigma^4}{(n-1)^2} = \text{var}(\delta_1(X)).$$

So we have a tradeoff: δ_0 has more bias, but less variance, than δ_1 .

Is there a clear overall winner between δ_0 and δ_1 , in terms of MSE?

$$\begin{aligned} \text{MSE}_\theta(\delta_0) &= [\text{bias}_\theta(\delta_0)]^2 + \text{var}_\theta(\delta_0) \\ &= \left(\frac{-\sigma^2}{n}\right)^2 + \frac{2(n-1)\sigma^4}{n^2} = \frac{\sigma^4}{n^2} [1 + 2(n-1)] = \frac{\sigma^4}{n^2} (2n-1) \end{aligned}$$

$$\text{MSE}_\theta(\delta_1) = [\text{bias}_\theta(\delta_1)]^2 + \text{var}_\theta(\delta_1) = (0)^2 + \frac{2\sigma^4}{n-1} = \frac{2\sigma^4}{n-1}.$$

So we see that $\text{MSE}_\theta(\delta_0) < \text{MSE}_\theta(\delta_1)$. ►

We can use R to do a simulation... e.g. sample of size 2.

Note also that we have gone through all of this for the conceptual interest and clarity I hope it will give, not for the practical importance of the difference between δ_0 and δ_1 !

3.3 Estimation examples and relative efficiency: mean versus median, Normal and double exponential distributions

Let's investigate some more examples and estimators through simulation.

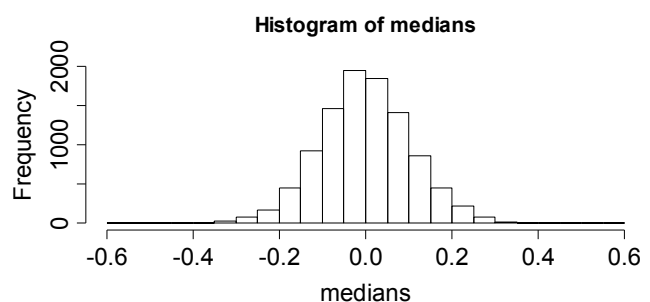
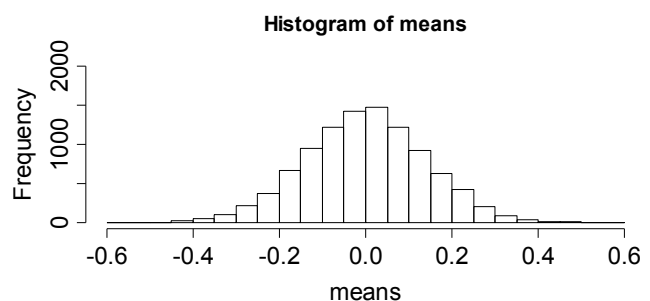
Example: Double exponential. Here the density is $f_\theta(x) = \frac{1}{2} e^{-|x-\theta|}$ for $-\infty < x < \infty$. We are interested in estimating the parameter θ , which is the mean (= median = center) of the distribution. We assume we are given a sample X_1, \dots, X_n (iid) from this distribution, and we want to compare the performance of the sample mean and the sample median as estimators of θ .

Here is a comparison of two simulated sampling distributions for the sample mean and median. This simulation took the example $n = 100$, that is, we are dealing with samples of size 100. The simulation was repeated 10,000 times, and the two histograms are based on 10,000 sample means and 10,000 sample medians.

In the simulations the true θ was taken to be 0.

Easy to see these estimators are unbiased; the interesting comparison is of the variances of their sampling distributions.

```
> var(means)
[1] 0.01962481
> var(medians)
[1] 0.01137893
```



So the sample median looks better here, with a substantial reduction in variance.

This can be summarized by saying that the relative efficiency of the median to the mean is about $.0196/.0114 = 1.72$. This means that, in order to achieve the same accuracy of estimation as one would achieve using the sample median with a sample of size n , if we used the sample mean instead, we would need a sample of size about $1.72 n$. ►

I'll try to say a bit more here about this concept of relative efficiency, without proving anything, but just saying how it works. First, it is easy to see by the symmetry of the above problem that both the sample mean and the sample median are unbiased estimators. So the mean squared error will simply be the variance of the sampling distribution. Second, it turns out that, as the sample size n grows large, estimators typically have sampling distributions that are asymptotically Normal, with variance that decreases inversely proportionally to n . That is, for large n the variance of the sampling distribution of an estimator will typically behave like c/n for some c . If we are comparing two estimators, such as the sample mean and the sample median as we are doing here, each estimator will have its own value of " c ," so that here, the MSE of the mean will be $\sim c_{\text{mean}}/n$ and the MSE of the median will be $\sim c_{\text{median}}/n$ as $n \rightarrow \infty$. From the simulation in the above example, we found for $n = 100$ that the variance of the sample mean and median were about 0.01962 and 0.01138, respectively. This indicates that

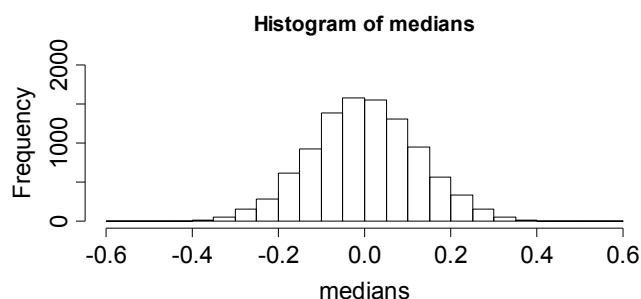
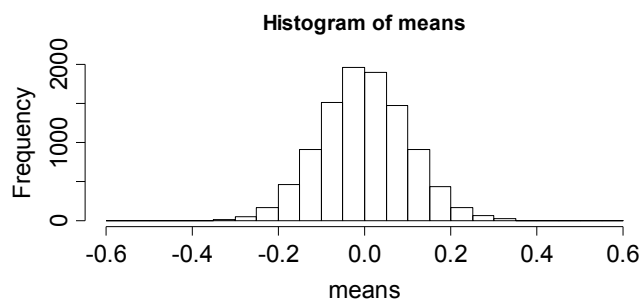
$$\frac{c_{\text{mean}}}{c_{\text{median}}} \approx \frac{0.01962}{0.01138} \approx 1.72.$$

Now suppose we want to determine the relationship between sample sizes

we need for the mean and for the median to have equal accuracy. Let n_{mean} and n_{median} denote sample sizes we are contemplating for the mean and for the median. To have equivalent MSE, these sample sizes should satisfy the relationship $\frac{c_{\text{mean}}}{n_{\text{mean}}} = \frac{c_{\text{median}}}{n_{\text{median}}}$, from which we obtain $\frac{n_{\text{mean}}}{n_{\text{median}}} = \frac{c_{\text{mean}}}{c_{\text{median}}} \approx 1.72$. That

is, for the sample mean to achieve the same accuracy as the sample median in this double-exponential distribution problem, we need the sample size to be 1.72 times as large. In other words, if you use a sample size 1000 from a double exponential distribution and use the sample mean to estimate the center of the distribution, it is as if you are wasting a substantial fraction of your data, since if you were using the sample median instead you could have achieved the same accuracy with a sample size of only $\frac{1000}{1.72} \approx 580$.

Example: Normal distribution. The simulation was done as above, with sample size $n = 100$, and 10,000 repetitions, except the population distribution was taken to be $N(0,1)$.




```
> var(means)
[1] 0.01010825
> var(medians)
[1] 0.01557692
```

In the Normal case, the sample mean is better! The relative efficiency is about $.01558/.01011 = 1.54$. ►

3.4 Maximum likelihood estimation

Example: Normal distribution with known variance $\{N(\theta, 1) : \theta \in \mathbb{R}\}$

$$L(\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(X_i - \theta)^2\right) \propto \exp\left(-\frac{1}{2} \sum_{i=1}^n (X_i - \theta)^2\right)$$

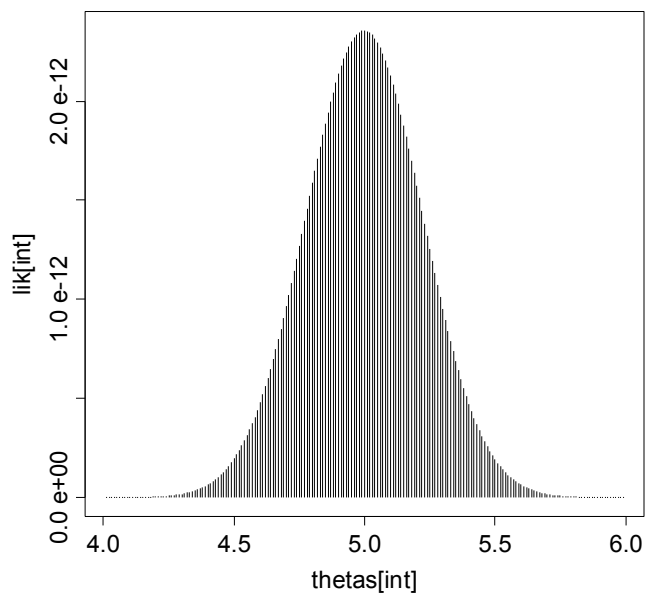
This likelihood is maximized at the θ that minimizes $\sum_{i=1}^n (X_i - \theta)^2$, that is, $\hat{\theta} = \bar{X}$. ►

```
normal.sample = rnorm(20, 5, 1)
> normal.sample
[1] 6.055599 4.888401 4.887066 4.587871 5.253812 3.795483 4.570845 7.409802
[9] 5.925742 4.484033 5.417041 4.511173 3.844954 4.744162 5.310769 6.361405
[17] 5.656658 4.260257 3.743060 4.272400

x = normal.sample
thetas = seq(0, 10, by = .01)

# Likelihood function for Normal
L <- function(th, x){
  return(prod(dnorm(x-th)))
}

# Plot of likelihood
lik <- thetas
m <- length(thetas)
for(i in 1:m){
  lik[i] <- L(thetas[i], x)
}
plot(thetas, lik, type="l")
int = (thetas > 4) & (thetas < 6)
plot(thetas[int], lik[int], type="h")
```



```
> mean(x)
[1] 4.999027
```

Example: Double exponential distribution with mean θ .

$f_{\theta}(x) = \frac{1}{2} e^{-|x-\theta|}$. Likelihood:

$$L(\theta) = \prod_{i=1}^n \frac{1}{2} \exp(-|X_i - \theta|) \propto \exp\left(-\sum_{i=1}^n |X_i - \theta|\right)$$

CLAIM: this is maximized at $\hat{\theta} = \text{median of } X_1, \dots, X_n$.

We want to find the θ that *minimizes* $\sum_{i=1}^n |X_i - \theta|$.

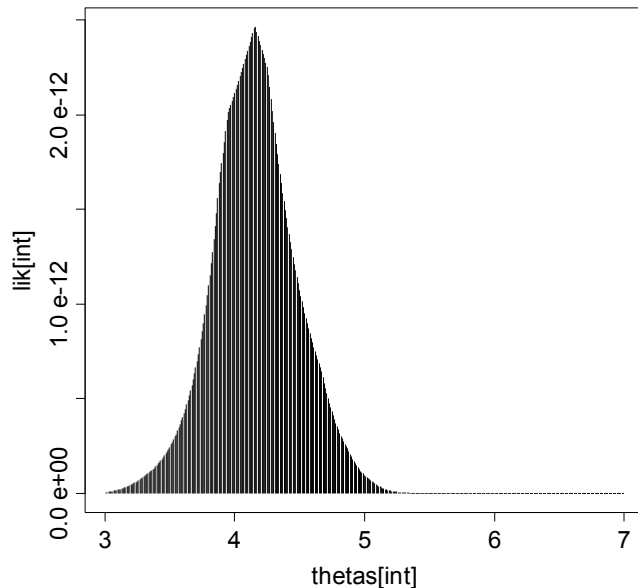
Note $\frac{d}{d\theta} |X_i - \theta| = \begin{cases} -1 & \text{for } \theta < X_i \\ +1 & \text{for } \theta > X_i \end{cases}$.

That is, $\frac{d}{d\theta} |X_i - \theta| = I\{\theta > X_i\} - I\{\theta < X_i\}$.

$$\begin{aligned} \frac{d}{d\theta} \sum_{i=1}^n |X_i - \theta| &= \sum_{i=1}^n (I\{\theta > X_i\} - I\{\theta < X_i\}) \\ &= \#(i : \theta > X_i) - \#(i : \theta < X_i) \\ &= \begin{cases} < 0 & \text{if } \#(i : \theta > X_i) < \#(i : \theta < X_i) \\ > 0 & \text{if } \#(i : \theta > X_i) > \#(i : \theta < X_i) \end{cases} \end{aligned}$$

This says $\frac{d}{d\theta} \sum_{i=1}^n |X_i - \theta| < 0$ for $\theta < \text{median}(X_1, \dots, X_n)$ and $\frac{d}{d\theta} \sum_{i=1}^n |X_i - \theta| > 0$ for $\theta > \text{median}(X_1, \dots, X_n)$, which is what we wanted to show. ▶

```
dexp.sample = 5 + rexp(19)*sign(runif(19)-0.5)
> sort(dexp.sample)
 [1] 0.3450312 0.9251110 2.2562507 2.7284197 3.0964596 3.1959126 3.2860848
 [8] 3.8820842 3.9458168 4.1561119 4.2554637 4.6684314 4.8761312 5.0624903
[15] 5.0863673 5.5371111 6.5535385 6.6418261 7.7053644
```



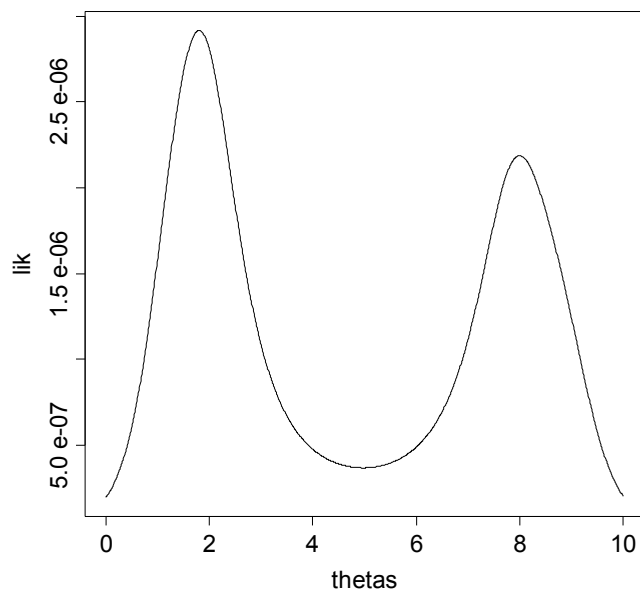
For variety, I chose this one to look pretty bad after several tries, since I kept getting medians so close to 5!

☺ → It is not a coincidence that the MLE for the Normal case is the sample mean and the MLE for the double exponential case is the median, both of which we found to be the better choice according to our simulations. Maximum likelihood estimators typically (under certain “regularity conditions,” etc.) perform about as well as one could hope for in a given problem. (There are results saying MLE’s are “asymptotically efficient” under certain conditions.)

Example: The laser pole. Here the family consists of the Cauchy densities with median θ allowed to vary. Here’s a little data set I made up.

```
x = c(1, 2, 7.8, 9.2)

# Likelihood function for Cauchy
L <- function(th, x){
  return(prod(dcauchy(x-th)))
}
lik <- thetas
m <- length(thetas)
for(i in 1:m){lik[i] <- L(thetas[i],x)}
plot(thetas,lik,type="l")
```



A bimodal likelihood!

Suggests θ is probably somewhere around 2, or somewhere around 8, with around 2 being the slightly more likely neighborhood for θ .

What is the MLE here? We'll determine which element of our vector thetas gives the highest likelihood. Since these are spaced pretty finely (.01 apart), this is a pretty good approximation to the maximizer over all possible theta values.

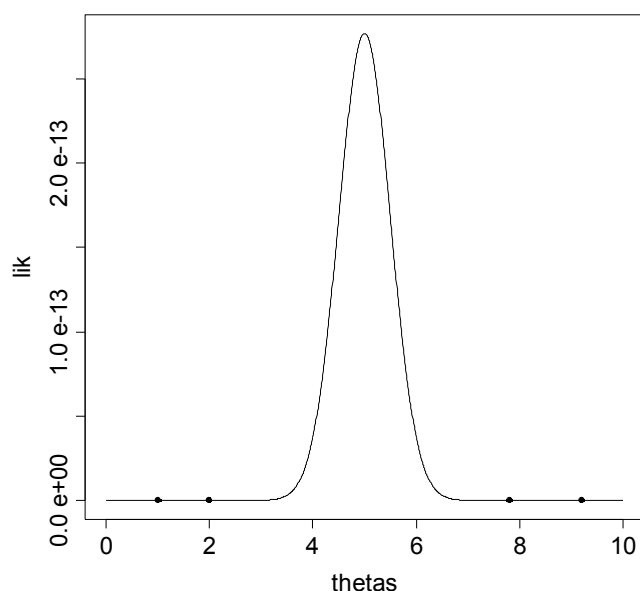
```
> match(max(lik), lik)
[1] 181
> thetas[181]
[1] 1.8
```

The MLE is about 1.8. Clearly this is neither a mean, nor a median. It is what it is.

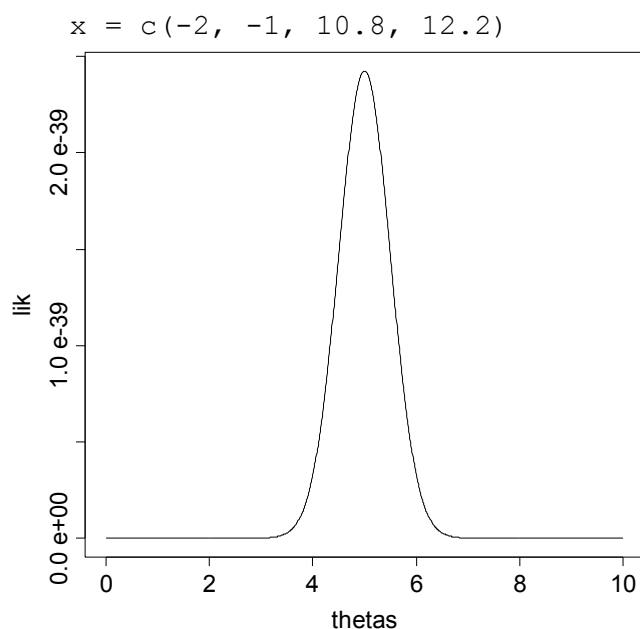
So sometimes (often) a computer and some sort of optimization procedure is required to calculate an MLE.

Note neither the Normal nor the double exponential would have this behavior. For example here is the normal likelihood for the same example $x = (1, 2, 7.8, 9.2)$.

[The little dots show where the data points are.]



The Normal likelihood confidently says that theta is around 5 (the mean, and the MLE). It would say the same thing, for example, if we subtract 3 from the lower two points and add 3 to the upper two points.



This is a bit of a problem. The $N(\theta, 1)$ model does not explain this data well. How would we discover this? It is good to compare different models. For example, the Cauchy model gave *much* higher likelihood to this data than the Normal. We might not know whether a much better model could be waiting to be discovered. But at least we can search for better and better models, and, for the models we have examined, get an idea about their relative plausibilities in explaining a data set.

3.5 Bayesian estimation and conjugate priors

This section will show a few examples of simple Bayesian inference problems that can be solved with paper and pencil, before getting back to our main thread that will use Markov chain Monte Carlo to compute answers in more general and realistic problems.

Example of Bayesian estimation: “Laplace’s rule of succession.”

Pierre Simon Laplace (1774) illustrated his calculation by asking the following question: Given that the sun has risen every day for the past 6000 years, what is the probability that it will rise tomorrow?

This may be a bit of an unfortunate illustration, but it is colorful at least.

The abstract form of this question is the same as what we have been considering: Suppose X_1, \dots, X_n are iid, with a $\text{Bern}(\theta)$ distribution. Suppose all of the n observed X_i ’s so far are 1’s. Estimate θ .

Our answer will depend on the prior distribution for θ . Suppose (as Laplace did) that we feel ignorant about θ and we choose the prior to be $\text{Unif}(0,1)$.

The likelihood is $L(\theta) = \theta^n$, and so the posterior is proportional to θ^n . The proportionality constant is $\int_0^1 \theta^n d\theta = 1/(n+1)$; that is, the posterior density is $(n+1)\theta^n$ for $0 \leq \theta \leq 1$.

If we want to summarize our posterior distribution by a point estimate that represents our “best guess,” we might choose to report the mean of the posterior distribution (that is, the “posterior mean”), which is:

$$\int_0^1 \theta ((n+1)\theta^n) d\theta = (n+1) \int_0^1 \theta^{n+1} d\theta = \frac{n+1}{n+2}$$

By the law of total probability this is also the probability the sun rises tomorrow given the observed X_i ’s.

Laplace would see the sun with probability $\frac{6000 \times 365 + 1}{6000 \times 365 + 2} = 0.99999954$. ►

The idea of a *conjugate prior* is that for many standard, simple statistical models, there is a family of prior distributions such that, if we choose our prior distribution to be a member of the family, then the posterior will also be a (different) member of the same family. In that case, the calculations in going from prior to posterior are typically very simple.

This is the way most Bayesian statistics used to be done. With the Markov chain Monte Carlo techniques we’re studying, we are much less restricted now, but it is still good to be acquainted with the idea of some paper-and-pencil problems. (BUGS also runs somewhat faster if we choose conjugate prior distributions.)

Example: Sampling from a Bernoulli distribution

Suppose X_1, \dots, X_n are iid, with a $\text{Bern}(\theta)$ distribution, that is, $X_i = \begin{cases} 1 & \text{with prob } \theta \\ 0 & \text{with prob } 1 - \theta \end{cases}$. The unknown parameter θ lies between 0 and 1.

As usual, the posterior is proportional to the prior times the likelihood.

The likelihood is $L(\theta) = \theta^S (1 - \theta)^{n-S}$, where $S = \sum_{i=1}^n X_i = \text{number of successes}$.

Suppose we choose our prior to have density that is also proportional to θ to some power, times $(1 - \theta)$ to some power.

There is such a family of distributions: the “Beta(α, β)” density is proportional to $\theta^{\alpha-1} (1 - \theta)^{\beta-1}$.

So, if we choose our prior distribution on θ to be Beta(α, β), then the posterior is proportional to

$$\left[\theta^{\alpha-1} (1-\theta)^{\beta-1} \right] \left[\theta^S (1-\theta)^{n-S} \right] = \theta^{\alpha+S-1} (1-\theta)^{\beta+n-S-1}.$$

That is, the posterior density is $\text{Beta}(\alpha + S, \beta + n - S)$. ▶

More about Beta distributions

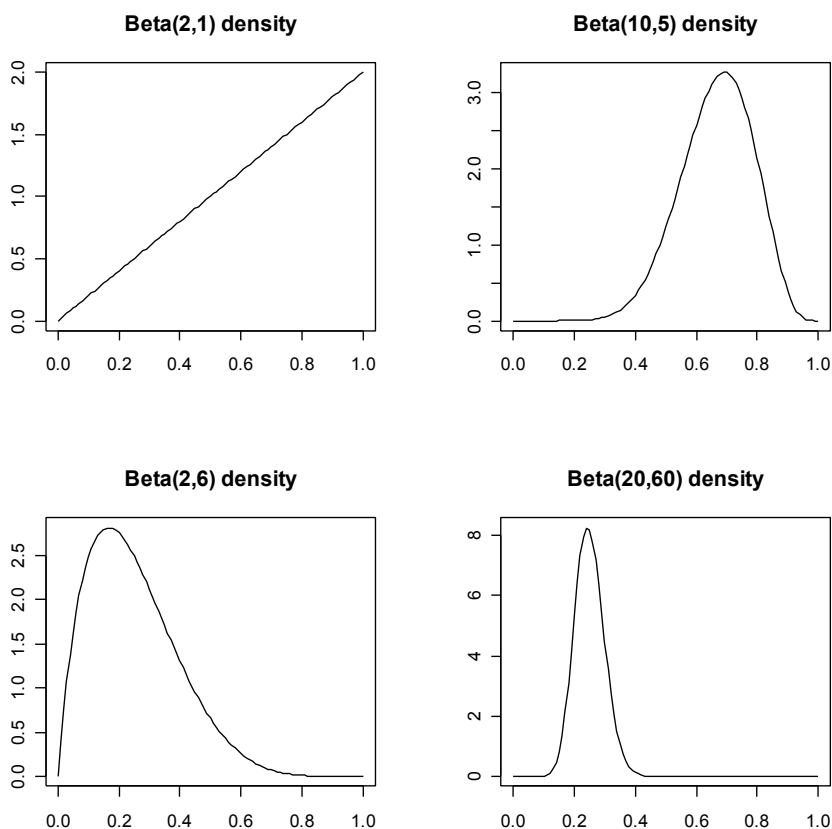
The normalizing constant (which is not needed in order to see that they Beta family is conjugate, but is needed in some other calculations):

$$\int_0^1 \theta^{\alpha-1} (1-\theta)^{\beta-1} d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \text{ where } \Gamma \text{ is the gamma function, defined by } \Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

[[If you want to see this worked out—good clean calculus fun—see the “appendix” below.]]

That is, the $\text{Beta}(\alpha, \beta)$ density is $f_{\alpha, \beta}(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$ for $0 < \theta < 1$.

The $\text{Beta}(1,1)$ density is the same as Uniform on the interval $(0,1)$. Here are some others:



Another pleasant calculation [appendix again] shows that the $\text{Beta}(\alpha, \beta)$ distribution has mean $\frac{\alpha}{\alpha + \beta}$, and variance $\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$.

E.g. the two densities on the top of the previous picture have mean $2/3$, and the two densities on the bottom of the previous picture have mean $1/4$.

Example: The "Laplace rule of succession" is a special case of our Beta-conjugate-prior story: We chose the uniform prior, which is $\text{Beta}(1,1)$. Then we observed n "successes" and 0 "failures". So our updating formula says the posterior distribution of θ is $\text{Beta}(1+n, 1+0)$, that is, $\text{Beta}(\alpha, \beta)$ with $\alpha = n+1$ and $\beta = 1$.

So the posterior mean is $\frac{\alpha}{\alpha + \beta} = \frac{n+1}{n+2}$. ►

Interpretation of posterior mean as a weighted average of information from prior and from data

In general if our prior is $\text{Beta}(\alpha, \beta)$ and we then observe n trials, out of which k are successes, our posterior distribution is $\text{Beta}(\alpha + k, \beta + n - k)$, and our posterior mean for θ is $\frac{\alpha + k}{\alpha + \beta + n}$.

This can be thought of as:

$$\frac{\alpha + k}{\alpha + \beta + n} = \frac{\alpha + \beta}{\alpha + \beta + n} \underbrace{\left(\frac{\alpha}{\alpha + \beta} \right)}_{\text{prior mean}} + \frac{n}{\alpha + \beta + n} \underbrace{\left(\frac{k}{n} \right)}_{\substack{\text{MLE} \\ \text{from} \\ \text{data}}}$$

That is, our posterior mean is a weighted average of our prior mean and the data mean.

If α and β are small and n is large, our prior mean gets little weight.

Let's discuss another example of the conjugate prior setup, more briefly.

Example: margin of error in a poll (as discussed in a homework problem...)

We take a random sample of size $n = 100$ and find that the number of "successes" is $X = 55$. If our prior distribution for θ is $U(0,1)$, then the posterior distribution for θ is $\text{Beta}(55+1, 45+1) = \text{Beta}(56, 46)$. We can use this posterior to answer our questions of interest. For example, a 95% probability interval for θ can be obtained by chopping off 2.5% of the probability from the left tail and 2.5% from the right tail, that is, finding the 2.5 and 97.5 percentiles of the $\text{Beta}(56, 46)$ distribution. Here is what R tells us:

```
> qbeta(.025, 56, 46)
[1] 0.4522192
> qbeta(.975, 56, 46)
[1] 0.6439984
```

That is, our 95% posterior probability interval for θ is (.452, .644). This agrees with what we got by "brute force" calculations on the homework problem earlier in the semester.

Example: Normal distribution

Suppose X_1, \dots, X_n are iid, with a $N(\theta, \sigma^2)$ distribution. Let us assume that σ^2 is known, and the unknown parameter of interest is the mean, θ .

In this case the conjugate prior is also the Normal distribution: assume $\theta \sim N(\mu_0, \sigma_0^2)$.

Let \bar{X}_n denote the sample mean – the mean of the observed data.

It is easy to see that the the posterior distribution for θ is again Normal, without needing detailed calculations, because

$$P(\theta | X_1, \dots, X_n) \propto \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[\frac{-1}{2\sigma_0^2}(\theta - \mu_0)^2\right] \prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-1}{2\sigma^2}(X_i - \theta)^2\right]\right) \\ \propto \exp[\text{quadratic function of } \theta].$$

[[We think of this as a function of θ , with μ_0 and σ_0^2 chosen by us, σ^2 known, and the data X_1, \dots, X_n fixed, observed numbers.]]

Since the posterior is Normal, we now just want to know the mean and variance. Some straightforward but a bit time-consuming algebra [complete a square in the exponent...] shows :

$$\text{posterior mean} = \frac{(1/\sigma_0^2)}{(1/\sigma_0^2) + (n/\sigma^2)} \mu_0 + \frac{(n/\sigma^2)}{(1/\sigma_0^2) + (n/\sigma^2)} \bar{X}_n, \\ \text{posterior variance} = \frac{1}{(1/\sigma_0^2) + (n/\sigma^2)}.$$

This is summarized more concisely and more memorably in terms of the concept of **precision**, simply defined to be the reciprocal of variance.

The prior precision is $\text{prec}_{\text{prior}} = 1/\sigma_0^2$.

The variance of \bar{X}_n is σ^2/n , so its precision is $\text{prec}_{\bar{X}_n} = n/\sigma^2$.

So the posterior mean is:

$$\frac{\text{prec}_{\text{prior}}}{\text{prec}_{\text{prior}} + \text{prec}_{\bar{X}_n}} \mu_0 + \frac{\text{prec}_{\bar{X}_n}}{\text{prec}_{\text{prior}} + \text{prec}_{\bar{X}_n}} \bar{X}_n.$$

That is, the posterior mean is a weighted average of prior mean and sample mean, where the relative weights are the corresponding precisions.

The posterior precision is:

$$\frac{1}{\text{posterior variance}} = (1/\sigma_0^2) + (n/\sigma^2) = \text{prec}_{\text{prior}} + \text{prec}_{\bar{X}_n}$$

That is, the posterior precision is simply the sum: prior precision + precision of the sample mean. ►

Example: As part of a small pilot study to evaluate the effectiveness of an experimental drug to lower blood pressure, the blood pressures of 12 patients with mild hypertension were measured before and after taking the drug. For $i = 1, 2, \dots, 12$, the difference $X_i = (\text{BP before drug}) - (\text{BP after drug})$ was recorded, and the average $(X_1 + \dots + X_{12})/12$ turned out to be $\bar{X} = 7.4$ mmHg (millimeters of mercury). Suppose it is assumed, based on past experience, that such differences are Normally distributed and have standard deviation 5.5 mmHg. So your model is that X_1, \dots, X_{12} are iid with distribution $N(\theta, 5.5^2)$. Suppose your prior distribution for θ is Normal with mean 0 and standard deviation 10.

(a) What is your posterior distribution for θ ?

- (b) The drug is considered worthwhile only if it produces a mean reduction θ of at least 5 mmHg. What is the posterior probability that the drug is worthwhile?

Solution:

(a) The prior mean is 0, prior precision is $1/10^2 = 1/100 = .01$.

The sample mean is 7.4, and its precision is $1/\text{var}(\bar{X}) = 1/(5.5^2/12) = 12/(5.5^2) = 0.3967$.

So the posterior distribution for θ has mean $\frac{(.01)(0) + (.3967)(7.4)}{.01 + .3967} = 7.218$,

and precision $.01 + .3967 = .4067$. [So the SD of the posterior is $1/\sqrt{.4067} = 1.568$.]

(b) The posterior probability that θ is at least 5 is the probability that a $N(7.218, 1.568^2)$ random variable is at least 5, which is

$$P\{N(7.218, 1.568^2) \geq 5\} = P\{N(0,1) \geq \frac{5 - 7.218}{1.568}\} = P\{N(0,1) \geq -1.4145\} \approx 0.92.$$

Appendix: Some calculations deferred from above

→ Gamma function is defined this way: $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$.

Q: For what α does this definition work? A: $\alpha > 0$.

Integration by parts gives this very useful relationship: $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ for $\alpha > 1$.

Like factorial. Since $\Gamma(1) = 1$, for $\alpha = n$ integer, get $\Gamma(n) = (n - 1)(n - 2) \cdots (1)\Gamma(1) = (n - 1)!$

→ Gamma(α, β) distribution has density $\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ for $x > 0$. Note β is just a scale parameter

(α is the “shape parameter”) so it’s often no loss to assume $\beta = 1$ in derivations, etc.

Gamma(α) distribution [with $\beta = 1$]: density function $\frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}$ for $x > 0$.

E.g. Gamma(1) is exponential, Gamma(k) is sum of k independent exponentials.

If $X_1 \sim \Gamma(\alpha_1)$ and $X_2 \sim \Gamma(\alpha_2)$ are independent, then $X_1 + X_2 \sim \Gamma(\alpha_1 + \alpha_2)$.

→ Beta distribution with parameters α and β has density proportional to $x^{\alpha-1}(1-x)^{\beta-1}$ on the interval $0 < x < 1$. Let’s find the normalizing constant.

Claim: $\int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}.$

Derivation: $\Gamma(\alpha)\Gamma(\beta) = \int_0^\infty x^{\alpha-1} e^{-x} dx \int_0^\infty y^{\beta-1} e^{-y} dy = \int_0^\infty \int_0^\infty x^{\alpha-1} y^{\beta-1} e^{-(x+y)} dx dy.$

Now let’s change variables, defining $u = \frac{x}{x+y}$ and $v = x+y$, or, in other words, $x = uv$ and

$y = (1-u)v$. The Jacobian needed to do this change of variables is

$$\left| \frac{dx}{du} \frac{dy}{dv} \right| = \left| \frac{\partial(x,y)}{\partial(u,v)} \right| = \left| \frac{\partial x}{\partial u} \frac{\partial y}{\partial v} \right| = \left| \begin{vmatrix} v & u \\ -v & 1-u \end{vmatrix} \right| = |v(1-u) + uv| = v, \text{ and we get that the last double}$$

integral is equal to

$$\begin{aligned} \int_0^1 \int_0^\infty (uv)^{\alpha-1} ((1-u)v)^{\beta-1} e^{-v} v dv du &= \int_0^1 u^{\alpha-1} (1-u)^{\beta-1} \underbrace{\int_0^\infty v^{\alpha+\beta-1} e^{-v} dv}_{\Gamma(\alpha+\beta)} du \\ &= \Gamma(\alpha+\beta) \int_0^1 u^{\alpha-1} (1-u)^{\beta-1} du \end{aligned}$$

So $\Gamma(\alpha)\Gamma(\beta) = \Gamma(\alpha+\beta) \int_0^1 u^{\alpha-1} (1-u)^{\beta-1} du$, which is what we wanted to show.

→ Mean and variance of the Beta distribution

Suppose $X \sim \text{Beta}(\alpha, \beta)$. Then

$$\begin{aligned} E(X) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x \cdot x^{\alpha-1} (1-x)^{\beta-1} dx = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^\alpha (1-x)^{\beta-1} dx \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+1+\beta)} \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} = \frac{\Gamma(\alpha+\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)} \frac{\alpha\Gamma(\alpha)}{\Gamma(\alpha)} = \frac{\alpha}{\alpha+\beta}, \end{aligned}$$

where we have used the relationship $\Gamma(c+1) = c\Gamma(c)$ twice – once for the choice $c = \alpha + \beta$ and once for the choice $c = \alpha$.

In a similar way, we can derive $E(X^2) = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}$ and get the variance

$$\text{var}(X) = E(X^2) - (EX)^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \text{ claimed above.}$$

→ Normal problem discussed in notes. Let $\tau = \sigma^{-2}$ denote the precision of the X distribution.

$X = (X_1, \dots, X_n) \sim N(\theta, \tau^{-1})$. Here we are assuming τ is known and θ is not; our goal is a posterior distribution for θ . We take θ to have a Normal prior, which is conjugate in this situation; let's say

$\theta \sim N(\theta_0, \tau_0^{-1})$. That is, $\tau_0 = \frac{1}{\sigma_0^2}$ is the prior precision.

$$\begin{aligned} p(x | \theta) &\propto \prod_{i=1}^n \left(\tau^{1/2} \exp\left(-\frac{1}{2}\tau(x_i - \theta)^2\right) \right) = \tau^{n/2} \exp\left(-\frac{\tau}{2} \left(\sum (x_i - \theta)^2\right)\right) \\ &= \tau^{n/2} \exp\left(-\frac{\tau}{2} \sum x_i^2 + \theta\tau n\bar{x} - \frac{\tau}{2} n\theta^2\right) \end{aligned}$$

So since τ is known, $p(x | \theta) \propto \exp\left(-\frac{n\tau}{2}(\theta^2 - 2\bar{x}\theta)\right)$.

Next we multiply this likelihood by the prior density, which is proportional to

$$\exp\left(-\frac{\tau_0}{2}(\theta - \theta_0)^2\right) \sim \exp\left(-\frac{\tau_0}{2}(\theta^2 - 2\theta_0\theta)\right), \text{ getting.}$$

$$\begin{aligned}
\text{Posterior} &\propto \exp\left(\frac{-1}{2}(\tau_0 + n\tau)\theta^2 + (\tau_0\theta_0 + n\tau\bar{x})\theta\right) \\
&= \exp\left(\frac{-1}{2}(\tau_0 + n\tau)\left\{\theta^2 - 2\frac{\tau_0\theta_0 + n\tau\bar{x}}{\tau_0 + n\tau}\theta\right\}\right) \\
&\propto \exp\left(\frac{-1}{2}(\tau_0 + n\tau)\left\{\theta - \frac{\tau_0\theta_0 + n\tau\bar{x}}{\tau_0 + n\tau}\right\}^2\right)
\end{aligned}$$

So the Posterior is Normal with mean $\frac{\tau_0\theta_0 + n\tau\bar{x}}{\tau_0 + n\tau}$ and precision $\tau_0 + n\tau$.

This is nice and easy to interpret:

- Prior has mean θ_0 and precision τ_0 .
- Data has mean \bar{x} , which has precision $n\tau$.
- Precision of posterior is the sum: precision of prior + precision of estimate from data.
- Mean of posterior is a weighted average of prior mean and data mean, where the weights are the corresponding precisions.

→ Just for completeness, let's consider the Normal distribution in the situation where θ is known, and we are interested in τ (i.e. σ^{-2}). Now $p(x | \tau) \propto \tau^{n/2} \exp\left\{\left(-\frac{1}{2}\sum x_i^2 + \theta n\bar{x} - \frac{1}{2}n\theta^2\right)\tau\right\}$.

E.g. say for simplicity (and without loss of generality), let's suppose θ is known to be 0, so that

$$p(x | \tau) \propto \tau^{n/2} \exp\left\{-\frac{1}{2}\sum x_i^2 \tau\right\}.$$

Say prior for τ is Gamma(α, β), having density $\frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau} \propto \tau^{\alpha-1} e^{-\beta\tau}$.

Then the posterior is proportional to

$$p(\tau | x) \propto \tau^{\alpha+n/2-1} \exp\left\{-\left(\beta + \frac{1}{2}\sum x_i^2\right)\tau\right\} \sim \text{Gamma}\left(\alpha + n/2, \beta + \frac{1}{2}\sum x_i^2\right).$$

E.g. the posterior mean for τ is $\frac{\alpha + \frac{n}{2}}{\beta + \frac{1}{2}\sum x_i^2}$.

Note if $\alpha = \beta = 0$, this corresponds to the estimate $\hat{\tau} = \frac{n}{\sum x_i^2}$, which is the usual MLE.

[$\hat{\sigma}^2 = \frac{\sum x_i^2}{n}$ is the usual MLE, unbiased also.] But the Gamma(0,0) density is proportional to τ^{-1} ,

which is improper (the integral is infinity, so it cannot be normalized to be a proper density, having integral 1). This is what motivates common choices like Gamma(.001,.001).

4 Bayesian data analysis and statistical inference using MCMC.

4.1 Subliminal math improvement: an example "from scratch" using random walk Metropolis

Here is a small but rather realistic example of a question of statistical inference. I got it from page 400 of David Moore's *The Basic Practice of Statistics*.

[[Apparently another study by the same authors: Hudesman, John, Warren Page, and Jussi Rautiainen. Use of Subliminal Stimulation to Enhance Learning Mathematics. *Perceptual and Motor Skills*. June, 1992: 1219-1224.]]

	Group(Treat=1,Control=2)	Pre-test	Post-test
18 students at CUNY who had failed a mathematics skills assessment test took a summer program designed to improve their skills. They were randomized into a "treatment" group (10 students) and a "control" group (8 students).	1	18	24
	1	18	25
	1	21	33
	1	18	29
	1	18	33
Each of the 18 students was exposed to a daily subliminal message flashed quickly on a screen, too quickly to be read, at least consciously.	1	20	36
	1	23	34
	1	23	36
	1	21	34
The message for the treatment group was: "Each day I am getting better in math."	1	17	27
	2	18	29
	2	24	29
The message for the control group was: "People are walking on the street."	2	20	24
	2	18	26
	2	24	38
	2	22	27
Each student took a Pre-test before the summer program and a Post-test, after. The data are shown at right.	2	15	22
	2	19	31

Our job is to speculate on whether the subliminal message had an effect on the test scores.

We will look at each person's improvement, that is the difference $\text{PostTest} - \text{PreTest}$. We imagine that we are seeing a sample from each of two population distributions of improvements – a treatment population and a control population. We formulate our questions as:

Is the treatment population mean greater than the control population mean? If so, how much?

Well, we don't know and we won't know for sure, but we can try to evaluate a probability that the treatment mean is greater than the control mean, and find a probability distribution for the difference between the two means. These will be posterior probabilities: assuming a prior distribution and a model, they will be probabilities conditional on the observed data.

→ First let's get the data into R and draw a relevant picture of the data... maybe that will make the answer obvious...

```
dat = read.csv("data/SubliminalMathImprovement-BPS-page400.csv")
improvement = dat[,3]-dat[,2]
trt = improvement[dat[,1]==1]
ctrl = improvement[dat[,1]==2]

### draw a picture
xlim = c(min(improvement),max(improvement))
par(mfrow=c(2,1))
hist(trt,100,col="red",xlim=xlim)
```

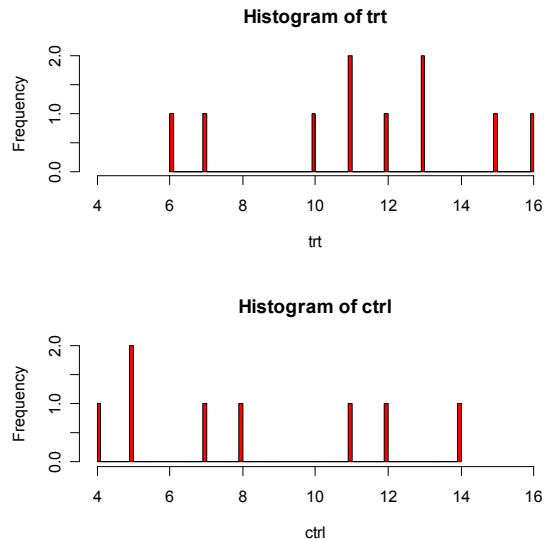
```
hist(ctrl,100,col="red",xlim=xlim)
```

Our picture is shown at the right.

OK, so it's not so obvious... On the average the treatment improvements do look a bit larger at least

It's time to create a statistical model that will allow us to formulate our question of interest as a question about a probability that some unknown values of some parameters lie in some set. Then we'll use MCMC to approximate this probability by Monte Carlo simulation.

We'll use a very common type of model: we imagine two populations, each with its own Normal distribution for the `improvement` variable.



Let's say for the treatment group, the improvements `[[in the vector trt]]` come from $N(\mu_1, \sigma_1^2)$, and for the control group, the improvements `[[in the vector ctrl]]` come from $N(\mu_2, \sigma_2^2)$.

So our parameter vector is $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2)$. That is, θ is a 4-dimensional vector.

Here is an outline of our tasks:

- Write a likelihood function `[[lik = function(th){...}]]`
- Write a prior `[[prior = function(th){...}]]`
- Multiply to get posterior `[[post = function(th){prior(th) * lik(th)}]]`
(OK, that last one won't much of a task.)
- Do Metropolis: Propose random changes to `th` and accept or reject them in a Metropolisisey way

→ Let's write a likelihood function. The likelihood is the probability (probability density here since we are dealing with continuous distributions) of the observed data, as a function of the unknown parameters.

```
lik = function(th){
  mu1=th[1]; sig1=th[2]; mu2=th[3]; sig2=th[4]
  prod(dnorm(trt,mean=mu1,sd=sig1))*prod(dnorm(ctrl,mean=mu2,sd=sig2))
}
```

```
# E.g.:
lik(c(10,3,10,3))
lik(c(11,3,9,3))
```

→ Next task: decide on a prior density for $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2)$, and write it as a function in R.

My first thought, pretty much pulled out of the air, was a prior like this:

- Distributions of μ_1 and μ_2 are both $N(10, 10^2)$.
- Distributions of σ_1 and σ_2 are both Exponential with mean 10 (rate = 0.1).
- All 4 variables are independent `[[nice and simple – just multiply densities to get joint density]]`

My basic thinking was simply to choose priors for the μ_i 's that did not seem to prejudge which is bigger, and to choose priors that are well "smeared out" and sort of uniformish over all even remotely plausible

values, with the aim of letting the data `[[the likelihood]]` produce any interesting features in the posterior distribution, rather than introducing them ourselves in the prior.

If we have doubts later or want to try with other prior distributions, just come back and change this function!

```
prior = function(th){
  mu1=th[1]; sig1=th[2]; mu2=th[3]; sig2=th[4]
  if(sig1 <= 0 | sig2 <= 0) return(0)
  dnorm(mu1,10,10)*dnorm(mu2,10,10)*dexp(sig1,rate=.1)*dexp(sig2,rate=.1)
}
```

→ Next, we multiply prior times likelihood to get the posterior.

```
post = function(th){prior(th) * lik(th)}
```

Well, that wasn't so hard.

As usual we realize that Bayes' rule says $\text{posterior}(\theta) \propto \text{prior}(\theta) \times \text{likelihood}(\theta)$ and we ignore the proportionality constant because our Metropolis MCMC method does not need it!

Now we can simply write `post(th)` to have R evaluate the posterior density at `th`.

We want to simulate a sample from the density `post` using Metropolis.

We'll run a Markov chain, say for `nit` iterations.

Choose a starting value for `th=c(mu1, sig1, mu2, sig2)`; call it `th0`.

Given a current state `th`, decide on a way to propose a "candidate" move, say to `cand`.

Evaluate `post(th)` and `post(cand)` and take the ratio `post(cand)/post(th)`.

We'll record all our results in a big matrix called `results`, of dimensions `nit` by 4.

The first row of `results` will be the `th0` vector, and each successive row of `results` will record the next `th` vector as we run the chain.

```
#Starting values
mu1 = 10; sig1 = 10; mu2 = 10; sig2 = 10
th0=c(mu1,sig1,mu2,sig2)

# Here is what does the MCMC (Metropolis method):
nit=10000
results = matrix(0, nrow=nit, ncol=4)
th = th0
results[1,] = th0
for(it in 2:nit){
  cand = th + rnorm(4,sd=.5)
  ratio = post(cand)/post(th)
  if(runif(1)<ratio) th=cand
  results[it,] = th
}

# Take a peek at what we got
edit(results)
```

OK, now we've got a bunch of parameter vectors, and we can use these to answer questions.

E.g. we can look at distributions using histograms, or calculate fractions of iterations that some statement about the parameters was true to approximate our posterior probability that the statement is true.

```
mu1s = results[,1]
sig1s = results[,2]
```

```

mu2s = results[,3]
sig2s = results[,4]

plot(mu1s)
plot(sig1s)
plot(mu1s-mu2s)

hist(mu1s-mu2s)
mean(mu1s-mu2s > 0)  # <-- our original question

```

So we've got an answer to our original question.

If we want to be a bit more careful we could take a look at how the chain ran and see if our starting values seemed to skew the early iterations of the chain noticeably. We could throw out some of the initial iterations as a "burn-in period."

```

# Look at bit closer at how the chain ran, and maybe throw out a "burn-in
period"
resultsPlot = function(results){
  old.par = par(no.readonly = TRUE)
  on.exit(par(old.par))
  nvar = dim(results)[2]
  par(mfrow = c(nvar,1))
  for(i in 1:nvar)plot(results[,i], ylab = paste("variable",i))
}

resultsPlot(results)
resultsPlot(results[1:2000,])

```

E.g. suppose we look at plots of the iterations of the various parameters and it looks as if the first few hundred iterations may be noticeably influenced by our starting values. Then we might throw away the first 500 iterations, like this:

```
res = results[500:nit,]
```

Then we could base our inferences on `res` instead of the whole `results` matrix.

```

mu1s = res[,1]
sig1s = res[,2]
mu2s = res[,3]
sig2s = res[,4]

hist(mu1s-mu2s)
mean(mu1s-mu2s > 0)

```

After doing all that, here are some highlights of what we got.

We got a big matrix $10,001 \times 4$ matrix of results, called `results`. Successive rows of the `results` matrix are successive states of the parameter vector $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2)$ as the Markov chain ran.

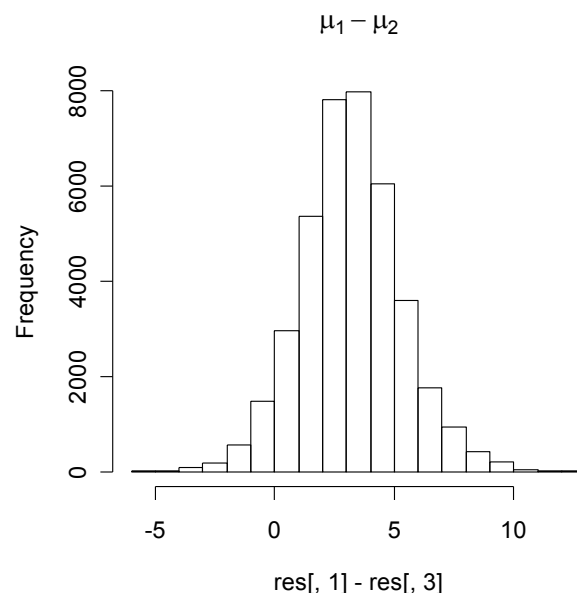
Here are the first 10 thetas, starting from the arbitrarily chosen starting point of $\theta = (10, 10, 10, 10)$:

```

> results[1:10,]
      [, 1]      [, 2]      [, 3]      [, 4]

```

Stat 238 notes, 9/2/09




```

[1,] 10.000000 10.000000 10.000000 10.000000
[2,] 10.000000 10.000000 10.000000 10.000000
[3,]  9.407660  9.256550 10.070058  9.452185
[4,]  9.664870  8.651573 10.234950 10.079506
[5,]  9.487590  9.273098 10.381989  9.328918
[6,]  9.487590  9.273098 10.381989  9.328918
[7,]  9.487590  9.273098 10.381989  9.328918
[8,]  9.132188  9.789791  9.945703  8.481567
[9,] 10.046112  9.955818  9.299973  8.849923
[10,]  9.872070 10.336362  9.422855  8.618168

```

And here are the last 10 states in this run of length 10001

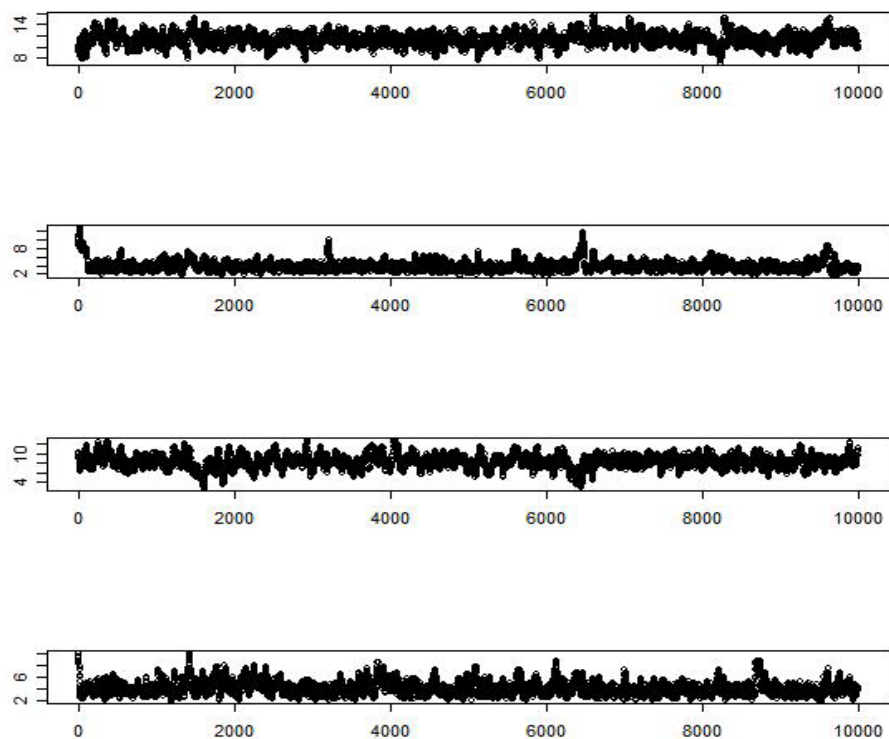
```

> results[9992:10001,]
      [,1]      [,2]      [,3]      [,4]
[1,] 10.483375  3.528636 10.113042  3.435419
[2,]  9.734556  3.795559  9.728825  3.711211
[3,] 10.246357  3.936654  9.609206  4.154013
[4,] 10.246357  3.936654  9.609206  4.154013
[5,] 10.246357  3.936654  9.609206  4.154013
[6,] 10.246357  3.936654  9.609206  4.154013
[7,]  9.545200  3.478542 10.457745  4.332268
[8,]  9.545200  3.478542 10.457745  4.332268
[9,]  9.958325  3.460580 10.465343  3.942501
[10,] 10.897010  2.960106 11.070092  4.116067

```

We can see that we are estimating the means (columns 1 and 3) to be somewhere around 10 and the standard deviations to be somewhere around 3 or 4 or so.

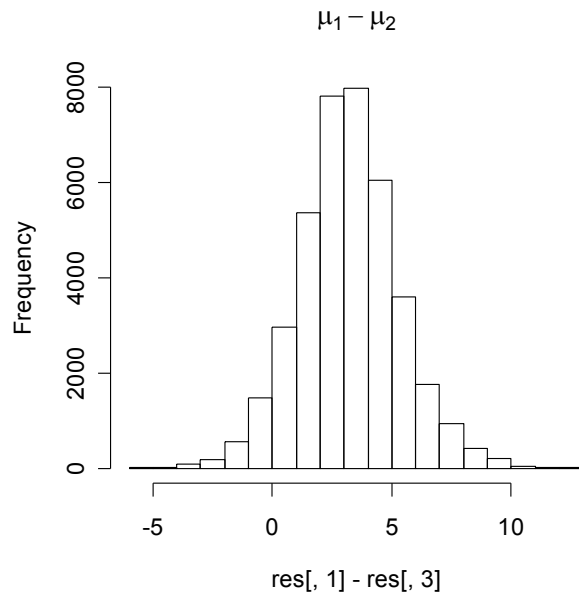
Here are some trace plots of the four parameters; the first plot is for μ_1 , then σ_1 , then μ_2 , then σ_2 :



We can see that there is a short initial "transient" that takes place because the chain starts far from the stationary distribution (see in particular the second and the fourth graphs, of the sigma parameters, which started at the high value of 10). So we might want to throw away some of the initial states of the chain, and consider those "burn-in". Here we'll throw away the first 500 states, calling what remains `res`.

```
res = results[500:nit,]
```

A picture for our posterior distribution for the difference $\mu_1 - \mu_2$, given by the command `hist(res[,1]-res[,3])`, looks like this:



What is our estimated posterior probability that $\mu_1 - \mu_2 > 0$?

```
> sum((res[,1]-res[,3] > 0))/dim(res)[1]
[1] 0.9408116
```

About 94%.

What if we want to change the prior?

E.g., could try the whole thing again with (improper) uniform priors on everything, simply by defining the prior density to be 1:

```
prior = function(th){
  mu1=th[1]; sig1=th[2]; mu2=th[3]; sig2=th[4]
  if(sig1 <= 0 | sig2 <= 0) return(0)
  return(1)
}
```

Well, that's a bit of a complicated way to write "1" but we should return the answer "0" if either sig1 or sig2 is negative.

4.2 BUGS.

BUGS, which stands for "Bayesian inference Using Gibbs Sampling," is a piece of software that makes running MCMC for Bayesian analysis easy.

...Say what the Gibbs sampler is...

The main message of this section is: Go get the BUGS software and start using it! If you want to be able to use it on your own computer, I will try to help you do that here. I can be most helpful for Windows users, and BUGS works best in Windows. I think it works fine on Macintosh or Unix systems with a Windows emulator too. And BUGS has Unix and Linux versions too, but I haven't got personal experience with those. Macintosh users: I don't think there is a "native" version for the Macintosh operating system; you need to be essentially running Windows on your Mac to use it. If you have trouble or don't have your own computer or whatever, you can go to the Statlab (at Rm 101 of Urban Hall, 140 Prospect St.) and use BUGS (and R) there.

[[I should also mention there is another very similar program you could try out if you'd like, called JAGS (for "Just Another Gibbs Sampler"), described at <http://www-fis.iarc.fr/~martyn/software/jags/>. I have heard Macintosh users (including students here) say they have been able to get it to work. There is a manual here: http://www.stat.yale.edu/~jtc5/238_2006/JAGS-manual.pdf]]

Although there are several ways to run BUGS and even several versions of the software itself, I want to go through one particular path to using BUGS, which works relatively well and I would recommend to you.

The easiest way for Windows users to get BUGS is to download the BRugs library (or "package"). This will install BUGS on your computer and also an R library that enables you to run BUGS from R, which is very convenient. Downloading the package is very easy; the download takes a bit longer than most other R packages, since the whole BUGS program is being downloaded, but it is the same simple procedure as for other packages. To review this, make sure you are connected to the web, and start R. Go on the packages menu, and then "Install package(s)..." Choose a CRAN mirror (such as "USA (PA 1)" which is in Pennsylvania, or whatever; it doesn't matter). After a few seconds you will see a long list of packages that you can download, and you just scroll down and select "BRugs," click "OK," and wait for the package to be downloaded and installed automatically!

To have easy access to a point-and-click version of BUGS after downloading the package, you may wish to find the program and put a shortcut on your desktop. You should be able to find a "winbugs.exe" program in a rather standard place – on my computer it is at

C:\Program Files\R\R-2.2.1\library\BRugs\OpenBUGS\winbugs.exe

Yours may differ in the version number of R. To put a shortcut on your desktop, navigate to that folder and right-click on the winbugs.exe file, and choose "Send to" and then "Desktop (create shortcut)."

There are several ways to use BUGS. You can start up the program by double-clicking on the shortcut you just made, and then do a lot of pointing and clicking – I may show you how that works in class.

For most purposes, I would recommend running BUGS from R as we'll see how to do in class.

Running BUGS from R.

Although it is actually helpful to go through the pointing and clicking once or twice, it is most convenient in the long run to run the software from R. The BRugs library helps make this work.

[[So again there seem to be several alternatives: bugs.R, rbugs, BRugs, and others, I think.]]

Here is a BUGS "model" file, which we might call `subliminal.bug` :

```
model{
  for(i in 1:n1){
    trt[i] ~ dnorm(mu1, tau1)
  }
}
```

```

for(i in 1:n2){
  ctrl[i] ~ dnorm(mu2, tau2)
}
mu1 ~ dnorm(10, .01)
mu2 ~ dnorm(10, .01)
tau1 ~ dunif(0.0,100)
tau2 ~ dunif(0.0,100)
sig1 <- 1.0/sqrt(tau1)
sig2 <- 1.0/sqrt(tau2)
}

```

And here is a way to use the model with BUGS, all from R.

```

dat = read.csv("data/SubliminalMathImprovement-BPS-page400.csv")
improvement = dat[,3]-dat[,2]
n1=10
n2=8
trt = improvement[1:n1]
ctrl = improvement[n1 + (1:n2)]

library(BRugs)  # ← Load the BRugs library!

# The next commands are like the pointing and clicking in WinBUGS:
modelCheck("subliminal.bug")
bugsData(c("trt","ctrl","n1","n2"))
modelData()
modelCompile()
inits = list(mu1 = 10, mu2 = 10, tau1 = .01, tau2 = .01)
bugsInits(list(inits), file="inits.txt")
modelInits(file="inits.txt")
# modelGenInits()
param = c("mu1","sig1","mu2","sig2")
samplesSet(param)
modelUpdate(10000)

# How do we get MCMC results defined in R so we can work with them?
# To get the next function you can do:
source("http://www.stat.yale.edu/~jtc5/238_2006/mybrugs.r")
r = getstuff()
edit(r)

x11()
par(mfrow=c(4,1))
for(j in 1:4){
  plot(r[[j]], ylab=names(r)[j])
}

for(j in 1:4){
  hist(r[[j]], n=100, col="red", main=names(r)[j])
}

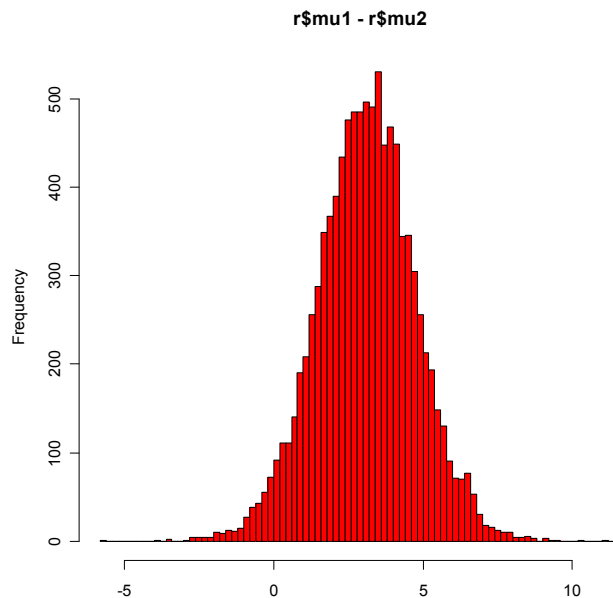
# Also in http://www.stat.yale.edu/~jtc5/238_2006/mybrugs.r
# I defined a "mybrugs" function that makes the above stuff about as easy
# as it can get - really just 4 steps to run BUGS:

dat = c("trt","ctrl","n1","n2")
inits = list(mu1 = 10, mu2 = 10, tau1 = .01, tau2 = .01)
param = c("mu1","sig1","mu2","sig2")
r = mybrugs("subliminal.bug", dat, inits, param, 10000)

```

```
# And now we can look at results, etc.:
```

```
myhist(r$mu1 - r$mu2)
```



```
> mean(r$mu1 > r$mu2)
[1] 0.9682968
```

Our posterior prob that $\mu_1 > \mu_2$ is about 96.8%.

4.3 Notes on BUGS and on running BUGS from R.

In more detail, here's how to set things up and run in R, following along with the “subliminal” example:

STEP 1: To start, we need to have the data in R. Here is how we did that in our example:

```
x = read.csv("SubliminalMathImprovement-BPS-page400.csv")
improvement = dat[,3]-dat[,2]
n1=10
n2=8
trt = improvement[1:n1]
ctrl = improvement[n1 + (1:n2)]
```

STEP 2: We need to make a "model" file for BUGS to use. This must simply be an ordinary text file. For example, using a text editor such as notepad is fine. [[E.g. if you use Microsoft Word, you need to save as text and not as a “.doc” file.]] We need to save it and remember the name we used and where it was saved. Sometimes it is most convenient to save it into R's current working directory, since then we can refer to the file just by its name without a path. The file I used (shown at right) is saved under the name "subliminal.bug" in my working directory (so R can find it just by the name... or, you can put the file in any folder as long as you give R a full path to the file).

→ Some notes about model files:

The BUGS "model" file specifies the joint distribution of a bunch of random variables. The distribution of a variable is allowed to depend on the values of other variables – these are the conditional distributions.

The model file consists of a collection of statements the following two types:

- `variable ~ distribution(parameter1,parameter2,...)`
[these parameters (↑) can involve other variables]
- `variable <- (expression involving other variables)`

The “variables” are the random variables in the model.

The first type of statement, with the “~”, specifies the probability distribution of a variable. BUGS refers to the variable on the left side of this type of statement as a “stochastic node.” The second type of statement, with the “<-”, is for deterministic relationships – when the value of one variable is determined by others. BUGS refers to the variable on the left side of this type of statement as a “logical node.”

We are allowed to use “loops” to write many such statements at once. You should not think of these loops as executing commands, but rather as writing out statements of the above two forms.

```
model{
  for(i in 1:n1){
    trt[i] ~ dnorm(mu1, tau1)
  }
  for(i in 1:n2){
    ctrl[i] ~ dnorm(mu2, tau2)
  }
  mu1 ~ dnorm(10,.01)
  mu2 ~ dnorm(10,.01)
  tau1 ~ dunif(0.0,100)
  tau2 ~ dunif(0.0,100)
  sig1 <- 1.0/sqrt(tau1)
  sig2 <- 1.0/sqrt(tau2)
}
```

The result of a BUGS “program” is simply a collection of such statements that collectively give the joint probability distribution of all of the random variables in the model, including the “prior” distributions of the parameters (which, as you know, in the Bayesian framework are also treated as random variables).

It is possible to get quite confused if you fall into the tempting trap of thinking of each BUGS statement as a command that executes some standard calculation. For example, you are not allowed to write a statement like `x <- x + 1` in BUGS, whereas this is common in most programming languages (including R!). A statement of the form “`x <- y`” in bugs is saying something like “`x` is defined to be a shorthand for `y`”. Thinking in this way, it becomes clear that it does not make sense to say “`x` is defined to be a shorthand for `x + 1`”!

STEP 3: Specify which variables in your R session are to be considered as “data” for your model file.

BUGS' aim in life is simply to simulate the conditional distribution of all the variables described in the model file, conditional on the values that we specify as “data”. That is the special role of the “data” variables – we condition on them. But you already knew that; that is what we do in Bayesian statistics. So you mentally go through your model file and ask yourself, for each variable you see, do we want BUGS to be simulating values for the variable, or do we want to specify fixed values for the variable? The ones that we want to specify fixed values for are the data variables; for example, in the subliminal model file above, those variables are `n1`, `n2`, `trt`, and `ctrl`. Then you would prepare the way for what comes next by making the following assignment in R:

```
dat = c("n1","n2","trt","ctrl")
```

STEP 4: Decide which variables you want to “monitor,” that is, for which variables in the MCMC do you to store values as the Markov chain runs? These will include the parameters you are interested in. In our example, these might include `mu1`, `mu2`, `sig1`, and `sig2`. So I would run this command next:

```
param = c("mu1","mu2","sig1","sig2")
```

STEP 5: Specify an initial state for variables in the Markov chain – typically you will want to give initial values for the parameters in your statistical problem (if you don’t want to specify these, you can have BUGS choose initial states randomly from the prior distribution). These are to be given as a list, for example:

```
inits = list(mu1 = 10, mu2 = 10, tau1 = .01, tau2 = .01)
```

Note: you are not suppose to specify values for “logical” variables like `sig1` and `sig2`.

STEP 6: Run BUGS. To do this in the easy way I am recommending, you need to load the BRugs library, and “source” the file of function definitions that I’ve provided on the web. Here is how those commands look:

```
library(BRugs)
source("http://www.stat.yale.edu/~jtc5/238_2006/mybrugs.r")
```

Of course, if you download that `mybrugs.r` file, you can refer to it by where it is on your computer instead of over the web. Now you use the function `mybrugs`, as follows:

```
r = mybrugs("subliminal.bug", dat, inits, param, 10000)
```

As you see, you need to tell `mybrugs` the name of the model file `[[here,"subliminal.bug"]]`, the data `[[we previously defined this as dat]]`, the initial values `[[inits]]`, the variables we want to monitor `[[params]]`, and the number of iterations `[[here, 10,000]]`. What you get back is a data frame with the MCMC iterations for the variables you specified in `param`, plus one more variable called `deviance`:

```
> names(r)
[1] "deviance" "mu1"      "mu2"      "sig1"     "sig2"
```

STEP 7: Look at your results in R, draw conclusions, get rich and famous..., etc. For example, here you could do:

```
hist(r$mu1 - r$mu2)
mean(r$mu1 > r$mu2)
```

4.4 JAGS: installing and running it

First a note on JAGS and BUGS. These two pieces of software are intended to do the same thing and they function very similarly. *If you have a Windows computer, you can use either BUGS or JAGS.* I have used BUGS in previously running this course, and detailed instructions for getting and using BUGS are contained in the class notes: Section 4.2 explains how to get BUGS, which is simply a matter of downloading the BRugs package for R, and section 4.3 goes into more detail about how to run BUGS from R, using the BRugs library and a function `mybrugs` that I wrote, which you also get from the web. I will focus on JAGS since it is available for all computers.

4.4.1 Website and documentation

The website for JAGS is <http://www.fis.iarc.fr/~martyn/software/jags/> and the main manual for JAGS can be obtained from that website, or you can get it with one click from

http://www.stat.yale.edu/~jtc5/238/readings/jags_1.0.3_user_manual.pdf

All of the detailed instructions I give below will be for Windows computers. I will point out any places I know of where Macs differ substantially from Windows, but I won’t be able to give details for Macs.


4.4.2 Installation


4.4.2.1 Installation instructions for Windows users

Version 1.0.3 of JAGS comes with a standard Windows setup program. You can download the program that you need at

<http://www-fis.iarc.fr/~martyn/software/jags/jags-1.0.3-setup.exe>

You install JAGS simply by double clicking on this setup program.


I will write sequences of actions separated by two colons; for example, “click File::Save as...” would mean “click the File menu, and then click ‘Save as...’”. I will use the symbol  to represent the “windows key” (also called the “start key”) on your keyboard, which has an icon like on it that resembles the picture. [[Note if you have a Windows computer that doesn’t have a Windows key (and there are some that don’t), you can use Ctrl-Esc instead.]] I will write things that you should type and enter in this font: type this stuff. So, e.g.,

::Run....: cmd::jags

means to hit the windows key, then click on “Run...”, then type `cmd`, then type `jags`. If you try this, in fact, you should see that it doesn’t work (it gives an error). Next we want to get to the point where the above sequence of keystrokes actually starts JAGS.

To do this, Right-click My Computer:: Properties::Advanced::Environment Variables. In the System variables window, click Path and then Edit. Go to the end of the Variable value string, and add this string of characters:

;C:\Program Files\JAGS\JAGS-1.0.3\bin

Then click OK, OK, OK. Now see if ::Run....: cmd::jags works; you will know it is successful if you see the words “Welcome to JAGS 1.0.3...”. If so you are done with installation and can run JAGS!

4.4.2.2 Installation for Mac users

I can’t give as detailed instructions here, since I don’t use or even have access to a Mac (although I helped some people compile JAGS on their Macs last year if they brought their Macs with them). However, this year, for the first time, there is an executable file that can be installed on the Mac (that is, you don’t have to compile the source code):

<http://www-fis.iarc.fr/~martyn/software/jags/JAGS-1.0.3u.dmg>

So I believe this should be very easy. If the install file doesn’t work, then you can do what people did last year, which is to compile the source code. We worked together to get that going, and everyone was able to do it. Please use the discussion group on the classesv2 server to post questions and answers!

4.4.3 Running JAGS

On either Windows or Mac, once you have gotten to the point where you can open a command shell and type `jags` and have JAGS start up, you are ready to go. I am writing these instructions referring to a Windows computer, but I believe most of it should work on Macs too, possibly with just minor modifications that I hope will be easy for Mac users.

I will talk about a number of ways to run JAGS. I would suggest that you try all of them in order. Even though I believe you find it most convenient to settle on the last way (doing everything from within R), it is good to have a bit of experience with the first way first.

Let's suppose we want to run a familiar example where we observe $X = 55$ successes out of 100 trials, and our model is $X \sim \text{Bin}(100, \theta)$. We want to do inference on θ , and let's say we take our prior distribution as $\theta \sim \text{Unif}(0, 1)$.

Contents of the file binomial.jags:

```
model{
  x ~ dbin(theta, 100)
  theta ~ dunif(0, 1)
}
```

It is important to put files in a definite place and be aware of where they are. For this example, let's say we are working in the folder

C:\Documents and Settings\jtc\My Documents\funstuff\stat238

We'll call this folder the "working directory." [By the way, I am using the terms "folder" and "directory" interchangeably.] We need to have a model file that expresses our model. For this example, we make a file named binomial.jags as shown above, and put it in our working directory.

4.4.3.1 From the command line, outside R

In addition to a model file, we also need a "data file" that contains the "data." The data are simply values for some of the variables mentioned in the model file. In this case our data is the value 55 for the variable x .

Contents of the file binomial-dat.r:

```
`x` <- 55
```

We will want to run MCMC to get a sample from the posterior distribution for θ . We can specify an initial value for our chain, that is, a value for θ to start the MCMC at. This would be done in another file, which we could name whatever we want, such as "inits.r". For example, we could start θ at the value 0.5 using the file shown at right.

Contents of the file binomial-inits.r:

```
`theta` <- 0.5
```


By the way, the funny ``` backquote that is used in the binomial-dat.r and binomial-inits.r files is not a single or double quotation mark, but rather a backquote... on my keyboard, this is located in the upper left, together with the tilde (`~`) symbol.

Also, one more note on making files like binomial-dat.r and binomial-inits.r. If we have the corresponding objects defined in R, then R can write the file for us. So, for example, if the variable x is already assigned the value 55 in your R session, then you can create the file binomial-dat.r above by this R command:

```
dump("x", file="binomial-dat.r")
```

and R will create a file binomial-dat.r in your working directory that contains the assignment ``x` <- 55` (except written out over more than one line, which is fine). If you want to dump more than one variable to a file, you can do it by giving a vector of the names of the objects you want to dump, like this:

```
dump(c("x", "theta"), file="dat.r").
```

You want to start a command shell going. In Windows, start a cmd window by doing ::Run....: cmd. Then we should change the working directory of the cmd window to the directory where we have stored the relevant files. So, to follow our example above, at the cmd prompt you would enter the command

```
cd C:\Documents and Settings\jtc\My Documents\funstuff\stat238
```

Then start JAGS by entering the command `jags`. At this point you see a welcome to JAGS, and the prompt changes to a period. That is, when you see the period JAGS is waiting for you to enter a command.

```

model in binomial.jags
data in binomial-dat.r
compile
parameters in binomial-inits.r
initialize
adapt 100
monitor theta
update 10000
coda *
exit

```

☺ **Using a script in JAGS:** Instead of typing the commands in separately, you can type them into a script file once, and run them all at once. To do this, suppose you have a file called `my_jags_script.txt` that contains the above jags commands. To run this script, simply issue the command

```
jags binomial_script.txt
```

in a command shell. Or, from within jags, use the command

```
run binomial_script.txt
```

Finally, you will want to collect the output of JAGS in R so you can do whatever you want to do with it, such as calculating quantiles, drawing histograms, etc. Here is how you do it:

```

library(coda)
stuff = read.openbugs()
summary(stuff)
r = as.data.frame(stuff[[1]])
names(r)
myhist(r$theta)

```

Note the function used above is called `read.openbugs`, although it works with JAGS and not just openbugs.

For reference beyond the example that I've gone through here, the relevant JAGS commands are described in detail in Chapter 3 (pages 12-16) of the JAGS manual.

4.4.3.2 Running JAGS and doing everything from inside R

I wrote a function that is supposed to help with all the above steps after the creation of the model file. That is, the function will help you create data and inits files, run JAGS, and collect the results back into R. As long as you are connected to the web, you can get this capability into your R session by running the following command:

```
source("http://www.stat.yale.edu/~jtc5/238/stat238rjags.r")
```

Set up the data and initial values in R:

```

x = 55          # <-- data
theta = 0.5     # <-- initial value

```

Run JAGS using the function `runjags`:

```

r = runjags(modelfile="binomial.jags", dat=c("x"),
monitor=c("theta"), nit=10000)

```

Now you can use `r` as above:

```

names(r)
myhist(r$theta)

```

Well, so that's it for this method of running JAGS through R. It's quite nice and concise.

If there were any problems or you want to look at some details of the run, you can do this:

```
file.show("jagslog.txt")
```

For a bit more explanation, I'll record the function template that indicates what the arguments are, etc., and make a few comments. Here is that function template:

```
runjags = function(modelfile, dat, monitor, nit, nburn = 100,
  initialize=monitor, datfile="dat.r", initsfile="inits.r",
  scriptfile="jags_script.txt", logfile="jagslog.txt", dic=TRUE)
```

Notes on the arguments of the `runjags` function:

- `modelfile` is a character string – the name of the model file.
- `dat` is a vector of names of the variables in the R session that should be considered data. You can think of these as the variables whose values are supposed to be passed in to the model file.
- `monitor` is a vector of names of the variables that will be sampled by the MCMC and whose values you want to monitor.
- `nit` is the number of MCMC iterations you want to run past the “burnin” or “adaptation” period.

The remaining arguments do not need to be set, since there are default values that are given if you do not set a value yourself.

- `nburn` is the number of iterations you want to run for a burnin or adaptation period. You might want to experiment with different values here if things don't seem to be working well.
- `initialize` is a vector of variable names in the R session. The MCMC will be initialized with whatever values these variables have when the `runjags` command is run, and any other variables that are being updated by the MCMC but were not included in the `initialize` vector will be initialized by drawing random values from their prior distributions.
- `datfile`, `initsfile`, `scriptfile`, and `logfile` are just names of files that are written during the process. You wouldn't need to change these unless you were worried about overwriting files by these names for some reason.
- `DIC=TRUE` causes JAGS to calculate and monitor the “deviance”. I think it is possible that JAGS may complain that it can't do this for some models, and in that case you could change to `DIC=FALSE`.

Note: A more concise way to express the above is

```
x = 55          # <-- data
theta = 0.5     # <-- initial value
dat = c("x")
monitor = c("theta")
r = runjags("binomial.jags", dat, monitor, 10000)
```

It is also nice sometimes to use the `mcmc.list` structure that `coda` and `read.openbugs` give. You can get that any time after the `runjags` run by doing `read.openbugs()` again [\[\[this is done in the runjags function too\]\]](#). For example, the `summary` function gives a nice summary. So, you can do things like this:

```
ro = read.openbugs()
summary(ro)
plot(ro)
window(ro, 2000, 2020)
plot(window(ro, 2000, 2020))
summary(window(ro, 2000, 2020))
```

The next thing is just like the “r” currently returned from `runjags` (maybe I should change `runjags` so that it returns the `mcmc.list` `ro` above):

```
rr = data.frame(ro[[1]])
```

→ Miscellaneous notes (080801)

I found out that JAGS seems much more fussy about names than BUGS.

In particular, names of model files must begin with a letter, not a number.

And, names of variables cannot contain the underscore character “_”.

Example of slight perturbations of a file m080801.jags, and how JAGS reacted to the command

. model in m080801.jags

```
model{
  for(i in 1:nt) {
    ns[i] ~ dpois(lamt)
  }
  for(i in 1:na){
    ns[nt+i] ~ dpois(lama)
  }
  lamt ~ dnorm(0,.0001)
  lama ~ dnorm(0,.0001)
}
```

```
model{
  for(i in 1:nt) {
    ns[i] ~ dpois(lamt)
  }
  for(i in 1:na){
    ns[nt+i] ~ dpois(lam_a)
  }
  lamt ~ dnorm(0,.0001)
  lam_a ~ dnorm(0,.0001)
}
```

The file on the left worked fine, but the file on the right didn't, giving error message:

_syntax error, unexpected NAME, expecting ',' or ')'

Parse error on line 1

Note also can see how line numbers in error messages are not useful! The error was not on line 1. (The first error is more like on line 6!)

Something about multiple chains:

```
data(line)
> class(line)
[1] "mcmc.list"
> nchain(line)
[1] 2
> niter(line)
[1] 200
> nvar(line)
[1] 3
> temp = mcmc.list(line[[1]],line[[2]],line[[1]])
> nchain(temp)
[1] 3
```

4.5 Regression, crying and IQ.

We'll begin with an example on Crying and IQ, involving infants in fact, and a heartwarming story...

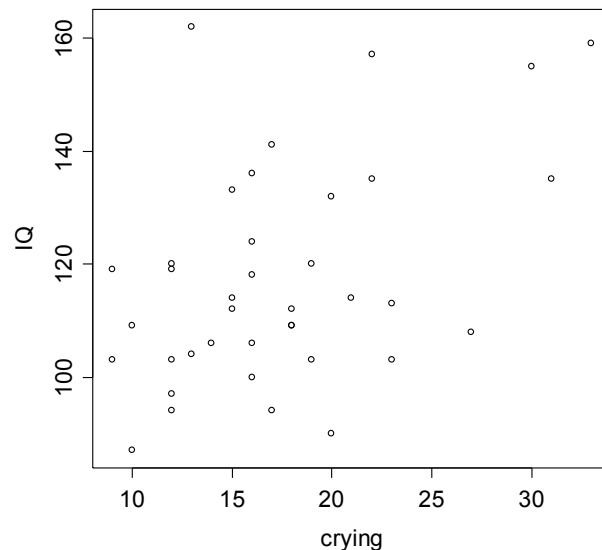
Example: Crying and IQ The heartwarming story:

Data on 38 infants (4 to 10 days old).

Researchers used a rubber band to snap infants in the foot.

Measured crying intensity.

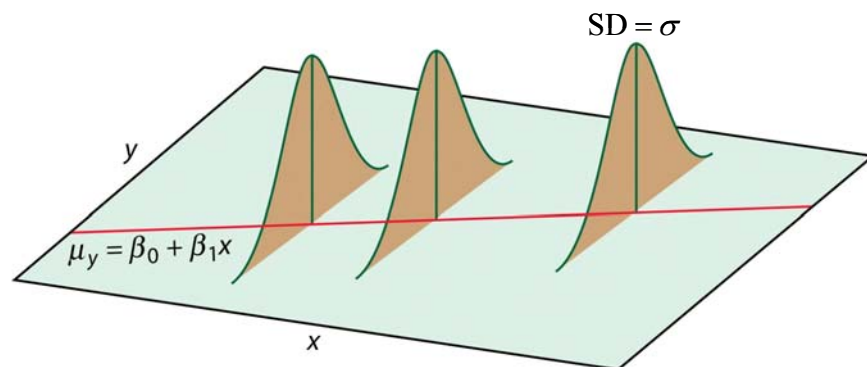
(“Number of peaks in the most active 20 seconds”)



Later measured IQ at age 3 years.

Our question: "Is there a real relationship?"

We need a probabilistic model.



$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2).$$

For this model, maximum likelihood would estimate β_0 and β_1 by the intercept and slope of the usual least squares regression line. This is a statistical motivation for doing least squares.

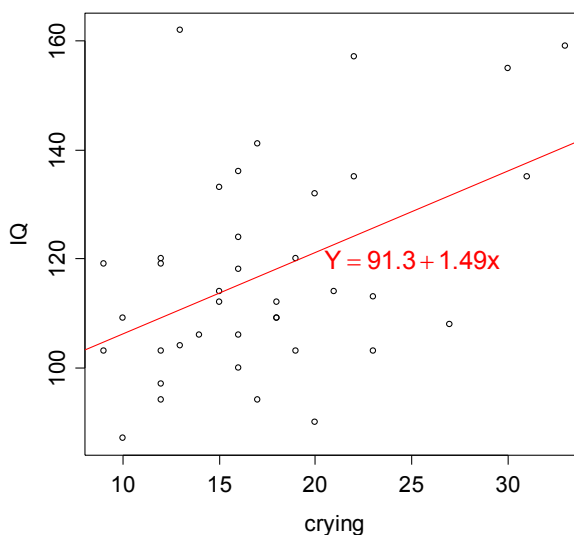
The least-squares regression line is shown at right.

It is the line that minimizes the sum of the squared residuals.

That is, we minimize the sum

$$\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 \text{ over all choices of } \beta_0 \text{ and } \beta_1.$$

Looks like: On average, gain about 1.5 IQ points per unit of crying intensity. Very scientific.



→ What is our posterior distribution for the slope of the "population" line? We'll use WinBUGS.

We follow the usual procedure: write a model file, set up the required variables in R, specify initial values for the MCMC and parameters to monitor.

```
library(BRugs)
source("http://www.stat.yale.edu/~jtc5/23
8_2006/mybrugs.r")
cry.dat = read.csv("crying.csv")

crying = cry.dat$crying
IQ = cry.dat$IQ
```

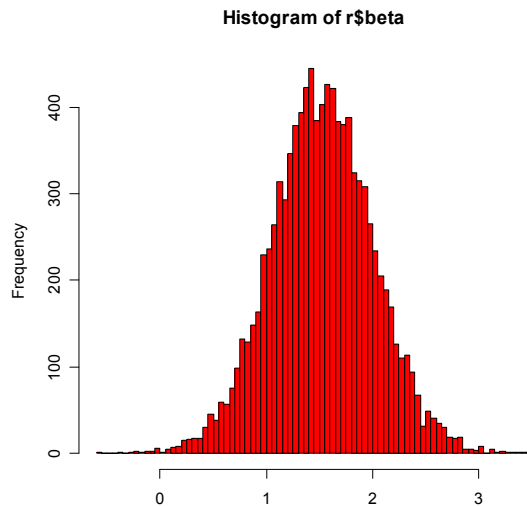
The model file cry.bug

```
model{
  for(i in 1:38){
    mu[i] <- alpha + beta * crying[i]
    IQ[i] ~ dnorm(mu[i], tau)
  }
  alpha ~ dnorm(0.0, 1.0E-4)
  beta ~ dnorm(0.0, 1.0E-4)
  tau ~ dexp(0.1)
}
```

```
plot(crying, IQ)

data = c("crying", "IQ")
inits = list(alpha=100, beta=0, tau=.01)
params = c("alpha", "beta", "tau")
r = mybrugs("cry.bug", data, inits, params, 10000)

names(r)
hist(r$beta, n=100, col="red")
```



```
> mean(r$beta)
[1] 1.528508
> quantile(r$beta, c(.025, .975))
      2.5%      97.5%
0.5770925 2.5000500
> mean(r$beta > 0)      # <--posterior prob that beta > 0
[1] 0.9981
```

4.6 More on correlation and regression

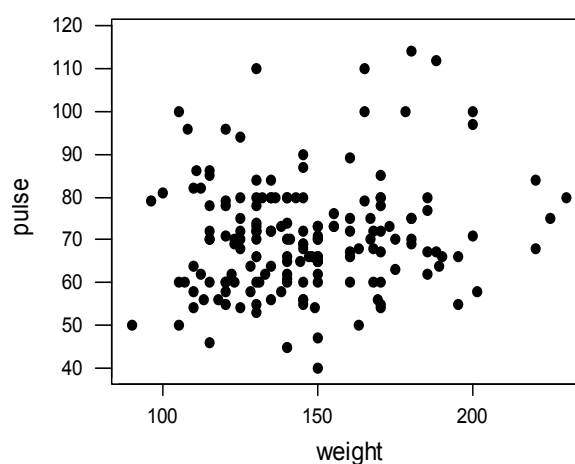
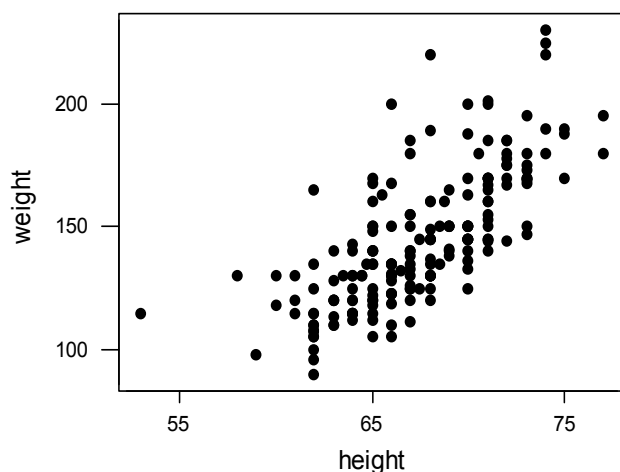
We'll start with a discussion of correlation and regression. Much of this was pioneered by Sir Francis Galton (1822-1911). Some highlights about him: Scientist and explorer; Cousin of Charles Darwin; Studied heredity, intelligence, eugenics...; IQ estimated at 200; Invented quincunx; Idea of branching processes; Statistical study of efficacy of prayer.

We'll address questions like:

- How strong a linear relationship is there between two variables?
 - E.g. when height increases, does weight also tend to increase?
 - E.g. How about the same question for weight and pulse?
- If we know the value of one variable for an individual, how can we best predict the value of another variable for that individual?

To look at data, it's always good to start with scatterplots: Plot two variables simultaneously. Put one variable on horizontal axis, other variable on vertical axis.

E.g.: Here are scatterplots for height versus weight, and pulse versus weight:



Correlation is a numerical measure of the strength of a linear relationship.

First a bit of mathematics... Here are various manifestations of the Cauchy-Schwarz inequality:

$$(1) \sum_{i=1}^n w_i z_i \leq \sqrt{\sum_{i=1}^n w_i^2} \sqrt{\sum_{i=1}^n z_i^2} \quad \left[\text{i.e., } w \cdot z \leq \|w\| \|z\| \right]$$

$$(2) E(WZ) \leq \sqrt{E(W^2)} \sqrt{E(Z^2)}$$

[Can see this from the Law of Large Numbers, and (1)]

$$(3) \text{cov}(X, Y) \leq \sqrt{\text{var}(X)} \sqrt{\text{var}(Y)}$$

[Get this from applying (2) to $W = X - E(X)$, $Z = Y - E(Y)$]

$$(4) \frac{|\text{cov}(X, Y)|}{\sqrt{\text{var}(X)} \sqrt{\text{var}(Y)}} \leq 1$$

Definition of correlation: $\rho(X, Y) = \frac{\text{cov}(X, Y)}{\text{SD}(X) \text{SD}(Y)}$

So by (4) we know that $-1 \leq \rho(X, Y) \leq 1$.

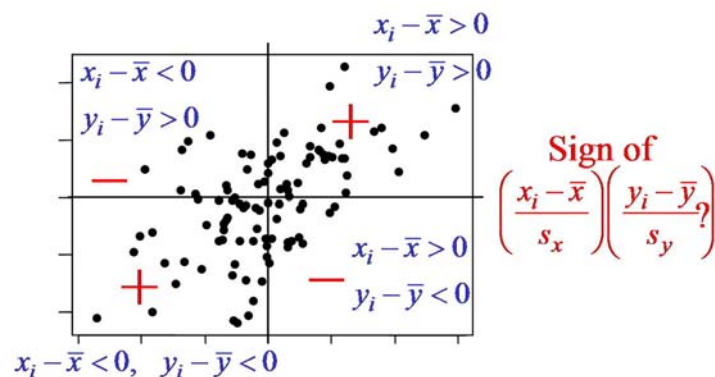
Given data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ we could

estimate $\text{cov}(X, Y)$ by $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$,

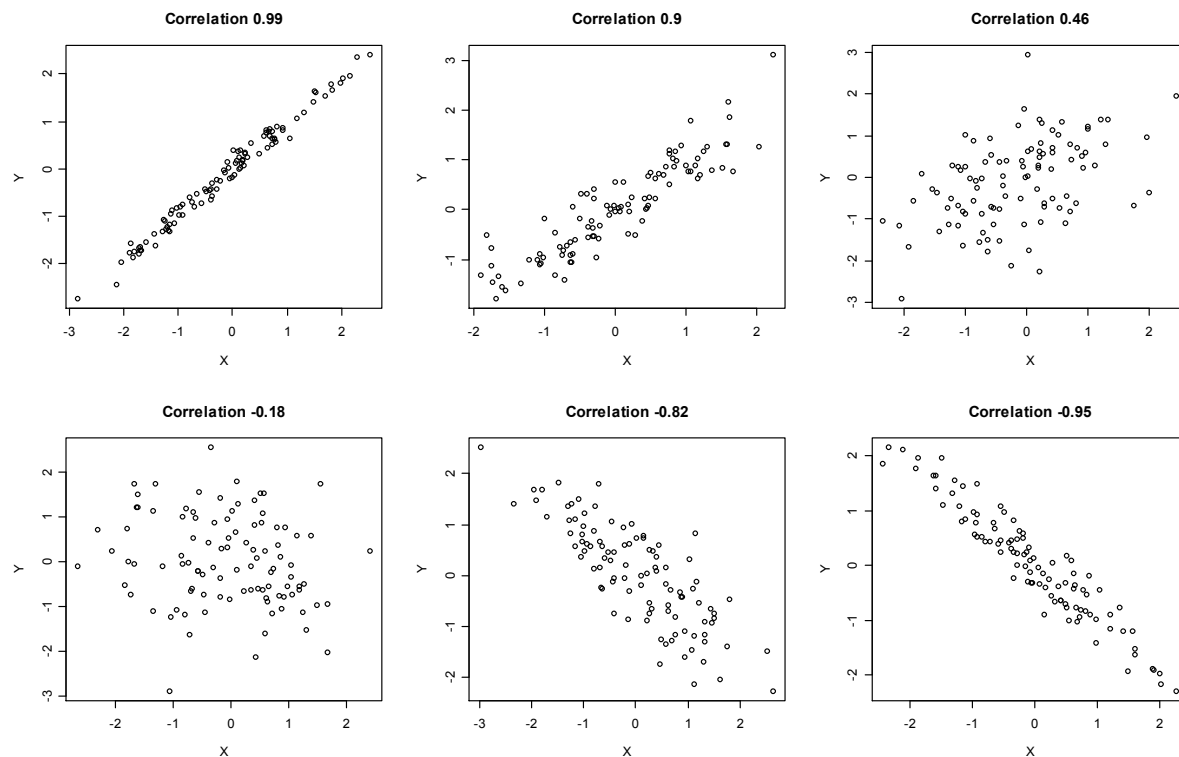
estimate $\text{SD}(X)$ by $\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$,

and $\text{SD}(Y)$ by $\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}$.

Plugging these in, we get the *definition of sample correlation*:



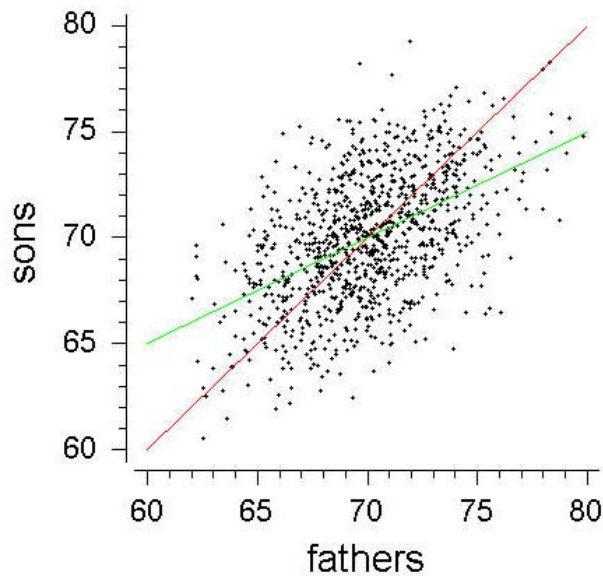
$$r(X,Y) = \frac{\widehat{\text{cov}}(X,Y)}{\widehat{\text{SD}}(X)\widehat{\text{SD}}(Y)} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}.$$



```
##### correlation picture #####
library(mvtnorm)
par(mfrow=c(2,3))
cors = c(.99,.9,.5,0,-.8,-.95)
for(i in 1:6){
  dat = rmvnorm(100,mean=c(0,0),sigma =
matrix(c(1,cors[i],cors[i],1),ncol=2))
  rho = round(cor(dat)[1,2],2)
  plot(dat,xlab="X",ylab="Y",main=paste("Correlation",rho))
}
```

Introduction to idea of regression

Data of Karl Pearson (a student of Galton) on heights of fathers and sons looked something like this:



Consider predicting the height of the son from the height of the father.

There are two interesting lines in the plot. The red line (slope 1) corresponds to the "natural guess" that the heights would be the same.

The green line (slope = correlation!) gives the best guess, in the sense that it minimizes the sum of squared errors.

This "regression line" also gives our estimate of the conditional expectation of Y given $X = x$. E.g. answers the question, what is the expected height of a son whose father is 75" tall? Since the slope of the regression line is less than 1, the answer is less than 75". Similarly, the expected height of the son of a short father is taller (closer to the overall mean height).

Galton called this phenomenon "regression toward mediocrity", which is where the name "regression" came from.

We'll talk about all this in more detail shortly.

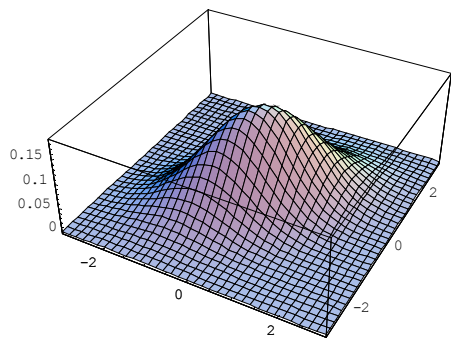
Bivariate Normal distributions

Density $f(x, y) = \exp(\text{quadratic function of } x \text{ and } y)$.

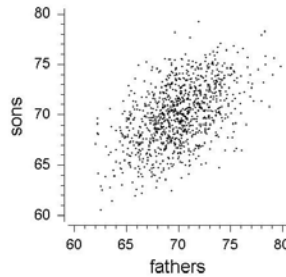
E.g. density where X and Y each have mean 0 and variance 1, and the correlation is ρ , is

$$f(x, y) \propto \exp\left(\frac{-1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right).$$

For general means and SD's μ_x, σ_x, μ_y , and σ_y replace x and y by $\frac{x-\mu_x}{\sigma_x}$ and $\frac{y-\mu_y}{\sigma_y}$.



density



sample data

Example:

Fathers: Mean = 70" SD = 3"

Sons: Mean = 70" SD = 3"

Correlation = 0.5

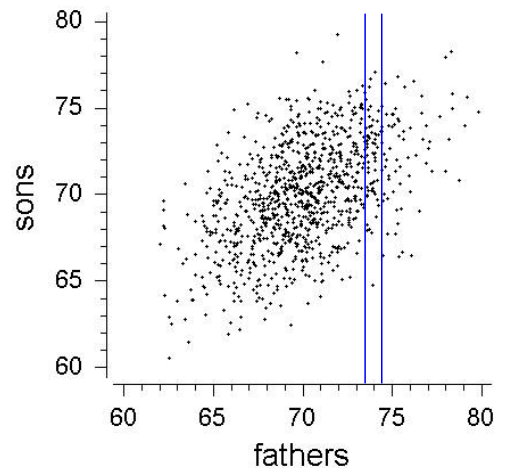
Question: Mean height of son if father is 74" ?

The "naive guess":

Father is $4/3$ SD's above mean, so guess son will be $4/3$ SD's above mean, or 74"

Look at the data: Can see the guess 74" is obviously too high!

[[For R demo see 061115.r also.]]



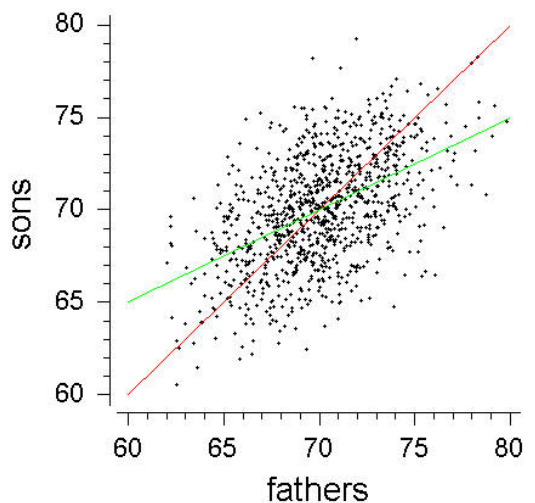
The linear regression guess:

If father is $4/3$ SD's above mean, guess that son will be, not $4/3$ SD's above mean, but rather:

correlation $\times 4/3 = 2/3$ SD's above mean.

That is, in our example, guess son's height to be $70 + (2/3) \times 3 = 72$ inches.

In the picture at right, the red line (slope 1) is the line of "naive guesses." The green line (slope = correlation = 0.5) is the least squares regression line, which gives the conditional expectation of Y (here, son's height) given the various possible values of X (father's height).



Example: A college wants to use SAT scores to predict 1st-year final exam scores
Historical data:

X = SAT scores: mean 650, SD 80

Y = final scores: mean 65, SD 10

correlation: $r = 0.4$

Question: Predict final score for student with $x = 750$.

Answer

Step 1: Standardize x

x is $\frac{750-650}{80} = 1.25$ SD's above the mean $E(X)$.

Naïve guess would guess Y to be 1.25 SD's above its mean, or $65 + 1.25 * 10 = 77.5$

Regression says: Guess Y to be $r \times 1.25 = 0.4 \times 1.25 = 0.5$ SD's above its mean, or $65 + 0.5 * 10 = 70$.

Equation of the regression line. This is just a formula for all the predictions as a linear function of x .
E.g. in the SAT example, it would be

$$\frac{\hat{y} - 65}{10} = (0.4) \left(\frac{x - 650}{80} \right)$$

In general:

$$\text{Population version: } \frac{E(Y | X = x) - E(Y)}{\sigma_Y} = \rho \left(\frac{X - E(X)}{\sigma_X} \right)$$

$$\text{Sample version: } \frac{\hat{y} - \bar{y}}{s_y} = r \left(\frac{x - \bar{x}}{s_x} \right)$$

The "regression fallacy"

Example:

In training, air force pilots make two practice landings with instructors and are rated on performance. The instructors discuss the ratings with the pilots after each landing. The boss looked at the data and noticed that the pilots who make poor landings the first time tend to do better the second, and those who make good landings the first time tend to do worse on the second try. The conclusion: criticism helps the pilots, while praise tends to make them do worse. As a result, instructors were ordered to criticize all landings, good or bad.

The "fallacy" is a failure to recognize that this type of observation is a natural consequence of statistical variability, which can lead to faulty attempted explanations.

Conditional distributions "within vertical strips"

We assume a bivariate Normal distribution, characterized by 5 parameters: $\mu_x = E(X)$, $\sigma_x = SD(X)$, $\mu_y = E(Y)$, $\sigma_y = SD(Y)$, **and** $\rho = \text{Correlation}(X, Y)$.

Fact: the conditional distribution of Y given $X = x$ is Normal, with:

- mean given by the " y " from the regression line $\left(\frac{y - \mu_y}{\sigma_y} \right) = \rho \left(\frac{x - \mu_x}{\sigma_x} \right)$,

that is, $E(Y | X = x) = \mu_y + \rho \frac{\sigma_y}{\sigma_x} (x - \mu_x)$, and

- SD given by $\sqrt{1 - \rho^2} \sigma_y$.

Note this makes some qualitative sense:

- SD within a strip is always $\leq \sigma_y$. $[\sigma_y]$ is the SD over all individuals, and the variable x can help in predicting y .]
- If $\rho = 1$ then SD in a strip is 0. [Perfect prediction of y from x .]
- if $\rho = 0$ then SD in a strip is same as σ_y . $[x]$ doesn't help in predicting y .]

Example: More on the college predicting final exams from SAT's.

	mean	SD	
SAT scores	650	80	$\rho = 0.4$
Final exams	65	10	

1. What percentage of students score over 75 on final exam?

Easy: 75 is $(75 - 65)/10 = 1$ SD above mean.

Answer is $1 - \Phi(1) \approx 0.16$ (16 %).

2. Among students who get 750 on SAT, what fraction get over 75 on final exam?

In strip at $x = 750$ (standard score = 1.25): these students have mean = 70 and

$$SD = \sqrt{1 - \rho^2} \sigma_y = \sqrt{1 - (0.4)^2} \times 10 = 9.165.$$

So we want the prob of a score > 75 from a $N(70, 9.165)$ distribution.

Standard score for 75 is $(75 - 70)/9.165 = 0.546$.

Answer: $1 - \Phi(0.546) \approx 0.29$. (Compare previous 0.16)

4.7 Notes on algebra and geometry of regression

We are working in \mathbb{R}^n . For $v, w \in \mathbb{R}^n$ define $v \cdot w = \sum_{i=1}^n v_i w_i$. The length of the vector v is

$\|v\| = \sqrt{v \cdot v} = \sqrt{\sum v_i^2}$, and the distance between v and w is $\|v - w\| = \sqrt{\sum (v_i - w_i)^2}$. We know $v \cdot w = \|v\| \|w\| \cos(\theta)$ where θ is the angle between v and w . If $v \cdot w = 0$, that is, $\theta = 90^\circ$, we say v and w are orthogonal.

A line is a one-dimensional subspace of \mathbb{R}^n ; a plane is a two-dimensional subspace of \mathbb{R}^n .

We can think of a data set $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ as n points in \mathbb{R}^2 , and this is how we draw a scatterplot.

☺ But it's also useful to think of a data set as two columns,

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \text{ and } y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \text{ that is, as two points in } \mathbb{R}^n !$$

In regression, we try to fit the given y_i 's by \hat{y}_i 's of the form:

$$\begin{aligned} \hat{y}_1 &= a + bx_1 \\ \hat{y}_2 &= a + bx_2 \\ &\vdots \\ \hat{y}_n &= a + bx_n \end{aligned}$$

That is,
$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = a \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + b \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \text{ or, more concisely } \hat{y} = a \vec{1} + b x.$$

The \hat{y} vector is a point in the plane \mathbb{L} spanned by the two vectors $\vec{1}$ and x .

Finding \hat{a} and \hat{b} so that $\sum (y_i - \hat{a} - \hat{b}x_i)^2 = \sum (y_i - \hat{y}_i)^2$ is minimized corresponds to finding the point \hat{y} in \mathbb{L} that is closest to y .

That is, \hat{y} is the orthogonal projection of y on the plane \mathbb{L} .

Let's find this "least-squares" solution for a and b .

Let's work with an orthonormal basis of \mathbb{L} ; this is a bit more convenient than the basis $\vec{1}$ and x . Define unit vectors (i.e. vectors having length 1)

$$v = \frac{\vec{1}}{\|\vec{1}\|} = \frac{1}{\sqrt{n}} \vec{1} \quad \text{and} \quad w = \frac{x - \bar{x} \vec{1}}{\|x - \bar{x} \vec{1}\|} = \frac{1}{\sqrt{\sum (x_i - \bar{x})^2}} (x - \bar{x} \vec{1}).$$

Note $v \perp w$.

Now we can write the projection simply as

$$\hat{y} = (y \cdot v)v + (y \cdot w)w,$$

so we just need to find expressions for $(y \cdot v)v$ and $(y \cdot w)w$.

$$(y \cdot v)v = \frac{1}{\sqrt{n}} (y \cdot \vec{1}) \frac{1}{\sqrt{n}} \vec{1} = \frac{1}{n} (y \cdot \vec{1}) \vec{1} = \frac{1}{n} \left(\sum_{i=1}^n y_i \right) \vec{1} = \bar{y} \vec{1}$$

$$(y \cdot w)w = \frac{1}{\|x - \bar{x} \vec{1}\|^2} (y \cdot (x - \bar{x} \vec{1})) (x - \bar{x} \vec{1}) = \underbrace{\frac{\sum y_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2}}_{\text{rewrite this}} (x - \bar{x} \vec{1}).$$

Letting r denote the correlation $r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$, we have

$$\frac{\sum y_i(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = r \frac{\sqrt{\sum (y_i - \bar{y})^2}}{\sqrt{\sum (x_i - \bar{x})^2}} = r \frac{SD(y)}{SD(x)}$$

So

$$\begin{aligned} \hat{y} &= (y \cdot v)v + (y \cdot w)w \\ &= \bar{y}\vec{1} + r \frac{SD(y)}{SD(x)}(x - \bar{x}\vec{1}) \end{aligned}$$

That is, $\frac{\hat{y} - \bar{y}\vec{1}}{SD(y)} = r \left(\frac{x - \bar{x}\vec{1}}{SD(x)} \right)$.

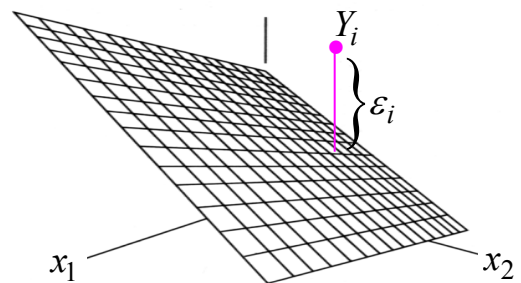
This is what we wanted to show: $\frac{\hat{y}_i - \bar{y}}{SD(y)} = r \left(\frac{x_i - \bar{x}}{SD(x)} \right)$ is the least squares solution! This gives the equation of the regression line.

4.8 Multiple linear regression

This is a very heavily used statistical technique. For example, it contains "ANOVA" (analysis of variance) as a special case, fitting lines and polynomials, and many kinds of analyses people do all the time. We use a linear model to predict or "explain" one variable in terms of others. In a linear model the mean of Y is a linear function of a number of x variables. E.g. if we have two explanatory variables,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \text{ where } \varepsilon \sim N(0, \sigma^2).$$

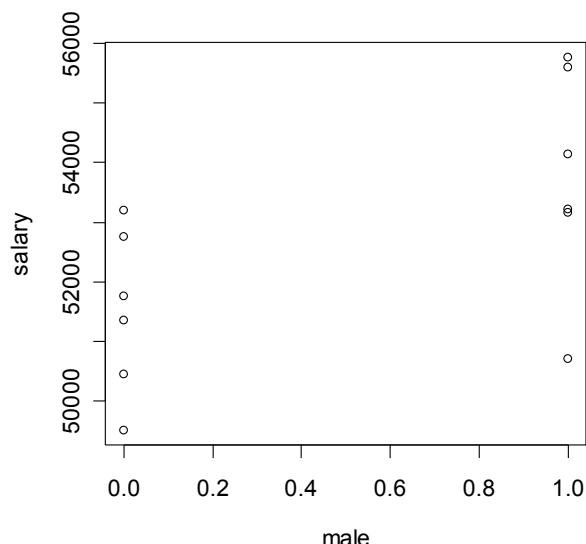
Given n data points: $Y_i \sim N(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}, \sigma^2)$ for $i = 1, \dots, n$.



A small salary data set

```
> salary.dat = read.csv("data/SalariesAndGender.csv")
> salary.dat
  salary experience gender
1   50710         1     m
2   49510         1     f
3   50440         2     f
4   53160         3     m
5   51340         3     f
6   53210         3     m
7   51760         3     f
8   54140         4     m
9   52750         4     f
10  55760         5     m
11  53200         5     f
12  55590         5     m
```

```
attach(salary.dat)
male = as.numeric(gender=="m")
plot(male,salary)
```



Male and female salaries look rather different. How convincing is this difference? We can use the same model file we used to analyze the effect of subliminal messages on math scores. Except, we'll choose a prior that is sufficiently spread out. E.g. for the salary variable, which is measured in tens of thousands, we want our prior variance to be more than hundreds of millions, and so our prior precision needs to be tiny. So here is the model file I used, which I saved as `TwoSample.bug`:

```
model{
  for(i in 1:n1){
    x[i] ~ dnorm(mu1, tau1)
  }
  for(i in (n1+1):(n1+n2)){
    x[i] ~ dnorm(mu2, tau2)
  }
  mu1 ~ dnorm(53000,1.0E-12)
  mu2 ~ dnorm(53000,1.0E-12)
  tau1 ~ dgamma(0.01,0.01)
  tau2 ~ dgamma(0.01,0.01)
}
```

And now we set up to run BUGS in R:

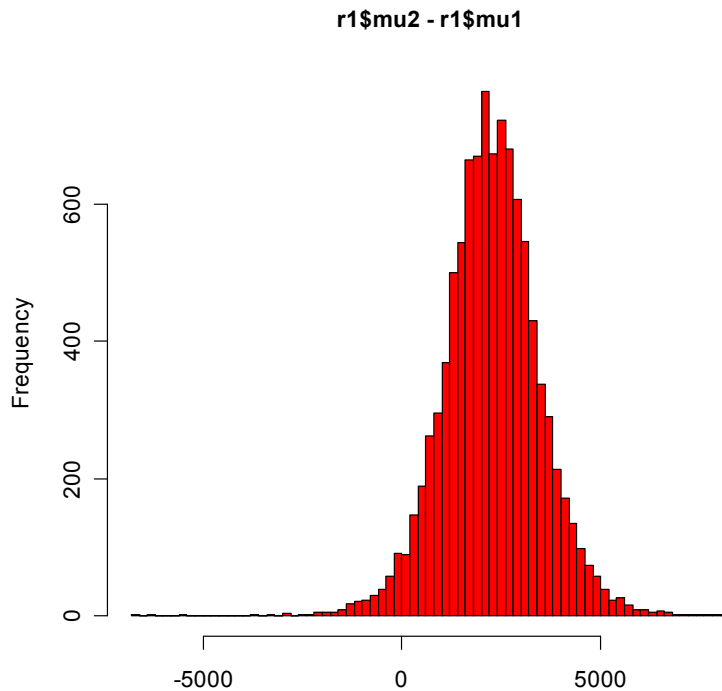
```
library(BRugs)
source("http://www.stat.yale.edu/~jtc5/238_2006/mybrugs.r")

x = c(salary[gender=="f"], salary[gender=="m"])
n1=n2=6
dat = c("x", "n1", "n2")
params = c("mu1", "mu2", "tau1", "tau2")
inits = list(mu1 = 53000, mu2=53000, tau1=1.0E-6, tau2=1.0E-6)
r1 = mybrugs("TwoSample.bug", dat, inits, params, 10000)
```

[[Some output:

```
> dicStats()
      Dbar Dhat  DIC    pD
x      213.8 209.3 218.3 4.504
total 213.8 209.3 218.3 4.504  ]]
```

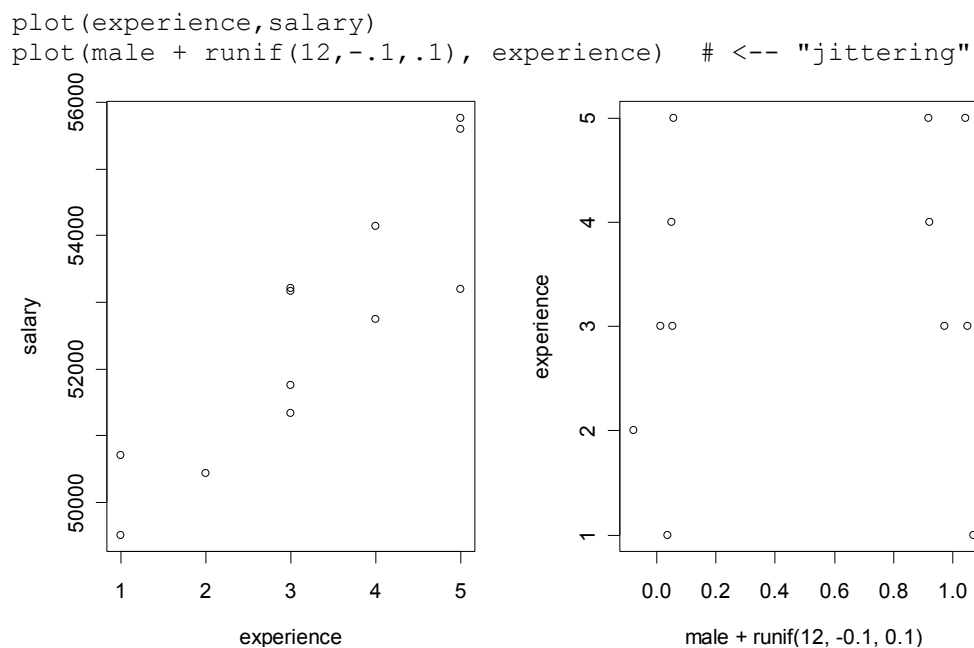
```
myhist(r1$mu2 - r1$mu1)
```



```
> mean(r1$mu2 - r1$mu1 > 0)
[1] 0.9686969
```

So we have a high posterior probability that male salaries are higher than female salaries.

Can this difference be explained by experience?

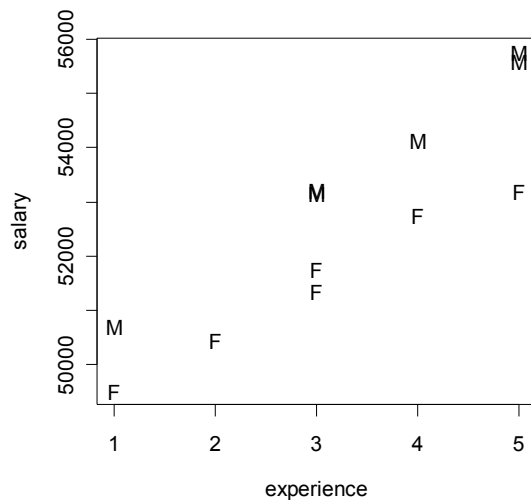


We see that experience is indeed important in determining salary. We would like to see if the effect of the variable `male` remains significant even when we include experience as an explanatory variable. The

observation that males have only perhaps slightly more experience than females suggests that the variable `male` may well remain significant.

To start with a plot, let's see all 3 variables in one plot. The binary variable `male` can be conveniently represented simply by using a different plotting variable for males and females.

```
plot(experience, salary, type="n")
labels = c("F", "M")[1+male]
points(experience, salary, pch=labels)
```



Now to do the regression using both `experience` and `male`. Here is a model file "salary1.bug":

```
model{
  for(i in 1:12){
    y[i] ~ dnorm(mu[i], tau)
    mu[i] <- a + b[1]*male[i] + b[2]*experience[i]
  }
  a ~ dnorm(50000, 1.0E-10)
  for(i in 1:2){b[i] ~ dnorm(0, 1.0E-10)}
  tau ~ dgamma(.01, .01)
}

y = salary
dat = c("y", "male", "experience")
params = c("a", "b", "tau")
inits = list(a=0, b=c(0,0), tau=1e-6)
r2 = mybrugs("salary1.bug", dat, inits, params, 10000)
```

Some output:

```
summaryStats(params)
```

```
> summaryStats(params)
      mean      sd   val2.5pc   median val97.5pc sample
a    4.819e+04 2.549e+02  4.753e+04 4.817e+04 4.908e+04  10000
b[1] 1.711e+03 1.842e+02  1.319e+03 1.711e+03 2.158e+03  10000
b[2] 1.104e+03 7.262e+01  8.320e+02 1.102e+03 1.306e+03  10000
tau  1.312e-05 6.148e-06 -1.012e-02 1.402e-05 1.015e-02  10000
```

```
> dicStats()
      Dbar Dhat   DIC    pD
y    170.4  166 174.7 4.393
```

```
total 170.4 166 174.7 4.393
```

Estimates look sensible... in 4 years gain 4 to 5 thousand dollars, and difference between male and female salaries looks well over 1000 dollars

Again could look at posterior distributions, etc.

```
> names(r2)
[1] "a"          "b.1."       "b.2."       "deviance" "tau"
> hist(r2$b.1.)
> hist(r2$b.2.)
> mean(r2$b.1. < 0)
[1] 0
> mean(r2$b.2. < 0)
[1] 0.00010001
```

Do male salaries increase faster than female salaries? In other words, is there an *interaction* between gender and experience?

Model: $\text{salary} = a + b_1 \cdot \text{male} + b_2 \cdot \text{experience} + b_3 \cdot \text{male} \cdot \text{experience}$.

Note this allows BUGS to fit two lines: one for males, another for females! The previous model allowed only the intercept to change, and this one allows both the slope and intercept to change.

We'll use the model file salary2.bug:

```
model{
  for(i in 1:12){
    y[i] ~ dnorm(mu[i],tau)
    mu[i] <- a + b[1]*male[i] + b[2]*experience[i] + b[3]*male[i]*experience[i]
  }
  a ~ dnorm(50000, 1.0E-10)
  for(i in 1:3){b[i] ~ dnorm(0, 1.0E-10)}
  tau ~ dgamma(.01,.01)
}
```

We run it as usual; it's enough to change the inits.

```
inits = list(a=0,b=c(0,0,0),tau=1e-6)
r3 = mybrugs("salary2.bug",dat,inits,params,10000)
```

Some results:

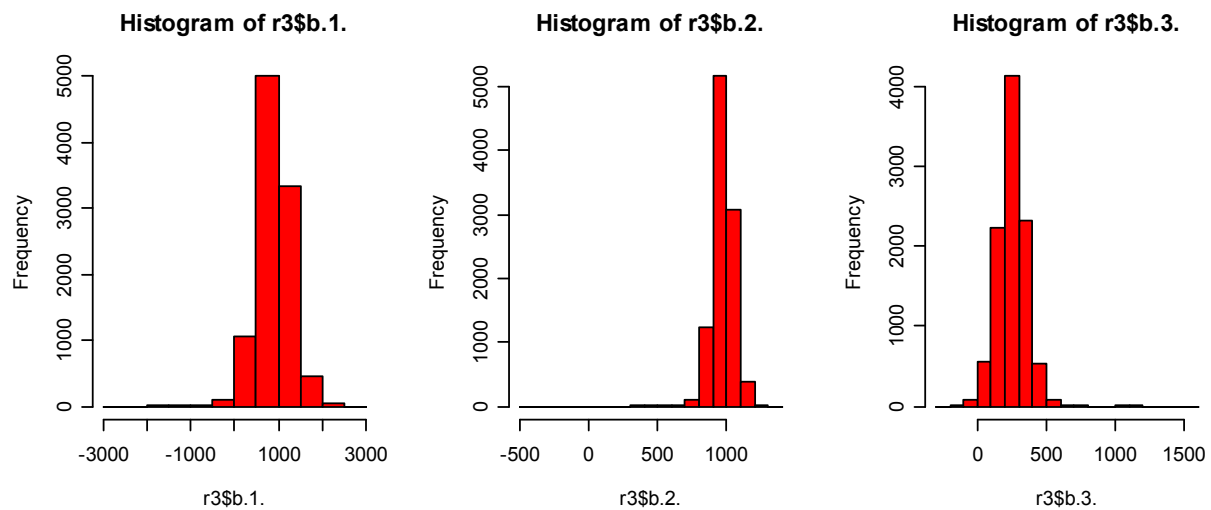
```
> s3 = summaryStats(params)
> s3
```

	mean	sd	val2.5pc	median	val97.5pc	sample
a	4.858e+04	2.694e+02	48090.00	4.861e+04	5.052e+04	10000
b[1]	8.997e+02	3.715e+02	-643.60	9.080e+02	1.639e+03	10000
b[2]	9.722e+02	8.237e+01	374.70	9.723e+02	1.121e+03	10000
b[3]	2.520e+02	1.073e+02	43.16	2.509e+02	7.788e+02	10000
tau	2.448e-05	1.228e-05	-0.01	3.081e-05	1.015e-02	10000

```
> dicStats()
      Dbar  Dhat  DIC    pD
y      163.1 157.4 168.7 5.644
total 163.1 157.4 168.7 5.644

> names(r3)
[1] "a"          "b.1."       "b.2."       "b.3."       "deviance" "tau"
```

```
> par(mfrow=c(3,1))
> myhist(r3$b.1.)
> myhist(r3$b.2.)
> myhist(r3$b.3.)
```



```
> mean(r3$b.1. < 0)
[1] 0.01090109
> mean(r3$b.2. < 0)
[1] 0.00050005
> mean(r3$b.3. < 0)
[1] 0.01080108
```

☺ The kind of analysis we have done is sometimes called "analysis of covariance".

Some nonBayesian R commands that do similar things:

```
t.test(salary[as.logical(male)], salary[!as.logical(male)])
lm1 = lm(salary ~ male)
lm2 = lm(salary ~ male + experience)
male.exp = male*experience
lm3 = lm(salary ~ male + experience + male.exp)
summary(lm1)
summary(lm2)
summary(lm3)
```

4.9 Model selection and the Deviance Information Criterion

Everything should be made as simple as possible, but not simpler.
Albert Einstein (1879-1955)

Entities should not be multiplied beyond necessity.
William of Ockham (c. 1285 - c. 1347)

These are two examples of quotations expressing the rough idea that, all things being equal, the simplest solution tends to be the best one.

How do we compare the fit of different models? This is a big area in statistics with many different proposed "answers", including AIC (for "An Information Criterion", due to Akaike), BIC ("Bayesian information criterion"), cross validation, ...

There is no single entirely convincing answer that covers all situations and purposes.

You get one model selection criterion by default from BUGS: the so-called Deviance Information Criterion, or DIC.

The *deviance* is defined to be $-2 \times \log(\text{likelihood})$. The bigger the deviance, the worse the fit of the model.

The crux of the idea of model selection is that a model whose maximum likelihood estimator gives a smaller deviance (i.e. higher likelihood) is not necessarily better. For example, in doing multiple regression, we could just keep on adding more predictors to the model and necessarily get a closer fit to the data. However, in alignment with William of Ockham and his razor, I think we all have a feeling that this is not a good idea.

The DIC (and other model selection criteria such as AIC and BIC) "penalizes" the likelihood that a model can achieve by some penalty that tends to increase for larger (higher dimensional) models. It incorporates a preference for "simpler" models by penalizing for complexity.

The MLE, and any good point estimate, should give about the lowest deviance a model can give on our data set. If we look at the deviances of the models randomly encountered in our MCMC run through the posterior distribution, they will be higher than this minimum. The penalty involves the difference between this average deviance and deviance at the point estimate.

There are in fact a number of variants in precisely how the DIC is defined. One of these is:

$$(\text{min deviance}) + 2 \times [(\text{average deviance over MCMC run}) - (\text{min deviance})]$$

or, equivalently,

$$(\text{average deviance}) + [(\text{average deviance}) - (\text{min deviance})].$$

This differs a bit from how BUGS calculates DIC for practical reasons, but this is the idea. BUGS does

$$(\text{deviance at the posterior mean}) + 2 \times \underbrace{[(\text{average deviance}) - (\text{deviance at the posterior mean})]}_{\text{Called "p}_D\text{" in dicStats display}}$$

An alternative definition that tends to be numerically quite close to this is:

$$(\text{average deviance}) + \left[\frac{1}{2} (\text{variance of deviance}) \right].$$

In all four of the formulas above, the quantity in squared brackets is a penalty for complexity, and is expected to be close to the number of parameters in the model.

The suggestion is to prefer models with smaller DIC.

→ There's an example in the R file on the web...

4.10 Gibbs sampler.

I've been meaning to tell you about this -- the "G" in BUGS: "Bayesian inference Using Gibbs Sampling." The Gibbs sampler is a special case of the Metropolis-Hastings method we have studied. Actually BUGS uses Metropolis-Hastings more generally, but in cases where it can use the Gibbs sampler, it will.

As with any MCMC method, the goal is to draw a sample that comes from a desired distribution. The idea of the Gibbs sampler is to iteratively perform the following operation:

Leaving all the variables except for one fixed, sample from the conditional distribution of that variable, given the fixed values of all the rest.

For example, if we were using the Gibbs sampler to draw an approximate sample from a joint density $f(x, y, z)$ for 3 variables (X, Y, Z) , we would do this:

- Start at an arbitrary initial state (x_0, y_0, z_0)
 - Draw a new x_1 from the conditional distribution of X given $Y = y_0$ and $Z = z_0$.
Now we are at (x_1, y_0, z_0) .
 - Draw a new y_1 from the conditional distribution of Y given $X = x_1$ and $Z = z_0$.
Now we are at (x_1, y_1, z_0) .
 - Draw a new z_1 from the conditional distribution of Z given $X = x_1$ and $Y = y_1$.
- Move to the state (x_1, y_1, z_1) .
 - Draw a new x_2 from the conditional distribution of X given $Y = y_1$ and $Z = z_1$.
Now we are at (x_2, y_1, z_1) .
 - Draw a new y_2 from the conditional distribution of Y given $X = x_2$ and $Z = z_1$.
Now we are at (x_2, y_2, z_1) .
 - Draw a new z_2 from the conditional distribution of Z given $X = x_2$ and $Y = y_2$.
- Move to the state (x_2, y_2, z_2) .
- Etc....

→ Gibbs sampler for the simple regression model.

Here the parameters are α , β , and τ , and we want to get a sample from the posterior, which has a density $f(\alpha, \beta, \tau) \propto \text{prior}(\alpha, \beta, \tau) \cdot \text{likelihood}(\alpha, \beta, \tau)$.

Say priors are $\alpha \sim N(r, u)$ [here u is the precision parameter], $\beta \sim N(s, v)$, and $\tau \sim G(g, h)$.
Posterior is proportional to

$$\exp\left(-\frac{u}{2}(\alpha - r)^2\right) \exp\left(-\frac{v}{2}(\beta - s)^2\right) \tau^{g-1} \exp(-h\tau) \tau^{n/2} \exp\left(-\frac{\tau}{2} \sum (y_i - \alpha - \beta x_i)^2\right).$$

Useful bit of algebra: Note if

$$f(z) \propto \exp\left(-\frac{a}{2}z^2 + bz\right) = \exp\left(-\frac{a}{2}\left(z^2 - 2\frac{b}{a}z\right)\right) \propto \exp\left(-\frac{a}{2}\left(z - \frac{b}{a}\right)^2\right) \text{ then } Z \sim N\left(\frac{b}{a}, a\right).$$

Given τ and β ,

$$\begin{aligned}
(\alpha \mid \tau, \beta) &\sim \exp\left(-\frac{u}{2}(\alpha - r)^2\right) \exp\left(-\frac{\tau}{2} \sum (y_i - \alpha - \beta x_i)^2\right) \\
&\propto \exp\left(-\frac{\alpha^2}{2}(u + \tau n) + \alpha(ur + \tau \sum (y_i - \beta x_i))\right) \\
\text{so } (\alpha \mid \tau, \beta) &\sim N\left(\frac{ur + \tau \sum (y_i - \beta x_i)}{u + \tau n}, u + \tau n\right).
\end{aligned}$$

Given τ and α ,

$$\begin{aligned}
(\beta \mid \tau, \alpha) &\sim \exp\left(-\frac{v}{2}(\beta - s)^2\right) \exp\left(-\frac{\tau}{2} \sum (y_i - \alpha - \beta x_i)^2\right) \\
&\propto \exp\left(-\frac{\beta^2}{2}(v + \tau \sum x_i^2) + \beta(vs + \tau \sum x_i(y_i - \alpha))\right), \\
\text{so } (\beta \mid \tau, \alpha) &\sim N\left(\frac{vs + \tau \sum x_i(y_i - \alpha)}{v + \tau \sum x_i^2}, v + \tau \sum x_i^2\right).
\end{aligned}$$

Given α and β ,

$$\begin{aligned}
(\tau \mid \alpha, \beta) &\sim \tau^{g-1} \exp(-h\tau) \tau^{n/2} \exp\left(-\frac{\tau}{2} \sum (y_i - \alpha - \beta x_i)^2\right) \\
&= \tau^{g+\frac{n}{2}-1} \exp\left(-\tau\left(h + \frac{1}{2} \sum (y_i - \alpha - \beta x_i)^2\right)\right), \\
\text{so } \tau &\sim G\left(g + \frac{n}{2}, h + \frac{1}{2} \sum (y_i - \alpha - \beta x_i)^2\right).
\end{aligned}$$

Here is an R program that performs the Gibbs sampler just as specified above:

```

x = data$crying
y = data$IQ
n = length(x)

# Hyperparameters
r=0; u=.000001
s=0; v=.000001
g=.001; h=.001

nit=10000
alphas = numeric(nit)
betas = numeric(nit)
taus = numeric(nit)

# initial values
alpha = alphas[1] = 0
beta = betas[1] = 0
tau = taus[1] = 1

for(i in 1:(nit-1)){
  alpha = rnorm(1,
    (u*r+tau*sum(y-beta*x)) / (u+tau*n),
    1/sqrt(u+tau*n))
  alphas[i+1] = alpha

  beta = rnorm(1,

```

```

      (v*s+tau*sum(x*(y-alpha))) / (v+tau*sum(x^2)),
      1/sqrt(v+tau*sum(x^2)))
    betas[i+1] = beta

    tau = rgamma(1, g+n/2, h+(1/2)*sum((y-alpha-beta*x)^2))
    taus[i+1]=tau
  }

```

The results look very much like what BUGS gives us. Indeed, you now understand how BUGS really does its calculations for this problem!

The R file on the web lets you watch the Gibbs sampler run, and shows how the behavior of the Gibbs sampler can be affected simply by adding a constant to the x variable (here, `crying`).

Note: Gibbs sampler as a special case of Metropolis-Hastings

Metropolis-Hastings can use any kind of proposal mechanism we dream up, as long as we accept and reject proposed moves according to the famous formula. For example, random walk Metropolis can go around mindlessly making “dumb” random walk proposals, and the proposal distribution is not particularly tuned to the desired distribution. We have seen at least that it is good at least to try to use a size or scale of proposed movements that are not too small or too large, but there’s not much more in the way of customization that we can think about with the random walk moves.

ID	AGE	CHD
1	20	0
2	23	0
3	24	0
4	25	0
5	25	1
6	26	0
7	26	0
8	28	0
9	28	0
10	29	0
.....		
91	60	0
92	60	1
93	61	1
94	62	1
95	62	1
96	63	1
97	64	0
98	64	1
99	65	1
100	69	1

We will see here is that the Gibbs sampler is a special case of Metropolis-Hastings, with a proposal distribution that tends to be “smarter” than generic random walk proposals. This is a rather typical tradeoff with MCMC methods: we can put more work into designing and executing clever proposals, in the hope that the better proposals will allow our chain to explore the space more effectively and produce a good sample from the desired distribution in fewer iterations.

Our claim in fact is that, if one takes Gibbs sampler moves as the proposal mechanism in the Metropolis-Hastings framework, then the Metropolis-Hastings acceptance probability is 1, and the moves are always accepted. This then is the Gibbs sampler method, which proposes Gibbs sampler type moves and doesn’t have to worry about whether or not to accept them.

To ease the notation, suppose we have just two coordinates and our desired density is $f(x, y)$. Say we’re updating y for fixed x . The Gibbs sampler proposes like this: $Q((x, y), (x, y')) = f(y' | x)$. Using this proposal, the Metropolis-Hastings acceptance probability works out to be

$$\min \left\{ 1, \frac{f(x, y') Q((x, y'), (x, y))}{f(x, y) Q((x, y), (x, y'))} \right\} = \min \left\{ 1, \frac{f(x, y') f(y | x)}{f(x, y) f(y' | x)} \right\} = 1.$$

4.11 Logistic Regression

- Coronary Heart Disease: a narrowing of the small blood vessels that supply blood and oxygen to the heart (coronary arteries).
- Alternate Names : Arteriosclerotic Heart Disease, CAD, CHD, Coronary Artery Disease

We have data on 100 people: ages, and whether or not they have CHD.
We want to investigate the relationship between age and CHD.

Model: Each person has a probability (p) of having CHD, which depends on the person’s age (x).
Depends how?

The *odds* for an event is $\frac{p}{1-p}$.

Many people talk in terms of odds *ratios*: smoking multiplies the odds of getting cancer by the factor of r , etc...

This suggests the log of the odds as a meaningful quantity.

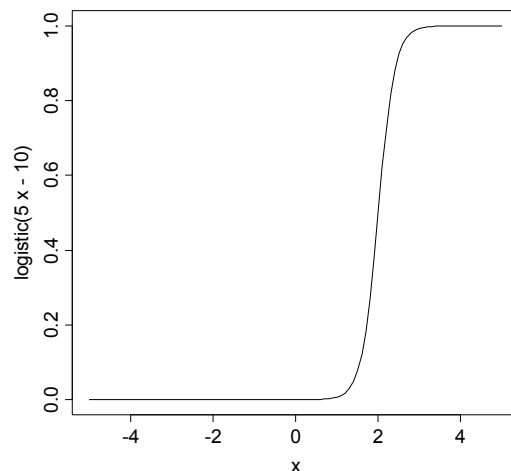
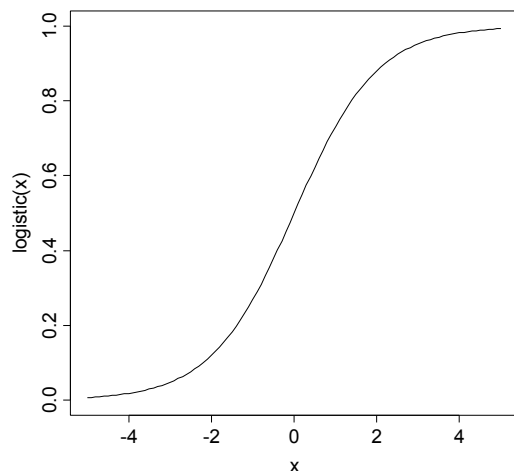
The “logit” function is the log odds: $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$.

In the logistic regression model, $\text{logit}(p)$ is a linear function of x , that is, $\text{logit}(p) = a + bx$.

Note saying $\log\left(\frac{p}{1-p}\right) = a + bx$ is equivalent to saying $p = \frac{e^{a+bx}}{1 + e^{a+bx}}$.

In other words, defining the *logistic* function to be the inverse of the logit, that is,

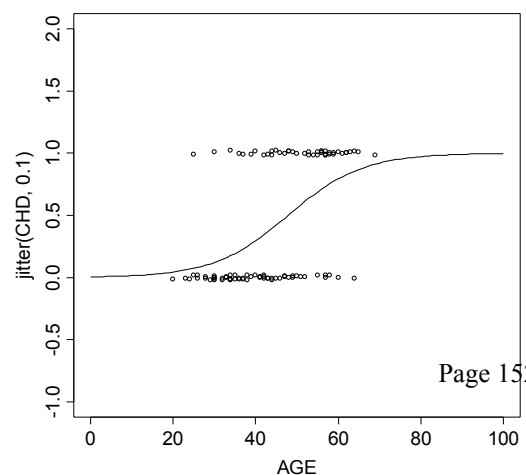
$\text{logistic}(y) = \frac{e^y}{1 + e^y} = \frac{1}{1 + e^{-y}}$, the model is: $p = \text{logistic}(a + bx)$.



E.g., if we increase $a + bx$ by 1, then p will increase along the logistic curve in a funny way... but the log odds will increase by 1. (That is, the odds will multiply by e .)

This is indeed a kind of regression. Recall in linear regression we modeled the mean of a variable Y as a linear function of another variable x . Here, we are modeling the probability of CHD as a function of age – this function, the logistic function, is nonlinear. In fact, here we are modeling the expectation of a variable as a function of another variable also, just as we were doing before with linear regression, since the probability is an expectation of the indicator variable (CHD) that we are observing! [You remember that the expectation of an indicator variable is the probability that the indicator variable is 1.]

How do we fit these nice curves? Least squares?



Well, sum of squared residuals has nothing particular to do with this model... We can use likelihood, etc., and fit this model to data in BUGS, using the same principles we have used many times already.

Here is a BUGS model file “logistic.bug” that expresses the simple logistic regression model:

```
model{
  for(i in 1:n){
    y[i] ~ dbern(p[i])
    logit(p[i]) <- a + b*x[i]
  }
  a ~ dnorm(0, .0001)
  b ~ dnorm(0, .0001)
}
```

Note: the unusual syntax in the line

```
logit(p[i]) <- a + b*x[i]
```

could be replaced by

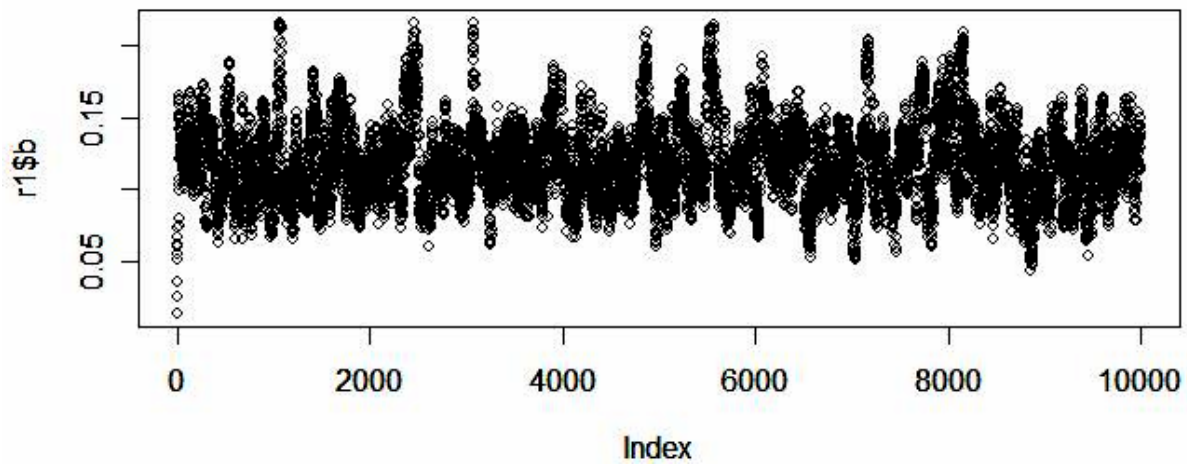
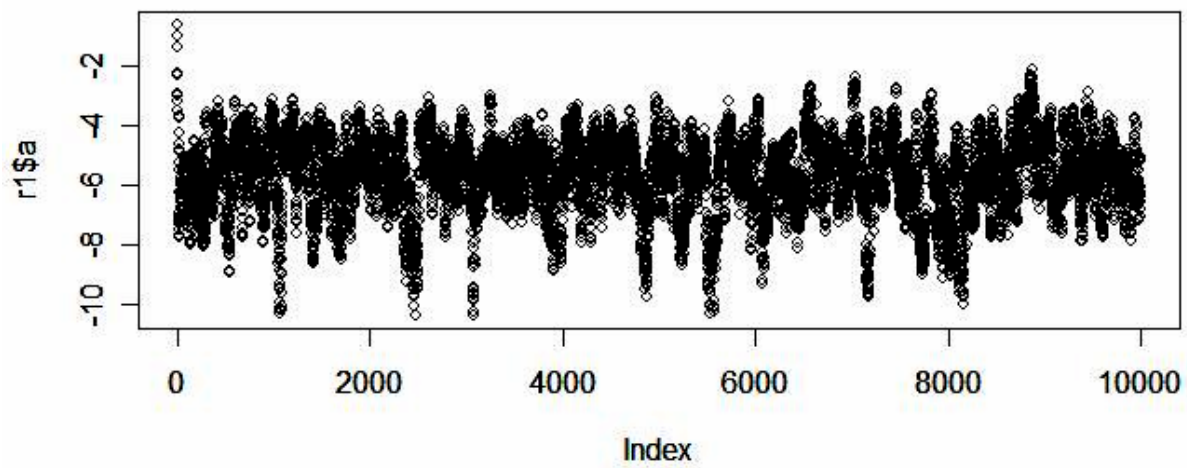
```
p[i] <- 1 / (1 + exp(-a-b*x[i])),
```

which is equivalent.

The R file for today has the details for running BUGS, which I hope are becoming familiar. I assume we have already created the data variables `age` and `chd` in R.

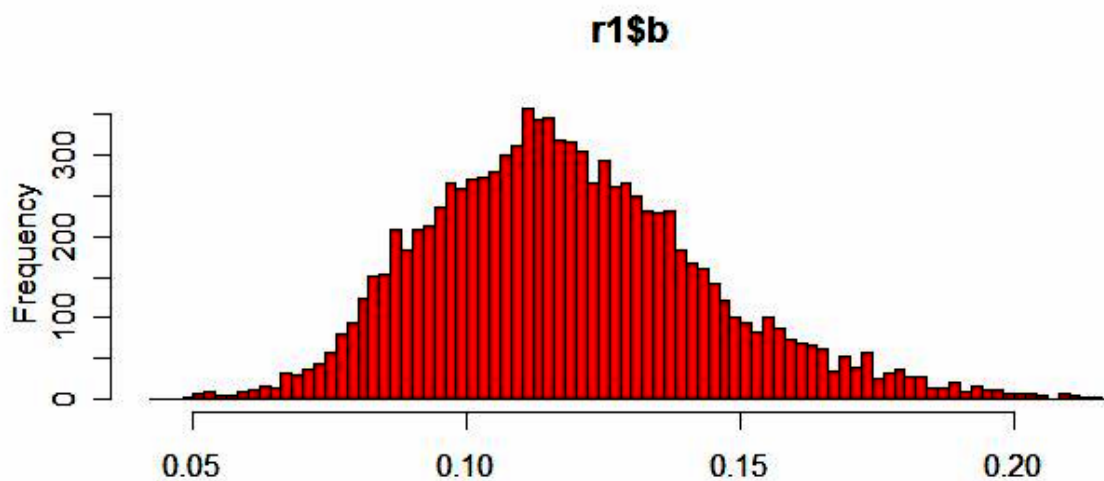
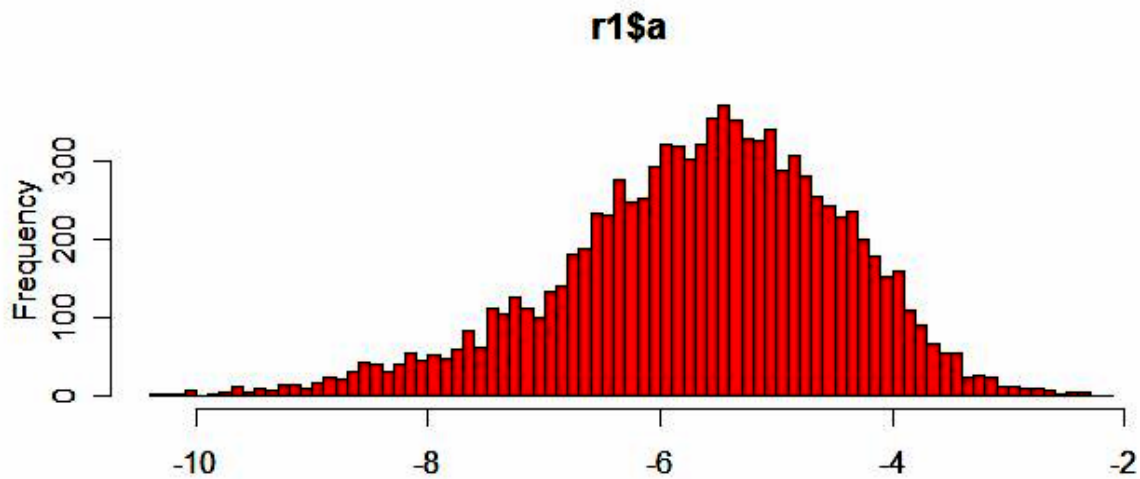
```
library(BRugs)
source("c:/_jtc/R/mybrugs.r")
y = chd
x = age
n = 100
dat = list("x", "y", "n")
inits = list(a=0, b=0)
params = c("a", "b")
r1 = mybrugs("logistic.bug", dat, inits, params, 10000)

par(mfrow=c(2,1))
plot(r1$a)
plot(r1$b)
```



Let's throw out the first 100 iterations as burnin, and make histograms for the posterior distributions of a and b .

```
myhist(r1$a)
myhist(r1$b)
```

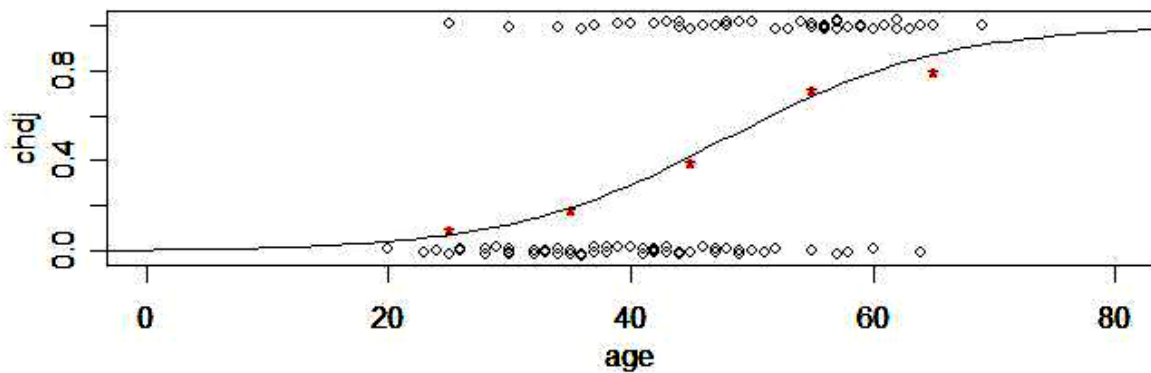


The posterior probability that $b > 0$ is approximated by

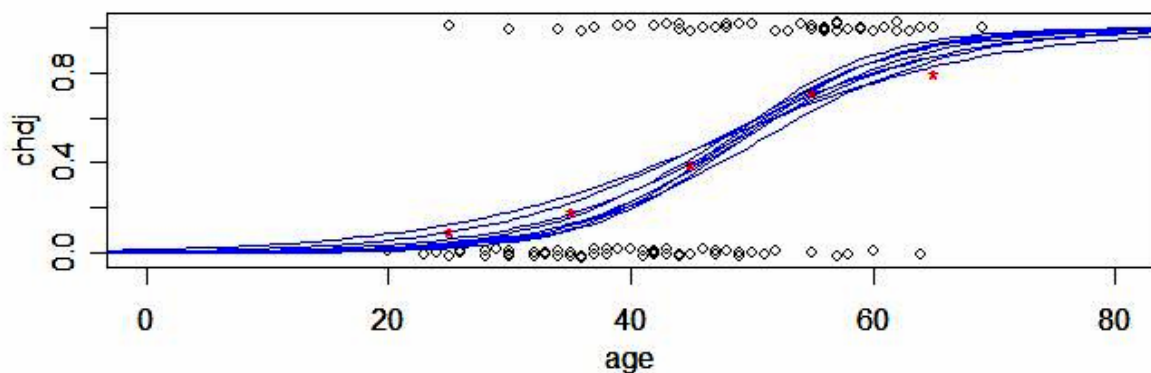
```
> mean(r1$b > 0)
[1] 1
```

I'll include a couple more plots we discussed in class, and the R code that produced them:

```
# the minimum deviance (max likelihood) solution found by the MC:
j = match(min(r1$dev), r1$dev)
chdj = jitter(chd,.1)
plot(age,chdj,xlim=c(0,80))
curve(logistic(r1$a[j] + r1$b[j]*x), add=T)
for(i in 2:6){
  ind = (age >= 10*i)&(age < 10*(i+1))
  points(10*i+5, mean(chd[ind]), pch="*", col="red")
}
```



```
# Several typical logistic fits from the posterior distribution:
plot(age, chdj, xlim=c(0,80))
curve(logistic(amed + bmed*x), add=T)
its = seq(5000,9000,by=500)
for(j in its){
  curve(logistic(r1$a[j] + r1$b[j]*x), add=T, col="blue")
}
for(i in 2:6){
  ind = (age >= 10*i)&(age < 10*(i+1))
  points(10*i+5, mean(chd[ind]), pch="*", col="red")
}
```



4.12 Hidden Markov models

These have been used with great success in a number of applications, including speech recognition and analysis of biological sequences (e.g. “parsing” a DNA sequence into genes, etc.).

Our friend the Frog is hopping in typical Markovian fashion among two lily pads, cunningly called #1 and #2. But now each lily pad has a die (singular of dice) sitting on it, and whenever the frog visits a lily pad, he rolls that pad's die once and calls out the result. However, all of this is happening inside a closed box, and we do not get to observe which lily pad, or "state," the frog is visiting at any time. Moreover, the two lily pads' dice are potentially "loaded" in different ways, giving different probabilities for the outcomes 1,2,3,4,5,6.

Again, we get to observe the outcomes of all the rolls of the dice. We do not get to observe which state the frog is in, we do not the transition matrix of the frog's Markov chain, and we do not know the two arbitrary distributions of the loaded dice. All we get to do is hear the froggy voice calling out one number after another.

For example, this might be a record of the results of the first 400 turns:

```

6 4 2 4 3 1 2 6 5 6 2 6 3 6 3 4 3 4 2 1 5 2 4 3 5 1 5 6 3 3
4 1 3 1 6 6 2 1 2 6 3 4 2 6 6 6 3 1 2 6 5 1 3 1 3 6 3 4 3 6
5 1 3 4 2 4 5 6 5 4 3 6 2 5 2 1 4 4 2 4 2 1 5 1 2 4 3 1 2 4
2 6 2 6 6 5 5 6 5 4 5 5 2 6 2 1 2 4 2 6 5 1 5 4 5 2 3 3 2 4
3 1 3 1 3 6 5 1 5 1 5 4 5 1 2 2 5 6 3 4 2 2 1 5 1 3 3 4 1 6
4 2 1 5 6 3 6 3 6 5 4 5 6 3 2 5 6 6 1 5 4 4 4 3 6 2 4 2 4 5
4 2 3 6 4 3 4 1 6 6 1 5 4 5 2 5 4 3 4 2 1 2 1 4 1 2 6 3 1 5
1 5 1 3 1 2 6 2 6 3 6 2 1 5 6 2 6 2 4 3 4 2 4 3 3 2 4 6 3 6
1 2 5 6 6 1 2 1 4 4 4 1 3 6 3 4 3 1 4 4 1 4 6 3 6 2 5 3 1 3
6 2 6 2 6 3 1 5 4 3 4 1 4 5 6 5 1 3 3 2 1 2 1 4 6 5 1 2 4 5
4 2 2 5 6 3 6 3 6 5 6 5 1 2 1 5 1 5 1 2 4 4 2 1 2 5 3 6 5 6
2 1 3 1 3 4 3 1 2 1 3 6 5 4 5 1 3 6 5 4 3 6 5 6 2 4 2 1 5 1
3 4 3 6 3 4 3 4 5 6 2 6 3 6 2 1 6 6 3 4 6 3 4 5 5 3 5 3 6 3
1 3 4 5 6 3 1 5 6 5

```

Or, here's another, maybe easier to look at:

```

> y
[1] 2 2 3 5 5 3 5 1 6 1 4 1 2 1 6 1 6 1 6 6 4 5 6 4 6
[26] 4 6 4 1 6 4 5 6 1 1 1 4 5 4 1 1 6 1 4 1 4 4 6 3 6
[51] 4 5 6 6 6 4 1 6 4 1 6 6 4 2 6 4 6 4 4 6 1 6 4 6 6
[76] 2 5 6 4 3 4 6 4 2 4 4 6 6 5 5 2 3 2 6 1 6 6 4 1 5
[101] 4 6 6 4 4 1 2 1 4 6 2 6 1 4 1 6 3 2 2 2 3 2 2 5 2
[126] 3 3 2 2 5 2 3 5 3 2 5 3 5 2 4 3 5 1 3 3 5 2 3 2 3
[151] 4 5 2 3 3 3 5 2 2 2 5 2 2 3 2 2 2 2 2 2 5 5 5 2 3
[176] 3 5 2 3 3 2 5 2 3 6 2 3 2 2 3 5 2 6 5 2 2 5 3 2 3
[201] 2 4 4 3 5 6 5 6 1 1 4 6 1 6 4 6 4 4 1 4 4 6 4 4 1
[226] 1 1 1 1 6 1 1 4 1 4 6 4 6 1 1 2 1 1 1 1 4 1 4 1 6
[251] 6 6 6 1 5 6 4 1 4 3 4 1 1 1 6 6 4 4 4 4 4 1 5 1 1
[276] 1 6 6 6 6 1 6 4 4 4 6 6 4 5 6 6 5 1 6 6 6 4 1 4 5
[301] 4 3 3 4 5 2 2 3 2 2 2 5 3 3 2 5 2 3 3 4 2 2 3 2 5
[326] 5 5 5 5 5 3 5 6 1 2 2 5 2 6 6 3 4 1 3 2 5 6 2 3 3
[351] 3 5 2 2 2 5 3 5 2 3 5 3 5 3 5 3 2 5 6 5 1 3 5 2 3
[376] 3 3 6 5 2 5 2 5 2 5 5 2 5 2 3 3 2 3 5 3 6 2 4 1 4

```

Aren't you curious? What are the two distributions of the dice, and what is the unobserved path of the frog?

We assume the sequence of states of the frog is a Markov chain. Let's call the two states are "1" and "2"

and let's say the transition matrix is
$$A = \begin{matrix} & \begin{matrix} 1 & 2 \end{matrix} \\ \begin{matrix} 1 \\ 2 \end{matrix} & \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} \end{matrix}.$$

So there is an unobserved ("hidden") Markov chain X_0, X_1, \dots, X_{399} that is making transitions according to the matrix A .

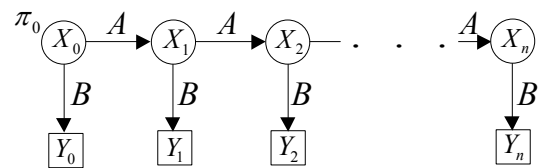
[And let's assume the chain has initial distribution $\pi_0 = (1/2 \ 1/2)$.]

The distribution of the 6 possible outcomes from die #1 is unknown, and the same for die #2. We can think of each distribution as a vector of the form $b = (b_1, b_2, \dots, b_6)$. Let's stack them into a matrix with

two rows:
$$B = \begin{pmatrix} b_{11} & b_{12} & b_{13} & b_{14} & b_{15} & b_{16} \\ b_{21} & b_{22} & b_{23} & b_{24} & b_{25} & b_{26} \end{pmatrix}$$

The unknown objects are the parameters p, q , and the matrix B , and also the unobserved states X_0, X_1, \dots, X_{399} of the hidden chain. We are interested in all of them.

Incidentally, here's a diagram of the hidden Markov model as a "graphical model" or "Bayesian network":



Here the matrix $B = \begin{pmatrix} b_{11} & b_{12} & b_{13} & b_{14} & b_{15} & b_{16} \\ b_{21} & b_{22} & b_{23} & b_{24} & b_{25} & b_{26} \end{pmatrix}$ governs transitions from the state X_k to the observation Y_k .

THE ANSWER – here is how I simulated the data:

```
B = rbind(c(.03,.3,.3,.03,.3,.04), c(.3,.03,.03,.3,.04,.3))
sim.hmm = function(n,p,q,B){
  A = matrix(c(1-p,p,q,1-q),ncol=2,byrow=TRUE)
  x=sample(2,1)
  for(i in 2:n){
    x[i] = sample(2,1,prob = A[x[i-1],])
  }
  y = numeric(n)
  for(i in 1:n){
    y[i] = sample(6,1,prob = B[x[i],])
  }
  return(list(x=x,y=y))
}
hmm2.dat = sim.hmm(400,.02,.02,B)
```

That is, the parameters I used were: $p = .02$, $q = .02$, $B = \begin{pmatrix} .03 & .3 & .3 & .03 & .3 & .04 \\ .3 & .03 & .03 & .3 & .04 & .3 \end{pmatrix}$.

And here are the hidden states: state 2 is underlined.

2235535161412161616645646464164561114541
 1614144636456664164166426464461646625643
 4642446655232616641546644121462614163222
 3225233225235325352435133523234523335222
 522322222555233523325236232235265225323
 2443565611461646441446441111161141464611
 211114141666615641434111664444151116666
 1644466456651666414543345223222533252334
 223255555356122526634132562333522253523
 5353532565135233365252525525233235362414

Of course we can't expect or hope to reconstruct the truth completely accurately! E.g. one would imagine any "sensible" predictor, having guessed that state 2 (underlined) has a low probability for 2 and state 1

has a higher probability for 2, would not guess that the last roll in the second underlined segment, a "3", comes from state 2, right?

There are many sets of parameters and reconstructions of the hidden chain that are roughly consistent with the data. We want to make guesses that accurately reflect our uncertainty while incorporating whatever information is in the data.

You will not find this model as a menu item in any common statistical package, and fitting this model to data is usually considered to be an advanced topic that requires specialized knowledge.

How can get BUGS to do this? Here is a BUGS model file that we could use:

```
model{
# prior for A:
  A[1,1] <- 1-p
  A[1,2] <- p
  A[2,1] <- q
  A[2,2] <- 1-q
  p ~ dunif(0,1)
  q ~ dunif(0,1)

# prior for B:
  for(j in 1:6){
    b1[j] ~ dexp(1)
    B[1,j] <- b1[j] / sum(b1[])
    b2[j] ~ dexp(1)
    B[2,j] <- b2[j] / sum(b2[])
  }
# model for x and y:
  x[1] ~ dcat(pi0[])
  for(i in 2:n){
    x[i] ~ dcat(A[x[i-1],])
  }
  for(i in 1:n){
    y[i] ~ dcat(B[x[i],])
  }
}
```

[[Note: on the web there is a model file that is expressed using different distributions but is in fact equivalent. The model file on the web uses Dirichlet distributions for the rows of the matrix B , as discussed in class.]]

We can run this in BUGS as usual.

```
n = 400
pi0 = c(.5,.5)
dat = c("n","pi0","y")
inits = list(p=.5,q=.5)
params = c("x","B","p","q")
r1 = mybugs("hmm.bug", dat, inits, params, 10000)
```

The R file on the web shows in detail some things we can do at this point. For example, a look at the results suggests throwing away an initial segment of the chain as “burn-in” since BUGS seems to have been wandering around, searching for the right region of parameter space during that time. We can throw away the first 1000 iterations like this:

```
r = r1[-(1:1000),]
```

At this point `r` is a big data frame with these columns:

```
> names(r1)
[1] "B.1.1." "B.1.2." "B.1.3." "B.1.4." "B.1.5." "B.1.6."
[7] "B.2.1." "B.2.2." "B.2.3." "B.2.4." "B.2.5." "B.2.6."
[13] "deviance" "p" "q" "x.1." "x.2." "x.3."
[19] "x.4." "x.5." "x.6." "x.7." "x.8." "x.9."
[25] "x.10." "x.11." "x.12." "x.13." "x.14." "x.15."
[31] "x.16." "x.17." "x.18." "x.19." "x.20." "x.21."

[403] ". . . . ."
[409] "x.388." "x.389." "x.390." "x.391." "x.392." "x.393."
[415] "x.394." "x.395." "x.396." "x.397." "x.398." "x.399."
[415] "x.400."
```

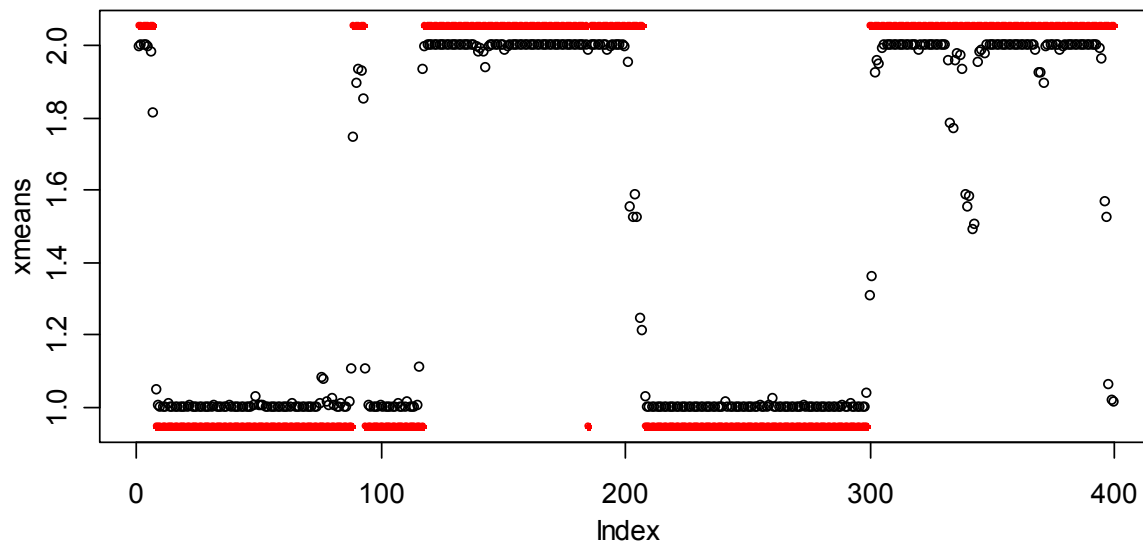
Let's take a look at some things as see how we did. For example, we could look at the posterior mean of the 12 elements of the matrix `B`, as follows.

```
> Bmeans = apply(r[,1:12], 2, mean)
> Bmeans = matrix(Bmeans, nrow=2, byrow=T)
> round(Bmeans,2)
      [,1] [,2] [,3] [,4] [,5] [,6]
[1,] 0.27 0.04 0.02 0.29 0.06 0.32
[2,] 0.02 0.36 0.29 0.03 0.27 0.04
```

Hey, that looks pretty good – one state with probabilities mainly on $\{1,4,6\}$, and another state with probabilities mainly on $\{2,3,5\}$! [[The states happen to have been labelled opposite to the labels I chose when I simulated the data, but this is unavoidable (until BUGS learns to read minds) and not a problem.]] Here is a plot of our estimated (mean) `x`'s and the true `x`'s (drawing lines a bit below 1 and a bit above 2 when the true state is 1 and 2)

```
xmeans = apply(r[,16:415], 2, mean)

plot(xmeans, ylim=c(.95,2.05))
points(1:400, 3-c(.95,2.05)[x], pch=20, col="red")
```



It worked quite well!

Suppose we decide to estimate the state as “1” or “2” depending on whether the posterior mean of the state is greater than 1.5 or less than 1.5:


```
> xest = 1+(xmeans <= 1.5)
> table(xest, x)
      x
xest   1   2
  1 194   2
  2   8 196
```

So we make just 10 mistakes out of the 400.

Here is our reconstruction of the hidden states according to their posterior probabilities:

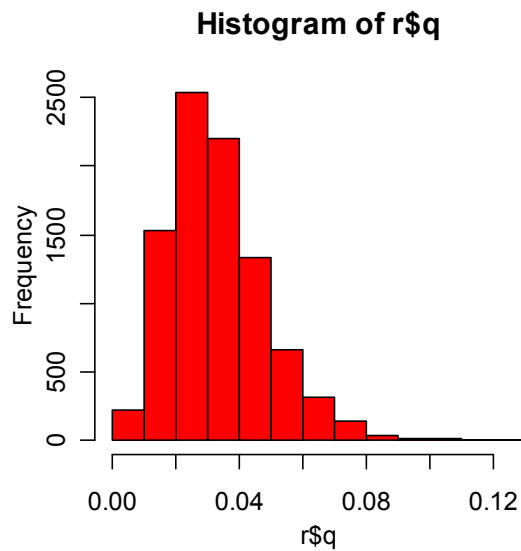
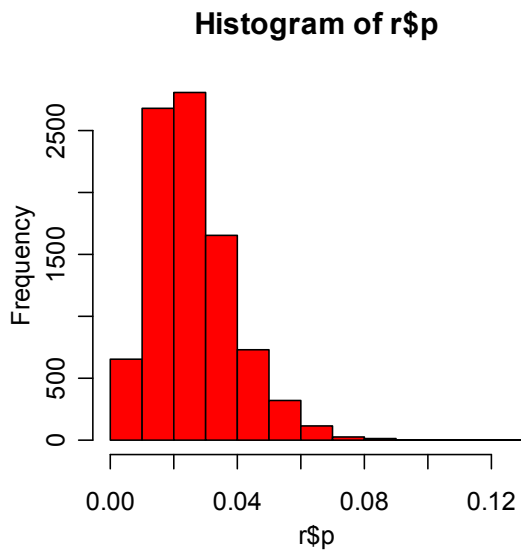
```
2 2 3 5 5 3 5 1 6 1 4 1 2 1 6 1 6 1 6 6 4 5 6 4 6 4 6 4 1 6 4 5 6 1 1 1 4 5 4 1
1 6 1 4 1 4 4 6 3 6 4 5 6 6 6 4 1 6 4 1 6 6 4 2 6 4 6 4 4 6 1 6 4 6 6 2 5 6 4 3
4 6 4 2 4 4 6 6 5 5 2 3 2 6 1 6 6 4 1 5 4 6 6 4 4 1 2 1 4 6 2 6 1 4 1 6 3 2 2 2
3 2 2 5 2 3 3 2 2 5 2 3 5 3 2 5 3 5 2 4 3 5 1 3 3 5 2 3 2 3 4 5 2 3 3 3 5 2 2 2
5 2 2 3 2 2 2 2 2 2 5 5 5 2 3 3 5 2 3 3 2 5 2 3 6 2 3 2 2 3 5 2 6 5 2 2 5 3 2 3
2 4 4 3 5 6 5 6 1 1 4 6 1 6 4 6 4 4 1 4 4 6 4 4 1 1 1 1 1 6 1 1 4 1 4 6 4 6 1 1
2 1 1 1 1 4 1 4 1 6 6 6 6 1 5 6 4 1 4 3 4 1 1 1 6 6 4 4 4 4 1 5 1 1 1 6 6 6 6
1 6 4 4 4 6 6 4 5 6 6 5 1 6 6 6 4 1 4 5 4 3 3 4 5 2 2 3 2 2 2 5 3 3 2 5 2 3 3 4
2 2 3 2 5 5 5 5 5 3 5 6 1 2 2 5 2 6 6 3 4 1 3 2 5 6 2 3 3 3 5 2 2 2 5 3 5 2 3
5 3 5 3 5 3 2 5 6 5 1 3 5 2 3 3 3 6 5 2 5 2 5 2 5 5 2 5 2 3 3 2 3 5 3 6 2 4 1 4
```

In this picture, the color of the number corresponds to our guessed state, and the underlining indicates the true state. You can see the 10 misclassified states, and it all looks pretty sensible.

An attractive aspect of our Bayesian approach is that, contained within the MCMC iterations is information that allows us to generate probabilistic estimates for all kinds of possible quantities of interest quite painlessly.

For example, for the posterior distribution of p and q (the probabilities of the frog changing jumping from one state to another, which were taken to be 0.02 in the simulation) we can simply draw histograms of the p and q values produced by BUGS.

```
> hist(r$p,col="red")
> hist(r$q,col="red")
```

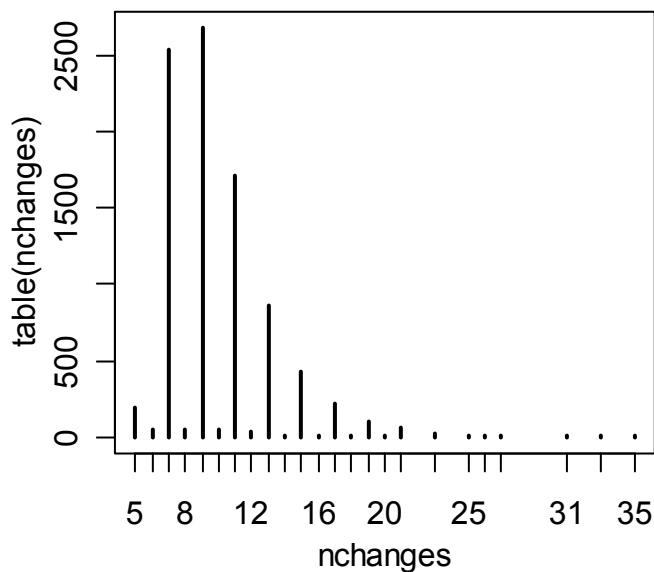


```
> mean(r$p < .02)
[1] 0.3702634
> mean(r$q < .02)
[1] 0.1943549
```

So the “true” values of 0.02 fall well within our posterior distributions, and are captured in, for example, 95% posterior probability intervals.

As another example of this, suppose we are interested in how many changes of state there were in the hidden Markov chain (that is, how many basketball shots were missed). Here is a picture of our posterior probabilities for the various possible number of changes.

```
nchange = function(x){
  sum(diff(x) != 0)
}
# A little example to see how it works
> temp=rbinom(7,1,.5)
> temp
[1] 1 0 1 0 1 0 1
> nchange(temp)
[1] 6
> # Now apply it to our simulated "x" sequences
> nchanges = apply(r[,16:415],1,nchange)
>
> plot(table(nchanges))
```



The actual number of changes was 8:

```
> nchange(hmm2.dat$x)
[1] 8
```

Hmm (interjection, not hidden markov model), that seems kind of curious. This is pretty much in the “middle” of our posterior distribution, although for some reason WinBUGS ended up being quite convinced that the number of changes is odd, and put small posterior probability on 8 changes. But that makes sense – the beginning of our data (y) is 2 2 3 5 5 3 5 which is quite clearly coming from state 1, whereas the last three y observations are 4 1 4, which look quite convincingly to be from state 2. It just happens that they represent an improbable occurrence that happened while in state 1! So we would expect to make a mistake of this sort – something weird happened in our simulation, and our best probabilistic guess is to guess that such a weird thing did not happen. [In fact the sequence of 3 states 4 1 4 is *one thousand times more likely* under state 2 than it is under state 1.]

Just for fun, here is another set of example data.

```
> hmm3.dat$y
[1] 6 4 2 4 3 1 2 6 5 6 2 6 3 6 3 4 3 4 2 1 5 2 4 3 5 1 5 6 3 3
[31] 4 1 3 1 6 6 2 1 2 6 3 4 2 6 6 6 3 1 2 6 5 1 3 1 3 6 3 4 3 6
[61] 5 1 3 4 2 4 5 6 5 4 3 6 2 5 2 1 4 4 2 4 2 1 5 1 2 4 3 1 2 4
[91] 2 6 2 6 6 5 5 6 5 4 5 5 2 6 2 1 2 4 2 6 5 1 5 4 5 2 3 3 2 4
[121] 3 1 3 1 3 6 5 1 5 1 5 4 5 1 2 2 5 6 3 4 2 2 1 5 1 3 3 4 1 6
[151] 4 2 1 5 6 3 6 3 6 5 4 5 6 3 2 5 6 6 1 5 4 4 4 3 6 2 4 2 4 5
[181] 4 2 3 6 4 3 4 1 6 6 1 5 4 5 2 5 4 3 4 2 1 2 1 4 1 2 6 3 1 5
[211] 1 5 1 3 1 2 6 2 6 3 6 2 1 5 6 2 6 2 4 3 4 2 4 3 3 2 4 6 3 6
[241] 1 2 5 6 6 1 2 1 4 4 4 1 3 6 3 4 3 1 4 4 1 4 6 3 6 2 5 3 1 3
[271] 6 2 6 2 6 3 1 5 4 3 4 1 4 5 6 5 1 3 3 2 1 2 1 4 6 5 1 2 4 5
[301] 4 2 2 5 6 3 6 3 6 5 6 5 1 2 1 5 1 5 1 2 4 4 2 1 2 5 3 6 5 6
[331] 2 1 3 1 3 4 3 1 2 1 3 6 5 4 5 1 3 6 5 4 3 6 5 6 2 4 2 1 5 1
[361] 3 4 3 6 3 4 3 4 5 6 2 6 3 6 2 1 6 6 3 4 6 3 4 5 5 3 5 3 6 3
[391] 1 3 4 5 6 3 1 5 6 5
```

What model do you think produced this data?

I did a short run of BUGS and here is the reconstruction it gave. The "truth" is given by the underlining, and the BUGS estimates are given by the colors, with red meaning state 2.

```

6 4 2 4 3 1 2 6 5 6 2 6 3 6 3 4 3 4 2 1 5 2 4 3 5 1 5 6 3 3 4 1 3 1 6 6 2 1 2 6
3 4 2 6 6 6 3 1 2 6 5 1 3 1 3 6 3 4 3 6 5 1 3 4 2 4 5 6 5 4 3 6 2 5 2 1 4 4 2 4
2 1 5 1 2 4 3 1 2 4 2 6 2 6 6 5 5 6 5 4 5 5 2 6 2 1 2 4 2 6 5 1 5 4 5 2 3 3 2 4
3 1 3 1 3 6 5 1 5 1 5 4 5 1 2 2 5 6 3 4 2 2 1 5 1 3 3 4 1 6 4 2 1 5 6 3 6 3 6 5
4 5 6 3 2 5 6 6 1 5 4 4 4 3 6 2 4 2 4 5 4 2 3 6 4 3 4 1 6 6 1 5 4 5 2 5 4 3 4 2
1 2 1 4 1 2 6 3 1 5 1 5 1 3 1 2 6 2 6 3 6 2 1 5 6 2 6 2 4 3 4 2 4 3 3 2 4 6 3 6
1 2 5 6 6 1 2 1 4 4 4 1 3 6 3 4 3 1 4 4 1 4 6 3 6 2 5 3 1 3 6 2 6 2 6 3 1 5 4 3
4 1 4 5 6 5 1 3 3 2 1 2 1 4 6 5 1 2 4 5 4 2 2 5 6 3 6 3 6 5 6 5 1 2 1 5 1 5 1 2
4 4 2 1 2 5 3 6 5 6 2 1 3 1 3 4 3 1 2 1 3 6 5 4 5 1 3 6 5 4 3 6 5 6 2 4 2 1 5 1
3 4 3 6 3 4 3 4 5 6 2 6 3 6 2 1 6 6 3 4 6 3 4 5 5 3 5 3 6 3 1 3 4 5 6 3 1 5 6 5

```

In fact, to simulate this sequence, I kept the B matrix the same, so that states "1" and "2" have the same meaning they had before, but I changed the jumping probabilities p and q from .02 to .98. So the state switches nearly every time.

This one seems harder for our eyes to see the pattern than it was in the last one, which had the long stretches favoring $\{2,3,5\}$ or favoring $\{1,4,6\}$, but for BUGS this example is just as easy as the previous example.

4.13 Epilogue

Where we have been: A Stat 238/538 retrospective

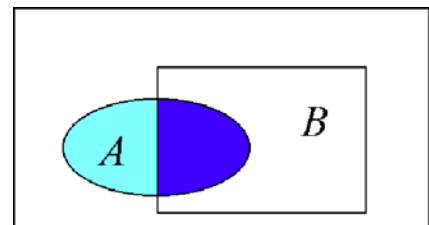
This was a course about two related subjects: Probability and Statistics. We were interested in Probability as a subject for its own sake, and also as the main tool in Statistics. We learned the basic definitions and rules about probability, what's a random variable, distribution, etc.

We started using R right from the beginning, as a calculator, to draw graphs, and to do simulations (e.g. the birthday problem, matching problem, etc.) through repeating random calculations many times.

A key definition was for conditional probability: $P(B | A) = \frac{P(A \cap B)}{P(A)}$

From this flowed the way we multiply probabilities, such as $P(ABC) = P(A)P(B | A)P(C | AB)$, the definition of independence, the law of total probability, and Bayes' rule, which we used throughout the course.

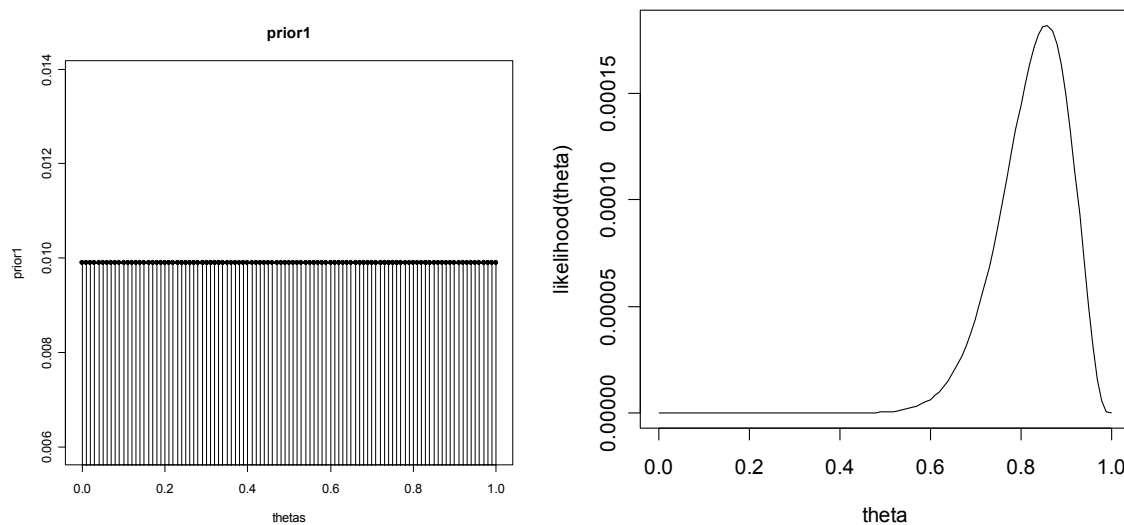
Definition: The *likelihood* is the probability of the observed data, as a function of the unknown parameter.



Bayes' rule: $\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$

In loose notation: $p(\theta | x) \propto p(\theta)p(x | \theta)$

The Statistical cat was let out of the bag with our first use of Bayes' rule to do a statistical inference problem about a clinical trial: $X \sim B(21, \Theta)$. We observed $X = 18$. Now what do we think about Θ ?



An example of a prior is above left. The Likelihood: $p(X=18|\theta) = \binom{21}{18} \theta^{18} (1-\theta)^3 \propto \theta^{18} (1-\theta)^3$ is above, right. The Posterior is below.

Important point: Our "inferences" in Bayesian statistics are simply calculations of probabilities. For example, for the drug our question of interest was the probability that the drug is better than the placebo, that is, the posterior probability that $\Theta > 0.5$, that is, $P\{\Theta > 0.5 | X = 18\}$.

So we did more about probability, again both for its own sake and since doing Statistics is all about working with probability models.

We discussed more distributions: uniform, exponential, geometric, Normal, Cauchy.

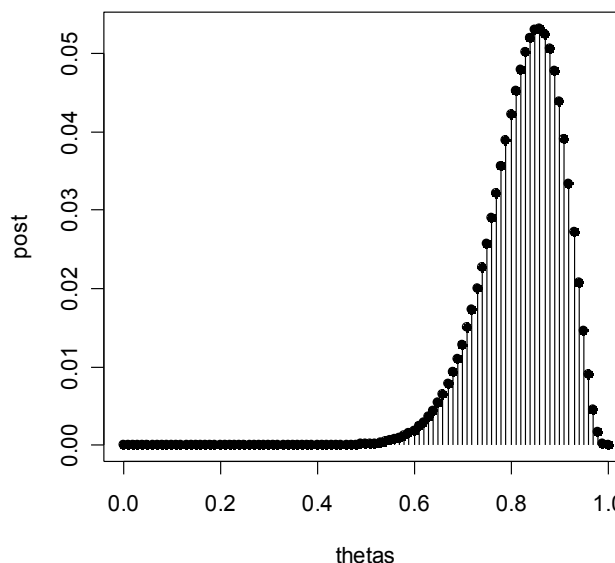
The Cauchy led to the "laser pole" problem, a nice example of a non-obvious statistical inference.

More about distributions. Expectation (or mean): definition, motivation for the definition from the Law of Large Numbers, Law of the Unconscious Statistician (how to find expectations of functions of random variables or vectors).

The indicator variable trick for calculating expectations involving counts.

Variance, standard deviation, and calculations such as:

Suppose X_1, X_2, \dots, X_n are iid – "independent and identically distributed" – with mean μ and variance σ^2 (and so SD = σ). Define $S = X_1 + \dots + X_n$ and $\bar{X} = S/n$.



$$E(S) = E(X_1) + \dots + E(X_n) = n\mu, \text{ so } E(\bar{X}) = \mu.$$

$$\text{var}(S) = \text{var}(X_1) + \dots + \text{var}(X_n) = n\sigma^2.$$

$$\text{var}(\bar{X}) = \text{var}\left(\frac{S}{n}\right) = \frac{\text{var}(S)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \quad \text{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

The \sqrt{n} in the denominator shows that the sample mean using a larger sample is less variable than the mean of a smaller sample.

This, together with Chebyshev's inequality, was the key to the Law of Large Numbers:

$$\bar{X}_n \rightarrow \mu \text{ "in probability" as } n \rightarrow \infty, \text{ that is, for each } \varepsilon > 0, \lim_{n \rightarrow \infty} P\{|\bar{X}_n - \mu| < \varepsilon\} = 1.$$

That was one of the two biggest classical theorems in probability. The other is the Central Limit Theorem, which says that for large n , \bar{X}_n is approximately Normally distributed.

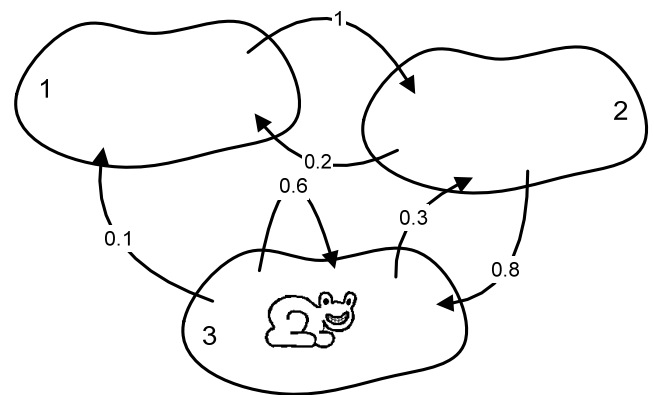
We studied the Normal distributions: density = exp(-quadratic function). We learned how to use the CLT to approximate probabilities involving sums and averages of random variables.

Our last big topic in probability theory was included both for its own interest and because it has become the main computational tool in Bayesian statistics: Markov chains.

We learned the definitions, saw several examples, learned to calculate π_t , the distribution at time t , through $\pi_t = \pi_0 P^t$.

The striking phenomenon was, e.g.,

```
> matpow(P, 50)
      [, 1]      [, 2]      [, 3]
[1,] 0.1176471 0.2941176 0.5882353
[2,] 0.1176471 0.2941176 0.5882353
[3,] 0.1176471 0.2941176 0.5882353
```



This says that, at large times the frog has probabilities (0.1176471 0.2941176 0.5882353) of being in the various states, no matter which state the frog started from at time 0.

Such a limiting probability is a stationary distribution π , satisfying the equation $\pi = \pi P$.

The Ergodic theorem said that, for an irreducible Markov chain having stationary distribution π , the long run fraction of time that the chain spends in state i converges to $\pi(i)$.

This is the key idea behind the use of Markov chains in simulation: to approximate probabilities for some desired density or pmf f , run a Markov chain whose stationary distribution is f for a long time. The resulting sequence X_0, \dots, X_n can be treated as a sample from μ . They are not independent of each other, but they can be used in approximating probabilities of the form $P_f(A)$ simply by counting to find the fraction of the X_t 's that fell in the set A .

To make this into a useful general tool, we needed a way, for any given density f , of constructing a Markov chain having f as its stationary density. This was the Metropolis method. Other variations also exist, like Gibbs sampling. An example of how Metropolis is done is as follows.

Given that we are at the current state $X_t = x$. Choose a *candidate* state, y , from a particular density centered on x . For example, we could take y from the uniform density $U(x - 1, x + 1)$.

- If $f(y) \geq f(x)$, accept the candidate; that is, take $X_{t+1} = y$.
- If $f(y) < f(x)$, accept the candidate with probability $\frac{f(y)}{f(x)}$, that is, take

$$X_{t+1} = \begin{cases} y & \text{with prob } f(y) / f(x) \\ x & \text{with prob } 1 - (f(y) / f(x)) \end{cases}$$

We learned why this recipe works, using the Markov chain theory developed in class. [The Metropolis chain is time reversible, making it easy to check that $f = fP$.]

We did several simple examples, and you did some for homework, including seeing how bad choices of the proposal distribution can slow the convergence of the Markov chain to its stationary distribution.

Then we turned back to Statistics and stayed there for the rest of the course. We studied some basic ideas of estimation: what is an estimator, what makes an estimator good and how to compare estimators (e.g. in terms of mean squared error), and again you did homework problems on this. You have also done homework on finding a maximum likelihood estimator. Again the likelihood function is a key to a principled approach to inference.

We went back to the Bayesian approach and saw some examples where posterior distributions, and point estimators such as posterior means, can be calculated simply with paper and pencil in closed form. These were the situations where we used "conjugate priors".

Most models that we want to use do not fall neatly into these special conjugate forms, and paper and pencil are not enough. But still, as in the **Important point** above, the Bayesian inference that we want is still conceptually the same: we typically just want a probability of some event involving Θ , conditional on the observed data X (or, more generally, we want an expectation of a random variable – e.g. the posterior mean is simply $E(\Theta | X)$). So here is our situation in a nutshell:

- To make our inference, we want to know a probability or an expectation.
- Probabilities and expectations can be approximated using simulation.
- We have this great MCMC tool for doing simulations.

We illustrated how this program can be carried through, doing everything from scratch in R. As our first example, we did an inference problem about the effect of subliminal messages on math test scores, and did a Metropolis method simulation to find that our posterior probability that the "treatment" mean was larger than the "control" mean was about 0.94 or so.

Next we learned how to use a new piece of software, BUGS, whose job is to help make implementing the above program easy and routine. To do inference using a given model on some given data, we just need to prepare a file of statements that specify the probability model, tell BUGS what the data is, and give it initial values for the Markov chain it is about to run for us. We saw how to run BUGS from R.

According to this approach, the process of statistical inference entails working with a probability model for the data. So learning more about inference and data analysis is largely a process of learning about how different types of data can be modeled. We went into some detail about the concepts behind linear regression models, including correlation, the "regression phenomenon" and "regression fallacy," and multiple regression. E.g. an example of multiple regression, done using BUGS, showed evidence from a given data set, that a company's salaries for males is higher on average than for females, and also that the salaries for males increase faster than those for females (that is, it is not just the intercepts that differ, but also the slopes).

We illustrated the process of inference with more examples, all using the same approach: logistic regression for the relationship between age and the probability of coronary heart disease, a mixture model for Martian male and female heights for your homework, and a hidden Markov model. If you have grasped these, you should feel happy with yourself; in a course taking a more standard approach, these would typically be considered separate topics, esoteric and difficult and beyond the scope of the course.

Thinking back on the hidden Markov model, for example, isn't it funny that such a simple-minded MCMC method can do the difficult detective work of extracting patterns out of data like this? The method starts from our initial guess, which might have no hint of the pattern we are looking for, and bumbles around randomly, stumbling upon improved likelihoods, moving toward the truth – that is, to the region where the bulk of the likelihood is – and sampling that region to give posterior distributions for unknown quantities that accurately reflect the precision and uncertainty of our knowledge.

This was not a condensed version of a standard year-long course, but rather took a different approach. Most people using Statistics will have learned some concepts that mirror some of ours, but are different, often backward in some sense from what we have done. Prime examples include confidence intervals and hypothesis tests. You can read about these in all kinds of standard books and on the web. Often the answers given by these different approaches will not differ a whole lot numerically (if they were done "well"), but they will differ conceptually. Often when asked to say what their confidence interval means, people trained in standard statistics will erroneously give the Bayesian interpretation because it is more natural. In a sense, Bayesian statistics uses a kind of backward probability result (Bayes' rule) and prior distributions to give conclusions that are conceptually natural, whereas nonBayesian statistics uses probability probability in the forward direction but makes contorted, turned-around conclusions.

Where to go next

I would recommend taking a look at the examples that are included with BUGS. They are good for browsing. In each example, some data is given with the context and scientific question described, and a model is built and analyzed with BUGS. You can actually run the examples by pointing and clicking, or from R. About courses: I have been imagining that for many of you, this will be your first and last statistics course. If you want to go on to something else next semester, you could consider Stat 230b or Stat 251/551; I can tell you more about those

* * * * *

I'll put a computational take-home exam on the web in several days and it will be due at the beginning of the "final exam" (really more like a second midterm). I'll send an email out to the class list I have on the classes server when the take-home exam is ready.

"Final exam" Time: Monday 12/18/09, 2:00 PM, Place: Not sure yet.

See you there and then!