Notes: Practical Statistics for Physics & Astronomy

R.B. Metcalf

Alma Mater Studiorum - Universitá di Bologna

February 15, 2018

Contents

1 What is Probability? 5
1.1 Frequentist 5
1.2 Bayesian 5
1.3 Quantum mechanical probability 5
1.4 the rules of probability 6

2 Some warm up problems 7
2.1 Rolling Dice 8
2.2 Birthday Paradox 9
2.3 Poker 10
2.4 The Monty Hall Problem 12

3 Probability distributions 13
3.1 properties of a probability distribution function (PDF) 13
3.2 mean, median, mode ... 13
3.3 moment generating function 15
3.3.1 changing of variables 15
3.4 Binomial and Bernoulli 15
3.4.1 drawing without replacement, the hypergeometric distribution 17
3.5 Poisson distribution 17
3.5.1 as a limit of the binomial distribution 18
3.6 Gaussian and normal 19
3.7 central limit theorem 20
3.7.1 The distribution of the sum of independent random variables 21
3.8 lognormal 23
3.9 Power law distribution 23
3.10 multivariant distributions 24
3.11 multinomial distributions 25
3.12 multivariant gaussian 25
3.12.1 conditional Gaussian distribution 27
3.12.2 marginalized Gaussian distribution 27
3.12.3 combining two multivariant Gaussians 28

3.13 χ^2 distribution . . . 28
3.14 student's t-distribution . . . 30

4 Sampling 31
4.1 estimating the mean . . . 31
4.2 estimating the variance . . . 34
4.3 estimating the mean when the variance is known . . . 35
4.4 median . . . 36
4.5 extreme values . . . 37
4.6 quintile estimation . . . 37

5 The Bayesian method 38
5.1 Posterior, likelihood, prior and evidence . . . 38
5.2 Parameter estimation . . . 39
5.2.1 example: . . . 39
5.2.2 example: . . . 40
5.2.3 example: . . . 42
5.3 Marginalization . . . 45
5.3.1 example: . . . 45
5.4 Choice of prior . . . 46
5.4.1 example: . . . 47
5.4.2 example: . . . 48
5.5 Model selection . . . 48
5.5.1 Occam's factor . . . 49
5.5.2 example: . . . 51
5.6 Calculating the evidence . . . 51
5.7 Example: A spectral line . . . 51
5.8 Example: The luminosity function . . . 51

6 Bayesian inference with Gaussian distributions & Gaussian approximations 52

7 Hypothesis testing & frequentist parameter fitting 53
7.0.1 mean of two populations are the same . . . 53
7.0.2 the variance of two populations are the same . . . 53
7.1 linear fitting, regression . . . 53
7.2 p-values, χ^2, degrees of freedom, etc. . . . 54
7.3 principle components . . . 54
7.4 nonlinear fitting . . . 54
7.5 hypothesis testing . . . 54
7.5.1 example: test isotropy of sources . . . 54
7.6 The early stop problem & shortcoming of frequenist hypothesis testing . . . 54

8 Frequentist non-parametric tests 55
8.1 Kolmogorov-Smirnov test . . . 55
8.2 rank statistics . . . 55
8.2.1 Wilcoxon test . . . 55
8.2.2 Spearman's correlation statistic . . . 55
8.3 bootstrap & jacknife sampling methods . . . 55

9 Frequentist vs Bayesian **55**
9.1 frequentist & Bayesian confidence levels . . . 55
9.1.1 Example: the highest redshift QSO . . . 55
9.2 numerical methods, Monte Carlo . . . 55
9.2.1 Creating a random number in a computer . . . 55

10 Estimators **56**
10.1 bias . . . 56
10.1.1 Example: Eddington-Malmquist bias . . . 56
10.2 maximum likelihood estimator . . . 56
10.3 least squares estimator . . . 56
10.3.1 weighting data . . . 56
10.4 minimum variance estimator . . . 56

11 The Fisher matrix and information **57**
11.1 the Fisher matrix . . . 57
11.1.1 the Gaussian case . . . 57
11.1.2 marginalized error estimates . . . 57
11.2 Rao-Cramer inequality . . . 57
11.3 information content of data . . . 57
11.3.1 different definitions of information . . . 57
11.3.2 Kullback–Leibler divergence (information gain) . . . 57
11.3.3 agreement/disagreement of data sets . . . 57
11.4 error forecasting . . . 57

12 Random Fields **58**
12.1 correlation function & power spectrum . . . 58
12.2 going between discrete and continuous variables . . . 58
12.3 Gaussian random fields . . . 58
12.4 Poisson noise . . . 58

13 Estimating correlation functions, power spectra & time-delays from irregularly sampled data **59**
13.1 power spectrum estimation . . . 59
13.2 data compression, Karhunen-Loeve modes . . . 59

14 Image reconstruction and map making **60**
14.1 Wiener filtering . . . 60
14.2 maximum entropy . . . 60
14.3 other filters & estimators . . . 60

15 Numerical methods for the Bayesian Inference Problem **61**
15.1 finding maximum likelihood . . . 61
15.2 Markov Chain Monte Carlo (MCMC) . . . 61
15.3 nested sampling & other strategies . . . 61
15.4 meaning & calculation of evidence . . . 61
15.5 displaying results graphically . . . 61

16 Classification and regression from a machine learning prospective 62

16.1 prediction vs inference 62

16.2 dividing objects into classes objectively: unsupervised learning 62

16.3 regression and classification: supervised learning 62

16.4 artificial neural networks & convolutional neural networks in astronomy 62

A Matrix basics 63

B notation 64

C Some Useful Integrals and mathematical definitions 64

1 What is Probability?

1.1 Frequentist

Imagine there is some event, instance or outcome of an experiment or observation called A. The probability of A is the fraction of times A occurs when the experiment or observation repeated in *exactly* the same way or circumstances and *infinite* number of times.

$$P(A) = \lim_{N\to\infty} \frac{\text{number of trials where A is true}}{N(\text{total number of trials})} \tag{1.1}$$

This is the traditional definition of probability as formally stated by Laplace in 1774 and used for centuries despite no one ever doing anything *exactly* the same twice or doing anything an *infinite* number of times.

Applying this definition to any physical phenomenon requires a partitioning of the world into things that are known and fixed on each repatriation of the observation and those things that are not known and change every repetition. If nature is deterministic and an experiment could be set up *exactly* the same way in all respects than the outcome would always be the same and probability would not apply. Of course even in classical physics it is not possible to know state of every atom and photon that might possibly influence your measurement apparatus (or brain). It is these things that change when repeating the observation.

This partitioning between known and unknown factors seems reasonable when we talk about the positions and momenta of particles in a gas or flipping a coin, but in many other common situations where probability is used it seems less well defined. Say someone tells you that there is a 30% probability that candidate A will win the election tomorrow. Of course an identical election will never be run again and was never run in the past. There are many factors, known and unknown, that could affect an election. This statement was probably based on polling data. By the above definition of probability, this means that if the election were held an infinite number of times in which the polling data were exactly the same the candidate would win a 3/10 of them.

***

1.2 Bayesian

Thomas Bayes (1701 - 1761) (and Jacob Bernoulli 1655-1705) had a different conception of what probability is although the idea was not put on a firm theoretical foundation until the 1940's and 50's by G. Polya, R.T. Cox and E.T. Jaynes.

In this school of thought, probability theory is an extension of formal logic and deductive reasoning to situations where the conclusions are not certainties but, given the prior knowledge, one outcome can be considered more plausible than another. Surprisingly from just a few axioms (3 to 6 depending on how you count) you can deduce the rules of probability and show that they are complete without ever mentioning randomness or repetition of experiments. These foundational proofs are very interesting, but outside the scope of this course (for those that are interested see chapter 2 of (Gregory, 2006) for an interesting discussion). One thing that is of importance here is that this definition allows one to define the probability of something that would not usually be considered a random variable. This is central to the Bayesian method of parameter estimation and model selection that we will get to later.

1.3 Quantum mechanical probability

The interpretation of probability as a measure certainty, or conversely ignorance,

1.4 the rules of probability

Suppose the A, B, ... are events that either occur or don't occur , that is they have values true or false (or 0 and 1 if you prefer). $P(A)$ is the probability of A occurring or being true. We can combine events in one of two ways. (A, B) means ”*A and B*”. It is true if both of them are true and false if both are false. $(A \cup B)$ means ”*A or B*” it is true if either A or B is true. It is true if both are true. $\overline{A}$ means ”not A”. Not that $(\overline{A \cup B}) = (\overline{A}, \overline{B})$ in the sense that there are no combinations of trues and falses for A and B that give different answers on either side of the equality. In the language of Boolean algebra, they have the same truth table and are there for the same statement.

$P(A, B)$ is often called the **joint probability** of events A and B. $P(A \cup B)$ is often called the **disjoint probability** of events A and B.

$P(A|B)$ is called a **conditional probability**. It means the probability of A *given* that B is true. You can imagine every probability being a conditional probability where it is ”conditioned” on everything that you assume about the state of the Universe. Some of these things are assumed to be irrelavent and are left out . Some might be relevant but it is taken for granted so they are left out. The probability that a coin comes up heads does not depend on the time of day. It does depend on the assumption that it is a fair coin - no more likely to be heads than tails - although it might not always be stated. This is a simple example of a **statistical model** for the experiment, in this case flipping a coin.

The two fundamental rules of probability theory are

$$\begin{array}{cl} P(A, B) = P(A)P(B|A) & \quad \textbf{product rule} \\ P(A) + P(\overline{A}) = 1 & \quad \textbf{sum rule} \end{array} \tag{1.2}$$

These rules are actually derivable from some basic requirements or ”desidariata” of how probabilities should behave, but for our purposes we can take them to be axioms. Form these two rules and logic rules we can derive all the necessary properties of probability.

There are several particularly useful results that follow from these rules. From the logical requirement that (A, B) is the same as (B, A) and the product rule we get

$$P(A|B) = \tfrac{P(A)P(B|A)}{P(B)} \qquad \textbf{Bayes' theorem} \tag{1.3}$$

Applying the sum rule to $(A \cup B)$ gives

$$P(A \cup B) = 1 - P(\overline{A \cup B}) \tag{1.4}$$
$$= 1 - P(\overline{A}, \overline{B}) \tag{1.5}$$
$$= 1 - P(\overline{A})P(\overline{B}|\overline{A}) \tag{1.6}$$
$$= 1 - P(\overline{A})\left[1 - P(B|\overline{A})\right] \tag{1.7}$$
$$= 1 - P(\overline{A}) - P(\overline{A})P(B|\overline{A}) \tag{1.8}$$
$$= P(A) + P(\overline{A})P(B|\overline{A}) \tag{1.9}$$
$$= P(A) + P(\overline{A}, B) \tag{1.10}$$
$$= P(A) + P(B)P(\overline{A}|B) \tag{1.11}$$
$$= P(A) + P(B)\left[1 - P(A|B)\right] \tag{1.12}$$
$$= P(A) + P(B) - P(B)P(A|B) \tag{1.13}$$
$$P(A \cup B) = P(A) + P(B) - P(B, A) \qquad \textbf{extended sum rule} \tag{1.14}$$

In words, the disjoint probability of two events is equal to the sum of their probabilities minus their joint probability.

It A and B are **independent** then the probability of A occurring must not depend on whether B has occurred so $P(A|B) = P(A)$ through the product rule this implies $P(B|A) = P(B)$ and

$$P(A,B) = P(A)P(B) \quad \text{independent events} \tag{1.15}$$

If two events are **mutually exclusive**, that is they cannot occur at the same time (the first flip of a coin cannot be both heads and tails) then $P(A,B) = 0$ and the extended sum rule becomes

$$P(A \cup B) = P(A) + P(B) \quad \text{mutually exclusive events} \tag{1.16}$$

Example: If you roll a die once the probability of getting a 6 *or* a 5 is $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$. If you roll a die twice the probability of getting a 6 *and then* a 5 is $\left(\frac{1}{6}\right)\left(\frac{1}{6}\right) = \frac{1}{36}$. The probability of getting a 6 *and* a 5 is twice this because, $\frac{1}{18}$, because there are two ways of doing this, a 6 first or a 5 first.

This second case can be calculated in an alternative way. In the first roll we must get a 5 or a 6. We have calculated that the probability of this is $\frac{1}{3}$. Once this is done in the second roll we most get whichever number we didn't get in the first roll, one number out of 6, probability $\frac{1}{6}$. The probability of these two independent events happening is then given by the product rule $\left(\frac{1}{3}\right)\left(\frac{1}{6}\right) = \frac{1}{18}$.

Now say we have a set of observations $\{A_I\}$ that are all mutually exclusive and together they include all possible outcome then

$$\begin{aligned} 1 &= P(A_1 \cup A_2 \cup A_3 \cup \ldots |B) + P(\overline{A_1 \cup A_2 \cup A_3 \cup \ldots}|B) && (1.17)\\ &= P(A_1|B) + P(A_2 \cup A_3 \cup \ldots |B) + 0 && (1.18)\\ &= P(A_1|B) + P(A_2|B) + P(A_3 \cup \ldots |B) && (1.19)\\ &= \sum_i P(A_i|B) && (1.20) \end{aligned}$$

This is the origin of the normalization requirement on any probability distribution function (PDF). Note that I have put a B in as a condition on all the probabilities, but this would hold without them.

Another important result along these lines is

$$\sum_i P(B|A_i)P(A_i) = \sum_i P(B, A_i) = \sum_i P(A_i|B)P(B) = P(B)\sum_i P(A_i|B) = P(B) \tag{1.21}$$

with the same requirements on the set $\{A_i\}$. This is the origin of what we will later call marginalization.

2 Some warm up problems

There are a large class of problems, classical statistical physics included, for which individual states are considered equally probable and the question is how many states out of all possible states have a certain property. The property could be the temperature, pressure or having a full house in your poker hand and states could be the position each atoms in a gas, the spin state of each atom in a metal or the identity of the five cards you are dealt in poker. Here are some very simple problems that illustrate some of the counting techniques used throughout statistics.

2.1 Rolling Dice

Say we roll a die 10 times. Lets consider the following questions:

What is the probability of getting at least one 6? This is an "or" question - What is the probability of the first roll being 6 or the second one being six or ... Lets call the proposition that the ith roll is a 6 A_i. The sum rule (1.14) applies, but since these are not mutually exclusive events the sum rule 1.16 does not. These are independent events since the outcome of any one does not effect the outcome of any other. We could successive apply the extended sum rule (1.14 and the product rule (1.15) to $P(A_1 \cup A_2 \cup \cdots \cup A_{10})$ to break it down into $P(A_i)$'s which we know is 1/6. However, a quicker way to the answer is to realize that the probability of at least one being 6 is 1 minus the probability that non are 6. This follows from the logical requirement that $\overline{A_1 \cup A_2 \cup \cdots \cup A_{10}} = \overline{A_1}, \overline{A_2}, \ldots, \overline{A_{10}}$. Using the original sum rule (1.2) we get symbolically

$$\begin{aligned} P(A_1 \cup A_2 \cup \cdots \cup A_{10}) &= 1 - P(\overline{A_1 \cup A_2 \cup \cdots \cup A_{10}}) && (2.1)\\ &= 1 - P(\overline{A_1}, \overline{A_2}, \ldots, \overline{A_{10}}) && (2.2)\\ &= 1 - P(\overline{A_1})P(\overline{A_2}) \ldots P(\overline{A_{10}}) && (2.3)\\ &= 1 - P(\overline{A})^{10} && (2.4)\\ &= 1 - \left(\frac{5}{6}\right)^{10} && (2.5)\\ &= 0.838\ldots && (2.6) \end{aligned}$$

We could also solve this problem by counting. How many combinations of rolls are there? The first roll has 6 possibilities, the second one 6, etc. so there are 6^{10} combinations. There are 5^{10} combinations with no 6s. So the fraction of the cases that have no 6s is $\left(\frac{5}{6}\right)^{10}$ so the probability of having 1 or more is $\left(\frac{5}{6}\right)^{10}$.

What is the probability of getting exactly one 6? Lets first try to solve this problem by pure symbolic logic and the rules of probability. The proposition could be stated as roll one is a 6 *and* all the others are not *or* roll two is a 6 *and* all the other are not *or* etc. Symbolically this is represented as

$$B_1 = (A_1, \overline{A_2}, .., \overline{A_{10}}) \cup (\overline{A_1}, A_2, .., \overline{A_{10}}) \cup \cdots \cup (\overline{A_1}, \overline{A_2}, .., A_{10}) \qquad (2.7)$$

Each of the propositions in the parenthesis are mutually exclusive so the sum rule (1.16) can be applied to B_1 to break it up into a sum

$$P(B_1) = P(A_1, \overline{A_2}, .., \overline{A_{10}}) + P(\overline{A_1}, A_2, .., \overline{A_{10}}) + \ldots \qquad (2.8)$$

Since each of the rolls are identical, the probabilities for reach of situation must be the same and each term must be the same

$$\begin{aligned} P(B_1) &= 10P(A_1, \overline{A_2}, .., \overline{A_{10}}) && (2.9)\\ &= 10P(A_1)P(\overline{A_2}, .., \overline{A_{10}}) && (2.10)\\ &= 10\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^{9} && (2.11)\\ &= 0.323\ldots && (2.12) \end{aligned}$$

where we use the same logic that got us from equation (2.2) to line (2.5) in the previous problem.

Now lets do this again by counting. There are 6^{10} possible combinations. If one roll is a 6 the other nine need to be less than 6. There are 5^9 combinations of nine numbers between 1 and 5. The 6 can come up on any of 10 rolls so there are in total 10^9 ways of rolling 10 times and getting one 6.

What is the probability of getting exactly four 6s? This can be confusing, but if we just look at it from a symbolic point of view we can avoid some common misunderstandings. Here we must find all the combinations of four As and six $\overline{A}$. The first A can go in one of ten slots and the second in one of the remaining 9, etc. giving $10 \times 9 \times 8 \times 7 = 10!/(10-4)!$. We have over counted here though because the order in which we place the As in the slots should not matter, it gives the same logical statement. How many orderings are there? For each selection of 4 slots there are four choices for the first one, three choices etc. - 4! orderings or **permutations**. So there are $\frac{10!}{4!(10-4)!}$ ways of having four As and six $\overline{A}$. The probability of all these combinations are equal and mutually exclusive (a roll cannot be both A and $\overline{A}$) so we can add their probabilities

$$\begin{aligned}
P(B_4) &= \frac{10!}{4!(10-4)!} P(A_1, A_2, A_3, A_4, \overline{A_5}, .., \overline{A_{10}}) && (2.13)\\
&= \frac{10!}{4!(10-4)!} P(A_1, A_2, A_3, A_4) P(\overline{A_5}, .., \overline{A_{10}}) && (2.14)\\
&= \frac{10!}{4!(10-4)!} P(A)^4 P(\overline{A})^6 && (2.15)\\
&= \frac{10!}{4!(10-4)!} \left(\frac{1}{6}\right)^4 \left(\frac{5}{6}\right)^6 && (2.16)\\
&= 0.05\ldots && (2.17)
\end{aligned}$$

A confusion with this problem often arises because it is often stated or implied that all the permutations of the 6s must be considered one combination because they are indistinguishable. This might lead one to consider any two repeated numbers that are not 6s as indistinguishable and try not to over count them. This quickly becomes a very complex calculation. Although it is true that the 6s are indistinguishable this misses the point. For the purposes of this problem each roll has a binary outcome. It is either a 6 or not a 6. 6s are indistinguishable, but so are not 6s. We could have considered a different problem – "What is the probability of getting 4 rolls that are more than 4?". The calculation would be exactly the same except that the probabilities $P(A)$ and $P(\overline{A})$ would be different, $\frac{1}{3}$ and $\frac{2}{3}$ instead of $\frac{1}{6}$ and $\frac{5}{6}$.

These dice throwing problems are a special case of the **binomial distribution** which we will discuss later in more detail.

2.2 Birthday Paradox

This is another widely known problem for which many people go down the wrong path and get confused. The "paradox" is that in a relatively small group of people there is a surprisingly high probability that two of them will have the same birthday.

Lets say there are n people at the party. There are 356 choices for the birthday of each person (not including leap years) so there are 356^n combinations of n birthdays. We will assume these are all equally likely. Instead of finding the number of combinations with repeat birthdays lets find the number of combinations with no repeats. There are 356 choices for the first person, then 355 choices for the second etc. until you get to the last person so the number of cases with no repeats

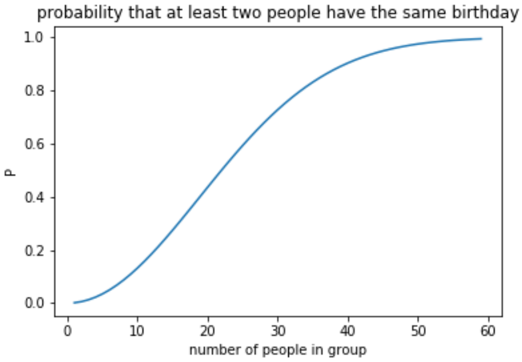

Figure 1: Probability of more than one person having the same birthday.

is $356 \times 354 \times ... \times (356 - n + 1) = 356!/(356 - n)!$. So the total probability is

$$P(\text{at least two the same}) = 1 - P(\text{no two the same}) = 1 - \frac{356!}{356^n(356 - n)!}. \tag{2.18}$$

If you try to calculate this number in your directly with a computer you will find that some of these numbers are too big to store. The scipy factorial function (scipy.special.factorial) will give infinity for 356 for example. But the quotient of these numbers is something reasonable. This problem often comes up in this kind of problem. We will need an approximation to complete the calculation. Taking the log of a quotient often helps you cancel some things out. And taking Stirling's approximation ($\ln N! = N \ln N - N$) often helps simplify factorials.

$$\begin{aligned}
\ln\left(\frac{N!}{N^n(N-n)!}\right) &= \ln N! - \ln(N-n)! - n\ln N & (2.19)\\
&= N\ln N - N - (N-n)\ln(N-n) - (N-n) - n\ln N & (2.20)\\
&= (n-n)\ln N - (N-n)\ln(N-n) - n & (2.21)\\
&= (N-n)\ln\left(\frac{N}{(N-n)}\right) - n & (2.22)
\end{aligned}$$

We can then take the exponential of this to get

$$P(\text{at least two the same}) \simeq 1 - \left(\frac{N}{N-n}\right)^{N-n} e^{-n} \tag{2.23}$$

This is plotted in figure1. For a group of 23 people there is a 50% chance that at least 2 of them will have the same birthday.

2.3 Poker

A deck of poker cards consists of 52 cards. There are four suits - diamonds (♢), hearts (♡), spades (♠) and clubs (♣). In each suit there are an ordered sequence of 13 cards (we will take the ace to be greater than the king). A poker hand consists of 5 cards. In ”five card stud” you are dealt five cards and you are not allowed to exchange any. This version of poker is almost never played because it relies too much on chance and not skill, but we will consider it here because it is simple.

What is the probability of getting a flush (five cards of the same suit) in five card stud? You might at first think this is just like the dice rolling problem and say it is $4(1/4)^5 \simeq 0.0039$, but this would be wrong because the draws are not independent. If your first card is a ♣ there will be fewer ♣ in the deck and the deck will be smaller so the probability of getting a club the second time will be $(13-1)/(52-1)$.

$$P(\text{flush}) = \frac{4}{4}\frac{12}{51}\frac{11}{50}\frac{10}{49}\frac{9}{48} = 0.00198\ldots \tag{2.24}$$

Significantly less probable than we would get if there where replacement.

What is the probability of a straight? This is getting five sequential cards, for example 8, 9, 10 ,J ,Q. The probability of drawing them all in a row must be the same as the probability of drawing them in any other order so we can calculate the probability of drawing them in order and then multiply by the number of permutations. First we need to draw a card below of 10 or lower or there wont be enough cards of higher value. That probability is $4 \times 9/52$. Then there are 4 cards of one higher value out of 51 remaining cards, etc.. Then for each case there are 5! permutations.

$$P(\text{straight}) = 5!\frac{36}{52}\frac{4}{51}\frac{4}{50}\frac{4}{49}\frac{4}{48} = 0.003546\ldots \tag{2.25}$$

Somewhat more likely than a flush which is why this hand is worth less. If we count the ace-low straight this is 0.00394.... This includes straight-flushes and royal-straight-flushes which are actually higher hands.

What is the probability of a full house? A full house is two of a kind (two 10's or two kings for example) and three of another kind (three aces or three twos).

Lets do this one a little differently. Lets count the total number of distinct five card hands and then count the number of distinct full houses. The probability will be the ratio of these since every hand is equally probable. Lets make this a little more abstract. There are N distinct objects (cards) we have N ways of choosing the first one. There are $N-1$ objects left when we pick the next one, etc. So there are $N \cdot (N-1) ... \cdot (N-n+1)$ distinct ways of choosing n objects out of N. This can also be written $N!/(N-n)!$. This counts combinations of objects in different orders as distinct (123 is different than 213). If we wish to count different permutations of the same objects as the same set then we need to divide by the number of permutations of n objects which is $n!$. So the number of these distinct sets is

$$\binom{N}{n} \equiv \frac{N!}{n!(N-n)!} \tag{2.26}$$

This is the **binomial coefficient**. In English this is often spoken as ”N choose n.” for obvious reasons. Lets use it on our problem.

There are $\binom{52}{5}$ distinct five card hands. There are four cards of each type, one for each suit, so there are $13 \cdot \binom{4}{2}$ distinct pairs of cards of the same kind. The three of a kind need to be different than the pair so there are $12 \cdot \binom{4}{3}$ of them. So the probability of a full house is

$$P(\text{full house}) = \frac{\binom{4}{2} \cdot \binom{4}{3} \cdot 13 \cdot 12}{\binom{52}{5}} = 0.00144\ldots \tag{2.27}$$

Very similar logic will lead you to the probabilities of getting two pair or four of a kind.

Calculating the probabilities for poker may seem frivolous, but the calculation of odds for gambling actually played a very important role in the development of statistics. Pascal and Fermat had a long correspondence in the 17th century in which they developed basic probability theory.

2.4 The Monty Hall Problem

This is a classic problem based on an old American TV game show. It was before my time, but apparently the host of the show was named Monty Hall. There are variations of this game show on Italian TV also. In this game the contestant is can choose between three doors. He knows that behind one of the doors is something nice like a new car and behind the other two are things that are not so nice like a chicken or an old shoe. The contestant chooses one door and then Monty eliminates one of the doors that were not chosen and shows that it has the shoe. The contestant then has a chance to change his choice or remain with his first choice. What should he do? Does it matter?

3 Probability distributions

In this section we will look at some frequently used probability distributions and probability distribution functions (PDFs) and what they are meant to represent. There are many, many named distributions that have been used to model many different things. I will discuss only a few of the most widely applicable distributions that come up very often in statistics. Most others distributions can be derived from these, are limiting cases of these or can be derived using the kind of arguments that I will use to derive them. In practical cases one might need to derive a statistical model that fits the question or the physical theory might dictate a probability distribution for an observable quantity that is not one of the classical distributions.

3.1 properties of a probability distribution function (PDF)

So far we have considered the probabilities of discrete events - the probability of getting a 5 or 6. If we consider a continuous variable x we can define the probability of being within an infinitesimal range x to $x + dx$ as $p(x)dx$. This probability must be positive.

$$p(x) \geq 0 \tag{3.1}$$

There are an infinite number of these bins across the range of x. A measurement of x will be in only one of them so we can apply the sum rule (1.17) to these bins. In the infinitesimal limit the some becomes an integral

$$\int_{-\infty}^{\infty} dx \; p(x) = 1 \tag{3.2}$$

All valid PDFs must satisfy these two requirements. Sometimes people call the PDF the **probability mass function**. They mean the same thing.

In the frequentist tradition x is called a **random variable**. A strict Bayesian might avoid using the term. He/she might say that there is an event were the value x is observed and we can attach a probability to this event given our prior knowledge and statistical model. There is no randomness about it. I will take a practical approach and ignore the linguistic distinctions as most scientists do.

3.2 mean, median, mode ...

Before we get started with the specific distributions, it will be useful to define some terms and quantities that are used to describe the properties of distributions.

cumulative distribution function - the function of x describing the probability of the measured value being $< x$:

$$F(x) = \int_{-\infty}^{x} dx' p(x') \tag{3.3}$$

By definition $F(-\infty) = 0$ and $F(+\infty) = 1$. The cumulative distribution for a discrete distribution is defined in the obvious way.

expectation value - The "average" of any function of the random variable. This is denote by $E[\dots]$ or $\langle \dots \rangle$. The expectation value of $f(x)$ is

$$E\left[f(x)\right] = \langle f(x) \rangle \begin{cases} \sum_x p(x) \; f(x) \\ \int_{-\infty}^{\infty} dx \; p(x) \; f(x) \end{cases} \tag{3.4}$$

mode - A point where a distribution has a maximum. **Unimodal** distributions have one mode and **multimodal** distributions have more than one.

median - The point in the distribution where $F(x) = 1/2$. The probability that x will be less than the median is equal to the probability that it will be more than the median. In a sample or data set the median is the data point that has equal numbers of data points larger than and less than it. For a set with an even number of points the arithmetic mean between the two points closest to having this property is often used.

mean - The mean is the expectation value of the random variable itself, $E[x]$. This will often be represented by μ.

moments - The nth moment of a distribution is $E[x^n]$.

central moments - The nth central moment is $E[(x-\mu)^n]$

variance - The variance is the second central moment $E[(x-\mu)^2]$. It is often denoted by $Var[x]$ or σ^2. This is a measure of the width of the distribution.

standard deviation - the square root of the variance. It is often denotes by σ. An equivalent measure of the width of the distribution in the same units as the random variable.

mean deviation $E[|x-\mu|]$. This is an alternative measure of the width of a distribution. It is often more robustly estimated from a small sample especially when the distribution has large "tails" (much of the probability lies far away from the peak or beyond $\sim \sigma$ from it.).

skewness - $E[(x-\mu)^3]/\sigma^3$. This is a unitless measure of the asymmetry of the distribution.

kurtosis - $E[(x-\mu)^4]/\sigma^4$. This is a measure of the relative importance of outliers (point differing from the mean by larger than several σ). If the kurtosis is larger than 1 the "tails" of the distribution are more important than for a Gaussian. This also reflects the "boxyness" of the distribution.

standardized variable - It is often useful to rescale a random variable with the standard deviation and mean of its distribution

$$X = \frac{(x-\mu)}{\sigma}. \tag{3.5}$$

This variable will always have a mean of 0 and a variance of 1.

Although the moments of a distribution are often used to describe a distribution, and it is true that two distributions with the same moments must be the same distribution, it is possible for a distribution to have no moments. An example of this that is of particular interest in physics and astronomy it the Cauchy or Lorentzian distribution:

$$p(x) = \frac{\gamma}{\pi\left[(x-x_o)^2+\gamma^2\right]} \quad \text{Cauchy-Lorentz distribution.} \tag{3.6}$$

Among other things, this is the natural profile of a spectral line because of the finite lifetime of the excited state. It is also the distribution of the ratio of two normally distributed variable with zero means (Try proving this.). Also if you have a point on a plane and you shoot rays out from it in random directions their intercepts with any line not going through the point will have this distribution (Try proving this!).

This distribution is normalized and it is symmetric around its mode at $x = x_o$, but the integrals that define all the moments, including the mean, are divergent. Later we will ask what would happen if we tried to estimate the mean or variance using a sample drawn from this distribution.

Note that

$$Var[x] = E\left[(x-\bar{x})^2\right] = E\left[x^2 - 2x\bar{x} + \bar{x}^2\right] \tag{3.7}$$

$$= E\left[x^2\right] - 2E\left[x\right]\bar{x} + \bar{x}^2 \tag{3.8}$$

$$= E\left[x^2\right] - \bar{x}^2. \tag{3.9}$$

3.3 moment generating function

The **moment generating function** (MGF) of a distribution is defined in the discrete and continuous cases as

$$m_x(t) = \langle e^{tx} \rangle = \left\{ \begin{array}{c} \sum_x e^{tx} p(x) \\ \int_{-\infty}^{+\infty} dx \; e^{tx} p(x) \end{array} \right. \tag{3.10}$$

From this we can easily see that the moments of a distribution can be calculated by taking the derivatives of the MGF

$$\left. \frac{d^n m_x(t)}{dt^n} \right|_{t=0} = \langle x^n \rangle \tag{3.11}$$

This can be very useful for cases where the MGF can be found analytically. With a change in sign of t this is the same thing as the Laplace transform

3.3.1 changing of variables

Say we have a variable x and the probability of it being between x and $x + dx$ is $p(x)dx$. Now say we have another variable y that is related to x by $x = f(y)$ where $f(y)$ is single valued and differentiated. Then for a change dy, x changes by $dx = \frac{d}{dy} f(y) dy$. The probability of being within these range of should not depend on which variable is used to measure the range so the must be the same

$$p(x)dx = p\left(f(y)\right) \frac{df}{dy} dy \tag{3.12}$$

In this way the pdf for one variable can be transformed into the pdf for another. For example if the PDF of x is $p(x)$, the PDF of $y = x^2$ is $\frac{1}{2} p(\sqrt{y})/\sqrt{y}$. We will see examples later.

This is really just the same as a change of variables in an integral of course. For a multivariant pdf variables can be changed in the usual way

$$p(x_1, x_2, \dots) dx_1 dx_2 \dots = p(y_1, y_2, \dots) \left| \frac{\partial x}{\partial y} \right| dy_1 dy_2 \dots \tag{3.13}$$

where $\left| \frac{\partial x}{\partial y} \right|$ is the determinant of the Jacobian matrix relating the volume element in one coordinate system to another.

For example if the probability of a galaxy existing at a point in three dimensional space is $p(x, y, z)dxdydz$ then the probability in spherical coordinates is

$$p\left(r \sin(\theta) \cos(\phi), r \sin(\theta) \cos(\phi), r \cos(\theta)\right) \; r^2 \sin(\theta) dr d\theta d\phi. \tag{3.14}$$

3.4 Binomial and Bernoulli

Say there is some experiment or observation and for each trial the probability of having some outcome A is p. The probability of not having this outcome will be $1 - p$. Each trial is statistically independent. In N trials what is the probability of having n A's?

Using the product rule for independent events we know that the probability of getting n A's in a row is p^n and the probability of having the other be not A is $(1 - p)^{N-n}$. This is the probability each every set of N with n A's. Now we need to count how many combinations there are. Since we are not concerned with the order the number is our friend the **binomial coefficient**

$$\binom{N}{n} \equiv \frac{N!}{n!(N - n)!} \tag{3.15}$$

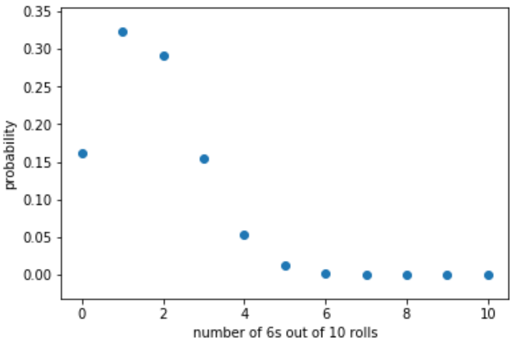

Figure 2: The binomial distribution for the number of 6s in ten rolls of a die or one roll of ten dice. $N = 10$, $k = 0 \dots 10$, $p = 1/6$

So we get the final result

$$p(n|N) = \binom{N}{n} p^n (1-p)^{N-n} \tag{3.16}$$

which is called the binomial distribution. The case of $N = 10$ and $p = 1/6$ is shown in figure 2. We can now calculate the number of getting any number of 6s out of any number of dice rolls.

We can also think of the binomial distribution as the solution to the problem of ”drawing with replacement”. Imagine a bag full of green and blue balls. Each trial you take one out record its color and put it back in the bag. The **Bernoulli distribution** is the special case of $N = 1$, an almost trivial case, but perhaps the first probability distribution written down.

The binomial distribution is important for calculating the distribution of any finite sample of observations and comes up a lot in statistics as we will see.

Note that the binomial coefficient gets its name because of the **binomial expansion**

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^n y^{n-k} \tag{3.17}$$

Using this expansion we can find the moment generating function

$$\begin{aligned} m_x(t) &= \sum_{n=0}^{N} e^{tn} \binom{N}{n} p^n (1-p)^{N-n} && (3.18)\\ &= \sum_{n=0}^{N} \binom{N}{n} (e^t p)^n (1-p)^{N-n} && (3.19)\\ &= \left(e^t p + 1 - p\right)^N && (3.20) \end{aligned}$$

From this we can find the mean and variance

$$\langle n \rangle = Np \tag{3.21}$$

$$\sigma^2 = \langle n^2 \rangle - \langle n \rangle^2 = np(1-p) \tag{3.22}$$

3.4.1 drawing without replacement, the hypergeometric distribution

Let us briefly consider the case where there are a finite number of objects of two types, we select them at random and we do not replace them before selecting the next. In this case each trial will not be independent of the ones before it (or the ones after it). We have a bag containing N balls with R of red ones and $N-R$ blue ones. The probability of getting r red ones out of n tries *without replacement* is

$$p(r|n,N,R) = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}} \tag{3.23}$$

Note that $p(r|1,N,R) = R/N$ and $p(r|N,N,R) = \delta_{Rr}$ as they should. The probability of a flush in 5 card stud would be $4 \times p(5|5,52,13)$ and in 7 card stud $4 \times (p(5|7,52,13) + p(6|7,52,13) + p(7|7,52,13))$ (see §2.3).

3.5 Poisson distribution

Lets say the probability of an event happening within t and $t+dt$ is a constant rdt. We want to know the probability of N of these events happening within a finite range of time.

First lets find the probability of *no* events happening within a finite range, t_o to $t+dt$. Lets call it $p(0|t_o, t+dt)$. The probability that no event happens between t and $t+dt$ is $1-rdt$. We can express the joint probability of no events happening in the range t_o to t and no events happening within t to $t+dt$ using the product rule for statistically independent events

$$p(0|t_o, t+dt) = p(0|t_o,t)\left[1-rdt\right] \tag{3.24}$$

Rearranging this we can obtain the differential equation

$$\frac{p(0|t_o,t+dt) - p(0|t_o,t)}{dt} = \frac{d}{dt}p(0|t_o,t) = -p(0|t_o,t)r \tag{3.25}$$

The solution to this is $p(0|t_0,t) = Ae^{-rt}$. We can find the normalization be requiring that $p(0|t_o,t_o) = 1$, there will always be no events in a range of zero length. The results is,

$$p(0|,t_o,t) = e^{-r(t-t_o)} \tag{3.26}$$

Now for a finite number of events. The probability of n events occurring at ordered times $t_1 \dots t_n$ all less than t (which will also be t_{n+1} in this notation) can also be found by the product rule:

$$p(0<t_1<t_2<\cdots<t_n<t) = p(0|0,t_1)rdt_1H(t_1<t_2)p(0|t_1,t_2)rdt_2H(t_2<t_3)\dots p(0|t_n,t)dt_nH(t_n<t) \tag{3.27}$$

$$= r^n e^{-rt}\prod_{i=1}^{n} dt_i H(t_i<t_{i+1}) \tag{3.28}$$

where

$$H(x<y) = \begin{cases} 1 & , \quad x \le y \\ 0 & , \quad x > y \end{cases} \tag{3.29}$$

Using the sum rule we know that the probability of n events occurring is the sum of the probabilities for all possible values of event times.

$$p(n|r,t) = \prod_I \int_0^t p(0 < t_1 < t_2 < \cdots < t_n < t) \tag{3.30}$$

$$= r^n e^{-rt} \int_0^t dt_n \cdots \int_0^{t_3} dt_2 \int_0^{t_2} dt_1 \tag{3.31}$$

$$= r^n e^{-rt} \int_0^t dt_n \cdots \int_0^{t_3} dt_2 t_2 \tag{3.32}$$

$$= r^n e^{-rt} \int_0^t dt_n \cdots \int_0^{t_4} dt_3 \frac{t_3^2}{2} \tag{3.33}$$

$$= \frac{(rt)^n}{n!} e^{-rt} \tag{3.34}$$

$$= \frac{(\nu)^n}{n!} e^{-\nu} \quad \text{Poisson Distribution} \tag{3.35}$$

This distribution has the following mean and variance

$$E[n] = \nu \tag{3.36}$$

$$Var[n] = \nu \tag{3.37}$$

The standard example of something that is Poisson distributed is the number of radio active decays within a fixed interval of time. If supernovae go off randomly the probability of seeing one during an hour of observations would be $re^{-r(1\,\text{hour})}$ where r would be the total rate of supernovae in the monitored galaxies. Another example is the counts of something, say stars or galaxies, within a volume, or cell, that are uniformly distributed in space. In this case r is the average number density of objects and t is the volume of the cell. It does not matter what the shape of the cell is. A common question is whether objects are uniformly distributed or clustered. This can be determined by comparing the number counts in cells to the predictions of a Poisson distribution. We will get back to this question later.

3.5.1 as a limit of the binomial distribution

Imagine a cube of space with volume, V, and a smaller cube within it with volume , v. Now imagine there are N uniformly distributed galaxies or stars in this volume. The number of galaxies in v will be n. n would be binomially distributed with the probability of one galaxy being in v equal to $p = \frac{v}{V}$.

Now lets take the limit of $N \to \infty$ and $p \to 0$ (or $V \to \infty$) while keeping the average density constant $\nu = N/V = Np$. Using Stirling's approximation one can show that $\frac{N!}{N-n)!} \simeq N^n$ to highest order.

$$\binom{N}{n} p^n (1-p)^{N-n} = \binom{N}{n} \left(\frac{\nu}{N}\right)^n \left(1 - \frac{\nu}{N}\right)^{N-n} \tag{3.38}$$

$$\simeq \frac{\nu^n}{n!} e^{-\nu} \tag{3.39}$$

where I have used $\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = e^x$. So the Poisson distribution is the binomial distribution in this limit.

A sometimes useful limit of the Poisson distribution is when $\nu \gg 1$ to treat n as continuous and replace $n!$ with the gamma function

$$p(n|\nu) \simeq \frac{\nu^n}{\Gamma(x+1)} e^{-\nu} \qquad \nu \gg 1 \tag{3.40}$$

3.6 Gaussian and normal

Gaussian and normal are two names for the same thing. It is a very widely used probability distribution. The usual justification for this is the central limit theorem although it is also justified as the maximum entropy distribution for a fixed variance. We will get to these justifications later.

The pdf for the Gaussian distribution is

$$p(x|\sigma,\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{3.41}$$

The mean is μ and the variance is σ^2.

A note on notations: To signify that a variable x is normally distributed with a mean of μ and a standard deviation of σ one can write $x \sim \mathcal{N}(\mu,\sigma)$. Sometimes, in an abuse of notation, $\mathcal{N}(\mu,\sigma)$ can stand for the actual pdf (3.41).

The *cumulative distribution function* is

$$F(x) = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{x}{\sqrt{2}}\right) \tag{3.42}$$

with the **error function** defined as

$$\text{erf}(z) \equiv \frac{2}{\sqrt{\pi}} \int_0^z e^{-u^2} du \tag{3.43}$$

The *moment generating function* is

$$\begin{aligned} m_{x-\mu}(t) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} dx\ e^{tx} e^{-\frac{x^2}{2\sigma^2}} & (3.44)\\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} dx\ \exp\left[-\left(\frac{x}{\sqrt{2}\sigma} - \frac{t\sigma}{\sqrt{2}}\right)^2 + \frac{t^2\sigma^2}{2}\right] & (3.45)\\ &= e^{\frac{1}{2}\sigma^2 t^2} & (3.46) \end{aligned}$$

The moments are

$$\mu_n = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} dx\ x^n e^{-\frac{x^2}{2\sigma^2}} = \begin{cases} \sigma^n (n-1)!! & n \text{ even} \\ 0 & n \text{ odd} \end{cases} \tag{3.47}$$

where !! is the **double factorial**,

$$!!n = n \cdot (n-2) \cdot (n-4) \ldots 1 \tag{3.48}$$

The probability of x being within $n\sigma$ of the mean is

$$\begin{aligned} p(\mu - n\sigma \le x \le \mu + n\sigma) &= 1 - F(\mu - n\sigma) - [1 - F(\mu + n\sigma)] & (3.49)\\ &= \frac{1}{2}\left[\text{erf}\left(\frac{n}{\sqrt{2}}\right) - \text{erf}\left(-\frac{n}{\sqrt{2}}\right)\right] & (3.50)\\ &= \text{erf}\left(\frac{n}{\sqrt{2}}\right) & (3.51) \end{aligned}$$

some specific values for this are

$$p(\mu - \sigma \leq x \leq \mu + \sigma) = 0.683 \tag{3.52}$$
$$p(\mu - 2\sigma \leq x \leq \mu + 2\sigma) = 0.954 \tag{3.53}$$
$$p(\mu - 3\sigma \leq x \leq \mu + 3\sigma) = 0.997 \tag{3.54}$$
$$p(\mu - 4\sigma \leq x \leq \mu + 4\sigma) = 0.999937 \tag{3.55}$$

3.7 central limit theorem

The Gaussian distribution plays an important role in statistics. The distribution of surprisingly large number of phenomena are observed to be well represented by a Gaussian distribution. The traditional explanation for this is the central value theorem. It holds that the sum of a large number of identically distributed independent random variables will be close to Gaussian distributed even if they are not individually Gaussian distributed. If the noise in a measurement can be considered the sum of many small unknown contributions than you would expect it to be Gaussian distributed.

Lets say we have N identically distributed variables x_i. We can define a set of standardized variables

$$z_i = \frac{x_i - \mu}{\sigma}. \tag{3.56}$$

With this scaling it is clear that $\langle z_i \rangle = 0$ and $\langle z_i^2 \rangle = 1$. The sum of these will be $Z = \sum_i z_i$. $\langle Z \rangle = 0$ and $\langle Z^2 \rangle = \sum_{ij} \langle z_i z_j \rangle = \sum_i \langle z_i^2 \rangle$ because each one is uncorrelated. So the standardized variable for the sum is

$$Y = \frac{1}{\sqrt{N}} Z = \frac{1}{\sqrt{N}} \sum_i z_i. \tag{3.57}$$

Now lets find the moment generating function for Y,

$$m_Y(t) = \langle \exp(tY) \rangle = \left\langle \exp\left(\frac{t}{\sqrt{N}} \sum_i z_i \right) \right\rangle = \left\langle \exp\left(\frac{t}{\sqrt{N}} z_i \right) \right\rangle^N \tag{3.58}$$
$$= \left\langle 1 + \frac{t}{\sqrt{N}} z_i + \frac{t^2}{N} \frac{z_i^2}{2} + \frac{t^3}{N^{3/2}} \frac{z_i^3}{3!} + \dots \right\rangle^N \tag{3.59}$$
$$= \left[1 + \frac{t}{\sqrt{N}} \langle z_i \rangle + \frac{t^2}{N} \frac{\langle z_i^2 \rangle}{2} + \frac{t^3}{N^{3/2}} \frac{\langle z_i^3 \rangle}{3!} + \dots \right]^N \tag{3.60}$$
$$= \left[1 + \frac{t^2}{2N} + \frac{t^3}{N^{3/2}} \frac{\langle z_i^3 \rangle}{3!} + \dots \right]^N \tag{3.61}$$
$$\simeq \lim_{N \to \infty} \left[1 + \frac{t^2}{2N} \right]^N \tag{3.62}$$
$$= e^{\frac{t^2}{2}} \tag{3.63}$$

This is the moment generating function for a Gaussian.

It is important to note that this theorem is strictly true only for a sum of an infinite number of variables with the same variance. You might not expect this to apply to our concept of noise coming from many small random contributions that are not all the same. If the variance of one of the variables where much larger than the others it would dominate the distribution of the sum for example. However the Gaussian distribution is widely and successfully used. We will later see another justification for it based on an entropy argument.

3.7.1 The distribution of the sum of independent random variables

Lets do a practical experiment to see how quickly the sum of variables will converge to a Gaussian distribution as the number of variables increases. To do this we will need the pdf of the sum of random variables. There is a general way of ding this that is useful. Lets take the sum of n random numbers to be $S = \sum_i x_i$. The pdf of each variable is $p_i(x_i)$ which may be different. We can marginalize over all the variables and use a Dirac delta function to force the sum of them to be S

$$
\begin{aligned}
P(S) &= \int_{-\infty}^{\infty} dx_1 \ldots \int_{-\infty}^{\infty} dx_n \; \delta(S - \sum_i x_i) \; p_1(x_1) \ldots p_n(x_n) & (3.64) \\
&= \int_{-\infty}^{\infty} dx_1 \ldots \int_{-\infty}^{\infty} dx_n \; \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} \exp\left[-ik(S - \sum_i x_i) \right] \; p_1(x_1) \ldots p_n(x_n) & (3.65) \\
&= \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} e^{-ikS} \prod_i \int_{-\infty}^{\infty} dx_i e^{+ikxi} p_i(x_i) & (3.66) \\
&= \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} e^{-ikS} \prod_i \tilde{p}_i(k) & (3.67)
\end{aligned}
$$

where $\tilde{p}_i(k)$ is the Fourier transform of $p_i(x_i)$. This means that $\prod_i \tilde{p}_i(k)$ is the Fourier transform of the pdf of S. In the special case where the distributions are all the same this will be $[\tilde{p}(k)]^n$. Note that in Fourier space the normalization requirement is $\tilde{p}(0) = 1$.

Lets look at a uniform distribution between $-L/2$ and $L/2$. The Fourier transform of this distribution is

$$
\tilde{p}(k) = \frac{1}{L} \int_{-L/2}^{L/2} dx \; e^{+ikx} = \frac{2}{Lk} \sin\left(\frac{kL}{2}\right) = \mathrm{sinc}\left(\frac{kL}{2}\right). \tag{3.68}
$$

So the pdf for the sum of n uniformly distributed variables, each over a range L/n is

$$
P_n(S) = \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} e^{-ikS} \mathrm{sinc}^n \left(\frac{kL}{2n}\right). \tag{3.69}
$$

Figure!3 shows this case for some small values of n. In this case each x_i has a maximum of 1 so S has a maximum of n. For this reason the tails of the distribution are cut off relative to the Gaussian. Even so you can see that the distribution becomes remarkably Gaussian even for $n = 5$.

This exercise can be done numerically for any distribution. It is not necessary to have an analytic expression for the Fourier transform of $p_i(x_i)$. An numerical DFT (discrete Fourier transformation) and inverse DFT will do the trick although care must be taken with the normalization convention that your software uses and a phase factor that comes in when n is even.

This technique can be used to study things like random walks and diffusion. The same idea is also used to derive halo mass functions in cosmology.

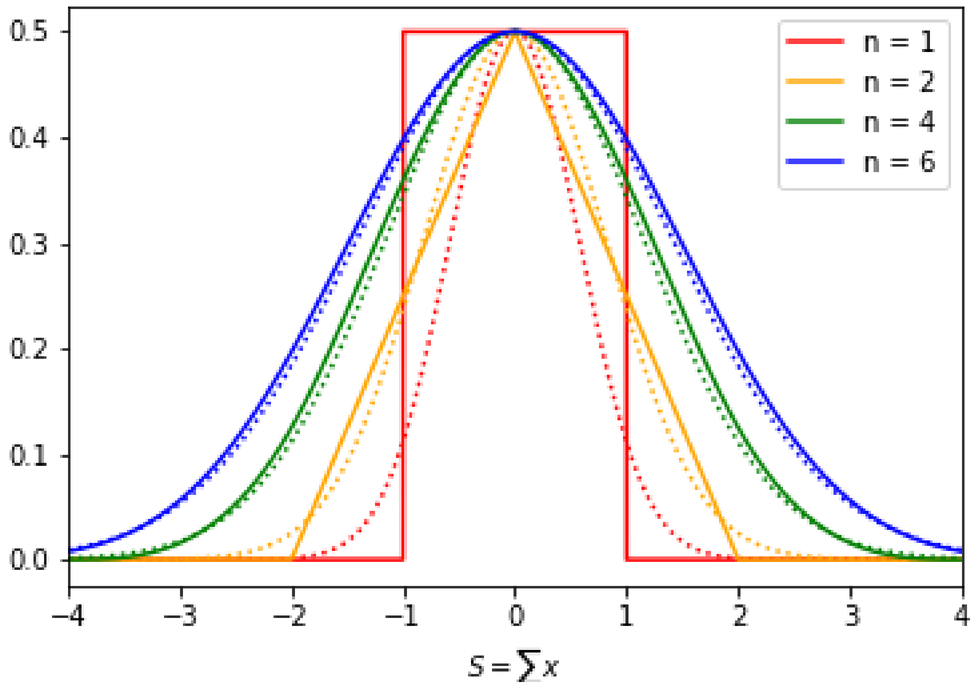

Figure 3: Probability distribution for the sum of n random variables that are uniformly distributed between -1 and 1. The normalizations have been changed so that there maximum is 0.5 in all cases. The dotted curves are for Gaussians with the same variance. You can see that the distribution converges to Gaussian remarkably quickly even for a very non Gaussian initial distribution.

3.8 lognormal

The lognormal distribution is simply the distribution where the log of the variable is normally distributed instead of the variable itself. This distribution is of particular interest in astronomy because photometric errors are often taken to be Gaussian in magnitudes which is the 2.5 times the log of the flux so the flux will be lognormally distributed. Since the inverse log of a real number cannot be negative the distribution is bounded from below by 0. The distribution is also used to model the distribution of matter many contexts. Another interpretation is that while the Gaussian is the right distribution for a sum of random variable, the lognormal is the right one for a product of many random variables.

The pdf comes from just changing variable from the Gaussian

$$p(y)dy = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(\ln(y)-\mu)^2}{2\sigma^2} \right\} \frac{dy}{y} \tag{3.70}$$

Some of its properties are

$$\begin{aligned} E[y] &= \exp(\mu + \frac{\sigma^2}{2}) \tag{3.71} \\ \mathrm{median[y]} &= \exp(\mu) \tag{3.72} \\ \mathrm{mode[y]} &= \exp(\mu - \sigma^2) \tag{3.73} \\ Var[y] &= [\exp(\sigma^2) - 1] \exp(2\mu + \sigma^2) \tag{3.74} \end{aligned}$$

If $\mu = 0$ and $\sigma \ll 1$ the distribution is Gaussian with a mean of 1 and a variance of σ^2. So if we take $y = 1 + \delta$ and $\mu = 0$ we have a model for fractional density fluctuations, δ, that will always be positive, will have a median of 0 and will tend to Gaussian when the variance is small. This is, for example, a good model for the Lyman-α absorption in quasar spectra. A multivariable version of this is possible by changing variable from the maltivariant Gaussian distribution (section 3.12) and a model for density fluctuations in the Universe.

3.9 Power law distribution

In astronomy it is common to model the distribution of many things (star masses, galaxy luminosities, planet masses, temperatures, densities of clouds, etc.) as a power law. The integral of a power law diverges either as $x \to 0$ or as $x \to \infty$ so some limits need to be fixed for the distribution to make sense. The normalized PDF is

$$p(x|x_{\min}, x_{\max}, \alpha) = x^\alpha \times \begin{cases} 0 & , \; x < x_{\min} \\ (\alpha+1)\left[x_{\max}^{\alpha+1} - x_{\min}^{\alpha+1}\right]^{-1} & , \; x_{\min} < x < x_{\max}, \alpha \neq -1 \\ \ln\left(\frac{x_{\max}}{x_{\min}}\right)^{-1} & , \; x_{\min} < x < x_{\max}, \alpha = -1 \\ 0 & , \; x > x_{\max} \end{cases} \tag{3.75}$$

The cumulative distribution is easily worked out

$$F(x|x_{\min}, x_{\max}, \alpha) = \begin{cases} 0 & , \; x < x_{\min} \\ \frac{\left[x^{\alpha+1} - x_{\min}^{\alpha+1}\right]}{\left[x_{\max}^{\alpha+1} - x_{\min}^{\alpha+1}\right]} & , \; x_{\min} < x < x_{\max}, \alpha \neq -1 \\ \frac{\ln\left(\frac{x}{x_{\min}}\right)}{\ln\left(\frac{x_{\max}}{x_{\min}}\right)} & , \; x_{\min} < x < x_{\max}, \alpha = -1 \\ 1 & , \; x > x_{\max} \end{cases} \tag{3.76}$$

3.10 multivariant distributions

A multivariant distribution is the probability distribution for the joint probability of two or more random variables. Lets number these variable x_1 through x_k. For discrete variable $p(x_1, x_2, \ldots, x_k)$ is the probability that the first variable has the value x_1 *and* the second variable has the value x_2, etc. There is the obvious extension to continous variables where $p(x_1, x_2, \ldots, x_k)dx_1dx_2 \ldots dx_k$ is the probability of all the variable simultaneously being within infinitesimal ranges near those values.

Now the expectation value implies a sum or integral over all the variables. For an arbitrary function $f(x_1, x_2, \ldots, x_k$

$$E[f(x_1, x_2, \ldots, x_k)] = \int \cdots \int dx_1 \ldots dx_k \; f(x_1, x_2, \ldots, x_k) \; p(x_1, x_2, \ldots, x_k) \tag{3.77}$$

$$= \prod_{i=1}^{k} \int dx_i \; f(x_1, x_2, \ldots, x_k) \; p(x_1, x_2, \ldots, x_k) \tag{3.78}$$

This is also written $\langle f(x_1, x_2, \ldots, x_k) \rangle$ or $\overline{f(x_1, x_2, \ldots, x_k)}$. The probability distribution is normalized so $E[1] = 1$.

The average and variance of each variable is defined in the same way as for a distribution of one variable. In this case there is also the **covariance** between two variable

$$C_{ij} = Cov[x_i x_j] \equiv E[(x_i - \bar{x}_i)(x_j - \bar{x}_j)] \tag{3.79}$$

If the covariance is greater than zero it means that they both tend to be high *and/or* low relative to their means simultaneously. If the covariance is negative one tends to be high while the other is low and vice versa.

C_{ij} is called the **covariance matrix**. You can see that by construction it is symmetric, $C_{ij} = C_{ji}$ and that the diagonal components $C_{ii} = E[(x_i - \bar{x}_i)^2]$ are positive which together mean its eigenvalues are positive or zero. Later we will talk about the covariance matrix of parameters and of data, two different covariance matrices which can be confusing. The covariance matrix is always positive definite (see appendix A).

Change the units for the variables will change the value of their covariance so a better measure of the degree of correlation it is the convenient to normalize the variance so that it is unitless,

$$\rho_{xy} \equiv \frac{C_{xy}}{\sigma_x \sigma_y} \tag{3.80}$$

Exy] satisfies all the requirements of an inner ("dot" or "scalor") product. One of the results of this is that covariance satisfies the **Cauchy–Schwarz inequality**

$$|Cov[xy]|^2 \leq Var[x]Var[y] \tag{3.81}$$

And a result of this is that $-1 \leq \rho_{xy} \leq 1$.

Another important relation is

$$C_{xy} = E[xy] - \bar{x}\bar{y} \tag{3.82}$$

which is an extension to the relation we already saw for the variance (3.9).

Two variables, x and y, are said to be **correlated variables** if $Cov[xy] \neq 0$. Otherwise they are correlated. Two variables that are **independent** variables are also uncorrelated, but uncorrelated variables are not necessarily independent. Variable with a negative covariance can be called **anticorrelated**.

3.11 multinomial distributions

The binomial distribution can be extended to the case where there are multiple possible outcomes of each trial. The probabilities are p_1, p_2, ... p_k and these are all the possible outcomes so $\sum_i p_i = 1$. The occurrence of each of theses is x_1, x_2, ... x_k. The probability of any sequence of these will be $\prod_i p_i^{x_i}$. There are $N!$ such sequences for N trials, but for each one with x_i there are $x_i!$ permutations that are the same. Thus

$$P(x_1, x_2, x_3, \ldots, x_k | N, \{p_i\}) = \frac{N!}{x_1!x_2!\ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k} = \frac{N!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_i^{x_i} \tag{3.83}$$

The mean and variance of the distribution are

$$E[x_i] = Np_i \tag{3.84}$$

$$Var[x_i] = Np_i(1 - p_i) \tag{3.85}$$

And the covariance is

$$Cov[x_i x_j] = -Np_i p_j \tag{3.86}$$

The negative value reflects the the property that if x_i is larger than its mean, for a fixed N, x_j is more likely to be below its mean and vice versa. If the units are not distributed exactly according their means then getting more in one mine implies there are less in others.

3.12 multivariant gaussian

The mutivariant Gaussian or normal distribution is by far the most often used multivariant distribution. It is a good approximation to many natural phenomena and is often used even when it is not as an approximation. It is also very useful when trying to understand some statistical argument or principle to put in multivariant Gaussian because often an analytic result can be obtained with it while it cannot in general. For these reasons it is essential for any good student of statistics to have a good intuitive understanding of and the ability to easily manipulate the multivariant normal distribution. I will do through some of its important properties and examples.

At this point it will be useful to the matrix notation. The n random variables will be grouped into a vector $\boldsymbol{x}$. The pdf of the multivariant gaussianis a generalization of the one dimensional Gaussian pdf.

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{C}) = \frac{1}{(2\pi)^{n/2}\sqrt{|\boldsymbol{C}|}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{C}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right] \tag{3.87}$$

$$\equiv \mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu}, \boldsymbol{C}\,\right) \tag{3.88}$$

where $\boldsymbol{C}$ is a n-by-n matrix and $\boldsymbol{\mu}$ is an n dimensional vector of parameters. This will define the function $\mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu}, \boldsymbol{C}\,\right)$. To signify that $\boldsymbol{x}$ is distributed in this way we write $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{C})$ just like for the one dimensional case.

The mean is

$$E[x_i] = \mu_i \quad \text{or} \quad E[\boldsymbol{x}] = \boldsymbol{\mu} \tag{3.89}$$

And the covariance is

$$Cov[x_i, x_j] = E[(x_i - \mu_i)(x_j - \mu_j)] = C_{ij} \quad \text{or} \quad Cov[\boldsymbol{x}, \boldsymbol{x}] = E[(\boldsymbol{x} - \boldsymbol{\mu})^T (\boldsymbol{x} - \boldsymbol{\mu})] = \boldsymbol{C} \tag{3.90}$$

So $\boldsymbol{C}$ is the correlation matrix as the choice of notation suggests.

For the **special case of a diagonal covariance matrix**, the diagonal elements are the σ^2's. The covariance matrix will take the form

$$\boldsymbol{C}^{-1} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots \\ 0 & \sigma_2^2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots \\ 0 & \frac{1}{\sigma_2^2} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \tag{3.91}$$

In this case there are no correlations between different variables. This form should be familiar from our discussion of the χ^2 distribution.

PROOF OF MEAN:

Lets calculate the means first

$$E[x_i] = \int_{-\infty}^{\infty} dx_i \ldots \int_{-\infty}^{\infty} dx_i \ldots \int_{-\infty}^{\infty} dx_n \; x_i \; p(|\boldsymbol{\mu}, \boldsymbol{C}) \tag{3.92}$$

$$\tag{3.93}$$

We can change variable to a set where $\boldsymbol{x'} = \boldsymbol{x} - \mu$ and all the others are unchanged. This will make $\boldsymbol{\mu}$ get substituted for $\boldsymbol{\mu'}$ which is the zero vector $\boldsymbol{\mu'} = 0$,

$$E[x_i] = \int_{-\infty}^{\infty} dx_i \ldots \int_{-\infty}^{\infty} dx_i' \ldots \int_{-\infty}^{\infty} dx_n \; (\mu_i + x_i') \; p(\boldsymbol{x'}|0, \boldsymbol{C}) \tag{3.94}$$

$$= \mu_i \int_{-\infty}^{\infty} dx_i \ldots \int_{-\infty}^{\infty} dx_i' \ldots \int_{-\infty}^{\infty} dx_n \; p(\boldsymbol{x'}|0, \boldsymbol{C}) + \int_{-\infty}^{\infty} dx_i \ldots \int_{-\infty}^{\infty} dx_i' \ldots \int_{-\infty}^{\infty} dx_n \; x_i' \; p(\boldsymbol{x'}|0, \boldsymbol{C}) \tag{3.95}$$

The first set of integrals must be 1 because the pdf is normalized. The second set must be zero because $p(\boldsymbol{x'}|0, \boldsymbol{C})$ is symmetric ($p(-\boldsymbol{x'}|0, \boldsymbol{C}) = p(\boldsymbol{x'}|0, \boldsymbol{C})$) and x_i' is antisymmetric.

PROOF OF VARIANCE:

$$Corr[\boldsymbol{x}, \boldsymbol{x}] = \int_{-\infty}^{\infty} d^n x \; (\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T \; p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{C}) \tag{3.96}$$

$$= \int_{-\infty}^{\infty} d^n x \; \boldsymbol{x}^T \boldsymbol{x} \; p(\boldsymbol{x}|0, \boldsymbol{C}) \tag{3.97}$$

Because $\boldsymbol{C}$ is a symmetric, positive definite matrix there exists a **eigendecomposition**

$$\boldsymbol{C} = \mathbf{M} \boldsymbol{\Sigma} \mathbf{M}^{-1} \tag{3.98}$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix whose elements are the eigenvalues and M is an **orthogonal matrix** which means that

$$\mathbf{M}^T = \mathbf{M}^{-1} \tag{3.99}$$

$$|\mathbf{M}| \equiv \det(\mathbf{M}) = 1 \tag{3.100}$$

The columns of $\boldsymbol{M}$ are the eigenvectors of $\boldsymbol{C}$.

Using this we can change variables into $\boldsymbol{y} = \boldsymbol{M}^{-1}\boldsymbol{x}$,

$$e^{\frac{1}{2}\boldsymbol{x}^T \boldsymbol{C} \boldsymbol{x}} \; d^n x = e^{\frac{1}{2}\boldsymbol{x}^T \boldsymbol{M} \boldsymbol{\Sigma} \boldsymbol{M}^T \boldsymbol{x}} \; d^n x = e^{\frac{1}{2}(\boldsymbol{M}^T \boldsymbol{x})^T \boldsymbol{\Sigma} (\boldsymbol{M}^T \boldsymbol{x})} \; d^n x = e^{\frac{1}{2}\boldsymbol{y}^T \boldsymbol{\Sigma} \boldsymbol{y}} \; |\boldsymbol{M}| d^n y = e^{\frac{1}{2}\boldsymbol{y}^T \boldsymbol{\Sigma} \boldsymbol{y}} \; d^n y \tag{3.101}$$

$$
\begin{aligned}
Corr[\boldsymbol{x},\boldsymbol{x}] &= \int_{-\infty}^{\infty} d^n x \; \boldsymbol{x}\boldsymbol{x}^T \; p(\boldsymbol{x}|0,\boldsymbol{C}) && (3.102)\\
&= \int_{-\infty}^{\infty} d^n y \; (\boldsymbol{M}\boldsymbol{y})(\boldsymbol{M}\boldsymbol{y})^T \; p(\boldsymbol{y}|0,\boldsymbol{\Sigma}) && (3.103)\\
&= \int_{-\infty}^{\infty} d^n y \; \boldsymbol{M}\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{M}^T \; p(\boldsymbol{y}|0,\boldsymbol{\Sigma}) && (3.104)\\
&= \boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}^T && (3.105)\\
&= \boldsymbol{C} && (3.106)
\end{aligned}
$$

3.12.1 conditional Gaussian distribution

Lets break the parameters, $\boldsymbol{x}$, into two set, $\boldsymbol{y}$ and $\boldsymbol{z}$. We will fix the parameters $\boldsymbol{z}$ and ask what the psf for the parameters $\boldsymbol{y}$ is, $p(\boldsymbol{y}|\boldsymbol{z})$. If the covariance matrix is diagonal then $p(\boldsymbol{y}|\boldsymbol{z})$ is clearly Gaussian. When the covariance is not diagonal the distribution of $\boldsymbol{y}$ is still Gaussian distributed but with a different covariance and mean.

Lets partition the covariance matrix into a part that involves only components of $\boldsymbol{y}$, $\mathbf{A}$, a part that involves only components of $\boldsymbol{z}$, $\mathbf{B}$ and a component that involves mixtures of the two, $\mathbf{D}$.

$$
\boldsymbol{x} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_z \end{bmatrix} \quad \boldsymbol{C} = \begin{bmatrix} \mathbf{C}_{yy} & \mathbf{C}_{zy} \\ \mathbf{C}_{zy}^T & \mathbf{C}_{zz} \end{bmatrix} \qquad (3.107)
$$

The conditional pdf is then

$$
p(\boldsymbol{y}|\boldsymbol{z}) = \mathcal{G}\left(\boldsymbol{y}\,|\boldsymbol{\mu}'_y, \boldsymbol{\Sigma}_{yy}\right) \qquad \begin{cases} \boldsymbol{\mu}'_y = \boldsymbol{\mu}_z + \boldsymbol{C}_{zy}\boldsymbol{C}_{zz}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_z) \\ \boldsymbol{\Sigma}_{yy} = \boldsymbol{C}_{yy} - \boldsymbol{C}_{zy}\boldsymbol{C}_{zz}^{-1}\boldsymbol{C}_{zy}^T \end{cases} \qquad (3.108)
$$

which means

$$
p(\boldsymbol{y}|\boldsymbol{z}) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{\mu}_y+\boldsymbol{C}_{zy}\boldsymbol{C}_{zz}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_z)\right)^T \left(\boldsymbol{C}_{yy}-\boldsymbol{C}_{zy}\boldsymbol{C}_{zz}^{-1}\boldsymbol{C}_{zy}^T\right)^{-1} \left(\boldsymbol{y}-\boldsymbol{\mu}_y+\boldsymbol{C}_{zy}\boldsymbol{C}_{zz}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_z)\right)\right] \qquad (3.109)
$$

3.12.2 marginalized Gaussian distribution

If we integrate over the parameters $\boldsymbol{z}$ we get the marginal distribution

$$
p(\boldsymbol{y}) = \int_{-\infty}^{\infty} d\boldsymbol{z} \; p(\boldsymbol{x}) = \int_{-\infty}^{\infty} d\boldsymbol{z} \; p(\boldsymbol{y},\boldsymbol{z}) = \int_{-\infty}^{\infty} d\boldsymbol{z} \; p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z}) \qquad (3.110)
$$

Using the same definitions

$$
p(\boldsymbol{y}) = \mathcal{G}\left(\boldsymbol{y}\,|\boldsymbol{\mu}_y, \boldsymbol{C}_{yy}\right) \qquad (3.111)
$$

So the correlation with $\boldsymbol{z}$ drop out.

The proof for the conditional and marginal distributions in the general case are rather long algebraicly. I wont go through it, but one step in it is an identity that will be useful in manipulating covariance matrices. This the matrix **completion of squares** formula

$$
\frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + \mathbf{b}^T\boldsymbol{x} = \frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{A}^{-1}\mathbf{b}\right)^T\boldsymbol{A}\left(\boldsymbol{x}-\boldsymbol{A}^{-1}\mathbf{b}\right) - \frac{1}{2}\mathbf{b}^T\boldsymbol{A}^{-1}\mathbf{b} \qquad (3.112)
$$

for a symmetric and invertable $\boldsymbol{A}$ which is the matrix equivalent of the scalar $ax^2 + bx = a(x + \frac{b}{2a})^2 - \frac{b^2}{4a}$.

3.12.3 combining two multivariant Gaussians

$$\mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu}_1,\boldsymbol{C}_1\right)\mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu}_2,\boldsymbol{C}_2\right) = \mathcal{G}\left(\boldsymbol{\mu}_1\,|\boldsymbol{\mu}_2,\boldsymbol{\Sigma}\right)\mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu}_c,\boldsymbol{\Sigma}\right) \tag{3.113}$$

$$\boldsymbol{\Sigma} = \boldsymbol{C}_1 + \boldsymbol{C}_2 \tag{3.114}$$

$$\boldsymbol{\mu}_c = \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{C}_1\boldsymbol{\mu}_1 + \boldsymbol{C}_2\boldsymbol{\mu}_2\right) \tag{3.115}$$

A particularly important application of this is for the distribution of the sum of two independent Gaussian distributed variables. Lets call them $\boldsymbol{x}$ and $\boldsymbol{x}'$ and their sum $\boldsymbol{s} = \boldsymbol{x} + \boldsymbol{x}'$.

$$\begin{aligned}
p(\boldsymbol{y}) &= \int_{-\infty}^{\infty} d^n x \int_{-\infty}^{\infty} d^n x'\; p(\boldsymbol{x},\boldsymbol{x}')\delta(\boldsymbol{s}-\boldsymbol{x}-\boldsymbol{x}') && (3.116)\\
&= \int_{-\infty}^{\infty} d^n x\; p(\boldsymbol{x},\boldsymbol{s}-\boldsymbol{x}) && (3.117)\\
&= \int_{-\infty}^{\infty} d^n x\; \mathcal{G}\left(\boldsymbol{x}\,|0,\boldsymbol{C}_1\right)\mathcal{G}\left(\boldsymbol{s}-\boldsymbol{x}\,|0,\boldsymbol{C}_2\right) && (3.118)\\
&= \int_{-\infty}^{\infty} d^n x\; \mathcal{G}\left(\boldsymbol{x}\,|0,\boldsymbol{C}_1\right)\mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{s},\boldsymbol{C}_2\right) && (3.119)\\
&= \int_{-\infty}^{\infty} d^n x\; \mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu}_c,\boldsymbol{\Sigma}\right)\mathcal{G}\left(\boldsymbol{s}\,|0,\boldsymbol{\Sigma}\right) && (3.120)\\
&= \mathcal{G}\left(\boldsymbol{s}\,|0,\boldsymbol{\Sigma}\right)\int_{-\infty}^{\infty} d^n x\; \mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu}_c,\boldsymbol{\Sigma}\right) && (3.121)\\
&= \mathcal{G}\left(\boldsymbol{s}\,|0,\boldsymbol{\Sigma} = \boldsymbol{C}_1 + \boldsymbol{C}_2\right) && (3.122)
\end{aligned}$$

In particular if

$$\boldsymbol{C}_1 = \sigma_1^2 \quad \text{and} \quad \boldsymbol{C}_1 = \sigma_2^2 \tag{3.123}$$

then

$$\begin{array}{l}
\boldsymbol{C}_1^{-1} = \frac{1}{\sigma_1^2} \quad \text{and} \quad \boldsymbol{C}_1^{-1} = \frac{1}{\sigma_2^2}\\
\boldsymbol{\Sigma} = \sigma_1^2 + \sigma_2^2\\
\boldsymbol{\Sigma}^{-1} = (\sigma_1^2 + \sigma_2^2)^{-1}
\end{array} \tag{3.124}$$

3.13 χ^2 distribution

The χ^2 distribution is not a multivariant distribution, but is closely related to the multivariate Gaussian. Consider a multivariate Gaussian distribution with uncorrelated variable, or a diagonal covariance. Lets define a new variable

$$z = \sum_i^N \frac{(x_i - \mu_i)^2}{\sigma_i^2}. \tag{3.125}$$

z is often called χ^2. This can be confusing because the random variable is not χ, but $z = \chi^2$. We want to change variables from $x_1, x_2, \dots$ to z. The Gaussian distribution is

$$\begin{aligned}
p(x_1, x_2, \dots x_N)dx_1 \dots dx_N &= \frac{1}{(2\pi)^{N/2}\prod_i \sigma_i} e^{-\frac{1}{2}\sum_i^N \frac{(x_i-\mu_i)^2}{\sigma_i^2}} dx_1 \dots dx_N && (3.126)\\
&= \frac{1}{(2\pi)^{N/2}\prod_i \sigma_i} e^{-\frac{1}{2}z} dx_1 \dots dx_N && (3.127)
\end{aligned}$$

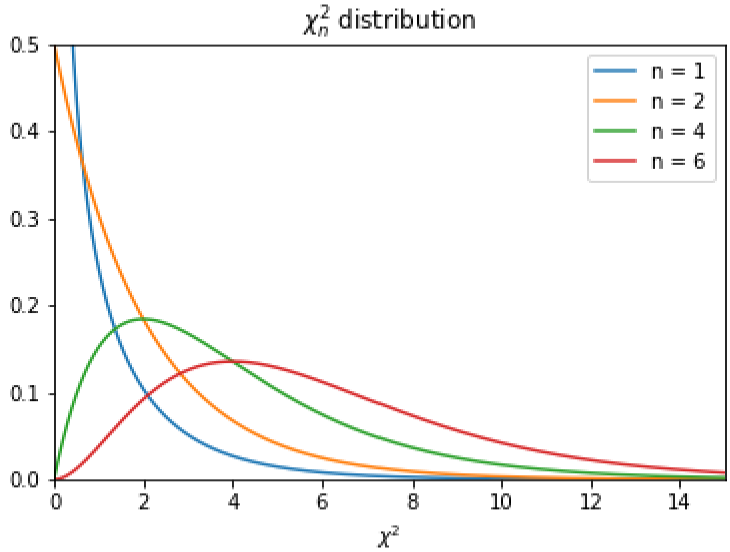

Figure 4: χ_n^2 distribution for some different degrees of freedom, n.

z can be seen as the square of the radial coordinate in N dimensional space

$$dx_1 \dots dx_N = r^{n-1} dr d\theta_1 d\theta_3 \dots = z^{n/2-1} dz d^n\Omega \tag{3.128}$$

Because the pdf is a function of only the z coordinate we can integrate, marginalize, over the angular coordinates which will result in a n dependent normalization constant. The final pdf is

$$p(z = \chi^2|n) = \begin{cases} \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} z^{\frac{n}{2}-1} e^{-\frac{z}{2}} & z \geq 0 \\ 0 & z < 0 \end{cases} \tag{3.129}$$

where the **gamma function** is defined as

$$\Gamma(x) \equiv \int_0^\infty dt\ e^{-t} t^{x-1}. \tag{3.130}$$

This is called the χ^2 distribution of n degrees of freedom. It will be very important for calculating the significance of Gaussian distributed data. The *mean* of this distribution is $E[x] = n$ and the variance $Var[x] = 2n$. For this reason the value of χ_n^2/n is often given and compared to 1. The *mode* is $x =$max$(n-2, 0)$ so $\chi_n^2/n = 1$ is not actually the most likely value. The *skewness* is $\sqrt{8/n}$ so as n increases the pdf becomes more symmetric. The pdf is plotted in figure 4.

The cumulative distribution function can be written down in terms of other special functions without much insight except in the special case of $n = 2$ where it is

$$F(x|2) = 1 - e^{-x/2} \tag{3.131}$$

Theorem 3.1 *If $x_1 \sim \chi_{n_1}^2$, $x_2 \sim \chi_{n_2}^2$ and $s = x_1 + x_2$ then $s \sim \chi_{n_1+n_2}^2$.*

This can be proven in a similar way to how it was shown that the some of squares of Gaussian distributed variables is $\sim \chi^2$.

3.14 student's t-distribution

Yet another distribution that comes up often is the student's t-distribution (or just the t-distribution). We will see that this is used to test if the means of two distributions are the same when the variance in each is not known. The pdf is

$$p_t(x|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{3.132}$$

This distribution has a mean and mode at zero. It is symmetric about this point. Variance is $\frac{\nu}{\nu-2}$ for $\nu > 2$. It resembles a Gaussian, but with more weight in the wings.

4 Sampling

In the last section we dealt with probability distributions and random variables. The means and variances were the means of variances evaluated by summing (or integrating) over all possible values of the random variables. A random variable is a purely theoretical construction and real data consists of a finite set of observed values. These are *sampled* from the distribution or are a sample of the possible data sets. This is where we move from the purely mathematical subject of probability theory to the practical (and more subjective) field of statistics.

A **statistic** is simply any function of points. The arithmetic mean and the sample variance are the simplest example of this. They are used extensively in frequentist statistics. In the case of normally distributed data the probability distribution of these statistics among all possible data sets can be derived analytically. Which makes them an important example and, before computers where widely used, one of the only practical statistics.

In this chapter we will look at some of the basic properties of a finite sample drawn from a random distribution.

4.1 estimating the mean

Say we have a finite sample drawn from a random distribution with pdf $p(x|\mu,\sigma)$ where μ is the mean and σ is the standard distribution. Lets say there are N samples denotes $x_1,\dots x_N$ and they are all independent draws from the distribution.

The **arithmetic mean** of this data is

$$\bar{x}_N \equiv \frac{1}{N}\sum_{i=0}^{N} x_i \tag{4.1}$$

which everyone knows. Confusingly this is usually called just the mean or average just like the mean or average of a distribution, $E[x]$. Although it is usually clear from the context which one is meant, these are distinct concepts. $E[x]$ is a sum over all possible values of x weighted by the pdf and $\bar{x}_N$ is an unweighted sum over a finite sample.

We can take the expectation value of the arithmetic mean

$$\begin{aligned} \langle \bar{x}_N \rangle &= \frac{1}{N}\sum_{i=0}^{N} \langle x_i \rangle & (4.2)\\ &= \frac{1}{N}\sum_{i=0}^{N} \mu & (4.3)\\ &= \mu & (4.4) \end{aligned}$$

This means that the arithmetic mean of a sample is an estimate of the mean of the distribution. This is the simplest example of an **unbiased estimator** (its average equals the quantity being estimated). It is not the only estimator of the mean and it is not always the best estimator of the mean.

For a finite sample the arithmetic mean will not equal the mean of the distribution. One might want to know how good an estimate it is. One way to quantify this is to calculate the variance of the arithmetic mean,

$$
\begin{aligned}
Var[\bar{x}_N] &= \left\langle [\bar{x}_N - \mu]^2 \right\rangle && (4.5)\\
&= \left\langle [\mathrm{Mean}(\{x\})]^2 \right\rangle - 2\mu\langle \mathrm{Mean}(\{x\})\rangle + \mu^2 && (4.6)\\
&= \left\langle [\mathrm{Mean}(\{x\})]^2 \right\rangle - \mu^2 && (4.7)\\
&= \left\langle \left[\frac{1}{N} \sum_{i=0}^{N} x_i \right]^2 \right\rangle - \mu^2 && (4.8)\\
&= \frac{1}{N^2} \sum_{i=0}^{N} \sum_{j=0}^{N} \langle x_i x_j \rangle - \mu^2 && (4.9)\\
&= \frac{1}{N^2} \left[\sum_{i=0}^{N} \langle x_i^2 \rangle + \sum_{i \neq j} \langle x_i x_j \rangle \right] - \mu^2 && (4.10)\\
&= \frac{1}{N^2} \left[\sum_{i=0}^{N} (\sigma^2 + \mu^2) + \sum_{i \neq j} \langle x_i \rangle \langle x_j \rangle \right] - \mu^2 && (4.11)\\
&= \frac{1}{N^2} \left[N(\sigma^2 + \mu^2) + N(N-1)\mu^2 \right] - \mu^2 && (4.12)\\
&= \frac{\sigma^2}{N} && (4.13)
\end{aligned}
$$

So you can see that the standard deviation of the mean will do down like $\propto 1/\sqrt{N}$ no matter what the underlying distribution is. Of course to calculate this variance we need to know the underlying variance, σ^2, which we sometimes do not.

So far we have not made any assumptions about how x is distributed. Since the arithmetic mean is a linear function of the data, if the data is normally distributed the arithmetic mean will be normally distributed.

$$
\text{if} \quad \boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \sigma) \quad \text{then} \quad \mathrm{Mean}(\{x\}) \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{\sigma}{\sqrt{N}} \right) \tag{4.14}
$$

It often happens that one is making repeated measurement of something, say the luminosity of a star, and the variance of the noise is not the same for each measurement because the conditions change or you are combining data from different instruments that have different noise levels. Neither the less the thing you want to know, the luminosity of the star, should be constant. The arithmetic mean (4.2) will on average equal μ, but what it one measurement one measurement has a lot of noise, σ_i is very large. This data point will be a less good estimate of the mean than the other points. Including it in the sum might make the estimate worse rather than better.

Consider the estimator

$$
\hat{\theta} = \sum_i w_i x_i \tag{4.15}
$$

which we can call the **weighted mean**. Clearly the average of this, $\left\langle \hat{\theta} \right\rangle$ will equal μ if

$$
\sum_i w_i = 1. \tag{4.16}
$$

We have the freedom to choose these weights subject to this constraint. A good idea is to minimize the variance of the estimator. This will make it the simplest case of a **minimum variance estimator**. The variance of the estimator will be

$$\begin{aligned}
\sigma_\theta^2 &= \left\langle \theta^2 \right\rangle - \mu^2 && (4.17)\\
&= \sum_{ij} w_i w_j \langle x_i x_j \rangle - \mu^2 && (4.18)\\
&= \sum_{ij} w_i w_j \langle x_i x_j \rangle - \mu^2 && (4.19)\\
&= \sum_{i} w_i^2 \langle x_i^2 \rangle + \sum_{i \neq j} w_i w_j \langle x_i \rangle \langle x_j \rangle - \mu^2 && (4.20)\\
&= \sum_{i} w_i^2 \left[\sigma_i^2 + \mu^2\right] + \mu^2 \sum_{i \neq j} w_i w_j - \mu^2 && (4.21)
\end{aligned}$$

To minimize the variance we will use the technique of **Lagrange multipliers** which you should know from calculus. We minimize the function

$$F(\boldsymbol{w}) = \sigma_\theta^2(\boldsymbol{w}) + \lambda \left(1 - \sum_i w_i\right) \tag{4.22}$$

that is

$$\frac{\partial F}{\partial w_k} = \frac{\partial \sigma_\theta^2}{\partial w_k} - \lambda = 0 \tag{4.23}$$

The derivative of the variance is

$$\begin{aligned}
\frac{\partial \sigma_\theta^2}{\partial w_k} &= 2w_k \left[\sigma_k^2 + \mu^2\right] + 2\mu^2 \sum_{i \neq k} w_i && (4.24)\\
&= 2w_k \left[\sigma_k^2 + \mu^2\right] + 2\mu^2 \left[\sum_{i=0}^{N} w_i - w_k\right] && (4.25)\\
&= 2w_k \left[\sigma_k^2 + \mu^2\right] + 2\mu^2 \left[1 - w_k\right] && \text{use constraint} \quad (4.26)\\
&= 2w_k \sigma_k^2 + 2\mu^2 && (4.27)
\end{aligned}$$

putting this into (4.23) gives

$$w_k = \frac{\lambda - 2\mu}{2\sigma_k^2} \tag{4.28}$$

Plugging this into the constraint (4.16) and solving for

$$\lambda = 2\mu + 2 \left[\sum_k \frac{1}{\sigma_k^2}\right]^{-1} \tag{4.29}$$

so

$$w_k = \left[\sum_i \frac{1}{\sigma_i^2}\right]^{-1} \frac{1}{\sigma_k^2} \tag{4.30}$$

So the estimator (4.15) is

$$\hat{\theta} = \frac{1}{\left[\sum_i \frac{1}{\sigma_i^2}\right]} \sum_i \frac{x_i}{\sigma_i^2}. \tag{4.31}$$

This is often called **inverse noise weighting**. You can see that a data point with a large σ_i^2 will be down weighted with respect to points that have small σ_i^2.

This can be generalized to the case where the data point are correlated as well, but I will leave that for later when we look at estimators and parameter estimation more generally.

4.2 estimating the variance

Lets go back to the case of N data points sampled from the same distribution. We might want to know the variance of the distribution. This could be the variance from noise so we can measure how well our apparatus is working or it could be that we are interested in the variance of the "signal" itself that is not constant. For example say we want to characteristic ocean waves from discrete measurements of the height of the water's surface. The variance in the height might be a good quantity to measure.

Known mean: If the mean of the underling distribution is known we can estimate the variance of that distribution with

$$S_N^2 = \frac{1}{N} \sum_i (x_i - \mu)^2 \tag{4.32}$$

You can easily show that $\left\langle S_N^2 \right\rangle = \sigma^2$.

Unkown mean: In most cases one does not know the average ahead of time. In this case the best estimator is

$$S_N^2 = \frac{1}{N-1} \sum_i \left(x_i - \bar{x}_N\right)^2. \tag{4.33}$$

Why is there an $N-1$ instead of an N in the denominator? Lets look at the average of it

$$\begin{aligned}
\left\langle S_N^2 \right\rangle &= \frac{1}{N-1} \sum_i \left\langle \left(x_i - \bar{x}_N\right)^2 \right\rangle && (4.34)\\
&= \frac{1}{N-1} \left[\sum_i \left\langle x_i^2 \right\rangle - 2 \left\langle \sum_i x_i \bar{x}_N \right\rangle + \sum_i \left\langle (\bar{x}_N)^2 \right\rangle \right] && (4.35)\\
&= \frac{1}{N-1} \left[\sum_i (\sigma^2 + \mu^2) - 2N \left\langle (\bar{x}_N)^2 \right\rangle + N \left\langle (\bar{x}_N)^2 \right\rangle \right] && (4.36)\\
&= \frac{1}{N-1} \left[\sum_i (\sigma^2 + \mu^2) - N \left\langle (\bar{x}_N)^2 \right\rangle \right] && (4.37)\\
&= \frac{1}{N-1} \left[N(\sigma^2 + \mu^2) - N \left(\frac{\sigma^2}{N} + \mu \right) \right] \qquad \text{using (4.13)} && (4.38)\\
&= \sigma^2 && (4.39)
\end{aligned}$$

So this estimator is unbiased. Note that this does not require that the x's be normally distributed. If there were an N in the denominator of (4.33) then $\left\langle s_N^2 \right\rangle = (N-1)\sigma/N$ which means it would be **biased**, but since the bias gets smaller as N increases it would be a simple example of an **asymptotically unbiased estimator**.

Theorem 4.1 *If $x_i \sim \mathcal{N}(\mu,\sigma)$ and S_n is given by (4.33) then $z = \frac{(n-1)S_n^2}{\sigma^2}$ is χ^2_{n-1} distributed.*

Proof:

$$(n-1)S_n^2 = \sum_i (x_i - \bar{x})^2 \tag{4.40}$$

$$= \sum_i \left[(x_i - \mu) - (\bar{x} - \mu)\right]^2 \tag{4.41}$$

$$= \sum_i \left[(x_i - \mu)^2 - 2(x_i - \mu)(\bar{x} - \mu) - (\bar{x} - \mu)^2\right] \tag{4.42}$$

$$= \sum_i \left[(x_i - \mu)^2\right] - 2n(\bar{x} - \mu)(\bar{x} - \mu) - n(\bar{x} - \mu)^2 \tag{4.43}$$

$$= \sum_i (x_i - \mu)^2 - n(\bar{x} - \mu)^2 \tag{4.44}$$

This is the difference of two χ^2 distributed quantities: $(\bar{x}-\mu)^2/(\sigma^2/n) \sim \chi^2_1$ and $\sum_i (x_i - \mu)^2 \sim \chi^2_n$. By theorem 3.1 the sum of the χ^2 distributed is χ^2 distributed. QED

Measuring the variance of a signal is closely related to measuring the correlation function or the power spectrum of a signal. We will return to that problem later.

4.3 estimating the mean when the variance is known

We have learned that the $\bar{x}$ is $\mathcal{N}(\mu, \sigma/\sqrt{n})$ distributed if the x_i's are normally distributed. So if we have a measurement and we know the noise, σ, we can put an error on our estimate of the mean $\pm\frac{\sigma}{\sqrt{n}}$. But often we do not know the σ's. We can estimate it with S_n^2, but this estimate is based on the same data as the estimate of $\bar{x}$ and so $\bar{x}$ will *not* be $\mathcal{N}(\mu, S_n/\sqrt{n})$ distributed.

Theorem 4.2 *If $x_i \sim \mathcal{N}(\mu,\sigma)$ then*

$$z = (\bar{x} - \mu)\sqrt{\frac{n}{S_n^2}} \tag{4.45}$$

is student-t distributed with $n-1$ degrees of freedom.

The t-distribution was introduced in section 3.14.

So if we wanted to measure the average level of some chemical in people's blood, for example, we might model the underlying distribution, human variation plus measurement error, to be Gaussian. We do not know the variance among people or perhaps the error in our chemical testing equipment. We estimate the mean with the arithmetic mean, $\bar{x}$, and we can calculate the probability of this estimate being within $\pm\delta x$ as

$$p(\mu - \delta x < \bar{x} < \mu + \delta x) = \int_{-\delta x/\sqrt{\frac{n}{S_n^2}}}^{+\delta x/\sqrt{\frac{n}{S_n^2}}} dt\ p_t(t|\nu = n-1) \tag{4.46}$$

$$= \sqrt{\frac{n}{S_n^2}} \int_{-\delta x}^{+\delta x} dx'\ p_t\left(x'\sqrt{\frac{n}{S_n^2}}\right|\nu = n-1\Bigg) \tag{4.47}$$

where $p_t(t|\nu)$ is given in section 3.14. Note that we calculate the probability that $\bar{x}$, a statistic of random data, will be within some range of μ, an unknown parameter. This is an example of frequentist hypothesis testing. We will return to this kind of problem later and examine it in detail.

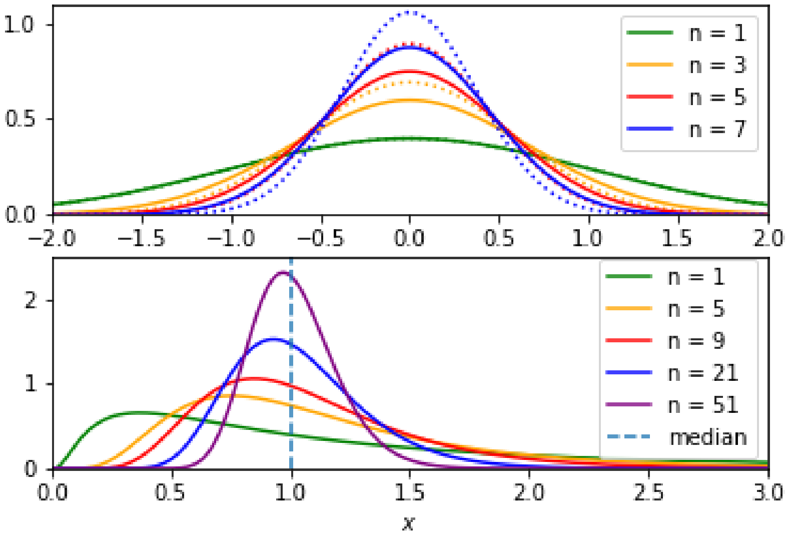

Figure 5: The probability of the sample median for normal (above) and lognormal (below) distributions. The $n = 1$ case is the original distribution. The dotted curves in the normal case are the distributions of the sample means based on the same n's.

4.4 median

It is often useful to estimate the median of a distribution. It can be a better representative value of a distribution than the mean when the distribution is highly skewed. A common example of this is the median income of a population. A small number of people with very high incomes can have a large effect on the mean income, but the median is a more robust representative value for a typical person in that population. Also,the median can often be more accurately estimated from a small number of observations than the mean. This is particularly true for a distribution with extended tails like a power-law or Lorentzian where the mean might not even be defined.

Consider the median of a sample. Lets assume there are an odd number of observation so the median is well defined. For the median to have value x one observation must be between x and $x + dx$. The probability of this is $p(x)dx$. In addition there must be $(N - 1)/2$ observed smaller (and larger) values out of the remaining $N - 1$ values. The probability of an observation being below x is the cumulative probability function $F(x)$. The probability of n independent observations out of $N - 1$ having being $< x$ is the binomial distribution $P_{\text{binom}}(n|N - 1, p = F(x))$. The probability of both of these things happening is the product of their probabilities (product rule for independent events). Any of the N values could be the median so there is a factor of N. The final pdf for the median is

$$p_{\text{med}}(x|N) = Np(x)P_{\text{multi}}\left(\frac{(N-1)}{2}\middle| F(x), N-1\right) \tag{4.48}$$

In the limit of large N this distribution becomes normal with variance

$$Var[x_{\rm m}] = \frac{1}{4(N+2)p(x_m)^2} \tag{4.49}$$

For $x \sim \mathcal{N}(\mu, \sigma)$ the sample mean has a smaller variance than the sample median by a factor of $\sim \frac{2}{\pi}\frac{(N+2)}{N}$. For a distribution with larger tails then Gaussian and with fewer samples the median will have a smaller variance than the mean.

4.5 extreme values

The distribution of the sample maximum (or minimum or the n-th largest value) can be found in the same way as the the median

$$\begin{align} p_{\max}(x|N) &= Np(x)P_{\rm multi}\left(N-1|F(x), N-1\right) \tag{4.50}\\ &= Np(x)P_{\rm multi}\left(0|1-F(x), N-1\right) \tag{4.51}\\ &= Np(x)F(x)^{N-1} \tag{4.52}\end{align}$$

4.6 quintile estimation

The q-**quintiles** of a distribution are the set of values that divide the full range into q regions of equal probability. They are the generalization of the median which would be the 2-quintile. The nth q-quintile is is at the point where $F(x) = n/q$. There are several slightly different ways to estimate this from a sample, but they all agree for large N and generally follow this approach. **Rank** the data (order them by value from least to greatest) and then take the data point whose rank is closest to $r = nN/q + 1/2$ to be an estimate of the nth q-quintile. This $1/2$ makes the ranks for the median ($q = 2$, $n = 1$) work out to the sample median we used before. There are other choices which have over properties (see wikapedia). If r is an integer then we can work out pdf in the same way as before.

$$\begin{align} p(x_n|N) &= Np(x_n)P_{\rm multi}\left(r-1|F(x_n), N-1\right) \tag{4.53}\\ p(x_n|N) &= Np(x_n)P_{\rm multi}\left(\frac{nN}{q} - \frac{1}{2}\middle| F(x_n), N-1\right) \tag{4.54}\end{align}$$

As we will see, when doing Monte Carlo calculations you might only have access to a sample taken from a distribution that you cannot write down analytically. It is often useful to estimate the quintile range the distribution or estimate a range that contains some fixed probability, say 68% or 95%. One might use (4.53) with an estimate of the true pdf to judge how well the range can be estimated.

5 The Bayesian method

The Bayesian approach gives us a general framework for constraining models for physical processes and for models that describe the probabilistic distribution of the data. It does this by attempting to calculate the probability of a model or specific values for model parameter given the data and any prior knowledge. The Bayesian interpretation of probability allows us to assign a probability to the possibility of a model being the true one. In contrast, the frequentist approach, that we will look at in the next section, prohibits assigning probability to the models; only data is probabilistic.

advantages of Bayesian approach

Bayesian inference

5.1 Posterior, likelihood, prior and evidence

All Bayesian analyses begin with Bayes's theorem. We saw this theorem in section 1.4 as a basic property of conditional probabilities. Let me point out that the theorem itself is a mathematical relation and thus its validity no matter what your interpretation of probability is or what your approach to statistical inference is. The difference between frequentist and Bayesian statistics fundamentally lies in what what you apply these probabilities to.

Let $\boldsymbol{D}$ be some amount of data. Let M_i be a model that attempts to explain this data. It is a member of a set of models $\{M_1, M_2 \dots\}$. These models might be totally different with different parameters (say General Relativity, Newtonian Gravity and MOND) or they might differ buy only the values of a model's parameters (the planet has unknown mass m). Lets let I represent everything else in the Universe that we will take to be fixed or irrelevant to our experiment (existence of the apperatise, the day of the week, the phase of the moon on a distant planet). We apply Bayes's theorem to this situation

$$P(M_i|\boldsymbol{D}, I) = \frac{P(\boldsymbol{D}|M_i, I)P(M_i|I)}{P(\boldsymbol{D}|I)} \tag{5.1}$$

$$= \frac{P(\boldsymbol{D}|M_i, I)P(M_i|I)}{\sum_i P(\boldsymbol{D}|M_i, I)P(M_i|I)} \tag{5.2}$$

The second line follows from $P(\boldsymbol{D}|I) = \sum_i P(\boldsymbol{D}|M_i, I)P(M_i|I)$ which is the probability that the data will occur is any one of the models is correct. I include I here only to emphasis that every probability has some implicit assumptions. Some of these assumptions could be incorporated into the model, but if they have no effect on the outcome of the experiment or they where never changed when the experiments where conducted they can be considered conditionals for all the probabilities. In the future the I will be considered implicit and not included.

In this context, each of the factors in Bayes's theorem have special names:

$P(M_i|\boldsymbol{D})$ is called the **posterior probability** for model M_i given the data. This is the goal of Bayesian inference.

$P(\boldsymbol{D}|M_i)$ is called the **likelihood**. It is the probability of getting the observed data given the model M_i. It is often denoted $\mathcal{L}(\boldsymbol{D}|M_i)$. This is the same probability as is used in frequentist methods. Often this is a Gaussian, but not always.

$P(M_i)$ is called the **prior**. It is the probability of the model prior to the data $\boldsymbol{D}$ being considered. This might take into account some previous experiment with data $\boldsymbol{D}'$ in which case it would be the posterior of that experiment $P(M_i|\boldsymbol{D}')$. It might also take into account that some models, or range of parameters, is not possible in which case $P(M_i) = 0$ for some i. For example, the mass of a planet cannot be negative.

$P(\boldsymbol{D}) = \sum_i P(\boldsymbol{D}|M_i, I)P(M_i|I)$ is called the **evidence**. Not that the evidence is not a function of M_i although it is implicitly dependent on the set of all model considered. Since the data does not change the evidence will be a constant for a fixed set of models. We will sometimes denote the evidence as $\mathcal{E}(\boldsymbol{D})$.

5.2 Parameter estimation

The most common use for Bayesian inference is parameter estimation. In this case we have a model that describes the data that is a function of parameters $\theta_1, \theta_2, \ldots$. The different models discussed above are actually the same model with different values. We will assume that these parameters take on a continuous range of values, although this is not necessary. The sum in the evidence then becomes an integral and the posterior is

$$P(\theta_1, \theta_2, \ldots |\boldsymbol{D}) = \frac{\mathcal{L}(\boldsymbol{D}|\theta_1, \theta_2, \ldots)p(\theta_1, \theta_2, \ldots)}{\left[\int d^n\theta \; \mathcal{L}(\boldsymbol{D}|\theta_1, \theta_2, \ldots)p(\theta_1, \theta_2, \ldots)\right]} \tag{5.3}$$

The posterior expresses the probability of a set of parameter values being correct *given that the model is the correct one.*

5.2.1 example:

Lets say you have a sample of water from a swamp next to a nuclear power plane. We want to now the level of radioactive contamination in this water. Let there be $N(t)$ unstable nuclei in our sample. The rate of decay is $\frac{dN}{dt} = \lambda N(t)$ where λ is the decay constant. Let's say we know what element we are dealing with and previous studies have measured the decay constant to a higher enough accuracy that we can consider it a known constant. The average rate of decay products going into a Geiger counter is then $r = \Omega\lambda N(t)$ assuming one product per decay. Ω is the angular area covered by the Geiger counter from the prospective of the sample which we will also assume is well enough measured that it can be considered known. If we can measure r we can easily find $N(t)$. We will measure the number of counts in the Geiger counter over a period of time that is small compared to $1/\lambda$ so that we can consider $N(t)$ to be small (Uranium 235 has a decay constant of 3.12×10^{-17} s^{-1} or $1/\lambda = 1.02$ Gyr so this isn't hard in most cases).

Since each nucleus has an constant probability of decay the number of counts, n, will be Poisson distributed (see section 3.5).

$$p(n|r) = \frac{(r\delta t)^n}{n!} \, e^{-r\delta t} \tag{5.4}$$

where δt is the time over which the measurement is done. In this case n is the data and r is the parameter we would like to measure. This Poisson distribution is the likelihood. We take the prior on the rate to be uniform between 0 and some large number r_{max}. We will see that the result will not depend on the value of r_{max} as long as it is much larger than the actual rate,

$$p(r) = \frac{\Theta(0 < r < r_{max})}{r_{max}} \tag{5.5}$$

We know that $p(n|r)$ is normalized to one for its sum over n from 0 to ∞, but to normalize the posterior by calculating the evidence we need to integrate $p(n|r)p(r)$ over r.

$$\begin{aligned}
\mathcal{E}(n) &= \int_{-\infty}^{\infty} dr\ p(n|r)p(r) = \frac{1}{r_{max}} \int_0^{r_{max}} dr\ \frac{(r\delta t)^n}{n!}\ e^{-r\delta t} && (5.6)\\
&= \frac{\delta t^{-1}}{n! r_{max}} \int_0^{\delta t r_{max}} dx\ x^n\ e^{-x} \qquad x = r\delta t && (5.7)\\
&\simeq \frac{\delta t^{-1}}{n! r_{max}} \int_0^{\infty} dx\ x^n\ e^{-x} \qquad r_{max} \gg 1/\delta t && (5.8)\\
&= \frac{\delta t^{-1}}{n! r_{max}} \Gamma\left(n+1\right) && (5.9)\\
&= \frac{1}{\delta t r_{max}} \qquad \text{because } \Gamma\left(n+1\right) = n! && (5.10)
\end{aligned}$$

So the posterior for the rate is

$$p(r|n) = \frac{\delta t}{n!}(\delta t r)^n e^{-r\delta t} \tag{5.11}$$

The average of this distribution is

$$\begin{aligned}
\langle r \rangle = \int_o^{\infty} dr\ rp(r|n) = \frac{\delta t}{n!} \int_o^{\infty} dr\ r(\delta t r)^n e^{-r\delta t} &= \frac{1}{\delta t n!} \int_o^{\infty} dx\ x^{n+1} e^{-x} && (5.12)\\
&= \frac{(n+1)!}{\delta t n!} = \frac{(n+1)}{\delta t} && (5.13)
\end{aligned}$$

and the variance is

$$Var\left[r\right] = \frac{(n+1)}{\delta t^2} \tag{5.14}$$

One might have expected that the rate should be $\sim n/\delta t$ and that the standard deviation should go like $\propto \sqrt{n}$. Why these extra 1s? We will see later that this small difference in expectation value for small n is related to our choice of prior.

5.2.2 example:

Lets say we have a very simple model for the alcohol content of wine coming out of a winery. The model is that it is constant. We will call the concentration θ. We know that our measurement apparatus has a Gaussian distributed error of σ when measuring the concentration. Say we measure one bottle and get d for the concentration. This kind of model is often written

$$d_i = \theta + n_i, \tag{5.15}$$

the data is some fixed value plus a noise component. The likelihood will be

$$\mathcal{L}(d|\theta) = \mathcal{G}\left(d\left|\theta, \sigma\right.\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(d-\theta)^2}{2\sigma^2}} \tag{5.16}$$

Now we need a prior for θ. It is common to use a uniform prior in this kind of problem. The argument for this being that without any measurements no particular concentration should be considered more probable than any other. So the prior will be

$$p(\theta) = \begin{cases} \frac{1}{\theta_{\max}-\theta_{\min}} & \theta_{\min} < \theta < \theta_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{5.17}$$

$$= \mathcal{C} \times \begin{cases} 1 & \theta_{\min} < \theta < \theta_{\max} \\ 0 & \text{otherwise} \end{cases} \qquad \mathcal{C} \equiv \frac{1}{\theta_{\max} - \theta_{\min}} \tag{5.18}$$

You might be concerned that the parameters $\theta_{\max}$ and $\theta_{\max}$ might effect the posterior, but we don't know their values. Note that if the likelihood constrains θ to a region that is much smaller than the range of $p(\theta)$ then it will not make any difference. Not also that the normalization of both the likelihood and the prior appear in both the numerator and denominator of the posterior so they drop out. If we take the range of the prior to be much larger than σ, the uniform prior will drop out and not appear.

So in that case the posterior is equal to the likelihood, $\mathcal{G}\left(d\,|\theta,\sigma\right)$ which obviously has a mode at $\theta = d$ and the average is $\langle\theta\rangle = d$.

Now lets consider a slightly more complicated case. We measure N bottles of wine coming out of the factory getting $d_1, d_2 \ldots d_n$ measurements, all with the same σ. Since these are statistically independent measurements the likelihood will be

$$\mathcal{L}(\boldsymbol{d}|\theta) = \mathcal{G}\left(d_1\,|\theta,\sigma\right)\mathcal{G}\left(d_2\,|\theta,\sigma\right)\ldots \tag{5.19}$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2}\sum_i \frac{(d_i-\theta)^2}{\sigma^2}\right) \tag{5.20}$$

which will also be the the posterior for a uniform prior.

$$\mathcal{L}(\boldsymbol{d}|\theta) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_i \left(d_i^2 - 2d_i\theta + \theta^2\right)\right) \tag{5.21}$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\left[\sum_i d_i^2 - 2\sum_i d_i\theta + n\theta^2\right]\right) \tag{5.22}$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\left[n\overline{d^2} + n(\theta-\bar{d})^2 - n(\bar{d})^2\right]\right) \tag{5.23}$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{n}{2\sigma^2}\left[\overline{d^2} - (\bar{d})^2\right]\right) \exp\left(-\frac{n}{2\sigma^2}(\theta-\bar{d})^2\right) \tag{5.24}$$

where

$$\overline{d^2} \equiv \frac{1}{n}\sum_i d_i^2. \tag{5.25}$$

To find the evidence we need to integrate this over θ.

$$\mathcal{E}(\boldsymbol{d}) = \frac{\mathcal{C}^n}{(2\pi)^{(n-1)/2}\sigma^{n-1}\sqrt{n}} \exp\left(-\frac{n}{2\sigma^2}\left[\overline{d^2} - (\bar{d})^2\right]\right) \tag{5.26}$$

All the constant factors will drop out of the posterior. The only part that is dependent on θ is proportional to a Gaussian. Since we already know the normalization of a Gaussian we don't even need to to the integration in this case. The posterior is

$$P(\theta|\boldsymbol{d}) = \sqrt{\frac{n}{2\pi\sigma^2}} \exp\left(-\frac{n}{2\sigma^2}(\theta-\bar{d})^2\right) = \mathcal{G}\left(\theta\,|\bar{d},\sigma^2/n\right). \tag{5.27}$$

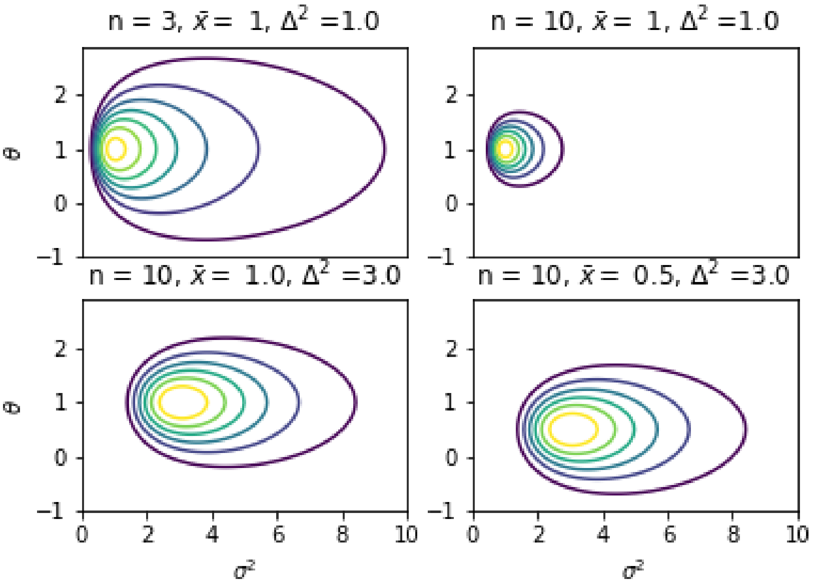

Figure 6: .

In section 4.1 we found that the sample mean of Gaussian random variables is Gaussian distributed with a variance of σ^2/n. We see here that this is also true for the posterior distribution of the estimated mean. The mean is $\langle\theta\rangle = \boldsymbol{d}$. No surprise here.

5.2.3 example:

Lets make it a little more complicated. It does not seem reasonable that the alcohol content is exactly constant in every bottle of wine so we should allow for it to change randomly with an unknown variance. We still have a normally distributed error in the measurements with standard deviation σ_n. In addition we will assume the distribution of the alcohol content among bottles is normally distributed with a mean of θ and a variance of σ_a. We would like to know the variance and in the future we would like to adjust the process to reduce the variance so that the product is more uniform. Some customers have been complaining.

Each data point is some constant plus (or minus) some random value plus random noise:

$$d_i = \theta + x_i + n_i \tag{5.28}$$

We can think of the likelihood as the probability of the actual alcohol is x and then the probability of the alcohol level x being measured as d,

$$\begin{aligned}\mathcal{L}(\boldsymbol{d}|\theta,\sigma_n^2,\sigma_a^2) &= \int_{-\infty}^{\infty} d^n x \; P(\boldsymbol{d},\boldsymbol{x}|\theta,\sigma_a^2) && (5.29)\\ &= \int_{-\infty}^{\infty} d^n x \; \left[\mathcal{G}\left(d_1\left|x_1,\sigma_n^2\right.\right)\mathcal{G}\left(d_2\left|x_2,\sigma_n^2\right.\right)\ldots\right]\left[\mathcal{G}\left(x_1\left|\theta,\sigma_a^2\right.\right)\mathcal{G}\left(x_2\left|\theta,\sigma_a^2\right.\right)\ldots\right] && (5.30)\\ &= \int_{-\infty}^{\infty} d^n x \; \mathcal{G}\left(\boldsymbol{d}\left|\boldsymbol{x},\sigma_n^2\right.\right)\mathcal{G}\left(\boldsymbol{x}\left|\theta,\sigma_a^2\right.\right) && (5.31)\\ &= \mathcal{G}\left(\boldsymbol{d}\left|\theta,\sigma_n^2+\sigma_a^2\right.\right) && (5.32)\end{aligned}$$

where we are using the results of section 3.12.3 to combine Gaussian pdfs. This is the same likelihood as we got in the first example except σ^2 is replaced with $\sigma_n^2+\sigma_a^2$,

$$\mathcal{L}(\boldsymbol{d}|\theta,\sigma_n^2,\sigma_a^2) = \frac{1}{\sqrt{(2\pi)^{n/2}(\sigma_n^2+\sigma_a^2)^n}}\exp\left(-\frac{n\left[\overline{d^2}-(\bar{d})^2\right]}{2(\sigma_n^2+\sigma_a^2)}\right)\exp\left(-\frac{n(\theta-\bar{d})^2}{2(\sigma_n^2+\sigma_a^2)}\right) \tag{5.33}$$

To make things simpler lets make the following substitutions

$$\begin{aligned}\Delta^2 &\equiv \overline{d^2}-(\bar{d})^2 && (5.34)\\ \sigma^2 &\equiv \sigma_n^2+\sigma_a^2 && (5.35)\end{aligned}$$

You can see that σ_n and σ_a enter into the likelihood only in the combination $\sigma_n^2+\sigma_a^2$. As a result you cannot constrain them separately unless the priors differentiates between them. This is possible. For example some previous calibration tests could put constraints on σ_n.

We will take the case where there are no previous constraints on either of the σ's. We can then use σ^2 as a parameter instead of σ_a^2. The likelihood is now

$$\mathcal{L}(\boldsymbol{d}|\theta,\sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\frac{n\Delta^2}{2\sigma^2}\right)\exp\left(-\frac{n(\theta-\bar{d})^2}{2\sigma^2}\right) \tag{5.36}$$

We will assume a uniform prior for both θ and σ^2 (we will talk later about using a Jeffreys prior for σ^2). Further more the variance cannot be less than zero

$$\begin{aligned}P(\theta,\sigma^2) &= \frac{\Theta(\theta_{\text{max}}<\theta<\theta_{\text{min}})}{(\theta_{\text{max}}-\theta_{\text{min}})}\frac{\Theta(0<\sigma^2<\sigma^2_{\text{max}})}{\sigma^2_{\text{max}}} && (5.37)\\ &= \mathcal{C}\Theta(\theta_{\text{max}}<\theta<\theta_{\text{min}})\Theta(0<\sigma^2<\sigma^2_{\text{max}}) && (5.38)\end{aligned}$$

where $\mathcal{C}$ is going to represent the normalization constant.

Now we need to find the evidence by integrating the likelihood over the parameters.

$$\begin{aligned}\mathcal{E}(\boldsymbol{d}) &= \mathcal{C}\int_0^{\sigma^2_{\text{max}}} d\sigma^2 \int_{\theta_{\text{min}}}^{\theta_{\text{max}}} d\theta \; \mathcal{L}(\boldsymbol{d}|\theta,\sigma^2) && (5.39)\\ &\simeq \mathcal{C}\int_0^{\sigma^2_{\text{max}}} d\sigma^2 \int_{-\infty}^{\infty} d\theta \; \mathcal{L}(\boldsymbol{d}|\theta,\sigma^2) && (5.40)\\ &= \frac{\mathcal{C}}{(2\pi)^{n/2}}\int_0^{\sigma^2_{\text{max}}} d\sigma^2 \int_{\theta_{\text{min}}}^{\theta_{\text{max}}} d\theta \; \frac{1}{\sigma^n}\exp\left(-\frac{n\Delta^2}{2\sigma^2}\right)\exp\left(-\frac{n(\theta-\bar{d})^2}{2\sigma^2}\right) && (5.41)\\ &= \frac{\mathcal{C}}{(2\pi)^{(n-1)/2}}\int_0^{\sigma^2_{\text{max}}} d\sigma^2 \; \frac{1}{\sigma^{n-1}}\exp\left(-\frac{n\Delta^2}{2\sigma^2}\right) && (5.42)\end{aligned}$$

In doing this we have taken the the range of the θ integral to go to infinity. This is justifiable if $|\sigma^2_{\rm max}| = |\sigma^2_{\rm min}| \gg \sigma^2$. We don't know this ahead of time, but can be justified in retrospect once constraints on σ are found. This can be considered a technical flaw that we will get back to later.

Now lets make the change of variables to

$$y = \sqrt{\frac{n\Delta^2}{2\sigma^2}} \quad \text{so} \quad d\sigma^2 = \frac{n\Delta^2}{y^3} dy \tag{5.43}$$

$$\mathcal{E}(\boldsymbol{d}) = \frac{2\mathcal{C}}{(2\pi)^{(n-1)/2}} \left(\frac{n\Delta^2}{2} \right)^{\frac{3-n}{2}} \int^{\infty}_{\sqrt{\frac{n\Delta^2}{2\sigma^2_{\rm max}}}} dy\ y^{n-4} e^{-y^2} \tag{5.44}$$

$$\simeq \frac{2\mathcal{C}}{(2\pi)^{(n-1)/2}} \left(\frac{n\Delta^2}{2} \right)^{\frac{3-n}{2}} \int^{\infty}_{0} dy\ y^{n-4} e^{-y^2} \tag{5.45}$$

$$= \frac{\mathcal{C}}{(2\pi)^{(n-1)/2}} \left(\frac{n\Delta^2}{2} \right)^{\frac{3-n}{2}} \Gamma\left(\frac{n-3}{2} \right) \tag{5.46}$$

Here we assumed that $\sigma^2_{\rm max} \gg n\Delta^2$ in the integration limits.

Now we can construct the posterior. The constant $\mathcal{C}$ in the prior and the evidence will cancel. We can then take the limits to go to infinity or at least so large that there is no need to put the $\Theta()$ parts of the prior in the posterior because the likelihood will constrain the parameters to be much less than this value. The posterior is

$$P(\theta, \sigma^2|\boldsymbol{d}) = \frac{1}{\sqrt{2\pi}\Gamma\left(\frac{n-3}{2}\right)} \left(\frac{\Delta^2}{2} \right)^{\frac{n-3}{2}} \left(\frac{n}{\sigma^2} \right)^{\frac{n}{2}} \frac{1}{n} \exp\left(-\frac{n\Delta^2}{2\sigma^2} \right) \exp\left(-\frac{n(\theta - \bar{d})^2}{2\sigma^2} \right) \tag{5.47}$$

This posterior is plotted in figure 6 for some values of n, $\bar{d}$ and Δ^2.

The mod of the posterior can be found by setting its derivatives with respect to the parameters to zero. It is often more convenient to take the log of the posterior first. Since the log is a monitonic function its maximum will be at the same place.

$$\ln P(\theta, \sigma^2|\boldsymbol{d}) = -\frac{n}{2} \ln(\sigma^2) - \frac{n}{2\sigma^2} \left[\Delta^2 + (\theta - \bar{d})^2 \right] + \text{constant terms} \tag{5.48}$$

$$\frac{\partial}{\partial\theta} \ln P(\theta, \sigma^2|\boldsymbol{d}) = -\frac{n}{\sigma^2}(\theta - \bar{d}) \tag{5.49}$$

$$\frac{\partial}{\partial\sigma^2} \ln P(\theta, \sigma^2|\boldsymbol{d}) = \frac{n}{2\sigma^2} \left(-1 + \frac{\Delta^2}{\sigma^2} + \frac{(\theta - \bar{d})^2}{\sigma^2} \right) \tag{5.50}$$

These are simultaneously zero at $\theta = \bar{d}$, $\sigma^2 = \Delta^2 = \overline{d^2} - \bar{d}^2$. These are almost, but not quite what we would have gotten with the arithmetic mean and variance we saw before in section 4. Specifically the $(N-1)^{-2}$ factor that we saw was needed to make the estimator unbiased has a been replaced with N^{-2}.

I chose to use σ^2 as a parameter, but I could just as well have chosen σ or $\sqrt{\sigma}$ as a parameter instead. The likelihoods would all be the same, but the evidence would be different since it would be an integral over a different variable. Since, by the chain rule,

$$\frac{\partial}{\partial\sigma^2} \ln P(\theta, \sigma^2|\boldsymbol{d}) = \frac{1}{2\sigma} \frac{\partial}{\partial\sigma} \ln P(\theta, \sigma|\boldsymbol{d}) = \frac{1}{4\sigma^3} \frac{\partial}{\partial\sigma^{1/2}} \ln P(\theta, \sigma^{1/2}|\boldsymbol{d}) \tag{5.51}$$

they will all be zero at the same spot the maximum of the posterior will give the same value. However the mean parameter values will not be the same, $\left\langle \sigma^2 \right\rangle \neq \left\langle \sigma \right\rangle^2$.

5.3 Marginalization

The situation often comes up where there are parameters of physical or statistical model that we are not interested in. For example we may not know what the variance is, but we are only interested in the mean. Or we may want to make a statement about the constraints on one or two parameters that is independent of what value all the other parameters have. In the Bayesian context these parameters that we are not interested in are called **nuisance parameters**. To remove them from the posterior we marginalize over them.

Lets say parameters $\alpha_1, \alpha_2, \ldots$ are the parameters we are interested in and parameters $\beta_1, \beta_2, \ldots$ are the ones we aren't interested.

$$P(\alpha_1, \alpha_2, \ldots |\boldsymbol{D}) = \int_{-\infty}^{\infty} d\beta_1 \int_{-\infty}^{\infty} d\beta_2 \ldots\ P(\alpha_1, \alpha_2, \ldots, \beta_1, \beta_2, \ldots |\boldsymbol{D})$$
$$= \frac{\int_{-\infty}^{\infty} d\beta_1 \int_{-\infty}^{\infty} d\beta_2 \ldots\ P(\boldsymbol{D}|\alpha_1, \alpha_2, \ldots, \beta_1, \beta_2, \ldots) P(\alpha_1, \alpha_2, \ldots, \beta_1, \beta_2, \ldots)}{\int_{-\infty}^{\infty} d\alpha_1 \int_{-\infty}^{\infty} d\alpha_2 \ldots \int_{-\infty}^{\infty} d\beta_1 \int_{-\infty}^{\infty} d\beta_2 \ldots\ P(\boldsymbol{D}|\alpha_1, \alpha_2, \ldots, \beta_1, \beta_2, \ldots) P(\alpha_1, \alpha_2, \ldots, \beta_1, \beta_2, \ldots)} \tag{5.52}$$

5.3.1 example:

As a simple example lets say we have the posterior (5.47). We are interested in the parameter θ, but we are not interested in the "noise" parameter σ^2. Lets marginalize over σ^2 so we have the distribution of θ alone.

We can ignore all the factors that don't have θ or σ^2 in them for the moment because they are just a normalization and we can recover the normalization at the end by integrating over θ. Lets make the substitution $A = n\Delta^2 + n(\theta - \bar{d})^2$ in which case the relevant parts of the posterior are

$$P(\theta|\Delta^2, \bar{d}) = \int_0^{\infty} d\sigma^2\ P(\theta, \sigma^2|\Delta^2, \bar{d}) \tag{5.53}$$
$$\propto \int_0^{\infty} d\sigma^2\ \frac{e^{-\frac{A}{2\sigma^2}}}{\sigma^n} \tag{5.54}$$
$$\propto -2 \int_{\infty}^{0} dx\ x^{n-3} e^{-\frac{A}{2}x^2} \qquad x = \frac{1}{\sigma} \tag{5.55}$$
$$\propto 2^{\frac{n-3}{2}} A^{-\left(\frac{n-2}{2}\right)} \Gamma\left(\frac{n-2}{2}\right) \qquad \text{integral in Appendix C} \tag{5.56}$$
$$\propto \left[\Delta^2 + \left(\theta - \bar{d}\right)^2\right]^{\frac{2-n}{2}} \tag{5.57}$$
$$\propto \left[1 + \frac{\left(\theta - \bar{d}\right)^2}{\Delta^2}\right]^{\frac{2-n}{2}} \tag{5.58}$$

If we compare this to the t-distribution (3.132) in section 3.14 we recognize that this $x = |\theta - \bar{d}|\sqrt{n-3}/\Delta$ is a t-distribution with $\nu = n - 3$ degrees of freedom. We can recover the normalization constant be comparing this to the standard form

$$P(\theta|\Delta^2, \bar{d}) = \frac{\Gamma\left(\frac{n-2}{2}\right)}{\sqrt{(n-3)\pi}\,\Gamma\left(\frac{n-3}{2}\right)} \left[1 + \frac{(\theta - \bar{d})^2}{\Delta^2}\right]^{\frac{2-n}{2}} \tag{5.59}$$

Because this is symmetric about $\bar{d}$ the mean is $\langle\theta\rangle = \bar{d}$. Since the variance of a t-distribution is $\frac{\nu}{\nu-2}$ and $\nu = n-3$ in this case

$$\left\langle x^2 \right\rangle = (n-3)\frac{\left\langle(\theta-\bar{d})^2\right\rangle}{\Delta^2} = \frac{\nu}{\nu-2} = \frac{n-3}{n-5} \tag{5.60}$$

so

$$\left\langle(\theta-\bar{d})^2\right\rangle = \frac{\Delta^2}{(n-5)}. \tag{5.61}$$

From this example we learn that if we model a series of observation to be independent and normally distributed with the same mean and variance and we give the mean and variance uniform priors the posterior distribution of the model mean, θ, (not to be confused conceptually with the sample mean $\bar{d}$) will be t-distributed. As was discussed in section 4.3, the distribution of $t = (\bar{x}-\mu)\sqrt{n/s^2}$ is t-distributed with $\nu = n-1$ degrees of freedom. That would be the frequentist method of putting a constraint on the distribution mean μ. The number of degrees of freedom are different! More on this later when we talk about the choice of prior.

5.4 Choice of prior

As its name suggests, the prior expresses the information one had about the parameters before using the current data to constrain them. This information might come from a previous experiment or observation in which case the prior would be the posterior of that experiment. The prior can also express the theoretically allowed range of a parameter. For example if mass or flux is a parameter the prior should be zero for negative values. Usually there is some boundaries one can put on the value of a parameter at least on theoretical grounds - the mass of a planet cannot be greater than a solar mass.

For the Bayesian parameter estimation problem the actual prior bounds on a parameters are often unimportant. This is because the likelihood will be so small at the boundaries of parameter space that they will not effect the integral in the evidence and the posterior will be zero at these points. In other cases posterior might be significant at the theoretically imposed boundaries to parameter space.

A **uniform prior** is often used in Bayesian analysis. This is the prior that is constant over a region of parameter space and zero outside of it. It is unnecessary to specify the limits when likelihood is zero at the boundaries because the normalization appears both in the prior and in the evidence so it cancels out of the posterior.

You might be tempted to always us a uniform prior. It has the appearance of being unprejudiced in the sense that it will not favor one parameter value over another without the data supporting it. This appearance is deceptive however. The prior imposes a metric on parameter space - the prior probability for a parameter being in an infinitesimal region is $p(\alpha)d\alpha$. What is a uniform prior for one set of parameters will not be a uniform prior for another set even though they might describe the same model. For example a uniform prior for σ^2 in the above example is equivalent to prior proportional to σ on the parameter σ because $d\sigma^2 = 2\sigma d\sigma$. With this in mind the uniform prior does not seem so nonprejudicial. It picks out one parameterization which might be an arbitrary choice. Another example of this is the choice of whether to use frequency or wavelength (or period) in some problems. There is no natural reason to choose either one, but a uniform prior on one choice will not be uniform for the other.

Another widely used prior is called **Jeffreys prior**. It is the prior

$$p(\alpha) = \frac{1}{\ln(\alpha_{max}/\alpha_{min})} \begin{cases} 1/\alpha & \alpha_{min} < \alpha < \alpha_{max} \\ 0 & \text{otherwise} \end{cases} \tag{5.62}$$

This prior gives equal weight to equal logorithmic ranges of α ($d\ln\alpha = d\alpha/\alpha$). If parameter, α, is replaced with parameter $\beta = \alpha^\gamma$ for any γ this prior will not change since $d\ln\alpha^\gamma = \gamma d\ln\alpha \propto d\alpha/\alpha$. Jeffreys prior is often used for a "*scale*" parameter as apposed to "*location*" parameters. The difference between these types of parameters is not always clear to me, but it is clear that a scale parameter cannot be less than zero. A location parameter can be shifted by a constant without fundamentally changing the problem. In the case of complete prior ignorance a uniform prior should be used for location parameters.

The normalization and value of Jeffreys prior is infinite if the range is extended to $0 < \alpha < \infty$. Similarly, the normalization of the uniform prior is formally zero for the range $-\infty < \alpha < \infty$. These ranges are routinely used when the posterior (likelihood times prior) has a well defined integral. These are examples of **improper prior** distributions that are not valid distributions by themselves, but make sense in a posterior.

Many researchers feel that the arbitrariness inherent in choosing a prior is a serious flaw in the Bayesian approach. This criticism, I think, only makes sense when the prior is not expressing the results of some previous experiment. Frequentist methods do not have a general way of including prior information which is an important advantage to the Bayesian method. In general, if the data strongly constrains the parameters beyond what was previously known the choice of prior should not strongly affect the resulting posterior. We will compare and contrast these methods further later.

5.4.1 example:

Going back the alcohol in wine example, we can now recagnise σ^2 as a scale parameter and θ as a location parameter. Previously we used a uniform prior for σ^2. Let's see how things change if we use Jeffreys prior for σ^2. The posterior (5.47) will change to

$$P(\theta, \sigma^2|\boldsymbol{d}) \propto \left(\frac{1}{\sigma^2}\right) \frac{1}{\sigma^n} \exp\left(-\frac{n\Delta^2}{2\sigma^2}\right) \exp\left(-\frac{n(\theta - \bar{d})^2}{2\sigma^2}\right) \tag{5.63}$$

where the σ^{-2} factor is from the prior. By integrating this we can determine the normalization

$$P(\theta, \sigma^2|\boldsymbol{d}) = \frac{n^{n/2}}{\sqrt{2^n \pi}\Gamma\left(\frac{n-1}{2}\right)} \left(\frac{\Delta^2}{2}\right)^{\frac{n-1}{2}} \frac{1}{\sigma^{n+2}} \exp\left(-\frac{n\Delta^2}{2\sigma^2}\right) \exp\left(-\frac{n(\theta - \bar{d})^2}{2\sigma^2}\right) \tag{5.64}$$

We can then marginalize over σ^2 as before to get the marginalized distribution for θ

$$P(\theta|\boldsymbol{d}) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \frac{1}{\Delta} \left[1 + \frac{(\theta - \bar{d})^2}{\Delta^2}\right]^{-\frac{n}{2}} \tag{5.65}$$

This is again a t-distribution, but now it is of $\nu = n - 1$ degrees of freedom. Recall that with the uniform prior we got a t-distribution of $n - 3$ degrees of freedom. From section 4.3 we know that the $t = (\bar{x} - \mu)\sqrt{n/s^2}$ is t-distributed with $\nu = n - 1$ degrees of freedom. So in a way this prior agrees with the frequentist result although keep in mind that these are really different quantities we are talking about. t is a function of the data. θ is a model parameter.

Note also that as n gets bigger the difference between $n - 1$ and $n - 3$ gets less significant and the difference between the posterior distributions for uniform and Jeffreys become insignificant. It is only when the likelihood is a weak constraint on the parameters relative to the prior that the prior will have a strong effect on the posterior.

5.4.2 example:

Going back to the example given in section 5.2.1 where we found the posterior for a rate of radioactive decay. We might now recognize the rate as a scale parameter and prefer to us Jeffreys prior rather than the uniform prior we used before. The posterior, after renormalizing, will go from (5.11) to

$$p(r|n) = \frac{\delta t}{(n-1)!}(\delta t r)^{n-1} e^{-r\delta t} \tag{5.66}$$

The mean and variance of this distribution are more in agreement with frequentist expectations

$$\langle r \rangle = \frac{n}{\delta t} \qquad Var\left[r\right] = \frac{n}{\delta t^2} \tag{5.67}$$

Again in the limit of large n the posteriors are the same for the two choices of prior.

5.5 Model selection

Lets consider a somewhat different problem than parameter estimation. Lets say here are competing models that describe the data, but these models do not just differ from each other by having different values for there parameters. The models might have completely different parameters or one model might be the same as the other except that it has additional parameters. Which model is more strongly supported by the data? This is called model selection.

Let us consider a set of all possible models that explain the data $M_1, M_2, ...$ We can write down the posterior for model M_i using Bayes' theorem as in the parameter estimation case

$$P(M_i|\boldsymbol{D}) = \frac{P(\boldsymbol{D}|M_i)P(M_i)}{P(\boldsymbol{D})} = \frac{P(\boldsymbol{D}|M_i)P(M_i)}{\sum_i P(\boldsymbol{D}|M_i)P(M_i)}. \tag{5.68}$$

It is difficult to imagine ever knowing *all* possible models so model selection is usually restricted to comparing the relative probability of two models, call them M_1 and M_2, by taking the ratio of their posteriors

$$O_{1,2} = \frac{P(M_1|\boldsymbol{D})}{P(M_2|\boldsymbol{D})} = \frac{P(\boldsymbol{D}|M_1)}{P(\boldsymbol{D}|M_2)}\frac{P(M_1)}{P(M_2)} = B_{1,2}\frac{P(M_1)}{P(M_2)}. \tag{5.69}$$

$O_{1,2}$ is called the **odds** of model 1 relative to model 2 and $B_{1,2}$, the ratio of the model likelihoods, is know as **Bayes's factor**. If the prior probabilities are equal, as they often are, then the odds is equal to Bayes' factor. Note that $P(\boldsymbol{D})$ cancels out so we avoid needing to know the probability of the data over all possible models. If the odds is large then model 1 is favored. If it is small then model 2 is favored. You can also take the log of the odds and then positive values would favor M_1 and negative N_2.

How can we calculate $P(\boldsymbol{D}|M)$? In the parameter estimation problem we stayed within one model. Because of this all the probabilities were conditional on this model being true although that was not explicitly shown. We can write Bayes' theorem again with the model conditionality explicitly shown

$$P(\boldsymbol{\theta}|\boldsymbol{D}, M) = \frac{P(\boldsymbol{D}|\boldsymbol{\theta}, M)P(\boldsymbol{\theta}|M)}{P(\boldsymbol{D}|M)}. \tag{5.70}$$

We can now see that $P(\boldsymbol{D}|M)$ is actually the evidence:

$$P(\boldsymbol{D}|M_i) = \int_{-\infty}^{\infty} d\boldsymbol{\theta}\; P(\boldsymbol{D}|\boldsymbol{\theta}, M_i)P(\boldsymbol{\theta}|M_i) \tag{5.71}$$

where the integral is over all of parameter space within model M_i. Bayes' factor is the ratio of evidences for two models.

5.5.1 Occam's factor

Situation often arises where one has a standard model that explains the data and an extension to the model that includes some additional parameters. For example, the standard ΛCDM model and ΛCDM plus dark energy with a equation of state parameter that is not -1 ($w = p/\rho \neq -1$) as it would be for a cosmological constant. Or the dark energy might be coupled to dark matter and there is a parameter describing the strength of this coupling. Or you have stellar evolution models that predicts the amount of lithium in a low mass star among other things. The standard model has no mixing in the atmosphere. The extended model has mixing regulated with an additional parameter.

In these situations the extended model will always have a set of parameter values the fit the data as well as or better than the standard model since the standard model is the extended model with additional degrees of freedom to fit the data. Usually the standard model is identical to the extended model with that additional parameters fixed to some value (perhaps 0 or for in the dark energy case $w = -1$). Lets label the likelihoods $\mathcal{L}_{st}(\boldsymbol{\theta}|\boldsymbol{D})$ for the standard model and $\mathcal{L}_{ex}(\boldsymbol{\theta}, \beta|\boldsymbol{D})$ for the extended model where β is the extra parameter. Lets denote the parameter values that maximize the standard model likelihood as $\hat{\boldsymbol{\theta}}_{st}$ and those that maximize the extended model likelihood as $(\hat{\boldsymbol{\theta}}_{ex}, \hat{\beta})$. Then

$$\mathcal{L}_{ex}(\hat{\boldsymbol{\theta}}_{ex}, \hat{\beta}) \geq \mathcal{L}_{st}(\hat{\boldsymbol{\theta}}_{st}) \tag{5.72}$$

Because of this one might be drawn to the conclusion that more complicated models are always as good or better than less complicated ones. This violates Occam's principle, or razor, that the best model is the simplest one that explains all the observations (William of Ackham ∼ 1300).

For a more concrete example, you can always fit a line to two data points perfectly. If you add another data point the line generally wont go through all the points. You could add a parameter and fit a quadratic function, a parabola, to the data and it would again go through all of the points. If you have n data points you can fit them perfectly with a nth order polynomial. But if your model includes random noise in the data you would not expect the correct model it to go through all the points perfectly. "Any theory that fits all the data is wrong, because some of the data is wrong." (I don't know who said this.) So when do you stop adding parameters? When does the model fit too well?

Although it is not immediately apparent, Bayesian model selection automatically incorporates Occam's razor to answer these questions. To demonstrate this lets consider an extended model with on extra parameter β. The prior on this parameter will be $\mathcal{N}(\beta_o, \sigma_\beta)$. The standard model will be the extended one with $\beta = \beta_o$. We will take the priors on the models to be equal ($P(M_1) = P(M_2)$). The odds between the models is

$$O_{2,1} = B_{2,1} = \frac{\int d\boldsymbol{\theta} \int d\beta \; \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, \beta) P(\boldsymbol{\theta}) P(\beta)}{\int d\boldsymbol{\theta} \; \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, \beta_o) P(\boldsymbol{\theta})} \tag{5.73}$$

Now lets consider two extreme cases. The first is where the prior, $P(\beta)$, is very broad compared to the likelihood so that $P(\beta) = \exp(-(\beta - \beta_o)^2/2\sigma_\beta^2)/\sqrt{2\pi\sigma_\beta^2} \simeq 1/\sqrt{2\pi\sigma_\beta^2}$ everywhere $\mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, \beta)$ is significant so

$$B_{2,1} \simeq \frac{1}{\sqrt{2\pi}\sigma_\beta} \frac{\int d\boldsymbol{\theta} \int d\beta \; \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, \beta) P(\boldsymbol{\theta})}{\int d\boldsymbol{\theta} \; \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, \beta_o) P(\boldsymbol{\theta})} \tag{5.74}$$

Now lets express the integrals here as the products of two factors

$$\int d\boldsymbol{\theta} \int d\beta \; \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, \beta) P(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{D}|\hat{\boldsymbol{\theta}}_{ex}, \hat{\beta}) \mathcal{V}_{\theta,\beta} \tag{5.75}$$

where $\mathcal{L}(\boldsymbol{D}|\hat{\boldsymbol{\theta}}_{ex},\hat{\beta})$ is the maximum likelihood and $\mathcal{V}_{\theta,\beta}$ is a measure of the volume in parameter space to which the parameters are constrained by the data in the extended model. Likewise $\mathcal{V}_\theta$ is a measure of the parameter space volume in the standard model. Writing them like this the Bayes' factor becomes

$$B_{2,1} \simeq \left[\frac{1}{\sqrt{2\pi}\sigma_\beta}\frac{\mathcal{V}_{\theta,\beta}}{\mathcal{V}_\theta}\right]\frac{\mathcal{L}(\boldsymbol{D}|\hat{\boldsymbol{\theta}}_{ex},\hat{\beta})}{\mathcal{L}(\boldsymbol{D}|\hat{\boldsymbol{\theta}}_{st},\beta_o)} \tag{5.76}$$

The part in square brackets is sometimes called Occam's factor. The ratio $\frac{\mathcal{V}_{\theta,\beta}}{\mathcal{V}_\theta}$ can be interpreted as a measure of the width in parameter space of the posterior in the β dimension in the extended model. So Occam's factor is small if the data constrains the parameter β to a much smaller range than was allowed by the prior ($\sim \sigma_\beta$). The other factor is the ratio of the likelihood at its maximum with and without the extra parameters. This factor will always be larger than one and thus favor the extended model. Occam's factor will always be smaller than one in this example where we have taken the prior to be very broad. For the odds to favor the extended model the fit must not just be better, but so much better that it overpowers Occam's factor to make the odds greater than 1.

Another extreme example is one where the prior on β is very narrow compared to its constraint from the likelihood. This could be the case if a previous experiment already constrained β much more strongly than the one we are considering here or it could be that the theory behind the model requires that this parameter be within a range within which it cannot significantly change the predictions for this data set. In this case

$$\begin{aligned}
O_{2,1} = B_{2,1} &= \frac{\int d\boldsymbol{\theta} \int d\beta \; \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\beta)P(\boldsymbol{\theta})P(\beta)}{\int d\boldsymbol{\theta} \; \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\beta_o)P(\boldsymbol{\theta})} &(5.77)\\
&\simeq \frac{\int d\boldsymbol{\theta} \; \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\beta_o)P(\boldsymbol{\theta}) \int d\beta \; P(\beta)}{\int d\boldsymbol{\theta} \; \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\beta_o)P(\boldsymbol{\theta})} &(5.78)\\
&\simeq \frac{\int d\boldsymbol{\theta} \; \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\beta_o)P(\boldsymbol{\theta})}{\int d\boldsymbol{\theta} \; \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta},\beta_o)P(\boldsymbol{\theta})} &(5.79)\\
&\simeq 1 &(5.80)
\end{aligned}$$

So if the extended model has a parameter that doesn't improve the fit to the data within its prior allowed range then this extended model will not be favored or disfavored over the simpler model. For this reason saying that Bayesian model selection accounts for Occam's razor is maybe a bit misleading. Occam's principle is that a simpler model should be favored, but here we see that a model with extra superfluous, irrelevant parameters is not disfavored. This is what we want however. We can always add extra irrelevant parameters to a model that have no effect on its predictions. These models are identical in terms of their physical predictions so the data should not favor one over the other.

One criticism of Bayesian model selection is that it depends on you having a well justified prior distribution for the parameters. Normalization or boundaries of allowed parameter space are important whereas in the parameter estimation case normalization of the prior cancels out and the boundaries are only important if the likelihood is significant at them. For this reason you can use a uniform or Jeffreys prior that extends to infinity for example. If you use an infinite uniform prior for a new parameter in the model selection problem you will always get an infinitely small odds. If you start from a state of ignorance what prior do you use? If you extent the prior to just some big number then the odds will depend on this sometimes arbitrary choice. For me this is a big uninstalling ambiguity in applying Bayesian model selection to practical problems.

5.5.2 example:

On/Off Poisson detection problems ***

5.6 Calculating the evidence

It is often difficult or impossible to obtain an analytic expression for the evidence, the normalization of the posterior. In practice it is usually calculated numerically by integrating the likelihood times the prior over the parameter space. This is usually a simple task if there are only 1, 2 or 3 parameters. One can simply grid the parameter space or use a standard integration routine.

Note that If one is doing parameter estimation one only needs the posterior and any factors in the prior and likelihood that do not depend on the parameters will cancel out. For this reason it is not necessary to normalize these probabilities individually, just the product of them. This can save some work, especially when the likelihood or prior are something strange that you don't know the normalization of. However, to get the evidence the likelihood and prior need to be properly normalized.

When the dimension of the parameter space is larger, $\gtrsim 3$, numerical integration can be much more difficult. We will return to this problem later when we talk about Monte Carlo techniques for Bayesian analysis.

5.7 Example: A spectral line

5.8 Example: The luminosity function

6 Bayesian inference with Gaussian distributions & Gaussian approximations

7 Hypothesis testing & frequentist parameter fitting

Hypothesis testing is the frequentist version of Bayesian model selection. It takes a distinctly different approach to the question of whether a theory or hypothesis is supported by the data or not. The Bayesian method always compares the probability of competing models while hypothesis testing seeks to disprove a hypothesis by showing that the observed data would not be likely if the hypothesis were true. In some cases it is easier to apply hypothesis testing and there are many specific hypothesis tests that are commonly used in practice so it is essential to understand how it works even for the most fanatical Bayesian.

The basic steps in any hypothesis test are as follows:

1. State the hypothesis as a well posed true or false question. The goal is the falsify this question.
2. Choose or invent a statistic that should be affected by the truth of the hypothesis.
3. Determine by analytic or numerical methods the probability distribution of the statistic.
4. Calculate the statistic with the data and determine if the measured value is improbable if the hypothesis is true.
5. If the statistic is in a region that is *sufficiently improbable* the hypothesis is ruled out. If it is not sufficiently improbable the hypothesis is consistent with this statistic.

To explain hypothesis testing let me tell a little fable. Someone brings you an unidentifiable animal. You say, "I think it is a dog." That is your hypothesis. You think about what a dog definitely has. Dogs have fur. That's your statistic. If the animal doesn't have fur you can say that the animal is not a dog. If it has fur you can say that this characteristic is consistent with it being a dog. You can't say it is a dog. There are other animals that have fur and there might be some other characteristic that is inconsistent with being a dog, say it has no claws. In most cases you can't even completely prove the hypothesis false, only unlikely. It might be a dog with a rare disease that made it lose it's fur or a rare genetically engineered dog that doesn't grow fur. Not that there are no specific alternative hypothesis, it either dog or not dog.

type I and type II errors

two sided and one sided

null hypothesis

We can never prove a hypothesis right. In fact in some cases a statistical test might show consistency with a hypothesis that is clearly ruled out by another statistical test.

7.0.1 mean of two populations are the same

7.0.2 the variance of two populations are the same

7.1 linear fitting, regression

Parameter covariance matrix. Spotting degeneracies

7.2 p-values, χ^2, degrees of freedom, etc.

7.3 principle components

7.4 nonlinear fitting

7.5 hypothesis testing

7.5.1 example: test isotropy of sources

7.6 The early stop problem & shortcoming of frequenist hypothesis testing

8 Frequentist non-parametric tests

8.1 Kolmogorov-Smirnov test

8.2 rank statistics

8.2.1 Wilcoxon test

8.2.2 Spearman's correlation statistic

8.3 bootstrap & jacknife sampling methods

9 Frequentist vs Bayesian

Differences, Exhaustive enumeration of models. decisions about priors and parameterization effect result information

models that differ in ranges that are not observed can effect frequentist bounds stop problem including prior information

9.1 frequentist & Bayesian confidence levels

9.1.1 Example: the highest redshift QSO

9.2 numerical methods, Monte Carlo

9.2.1 Creating a random number in a computer

10 Estimators

10.1 bias

10.1.1 Example: Eddington-Malmquist bias

10.2 maximum likelihood estimator

10.3 least squares estimator

10.3.1 weighting data

10.4 minimum variance estimator

11 The Fisher matrix and information

11.1 the Fisher matrix

11.1.1 the Gaussian case

11.1.2 marginalized error estimates

11.2 Rao-Cramer inequality

11.3 information content of data

11.3.1 different definitions of information

11.3.2 Kullback–Leibler divergence (information gain)

11.3.3 agreement/disagreement of data sets

11.4 error forecasting

12 Random Fields

12.1 correlation function & power spectrum

12.2 going between discrete and continuous variables

12.3 Gaussian random fields

12.4 Poisson noise

13 Estimating correlation functions, power spectra & time-delays from irregularly sampled data

13.1 power spectrum estimation

13.2 data compression, Karhunen-Loeve modes

14 Image reconstruction and map making

14.1 Wiener filtering

14.2 maximum entropy

14.3 other filters & estimators

15 Numerical methods for the Bayesian Inference Problem

15.1 finding maximum likelihood

15.2 Markov Chain Monte Carlo (MCMC)

15.3 nested sampling & other strategies

15.4 meaning & calculation of evidence

15.5 displaying results graphically

16 Classification and regression from a machine learning prospective

Why is this not showing up?

16.1 prediction vs inference

16.2 dividing objects into classes objectively: unsupervised learning

16.3 regression and classification: supervised learning

16.4 artificial neural networks & convolutional neural networks in astronomy

A Matrix basics

$$(\boldsymbol{ABC}\dots)^T = \dots\boldsymbol{C}^T\boldsymbol{B}^T\boldsymbol{A}^T \tag{A.1}$$

$$(\boldsymbol{ABC}\dots)^{-1} = \dots\boldsymbol{C}^{-1}\boldsymbol{B}^{-1}\boldsymbol{A}^{-1} \tag{A.2}$$

$$(\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T \tag{A.3}$$

$$(\boldsymbol{A}+\boldsymbol{B})^T = \boldsymbol{A}^T + \boldsymbol{B}^T \tag{A.4}$$

Some properties of the determinant

$$|\boldsymbol{A}| = \prod_i \lambda_i \tag{A.5}$$

$$|\boldsymbol{A}^{-1}| = 1/|\boldsymbol{A}| \tag{A.6}$$

$$|\boldsymbol{BA}| = |\boldsymbol{B}||\boldsymbol{A}| \tag{A.7}$$

$$|c\boldsymbol{A}| = c^n|\boldsymbol{A}| \tag{A.8}$$

$$|\boldsymbol{A}^T| = |\boldsymbol{A}| \tag{A.9}$$

Some properties of the trace

$$\operatorname{tr}(\boldsymbol{A}) = \sum_i A_{ii} \tag{A.10}$$

$$\operatorname{tr}(\boldsymbol{A}) = \sum_i \lambda_{ii} \tag{A.11}$$

$$\operatorname{tr}(\boldsymbol{A}^T) = \operatorname{tr}(\boldsymbol{A}) \tag{A.12}$$

$$\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA}) \tag{A.13}$$

$$\operatorname{tr}(\boldsymbol{A}+\boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}) + \operatorname{tr}(\boldsymbol{B}) \tag{A.14}$$

$\boldsymbol{A}$ is an **orthogonal matrix** if and only if

$$\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I} \tag{A.15}$$

An orthogonal matrix has the following properties

$$\boldsymbol{A}^T = \boldsymbol{A}^{-1} \tag{A.16}$$

$$|\boldsymbol{A}| = \pm 1 \tag{A.17}$$

The $|\lambda_i| = 1$ for all eigenvalues and the magnitude of all eigenvactors are 1.

$\boldsymbol{C}$ is a **positive definite matrix** if

$$\boldsymbol{x}^T\boldsymbol{C}\boldsymbol{x} > 0 \;\; \forall\boldsymbol{x}. \tag{A.18}$$

It has the following properties

- all eigenvalues at positive
- $\operatorname{tr}(\boldsymbol{C}) > 0$
- all diagonal elements are positive, $\boldsymbol{C}_{ii} > 0, \forall i$
- $\boldsymbol{C}$ is invertible

The covariance matrix is always positive definite.

”A and B”	A, B
”A or B”	$A \cup B$
continuous random variables	x,y,x_i,y_i
vector of random variables	$\boldsymbol{x}$ or $\vec{x}$
discrete random numbers	n,m
parameters	α , β
estimator of parameter α	θ_α or $\hat{\alpha}$
data	D or d_i
indexes data or for multiple random numbers	i, j
statistical and/or theoretical model	M
Gaussian or Normal pdf	$\mathcal{G}\left(\boldsymbol{x}\,\vert\boldsymbol{\mu},\boldsymbol{C}\,\right)$
$\boldsymbol{x}$ is normally distributed	$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\sigma})$
x is χ^2 distributed with n degrees of freedom	$x \sim \chi^2_n$
arithmetic mean of N samples	$\bar{x}_N$
likelihood of data given model	$\mathcal{L}(\boldsymbol{D}\vert M_i)$ or $P(\boldsymbol{D}\vert M_i)$
Bayesian evidence of data	$\mathcal{E}(\boldsymbol{D})$
Heaviside function, 1 when B is true, 0 otherwise	$\Theta(B)$

Table 1: notation

B notation

Notation may vary but in general I will follow the guide in table 1

C Some Useful Integrals and mathematical definitions

$$\int_{-\infty}^{\infty} dx\ e^{-\frac{x^2}{2}} = \sqrt{2\pi} \tag{C.1}$$

$$\begin{aligned}\int_{-\infty}^{\infty} dx\ e^{-(ax^2+bx+c)} &= e^{-c}\int_{-\infty}^{\infty} dx\ e^{-\left(\sqrt{a}x+\frac{b}{2\sqrt{a}}\right)^2+\frac{b^2}{4a}} = e^{-c+\frac{b^2}{4a}}\int_{-\infty}^{\infty}\frac{dy}{\sqrt{a}}\ e^{y^2} \\ &= \sqrt{\frac{\pi}{a}}e^{-c+\frac{b^2}{4a}}\end{aligned} \tag{C.2}$$

$$\int_0^{\infty} dx\ x^n e^{-\frac{1}{2}Ax^2} = 2^{\frac{n-1}{2}}A^{-\frac{n+1}{2}}\Gamma\left(\frac{n+1}{2}\right) \qquad n > -1 \tag{C.3}$$

The Gamma function

$$\begin{array}{lll} \int_0^\infty dx\ x^n e^{x^2} & = \frac{1}{2}\Gamma\left(\frac{n+1}{2}\right) & \\ \Gamma(n) & = (n-1)! & n = 1, 2, \ldots \\ \Gamma\left(\frac{1}{2}+n\right) & = \frac{(2n)!}{4^n n!}\sqrt{\pi} & n = 0, 1, 2, \ldots \end{array} \tag{C.4}$$

Stirling's approximation

$$\ln N! \simeq N\ln N - N \text{ for } N \gg 1 \tag{C.5}$$

or more accurately

$$N! \simeq \sqrt{2\pi N}\left(\frac{N}{e}\right)^N \text{ for } N \gg 1 \tag{C.6}$$

Index

χ^2 distribution, 28

anticorrelated, 24
arithmetic mean, 31
asymptotically unbiased estimator, 35

Bayes' theorem, 6
Bayes's factor, 48
Bayesian inference, 38
Bayesian model selection, 48
Bernoulli distribution, 16
biased, 35
binomial coefficient, 11, 15
binomial distribution, 9, 15
binomial expansion, 16

Cauchy distribution, 14
Cauchy–Schwarz inequality, 24
central limit theorem, 20
central moments, 14
completion of squares, 27
conditional probability, 6
correlated variables, 24
covariance, 24
covariance matrix, 24
cumulative distribution function, 13

disjoint probability, 6
double factorial, 19

eigendecomposition, 26
error function, 19
estimator, 31
evidence, 39
expectation value, 13
extended sum rule, 6

Fisher information matrix, 57

gamma function, 29
Gaussian distribution, 19

hypergeometric distribution, 17
hypothesis testing, 53

improper prior, 47
independent, 7, 24
inverse noise weighting, 34

Jeffreys prior, 47
joint probability, 6

kurtosis, 14

Lagrange multipliers, 33
likelihood, 38
lognormal distribution, 23
Lorentzian profile, 14

marginalization, 45
mean, 14
mean deviation, 14
median, 14, 36
minimum variance estimator, 33
mode, 13
moment generating function (MGF), 15
moments, 14
multimodal, 13
multinomial distribution, 25
multivariant distribution, 24
multivariant Gaussian, 25
mutually exclusive, 7

normal distribution, 19
nuisance parameters, 45
null hypothesis, 53

Occam's factor, 49
odds, 48
orthogonal matrix, 26, 63

permutations, 9
Poisson distribution, 17
positive definite matrix, 63
posterior probability, 38
prior, 38
probability distribution function (PDF), 13
probability mass function, 13
product rule, 6

quintiles, 37

random variable, 13
rank, 37

Rao-Cramer inequality, 57

skewness, 14
standard deviation, 14
standardized variable, 14
statistic, 31, 53
statistical model, 6
student's t-distribution, 30, 35
sum rule, 6

t-distribution, 30, 35

unbiased, 31
uniform prior, 46
unimodal, 13

variance, 14

weighted mean, 32

References

Gregory P., 2006, Bayesian Logical Data Analysis for the Physical Sciences. Cambridge University Press