Notes: Practical Statistics for Physics & Astronomy

R. Benton Metcalf

Alma Mater Studiorum - Universitá di Bologna

March 16, 2018

Contents

1 What is Probability?

1.1 Frequentist interpretation of probability

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 the same way or circumstances an *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 almost universally used for centuries despite no one ever having done anything *exactly* the same twice let along 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 the 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 an 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 30% of them. This seems like a completely unverifiable claim. If scientific knowledge must be reproducible to be considered true then it would seem that any such argument should be considered unscientific. And yet probability through statistics is at the foundation of all quantitative measurements.

Lets be a bit more practical. Lets say we don't need an infinite number of trials, but just a very *large number* of them. Lets say we flip a coin a *very large number* of times. If we did it say one billion times we would not expect that *exactly* 500 million times it would be heads. We would expect that roughly half, but not exactly half of the times it would be heads even if the probability of getting heads in each flip is 1/2. We might try to quantify how close the number of heads should be to 500 million, but in doing so we would need to use a probabilistic argument that would use the very concept we are trying to define.

Many statisticians and philosophers have found this definition of probability problematic. Despite this it is the definition usually used by scientists when they are forced to addressing this subject.

1.2 classical interpretation of probability

The classical interpretation of probability relies on identifying events that are equally likely or probable. This is often the argumentation used in classical statistical mechanics where each micro-state of the system is taken to be equally probable. If one then says that the probability of being in either of two mutually exclusive states is the sum of their probabilities and that the some of the probabilities of being in all possible states is one then you can find a numerical value for the

probability of each state. A macro-state (one with temperature equal to some value or total energy equal to some value) corresponds to many micro-states so by adding up their probabilities you can find the probability of macro states which will not necessarily be equal.

The biggest criticism of this interpretation is that it doesn't really say what probability is, it just tells you how to calculate it in a restricted class of problems. What does it mean that two states are equally probable? What does the probability of a macro-state mean? Another problem is that not all events that we commonly apply probability to can reduced in this way to a collections of equally probable mutually exclusive events.

1.3 Subjective or Bayesian interpretation of probability

Thomas Bayes (1701 - 1761) (and initially by 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. It did not make its way into common use in science, in the form of Bayesian statistics, until relatively recently (80s and 90s).

In this school of thought, probability theory is an extension of formal logic to situations where the truth or falsehood of a proposition (e.g. "It will rain tomorrow." or "The mass of the Earth is between 5.972×10^{24} kg and 5.978×10^{24} kg.") cannot be deduced conclusively by deductive reasoning. A proposition has a probability function that depends on the evidence for and against its truth. When deductive reasoning can be applied conclusively this function is either zero (false) or one (true). In this way Boolean logic is a limiting case of probability theory. Surprisingly from just the following requirements (or *desiderata*) on the probability function of a proposition you can deduce the rules of probability and show that they are complete without ever mentioning randomness or repetition of experiments.

1. Degrees of plausibility are represented by real numbers.
2. The measure of plausibility must exhibit qualitative agreement with rationality. This means that as new information supporting the truth of a proposition is supplied, the number which represents the plausibility will increase continuously and monotonically. Also, to maintain rationality, the deductive limit must be obtained where appropriate.
3. Consistency
   (a) *Structured consistency* : If the conclusion can be reasoned out in more than one way, every possible way must lead to the same result.
   (b) *Propriety*: The theory must take account of all information that is relevant to the question.
   (c) *Jaynes consistency*: Equivalent states of knowledge must be represented by equivalent plausibility assignments. For example, if $A, B|C = B|C$, then the plausibility of $A, B|C$ must equal the plausibility of $B|C$

(taken from Gregory (2006)).

These foundational proofs are very interesting, but outside the scope of this course (for those that are interested see chapter 2 of Gregory (2006) or, more comprehensively, Jaynes (2003)). 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* or a repeated even. It also establishes the accumulation of supporting evidence as central to the meaning of probability. Probability is a measure of knowledge, or ignorance, of an event and not a property of the event itself. These principles are central to the Bayesian method of parameter estimation and model selection that we will study later.

A	B	A, B	$\overline{A, B}$	$\overline{A} \cup \overline{B}$	$A \cup B$	$\overline{A \cup B}$	$\overline{A}, \overline{B}$
F	T	F	T	T	T	F	F
F	F	F	T	T	F	T	T
T	T	T	F	F	T	F	F
T	F	F	T	T	T	F	F

Table 1: The truth table for binary logical expressions.

1.4 Quantum mechanical probability

Probability in standard quantum mechanics is a fundamentally different thing than the probability that was in use before. In the frequentist interpretation of probability it is assumed that there are some "hidden variable" that are different every trial. In quantum mechanics it can be proven that such hidden variables do not exist or do not take on deterministic values for example with Bell's inequalities. When a measurement is made the square of the wave function gives the probability of an observation, but up to that point the outcome was impossible to determine, not just difficult to determine. This makes probability a property of physical systems and not solely a property of the observer. This seems to imply an intimate connection between physical laws and knowledge.

This is obviously a subject for a different course (or a *Star Trek* episode) so I will go no further.

1.5 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". Note that $\overline{A \cup B} = \overline{A}, \overline{B}$ and $\overline{A, B} = \overline{A} \cup \overline{B}$ 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. See table 1. In the language of Boolean algebra, they have the same truth table and are therefor equivalent statements. There probabilities must also be the same.

$P(A, B)$ is 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

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

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. From 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

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

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

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

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

If A and B are **independent** then the probability of A occurring does 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.

1.6 Exercises

1. We know the probability of a person having red hair is $P(R)$, the probability of a person having blue eyes is $P(B)$ and that the probability of a red headed person having blue eyes is $P(B|R)$.

   (a) What is the probability that a blue eyed person will have red hair?

   (b) What is the probability that a person will have both blue eyes and red hair?

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 event 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

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

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

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

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

each term must be the same

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

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

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

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

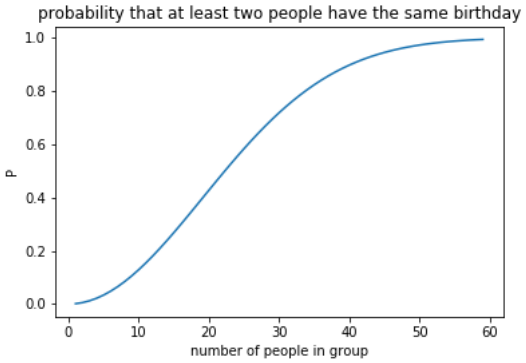

Figure 1: Probability of more than one person having the same birthday.

probability that two of them will have the same birthday.

Lets say there are n people at the party. There are 365 choices for the birthday of each person (not including leap years) so there are 365^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 364 choices for the second etc. until you get to the last person so the number of cases with no repeats is $365 \times 364 \times ... \times (365 - n + 1) = 365!/(365 - n)!$. So the total probability is

$$P(\text{at least two the same}) = 1 - P(\text{no two the same}) = 1 - \frac{365!}{365^n (365 - n)!}. \tag{2.18}$$

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 ($\ln N! \simeq N \ln N - N$) often helps simplify factorials.

$$\ln\left(\frac{N!}{N^n (N-n)!}\right) = \ln N! - \ln(N-n)! - n \ln N \tag{2.19}$$

$$= N \ln N - N - (N-n)\ln(N-n) - (N-n) - n \ln N \tag{2.20}$$

$$= (N-n)\ln N - (N-n)\ln(N-n) - n \tag{2.21}$$

$$= (N-n)\ln\left(\frac{N}{(N-n)}\right) - n \tag{2.22}$$

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 figure 1. 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 Exercises

1. **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 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, but does not open it. Monty then eliminates one of the doors that were not chosen and shows that it has the shoe or chicken. The contestant then has a chance to change his choice or remain with his first choice.

   What are the probabilities of getting the prize for each choice?

   (a) Stay with the first choice :

   (b) Change doors :

2. If you roll a die 10 times what is the probability of getting one 1, two 2s, three 3s and four 4s?

3. You have a bag of 100 blue and yellow balls. 60 of them are blue and 40 of them are yellow.

   (a) What is the probability of drawing 5 yellow balls in a row out of the bag without looking?

   (b) What is the probability of 6 draws out of 10 being yellow?

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

$$m_x(t) = \langle e^{tx} \rangle = \begin{cases} \sum_x e^{tx} p(x) \\ \int_{-\infty}^{+\infty} dx\ e^{tx} p(x) \end{cases} \tag{3.10}$$

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. If t is replaced with it it is the Fourier transform.

3.4 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 $dx = \left[\frac{d}{dy} f(y)\right] dy$. The probability of being within this range of should not depend on which variable is used to measure the range so it must be that

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

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

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

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

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

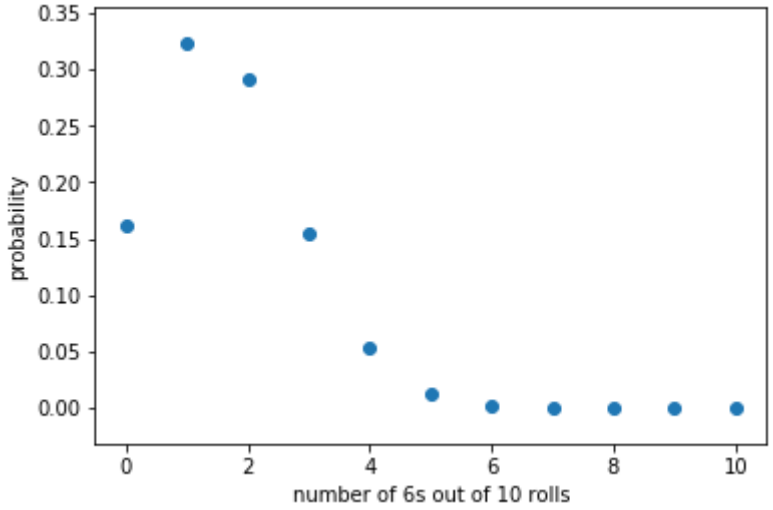

Figure 2: The binomial distribution for the number of 6s in ten rolls of a die or one roll of ten dice. $N = 10$, $k = 0 \ldots 10$, $p = 1/6$

3.5 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 just like for the dice rolls we discussed earlier. This is the probability each 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**

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

So we get the final result

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

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 $N = 1$

$$p(n) = \begin{cases} p & , \quad n = 1 \\ (1-p) & , \quad n = 0 \end{cases} \tag{3.17}$$

, 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**

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^n y^{n-k} \tag{3.18}$$

Using this expansion we can find the moment generating function

$$m_x(t) = \sum_{n=0}^{\infty} e^{tn} \binom{N}{n} p^n (1-p)^{N-n} \tag{3.19}$$

$$= \sum_{n=0}^{\infty} \binom{N}{n} (e^t p)^n (1-p)^{N-n} \tag{3.20}$$

$$= \left(e^t p + 1 - p\right)^N \tag{3.21}$$

From this we can find the mean and variance

$$\langle n \rangle = Np \tag{3.22}$$

$$\sigma^2 = \langle n^2 \rangle - \langle n \rangle^2 = Np(1-p) \tag{3.23}$$

3.5.1 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

$$p(r|n,N,R) = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}} \tag{3.24}$$

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).

3.6 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

$$p(0|t_o, t+dt) = p(0|t_o, t)\,[1 - rdt] \tag{3.25}$$

Rearranging this we can obtain the differential equation

$$\frac{p(0|t_o, t+dt) - p(0|t_o, t)}{dt} = \frac{d}{dt} p(0|t_o, t) = -p(0|t_o, t) r \tag{3.26}$$

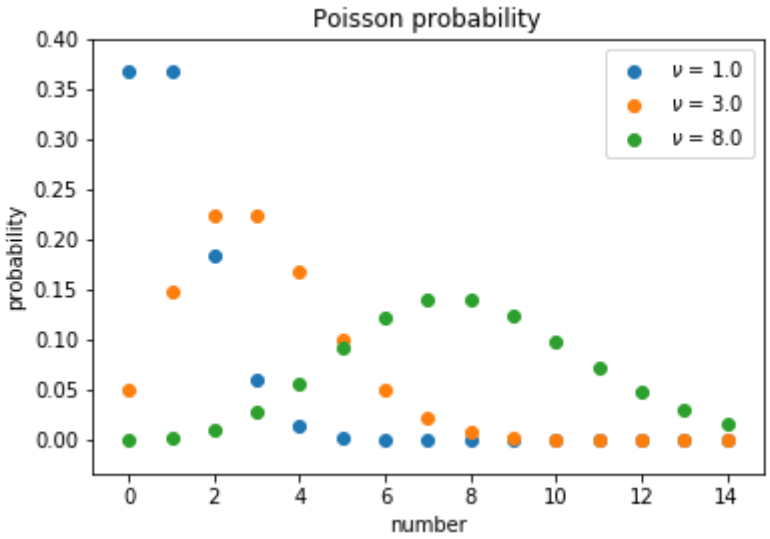

Figure 3: The Poisson distribution for several rates ν.

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,

$$p(0|t_o, t) = e^{-r(t-t_o)} \tag{3.27}$$

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:

$$p(0 < t_1 < t_2 < \cdots < t_n < t) = p(0|0, t_1) r dt_1 \Theta(t_1 < t_2) p(0|t_1, t_2) r dt_2 \Theta(t_2 < t_3) \dots p(0|t_n, t) dt_n \Theta(t_n < t) \tag{3.28}$$

$$= r^n e^{-rt} \prod_{i=1}^{n} dt_i \Theta(t_i < t_{i+1}) \tag{3.29}$$

where

$$\Theta(x < y) = \begin{cases} 1 & , \quad x \leq y \\ 0 & , \quad x > y \end{cases} \tag{3.30}$$

Using the sum rule we know that the probability of n events occurring is the sum of the probabilities

for all possible values for the event times.

$$p(n|r,t) = \prod_i \int_0^{t_i} p(0 < t_1 < t_2 < \cdots < t_n < t) \tag{3.31}$$

$$= r^n e^{-rt} \int_0^t dt_n \cdots \int_0^{t_3} dt_2 \int_0^{t_2} dt_1 \tag{3.32}$$

$$= r^n e^{-rt} \int_0^t dt_n \cdots \int_0^{t_3} dt_2 t_2 \tag{3.33}$$

$$= r^n e^{-rt} \int_0^t dt_n \cdots \int_0^{t_4} dt_3 \frac{t_3^2}{2} \tag{3.34}$$

$$= \frac{(rt)^n}{n!} e^{-rt} \tag{3.35}$$

$$= \frac{(\nu)^n}{n!} e^{-\nu} \quad \text{Poisson Distribution} \tag{3.36}$$

where $\nu \equiv rt$. This distribution has the following mean and variance

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

$$Var[n] = \nu \tag{3.38}$$

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 $r(1\,\text{hour})e^{-r(1\,\text{hour})}$ 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.

3.6.1 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.

$$\binom{N}{n} p^n (1-p)^{N-n} = \binom{N}{n} \left(\frac{\nu}{N}\right)^n \left(1 - \frac{\nu}{N}\right)^{N-n} \tag{3.39}$$

$$= \frac{N!}{n!(N-n)!} \left(\frac{\nu}{N}\right)^n \left(1 - \frac{\nu}{N}\right)^{N-n} \qquad \text{using } \frac{N!}{(N-n)!} \simeq N^n \tag{3.40}$$

$$\simeq \frac{\nu^n}{n!} e^{-\nu} \tag{3.41}$$

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

$$p(n|\nu) \simeq \frac{\nu^n}{\Gamma(x+1)} e^{-\nu} \qquad \nu \gg 1 \tag{3.42}$$

3.7 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

$$p(x|\sigma,\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{3.43}$$

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 (3.43). I will use $\mathcal{G}(x\,|\mu,\sigma)$ to signify this Gaussian function.

The *cumulative distribution function* is

$$F(x) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right) \tag{3.44}$$

with the **error function** defined as

$$\mathrm{erf}(z) \equiv \frac{2}{\sqrt{\pi}} \int_0^z e^{-u^2} du \tag{3.45}$$

Note that $\mathrm{erf}(-z) = -\mathrm{erf}(z)$.

The *moment generating function* is

$$m_{x-\mu}(t) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} dx\ e^{tx} e^{-\frac{x^2}{2\sigma^2}} \tag{3.46}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} dx\ \exp\left[-\left(\frac{x}{\sqrt{2}\sigma} - \frac{t\sigma}{\sqrt{2}}\right)^2 + \frac{t^2\sigma^2}{2}\right] \tag{3.47}$$

$$= e^{\frac{1}{2}\sigma^2 t^2} \tag{3.48}$$

The moments are

$$\mu_n = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} dx\ x^n e^{-\frac{x^2}{2\sigma^2}} = \begin{cases} \sigma^n (n-1)!! & n \text{ even} \\ 0 & n \text{ odd} \end{cases} \tag{3.49}$$

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

$$!!n = n \cdot (n-2) \cdot (n-4) \ldots 1 \tag{3.50}$$

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

$$p(\mu - n\sigma \leq x \leq \mu + n\sigma) = 1 - F(\mu - n\sigma) - [1 - F(\mu + n\sigma)] \tag{3.51}$$

$$= \frac{1}{2}\left[\text{erf}\left(\frac{n}{\sqrt{2}}\right) - \text{erf}\left(-\frac{n}{\sqrt{2}}\right)\right] \tag{3.52}$$

$$= \text{erf}\left(\frac{n}{\sqrt{2}}\right) \tag{3.53}$$

some specific values for this are

$$p(\mu - \sigma \leq x \leq \mu + \sigma) = 0.683 \tag{3.54}$$

$$p(\mu - 2\sigma \leq x \leq \mu + 2\sigma) = 0.954 \tag{3.55}$$

$$p(\mu - 3\sigma \leq x \leq \mu + 3\sigma) = 0.997 \tag{3.56}$$

$$p(\mu - 4\sigma \leq x \leq \mu + 4\sigma) = 0.999937 \tag{3.57}$$

3.8 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 limit 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

$$z_i = \frac{x_i - \mu}{\sigma}. \tag{3.58}$$

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 $\langle Z^2 \rangle = \sum_{ij} \langle z_i z_j \rangle = \sum_i \langle z_i^2 \rangle = N$ because each one is uncorrelated. So the standardized variable for the sum is

$$Y = \frac{1}{\sqrt{N}} Z = \frac{1}{\sqrt{N}} \sum_i z_i. \tag{3.59}$$

This will again have mean zero and variance 1. Now lets find the moment generating function for

Y,

$$m_Y(t) = \langle \exp(tY) \rangle = \left\langle \exp\left(\frac{t}{\sqrt{N}} \sum_i z_i\right)\right\rangle = \left\langle \exp\left(\frac{t}{\sqrt{N}} z_i\right)\right\rangle^N \tag{3.60}$$

$$= \left\langle 1 + \frac{t}{\sqrt{N}} z_i + \frac{t^2}{N}\frac{z_i^2}{2} + \frac{t^3}{N^{3/2}}\frac{z_i^3}{3!} + \dots \right\rangle^N \tag{3.61}$$

$$= \left[1 + \frac{t}{\sqrt{N}} \langle z_i \rangle + \frac{t^2}{N}\frac{\langle z_i^2 \rangle}{2} + \frac{t^3}{N^{3/2}}\frac{\langle z_i^3 \rangle}{3!} + \dots \right]^N \tag{3.62}$$

$$= \left[1 + \frac{t^2}{2N} + \frac{t^3}{N^{3/2}}\frac{\langle z_i^3 \rangle}{3!} + \dots \right]^N \tag{3.63}$$

$$\simeq \lim_{N\to\infty} \left[1 + \frac{t^2}{2N} \right]^N \tag{3.64}$$

$$= e^{\frac{t^2}{2}} \tag{3.65}$$

This is the moment generating function for a Gaussian as we saw earlier.

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. It can also be shown that many distributions tend toward Gaussian in some limit that is commonly encountered.

3.8.1 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 way of doing this that is of general use. Lets take the sum of n random numbers to be $S = \sum_i x_i$. The pdf of variable x_i is $p_i(x_i)$, each one 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

$$p(S) = \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx_n \; \delta(S - \sum_i x_i) \; p_1(x_1) \dots p_n(x_n) \tag{3.66}$$

$$= \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx_n \; \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} \exp\left[-ik(S - \sum_i x_i) \right] \; p_1(x_1) \dots p_n(x_n) \tag{3.67}$$

$$= \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} e^{-ikS} \prod_i \int_{-\infty}^{\infty} dx_i e^{+ikxi} p_i(x_i) \tag{3.68}$$

$$= \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} e^{-ikS} \prod_i \tilde{p}_i(k) \tag{3.69}$$

$$= \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} e^{-ikS} \tilde{p}_S(k) \tag{3.70}$$

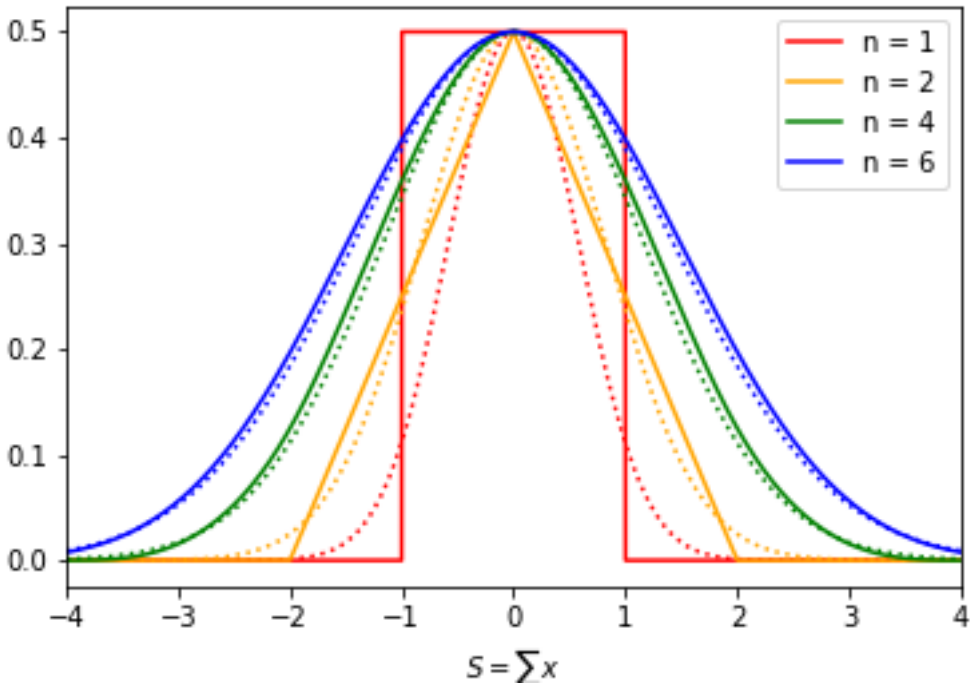

Figure 4: Probability distribution for the sum of n random variables that are uniformly distributed between -1 and 1. The normalizations have been changed so that there maximum is 0.5 in all cases. The dotted curves are for Gaussians with the same variance. You can see that the distribution converges to Gaussian remarkably quickly even for a very non Gaussian initial distribution.

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

$$\tilde{p}(k) = \frac{1}{L}\int_{-L/2}^{L/2} dx\ e^{+ikx} = \frac{2}{Lk}\sin\left(\frac{kL}{2}\right) = \text{sinc}\left(\frac{kL}{2}\right). \tag{3.71}$$

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

$$p_n(S) = \int_{-\infty}^{\infty} \frac{dk}{(2\pi)} e^{-ikS}\text{sinc}^n\left(\frac{kL}{2n}\right). \tag{3.72}$$

Figure 4 shows this case for some small values of n with $L = 2$. 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 which extends to infinity. Even so you can see that the distribution becomes remarkably Gaussian even for $n = 5$ or 6.

This exercise can be done numerically for any distribution. It is not necessary to have an analytic expression for the Fourier transform of $p_i(x_i)$. Any 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 for finding the distribution of the sum of variables can be used to study things like random walks and diffusion. The same idea is also used to derive halo mass functions in cosmology.

3.9 connection between Poisson and Gaussian distributions

You can see from the figure 3 of the Poisson distribution that as the average gets larger the Poisson pdf gets more symmetric and looks more Gaussian. Lets make this connection more precise. The Poisson distribution is

$$p(n|\nu) = \frac{(\nu)^n}{n!}\, e^{-\nu} \tag{3.73}$$

Lets make the substitution $n = \nu(1+\delta)$ which also means $\delta = (n-\nu)/\nu$. Lets take the limit where $\nu \gg 1$ while $\delta \ll 1$ which also means $n \gg 1$. Lets again use the Stirling's approximation

$$n! \to \sqrt{2\pi n} e^{-x} n^n \tag{3.74}$$

Making this substitution we get the probability

$$p(n) = \frac{\nu^{\nu(1+\delta)} e^{-\nu}}{\sqrt{2\pi} e^{-\nu(1+\delta)} \left[\nu(1+\delta)\right]^{\nu(1+\delta)+1/2}} \tag{3.75}$$

$$= \frac{e^{\nu\delta}(1+\delta)^{-\nu(1+\delta)-1/2}}{\sqrt{2\pi\nu}} \tag{3.76}$$

Lets look at the lowest order terms of the numerator

$$\ln\left[(1+\delta)^{-\nu(1+\delta)-1/2}\right] = -(\nu(1+\delta)+1/2)\ln(1+\delta) \tag{3.77}$$

$$= -\left(\nu+\nu\delta+1/2\right)\left(\delta - \frac{\delta^2}{2} + \ldots\right) \qquad \nu \gg 1 \tag{3.78}$$

$$\simeq -\left(\nu+\nu\delta\right)\left(\delta - \frac{\delta^2}{2} + \ldots\right) \tag{3.79}$$

$$\simeq -\nu\delta - \frac{\nu\delta^2}{2} + \ldots \tag{3.80}$$

Putting this back into the above

$$p(\delta) = p(n)\nu \tag{3.81}$$

$$\simeq \sqrt{\frac{\nu}{2\pi}}\, e^{-\frac{\nu\delta^2}{2}}. \tag{3.82}$$

So if ν is large the excursion from the mean, δ, is Gaussian distributed with a variance of $1/\nu$. In practice this can be a good enough approximation for moderate values of ν, say greater than 20. The photon noise or **shot noise** in astronomical images is Poisson distributed, but if the photon count is high it is essentially Gaussian distributed.

3.10 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 in many contexts. Another interpretation is that while the Gaussian is the right distribution for a sum of many 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

$$p(y)dy = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(\ln(y)-\mu)^2}{2\sigma^2}\right\} \frac{dy}{y} & , \quad y > 0 \\ 0 & , \quad y \leq 0 \end{cases} \tag{3.83}$$

Some of its properties are

$$E[y] = \exp(\mu + \frac{\sigma^2}{2}) \tag{3.84}$$

$$\text{median[y]} = \exp(\mu) \tag{3.85}$$

$$\text{mode[y]} = \exp(\mu - \sigma^2) \tag{3.86}$$

$$Var[y] = [\exp(\sigma^2) - 1]\exp(2\mu + \sigma^2) \tag{3.87}$$

If $\mu = 0$ and $\sigma \ll 1$ the distribution is approximately 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 multivariant Gaussian distribution (section 3.14). This is sometimes also used as a model for density fluctuations in the Universe.

3.11 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

$$p(x|x_{\min}, x_{\max}, \alpha) = x^\alpha \times \begin{cases} 0 & , \quad x < x_{\min} \\ (\alpha+1)\left[x_{\max}^{\alpha+1} - x_{\min}^{\alpha+1}\right]^{-1} & , \quad x_{\min} < x < x_{\max}, \alpha \neq -1 \\ \ln\left(\frac{x_{\max}}{x_{\min}}\right)^{-1} & , \quad x_{\min} < x < x_{\max}, \alpha = -1 \\ 0 & , \quad x > x_{\max} \end{cases} \tag{3.88}$$

The cumulative distribution is easily worked out

$$F(x|x_{\min}, x_{\max}, \alpha) = \begin{cases} 0 & , \quad x < x_{\min} \\ \frac{\left[x^{\alpha+1} - x_{\min}^{\alpha+1}\right]}{\left[x_{\max}^{\alpha+1} - x_{\min}^{\alpha+1}\right]} & , \quad x_{\min} < x < x_{\max}, \alpha \neq -1 \\ \frac{\ln\left(\frac{x}{x_{\min}}\right)}{\ln\left(\frac{x_{\max}}{x_{\min}}\right)} & , \quad x_{\min} < x < x_{\max}, \alpha = -1 \\ 1 & , \quad x > x_{\max} \end{cases} \tag{3.89}$$

as is the mean and variance.

3.12 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 $f(x_1, x_2, \ldots, x_k)$

$$E[f(x_1, x_2, \ldots, x_k)] = \int \cdots \int dx_1 \ldots dx_k \; f(x_1, x_2, \ldots, x_k) \; p(x_1, x_2, \ldots, x_k) \tag{3.90}$$

$$= \prod_{i=1}^{k} \int dx_i \; f(x_1, x_2, \ldots, x_k) \; p(x_1, x_2, \ldots, x_k) \tag{3.91}$$

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

$$C_{ij} = Cov[x_i x_j] \equiv E[(x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle)] \tag{3.92}$$

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 to better measure the degree of correlation it is convenient to normalize the variance so that it is unitless,

$$\rho_{xy} \equiv \frac{C_{xy}}{\sigma_x \sigma_y} \tag{3.93}$$

$Cov[xy]$ satisfies all the requirements of an inner (or "dot" or "scalar") product. One of the results of this is that covariance satisfies the **Cauchy–Schwarz inequality**

$$|Cov[xy]|^2 \leq Var[x]Var[y] \tag{3.94}$$

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

Another important relation is

$$C_{xy} = E[xy] - \bar{x}\bar{y} \tag{3.95}$$

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 uncorrelated. 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**.

3.13 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 $\prod_i p_i^{x_i}$ (Look back at the dice throwing example again.). 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

$$P(x_1, x_2, x_3, \ldots, x_k | N, \{p_i\}) = \frac{N!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k} = \frac{N!}{\prod_{i=1}^k x_i!} \prod_{i=1}^k p_i^{x_i} \tag{3.96}$$

The mean and variance of the distribution are

$$E[x_i] = N p_i \tag{3.97}$$

$$Var[x_i] = N p_i (1 - p_i) \tag{3.98}$$

And the covariance is

$$Cov[x_i x_j] = -N p_i p_j \tag{3.99}$$

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 to their means then getting more in one bin implies there are less in others.

3.14 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. It is also very useful when trying to understand some statistical argument or principle to put in a 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 go through some of its important properties and examples.

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

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{C}) = \frac{1}{(2\pi)^{n/2}\sqrt{|\boldsymbol{C}|}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{C}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right] \tag{3.100}$$

$$\equiv \mathcal{G}(\boldsymbol{x}\,|\boldsymbol{\mu}, \boldsymbol{C}\,) \tag{3.101}$$

where $\boldsymbol{C}$ is a n-by-n matrix and $\boldsymbol{\mu}$ is an n dimensional vector of parameters. $|\boldsymbol{C}|$ is the determinant of $\boldsymbol{C}$. This will define the function $\mathcal{G}(\boldsymbol{x}\,|\boldsymbol{\mu}, \boldsymbol{C}\,)$. 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.

Theorem 3.1 *The means of the multivariant Gaussian are*

$$E[x_i] = \mu_i \quad or \quad E[\boldsymbol{x}] = \boldsymbol{\mu} \tag{3.102}$$

Theorem 3.2 *And the covariances of the multivariant Gaussian are*

$$Cov[x_i, x_j] = E[(x_i - \mu_i)(x_j - \mu_j)] = C_{ij} \quad or \quad Cov[\boldsymbol{x}, \boldsymbol{x}] = E[(\boldsymbol{x} - \boldsymbol{\mu})^T (\boldsymbol{x} - \boldsymbol{\mu})] = \boldsymbol{C} \tag{3.103}$$

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

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

$$\boldsymbol{C}^{-1} = \begin{pmatrix} \sigma_1^2 & 0 & \dots \\ 0 & \sigma_2^2 & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \dots \\ 0 & \frac{1}{\sigma_2^2} & \dots \\ \vdots & \vdots & \ddots \end{pmatrix} \tag{3.104}$$

In this case there are no correlations between different variables.

PROOF OF MEAN: (theorem 3.1)

Lets calculate the means first

$$E[x_i] = \int_{-\infty}^{\infty} dx_i \dots \int_{-\infty}^{\infty} dx_i \dots \int_{-\infty}^{\infty} dx_n \; x_i \; p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{C}) \tag{3.105}$$

$$\tag{3.106}$$

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 $mu'_i = 0$,

$$E[x_i] = \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx'_i \dots \int_{-\infty}^{\infty} dx_n \; (\mu_i + x'_i) \; p(\boldsymbol{x'}|\boldsymbol{\mu'}, \boldsymbol{C}) \tag{3.107}$$

$$= \mu_i \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx'_i \dots \int_{-\infty}^{\infty} dx_n \; p(\boldsymbol{x'}|\boldsymbol{\mu'}, \boldsymbol{C}) + \int_{-\infty}^{\infty} dx_1 \dots \int_{-\infty}^{\infty} dx'_i \dots \int_{-\infty}^{\infty} dx_n \; x'_i \; p(\boldsymbol{x'}|\boldsymbol{\mu'}, \boldsymbol{C}) \tag{3.108}$$

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: (theorem 3.2)

$$Corr[\boldsymbol{x}, \boldsymbol{x}] = \int_{-\infty}^{\infty} d^n x \; (\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T \; p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{C}) \tag{3.109}$$

$$= \int_{-\infty}^{\infty} d^n y \; \boldsymbol{y}^T \boldsymbol{y} \; p(\boldsymbol{y}|0, \boldsymbol{C}) \qquad\qquad \boldsymbol{y} = \boldsymbol{x} - \boldsymbol{\mu} \tag{3.110}$$

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

$$\boldsymbol{C} = \mathbf{M}\boldsymbol{\Sigma}\mathbf{M}^{-1} \tag{3.111}$$

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

$$\mathbf{M}^T = \mathbf{M}^{-1} \tag{3.112}$$

$$|\mathbf{M}| \equiv \det(\mathbf{M}) = 1 \tag{3.113}$$

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

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

$$e^{\frac{1}{2}\boldsymbol{x}^T\boldsymbol{C}\boldsymbol{x}} \; d^n x = e^{\frac{1}{2}\boldsymbol{x}^T\boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}^T\boldsymbol{x}} \; d^n x = e^{\frac{1}{2}(\boldsymbol{M}^T\boldsymbol{x})^T\boldsymbol{\Sigma}(\boldsymbol{M}^T\boldsymbol{x})} \; d^n x = e^{\frac{1}{2}\boldsymbol{y}^T\boldsymbol{\Sigma}\boldsymbol{y}} \; |\boldsymbol{M}| d^n y = e^{\frac{1}{2}\boldsymbol{y}^T\boldsymbol{\Sigma}\boldsymbol{y}} \; d^n y \tag{3.114}$$

$$Corr[\boldsymbol{x},\boldsymbol{x}] = \int_{-\infty}^{\infty} d^n x\ \boldsymbol{x}\boldsymbol{x}^T\ p(\boldsymbol{x}|0,\boldsymbol{C}) \tag{3.115}$$

$$= \int_{-\infty}^{\infty} d^n y\ (\boldsymbol{M}\boldsymbol{y})(\boldsymbol{M}\boldsymbol{y})^T\ p(\boldsymbol{y}|0,\boldsymbol{\Sigma}) \tag{3.116}$$

$$= \int_{-\infty}^{\infty} d^n y\ \boldsymbol{M}\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{M}^T\ p(\boldsymbol{y}|0,\boldsymbol{\Sigma}) \tag{3.117}$$

$$= \boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}^T \tag{3.118}$$

$$= \boldsymbol{C} \tag{3.119}$$

The transformed or rotated variables $\boldsymbol{y} = \mathbf{M}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$ will be uncorrelated. This is the basis of the **principle components** decomposition of the data that we will get to later.

3.14.1 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 $\boldsymbol{y}$, $\mathbf{C}_{yy}$, a part that involves only components of $\boldsymbol{z}$, $\mathbf{C}_{zz}$ and a component that involves mixtures of the two, $\mathbf{C}_{xy}$.

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_z \end{bmatrix} \quad \boldsymbol{C} = \begin{bmatrix} \mathbf{C}_{yy} & \mathbf{C}_{zy} \\ \mathbf{C}_{zy}^T & \mathbf{C}_{zz} \end{bmatrix} \tag{3.120}$$

The conditional pdf is then

$$p(\boldsymbol{y}|\boldsymbol{z}) = \mathcal{G}\left(\boldsymbol{y}\,|\boldsymbol{\mu}'_y, \boldsymbol{\Sigma}_{yy}\right) \qquad \begin{cases} \boldsymbol{\mu}'_y = \boldsymbol{\mu}_y + \boldsymbol{C}_{zy}\boldsymbol{C}_{zz}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_z) \\ \boldsymbol{\Sigma}_{yy} = \boldsymbol{C}_{yy} - \boldsymbol{C}_{zy}\boldsymbol{C}_{zz}^{-1}\boldsymbol{C}_{zy}^T \end{cases} \tag{3.121}$$

which means

$$p(\boldsymbol{y}|\boldsymbol{z}) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{\mu}_y - \boldsymbol{C}_{zy}\boldsymbol{C}_{zz}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_z)\right)^T \left(\boldsymbol{C}_{yy} - \boldsymbol{C}_{zy}\boldsymbol{C}_{zz}^{-1}\boldsymbol{C}_{zy}^T\right)^{-1} \left(\boldsymbol{y} - \boldsymbol{\mu}_y - \boldsymbol{C}_{zy}\boldsymbol{C}_{zz}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_z)\right)\right] \tag{3.122}$$

3.14.2 marginalized Gaussian distribution

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

$$p(\boldsymbol{y}) = \int_{-\infty}^{\infty} d\boldsymbol{z}\ p(\boldsymbol{x}) = \int_{-\infty}^{\infty} d\boldsymbol{z}\ p(\boldsymbol{y},\boldsymbol{z}) = \int_{-\infty}^{\infty} d\boldsymbol{z}\ p(\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z}) \tag{3.123}$$

Using the same definitions (without proof) this is

$$p(\boldsymbol{y}) = \mathcal{G}\left(\boldsymbol{y}\,|\boldsymbol{\mu}_y, \boldsymbol{C}_{yy}\right) \tag{3.124}$$

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 is the matrix **completion of squares** formula

$$\frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + \mathbf{b}^T\boldsymbol{x} = \frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{A}^{-1}\mathbf{b}\right)^T\boldsymbol{A}\left(\boldsymbol{x} - \boldsymbol{A}^{-1}\mathbf{b}\right) - \frac{1}{2}\mathbf{b}^T\boldsymbol{A}^{-1}\mathbf{b} \tag{3.125}$$

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

3.14.3 combining two multivariant Gaussians

$$\mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu}_1, \boldsymbol{C}_1\right)\mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu}_2, \boldsymbol{C}_2\right) = \mathcal{G}\left(\boldsymbol{\mu}_1\,|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}\right)\mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}\right) \tag{3.126}$$

$$\boldsymbol{\Sigma} = \boldsymbol{C}_1 + \boldsymbol{C}_2 \tag{3.127}$$

$$\boldsymbol{\mu}_c = \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{C}_1\boldsymbol{\mu}_1 + \boldsymbol{C}_2\boldsymbol{\mu}_2\right) \tag{3.128}$$

A particularly important application of this is for the distribution of the sum of two independent Gaussian distributed variables.

Theorem 3.3 *If* $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{C}_1)$ *and* $\boldsymbol{x}' \sim \mathcal{N}(0, \boldsymbol{C}_2)$ *and their sum is* $\boldsymbol{s} = \boldsymbol{x} + \boldsymbol{x}'$ *then* $\boldsymbol{s} \sim \mathcal{N}(0, \boldsymbol{C}_1 + \boldsymbol{C}_2)$.

Lets call them $\boldsymbol{x}$ and $\boldsymbol{x}'$ and their sum $\boldsymbol{s} = \boldsymbol{x} + \boldsymbol{x}'$.

$$\begin{aligned}
p(\boldsymbol{s}) &= \int_{-\infty}^{\infty} d^n x \int_{-\infty}^{\infty} d^n x'\; p(\boldsymbol{x}, \boldsymbol{x}')\delta({}^D\boldsymbol{s} - \boldsymbol{x} - \boldsymbol{x}') && (3.129)\\
&= \int_{-\infty}^{\infty} d^n x\; p(\boldsymbol{x}, \boldsymbol{s} - \boldsymbol{x}) && (3.130)\\
&= \int_{-\infty}^{\infty} d^n x\; \mathcal{G}\left(\boldsymbol{x}\,|0, \boldsymbol{C}_1\right)\mathcal{G}\left(\boldsymbol{s} - \boldsymbol{x}\,|0, \boldsymbol{C}_2\right) && (3.131)\\
&= \int_{-\infty}^{\infty} d^n x\; \mathcal{G}\left(\boldsymbol{x}\,|0, \boldsymbol{C}_1\right)\mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{s}, \boldsymbol{C}_2\right) && (3.132)\\
&= \int_{-\infty}^{\infty} d^n x\; \mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}\right)\mathcal{G}\left(\boldsymbol{s}\,|0, \boldsymbol{\Sigma}\right) && (3.133)\\
&= \mathcal{G}\left(\boldsymbol{s}\,|0, \boldsymbol{\Sigma}\right)\int_{-\infty}^{\infty} d^n x\; \mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}\right) && (3.134)\\
&= \mathcal{G}\left(\boldsymbol{s}\,|0, \boldsymbol{\Sigma} = \boldsymbol{C}_1 + \boldsymbol{C}_2\right) && (3.135)
\end{aligned}$$

In particular if

$$\boldsymbol{C}_1 = \sigma_1^2 \quad \text{and} \quad \boldsymbol{C}_1 = \sigma_2^2 \tag{3.136}$$

then

$$\begin{aligned}
&\boldsymbol{C}_1^{-1} = \tfrac{1}{\sigma_1^2} \quad \text{and} \quad \boldsymbol{C}_1^{-1} = \tfrac{1}{\sigma_2^2}\\
&\boldsymbol{\Sigma} = \sigma_1^2 + \sigma_2^2\\
&\boldsymbol{\Sigma}^{-1} = (\sigma_1^2 + \sigma_2^2)^{-1}\\
&\boldsymbol{\mu}_c = \tfrac{\mu_1\sigma_1^2 + \mu_2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}
\end{aligned} \tag{3.137}$$

3.15 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 equivalently

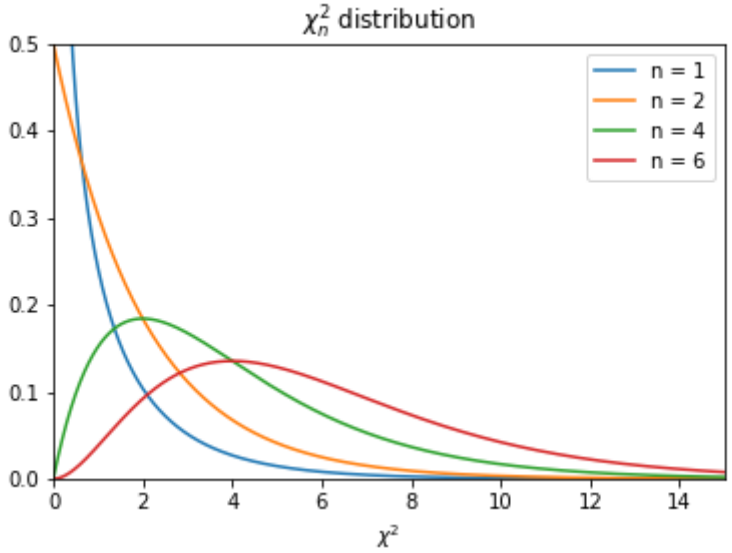

Figure 5: χ^2_n distribution for some different degrees of freedom, n.

a diagonal covariance. Lets define a new variables

$$z = \sum_i^N \frac{(x_i - \mu_i)^2}{\sigma_i^2}. \tag{3.138}$$

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 $x_1, x_2, \ldots$ to z. The Gaussian distribution is

$$p(x_1, x_2, \ldots x_N)dx_1 \ldots dx_N = \frac{1}{(2\pi)^{N/2}\prod_i \sigma_i} e^{-\frac{1}{2}\sum_i^N \frac{(x_i-\mu_i)^2}{\sigma_i^2}} dx_1 \ldots dx_N \tag{3.139}$$

$$= \frac{1}{(2\pi)^{N/2}\prod_i \sigma_i} e^{-\frac{1}{2}z} dx_1 \ldots dx_N \tag{3.140}$$

$$\frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2}z} dx'_1 \ldots dx'_N \qquad x' = \frac{x-\mu}{\sigma} \tag{3.141}$$

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

$$dx'_1 \ldots dx'_N = r^{n-1} dr d\theta_1 d\theta_3 \cdots = z^{n/2-1} dz d^n\Omega \tag{3.142}$$

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

$$p(z = \chi^2|n) = \begin{cases} \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} z^{\frac{n}{2}-1} e^{-\frac{z}{2}} & z \geq 0 \\ 0 & z < 0 \end{cases} \tag{3.143}$$

where the **gamma function** is defined as

$$\Gamma(x) \equiv \int_0^\infty dt\ e^{-t} t^{x-1}. \tag{3.144}$$

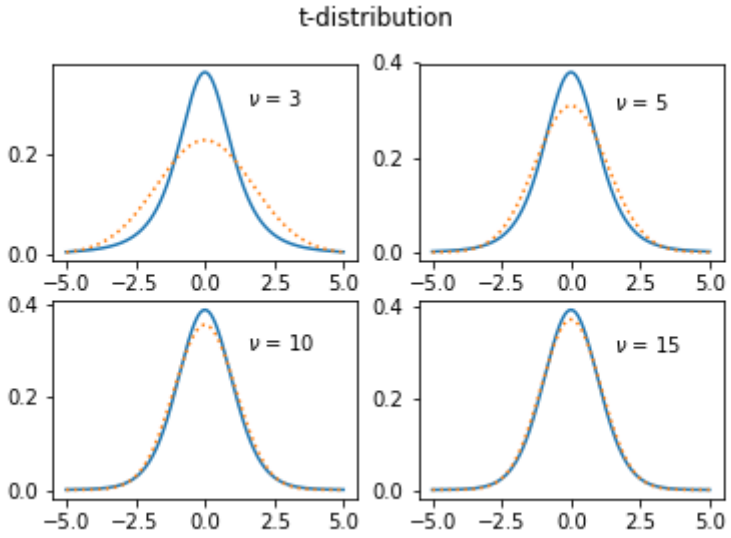

Figure 6: Student's t distribution for some different degrees of freedom, ν. The dotted curves are Gaussians with the same variances for comparison.

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^2_n/n is often given and compared to 1. The *mode* is $x =\max(n-2, 0)$ so $\chi^2_n/n = 1$ 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 5.

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

$$F(x|2) = 1 - e^{-x/2} \tag{3.145}$$

Theorem 3.4 *If* $x_1 \sim \chi^2_{n_1}$, $x_2 \sim \chi^2_{n_2}$ *and* $s = x_1 + x_2$ *then* $s \sim \chi^2_{n_1+n_2}$.

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

3.16 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

$$p_t(x|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{3.146}$$

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, see figure 6.

3.17 Exercises

1. Prove theorem 3.4.

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 a sample or data 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 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

$$\langle \bar{x}_N \rangle = \frac{1}{N}\sum_{i=0}^{N} \langle x_i \rangle \tag{4.2}$$

$$= \frac{1}{N}\sum_{i=0}^{N} \mu \tag{4.3}$$

$$= \mu \tag{4.4}$$

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 always 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,

$$Var[\bar{x}_N] = \left\langle [\bar{x}_N - \mu]^2 \right\rangle \tag{4.5}$$

$$= \left\langle [\text{Mean}(\{x\})]^2 \right\rangle - 2\mu\langle \text{Mean}(\{x\})\rangle + \mu^2 \tag{4.6}$$

$$= \left\langle [\text{Mean}(\{x\})]^2 \right\rangle - \mu^2 \tag{4.7}$$

$$= \left\langle \left[\frac{1}{N} \sum_{i=0}^{N} x_i \right]^2 \right\rangle - \mu^2 \tag{4.8}$$

$$= \frac{1}{N^2} \sum_{i=0}^{N} \sum_{j=0}^{N} \langle x_i x_j \rangle - \mu^2 \tag{4.9}$$

$$= \frac{1}{N^2} \left[\sum_{i=0}^{N} \langle x_i^2 \rangle + \sum_{i \neq j} \langle x_i x_j \rangle \right] - \mu^2 \tag{4.10}$$

$$= \frac{1}{N^2} \left[\sum_{i=0}^{N} (\sigma^2 + \mu^2) + \sum_{i \neq j} \langle x_i \rangle \langle x_j \rangle \right] - \mu^2 \tag{4.11}$$

$$= \frac{1}{N^2} \left[N(\sigma^2 + \mu^2) + N(N-1)\mu^2 \right] - \mu^2 \tag{4.12}$$

$$= \frac{\sigma^2}{N} \tag{4.13}$$

So you can see that the standard deviation of the mean will go 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 know.

So far we have not made any assumptions about how x is distributed except that the first 2 moments exist. Since the arithmetic mean is a linear function of the data, if the data is normally distributed the arithmetic mean will be normally distributed.

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

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 if 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

$$\sigma_\theta^2 = \left\langle \theta^2 \right\rangle - \mu^2 \tag{4.17}$$

$$= \sum_{ij} w_i w_j \langle x_i x_j \rangle - \mu^2 \tag{4.18}$$

$$= \sum_{ij} w_i w_j \langle x_i x_j \rangle - \mu^2 \tag{4.19}$$

$$= \sum_i w_i^2 \left\langle x_i^2 \right\rangle + \sum_{i \neq j} w_i w_j \langle x_i \rangle \langle x_j \rangle - \mu^2 \tag{4.20}$$

$$= \sum_i w_i^2 \left[\sigma_i^2 + \mu^2 \right] + \mu^2 \sum_{i \neq j} w_i w_j - \mu^2 \tag{4.21}$$

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

$$\frac{\partial \sigma_\theta^2}{\partial w_k} = 2 w_k \left[\sigma_k^2 + \mu^2 \right] + 2\mu^2 \sum_{i \neq k} w_i \tag{4.24}$$

$$= 2 w_k \left[\sigma_k^2 + \mu^2 \right] + 2\mu^2 \left[\sum_{i=0}^{N} w_i - w_k \right] \tag{4.25}$$

$$= 2 w_k \left[\sigma_k^2 + \mu^2 \right] + 2\mu^2 \left[1 - w_k \right] \qquad \text{use constraint} \tag{4.26}$$

$$= 2 w_k \sigma_k^2 + 2\mu^2 \tag{4.27}$$

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 points 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:

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

$$= \frac{1}{\sigma^2}\sum_i \left[(x_i - \mu) - (\bar{x} - \mu)\right]^2 \tag{4.41}$$

$$= \frac{1}{\sigma^2}\sum_i \left[(x_i - \mu)^2 - 2(x_i - \mu)(\bar{x} - \mu) - (\bar{x} - \mu)^2\right] \tag{4.42}$$

$$= \frac{1}{\sigma^2}\sum_i \left[(x_i - \mu)^2\right] - 2n(\bar{x} - \mu)(\bar{x} - \mu) - n(\bar{x} - \mu)^2 \tag{4.43}$$

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

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 3.4 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 unknown

We have learned that $\bar{x}$ 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 3.16.

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

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

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

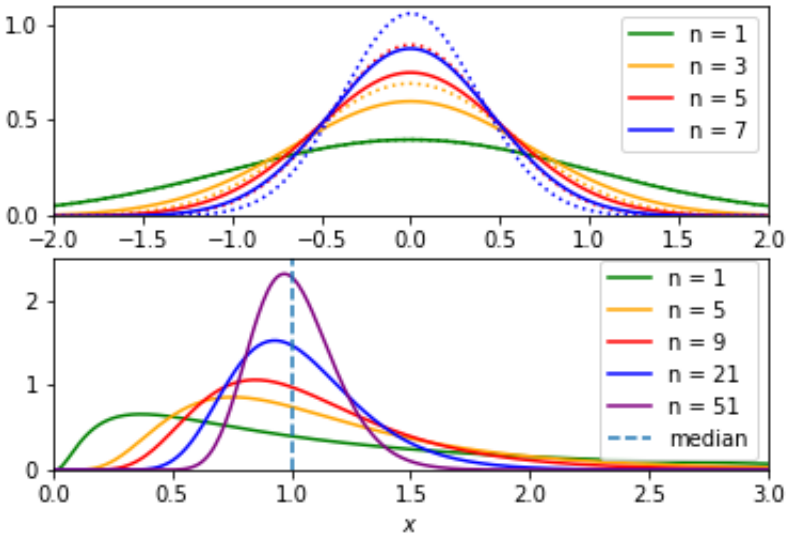

Figure 7: The probability of the sample median for normal (above) and lognormal (below) distributions. The $n = 1$ case is the original distribution. The dotted curves in the normal case are the distributions of the sample means based on the same n's.

where $p_t(t|\nu)$ is given in section 3.16. 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.

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 or there are a few large extreme outliers. 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. Running median filtering is also a common way to subtract a background in say a spectrum and usually preforms better than a running mean filter.

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

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

$$= N\binom{N-1}{\frac{N-1}{2}} p(x)F(x)^{\frac{N-1}{2}}\left[1-F(x)\right]^{\frac{N-1}{2}} \tag{4.49}$$

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

$$Var[x_{\rm m}] = \frac{1}{4(N+2)p(x_m)^2} \tag{4.50}$$

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 small sample sizes 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

$$p_{\rm max}(x|N) = Np(x)P_{\rm multi}\left(N-1|F(x), N-1\right) \tag{4.51}$$

$$= Np(x)P_{\rm multi}\left(0|1-F(x), N-1\right) \tag{4.52}$$

$$= Np(x)F(x)^{N-1} \tag{4.53}$$

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 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.

$$p(x_n|N) = Np(x_n)P_{\rm multi}\left(r-1|F(x_n), N-1\right) \tag{4.54}$$

$$p(x_n|N) = Np(x_n)P_{\rm multi}\left(\frac{nN}{q}-\frac{1}{2}\middle| F(x_n), N-1\right) \tag{4.55}$$

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 (4.54) with an estimate of the true pdf to judge how well the range can be estimated.

4.7 Exercises

1. Find the minimum variance weighting for an estimator of the the variance in the form

$$S_w^2 = \sum_i w_i (x_i - \bar{x})^2 \tag{4.56}$$

where the variance, σs, of each measurement are not equal.

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

$$\mathrm{tr}(\boldsymbol{A}) = \sum_i A_{ii} \tag{A.10}$$
$$\mathrm{tr}(\boldsymbol{A}) = \sum_i \lambda_{ii} \tag{A.11}$$
$$\mathrm{tr}(\boldsymbol{A}^T) = \mathrm{tr}(\boldsymbol{A}) \tag{A.12}$$
$$\mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{BA}) \tag{A.13}$$
$$\mathrm{tr}(\boldsymbol{A}+\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{A}) + \mathrm{tr}(\boldsymbol{B}) \tag{A.14}$$

$\boldsymbol{A}$ 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

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

It has the following properties

- all eigenvalues at positive
- $\mathrm{tr}(\boldsymbol{C}) > 0$
- 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	$\mathcal{G}\left(\boldsymbol{x}\,|\boldsymbol{\mu},\boldsymbol{C}\right)$
$\boldsymbol{x}$ 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	$\mathcal{L}(\boldsymbol{D}\|M_i)$ or $P(\boldsymbol{D}\|M_i)$
Bayesian evidence of data	$\mathcal{E}(\boldsymbol{D})$
Heaviside function, 1 when B is true, 0 otherwise	$\Theta(B)$
factorial	$N! = N(N-1)(N-2)\ldots 1$
double factorial	$N!! = N(N-2)(N-4)\ldots$
expectation value of $f(x)$	$\langle f(x)\rangle$ or $E[f(x)]$

Table 2: notation

B Notation

Notation may vary but in general I will follow the guide in table 2

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

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

Error function

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-u^2} du = \frac{1}{\sqrt{\pi}} \int_{-z}^z e^{-u^2} du \tag{C.5}$$

$$\lim_{N\to\infty} \left[1 + \frac{t^2}{2N}\right]^N = e^{\frac{t^2}{2}} \tag{C.6}$$

Stirling's approximation

$$\ln N! \simeq N \ln N - N \text{ for } N \gg 1 \tag{C.7}$$

or more accurately

$$N! \simeq \sqrt{2\pi N} \left(\frac{N}{e}\right)^N \text{ for } N \gg 1 \tag{C.8}$$

Index

References

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

Jaynes E., 2003, Probability Theory - The Logic of Science