Notes on Probability Theory

ASTP 611-01: Statistical Methods for Astrophysics*

Fall Semester 2017

Contents

Tuesday, September 12, 2017

1 Fundamentals of Probability

1.1 Probability and Logic

See Gregory, Chapters 1 and 2; see also section 1 of ***http://ccrg.rit.edu/~whelan/courses/2017_1sp_STAT_489/notes_parameters.pdf***

There are numerous interpretations of probability, but one which applies well to observational science is that of an extended logic. Let A be a proposition which could be either true or false, e.g., "The orbital period of Mars is between 686 and 687 days," "The student in question receives an A in stat methods," or "My detector will collect 427 photons in the next two hours." We may know, given the information at hand, that A is definitely true or definitely false, or we may be uncertain about the answer, either because our knowledge of the situation is incomplete, or because it refers to the outcome of an experiment with a random element, which has not occurred yet. The probability of the proposition A (which we also call an "event") is a number between 0 and 1 which quantifies our degree of certainty, given the information at hand. We write this as $P(A|I)$, where I represents some state of knowledge, to emphasize that the probability we assign always depends on the information we have, the assumption that a model is correct, etc. If A is definitely true, in the context of I, then $P(A|I) = 1$. If it's definitely false, $P(A|I) = 0$.

If A represents the outcome of an experiment which we could somehow arrange to repeat under identical circumstances, then $P(A|I)$ will be approximately equal to the long-term frequency of the event A. I.e., if we do some large number N of repetitions of the experiment, at the beginning of which we recreate the situation described by I, the approximate number of experiments in which A will turn out to be true is $N \times P(A|I)$. In the classical or "frequentist" approach to statistics, this is the only sort of event to which we're allowed to assign a probability, but in the more general "Bayesian" framework we are free to assign probabilities to any logical proposition.

Several basic operations can be used to combine logical propositions:

- Negation. $\overline{A}$ is true if A is false, and vice-versa. In words, we can think of $\overline{A}$ as "not A". (Other notations include A' and $\neg A$.)
- Intersection. A, B is true if A and B are both true. In words, this is "A and B". (Other notations include AB, $A \cap B$ and $A \wedge B$.) The advantage of the comma is that $P(A, B|I)$ is the probability that both A and B are true, given I.
- Union. $A + B$ is true if either A or B (or both) is true. In words, this is "A or B". (Other notations include $A \cup B$ and $A \vee B$.) Note the unfortunate aspect of this notation that $+$ is to be read as "or" rather than "and".

The relationships between logical propositions are easier to see with so-called *truth tables*, where you make a list of all of the possible combinations of truth and falsehood for different propositions:

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

Two propositions are considered to be equivalent if one is true whenever the other is true and false whenever the other is false. For instance, we can show that $A = A, B + A, \overline{B}$ with the following truth table:

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

There are basic rules of probability corresponding to these logical operations:

- $P(A|I) + P(\overline{A}|I) = 1$
- The product rule: $P(A, B|I) = P(A|B, I)P(B|I)$
- The sum rule: if A and B are mutually exclusive, i.e., if $P(A, B|I) = 0$, then $P(A + B|I) = P(A|I) + P(B|I)$.

Note that in this approach, where all probabilities are conditional, the product rule is really what's fundamental. Classical approaches to probability instead define the conditional probability as $P(A|B) = \frac{P(A,B)}{P(B)}$, and therefore only entertain consideration of the conditional probability $P(A|B)$ if B is not only something to which they're allowed assign a probability, but for which that probability is nonzero.

It is possible to show that any extension of logic which allows for a quantitative characterization of the plausibility of logical statements, obeying a few desiderata, is necessarily equivalent to probability theory. I encourage you to read Chapter Two of Gregory (which is based on a presentation in Jaynes) for details.

1.2 Bayes's Theorem

Because the logical "and" and "or" operations are symmetrical, i.e., A, B is equivalent to B, A and $A + B$ is equivalent to $B + A$, we can write the product rule in two different ways:

$$P(A, B|I) = P(A|B, I)P(B|I) = P(B|A, I)P(A|I) \tag{1.1}$$

this can be rearranged into *Bayes's Theorem*, which says that

$$P(A|B, I) = \frac{P(B|A, I)P(A|I)}{P(B|I)} \tag{1.2}$$

which is incredibly useful when you naturally know $P(B|A, I)$ but would like to know $P(A|B, I)$. For instance, suppose A refers to "I have terrible-disease-of-the-year (TDY)", B refers to "I test positive for TDY", and I represents the information that I had no extra risk factors or symptoms for TDY but was routinely tested, 0.1% of people in such a group have TDY, the test has a 2% false positive rate (2% of people without TDY will test positive for it) and a 1% false negative rate (1% of people with TDY will test negative for it). This information tells us that:

- $P(A|I) = 0.001$ so $P(\overline{A}|I) = 0.999$.
- $P(\overline{B}|A, I) = 0.01$ so $P(B|A, I) = 0.99$.
- $P(B|\overline{A}, I) = 0.02$ so $P(\overline{B}|\overline{A}, I) = 0.98$.

Additionally, since $B = B, A + B, \overline{A}$,

$$\begin{aligned} P(B|I) &= P(B, A|I) + P(B, \overline{A}|I) \\ &= P(B|A, I)P(A|I) + P(B|\overline{A}, I)P(\overline{A}|I) \\ &= 0.99 \times 0.001 + 0.02 \times 0.999 = 0.00099 + 0.01998 \\ &= 0.02097 \end{aligned} \tag{1.3}$$

We can then use Bayes's theorem to show that

$$P(A|B, I) = \frac{0.00099}{0.02097} \approx 0.04721 \tag{1.4}$$

I.e., if I test positive for TDY, I have about a 4.7% chance of actually having the disease. This is a lot less than $P(B|A, I)$, which is 99%!

In the context of observational science, Bayes's theorem is most commonly applied to a situation where H is a hypothesis which I'd like to evaluate and D is a particular set of data I've collected. It's usually straightforward to work out $P(D|H,I)$, the probability of observing a particular set of data values given a model, but I generally want to answer the question, what is my degree of belief in the hypothesis H after the observation. The answer, according to Bayes's Theorem, is

$$P(H|D,I) = \frac{P(D|H,I)P(H|I)}{P(D|I)} \tag{1.5}$$

As a bit of terminology:

- $P(H|I)$ is called the *prior probability* of hypothesis H.
- $P(H|D,I)$ is called the *posterior probability* of H. (This is a bit of a misnomer since it implies they are chronological; really we're considering the probabilities we assign with and without the knowledge of D.)
- $P(D|H,I)$ is called the *sampling distribution* if we view it as a function of D or the *likelihood* if we view it as a function of H.
- $P(D|I)$ is called the evidence, but as we'll see, it's usually considered a normalization constant, since it doesn't depend on H.

1.3 Bayesian and Frequentist Inference

Now a brief preview of how probabilities can be used to make statements about different hypotheses in light of observational data. We have a hypothesis H (e.g., a model, with certain values for its parameters) and it makes predictions about data D, i.e., the results of an observation or experiment. If these predictions had no randomness or uncertainty, i.e., H always led to a particular D, we could know for sure that e.g., $\overline{D}$ implied $\overline{H}$. Instead, we need to take the probability $P(D|H,I)$ for the outcome of an experiment given a hypothesis, and use it to say something about the hypothesis.

The Bayesian approach uses Bayes's theorem (1.5) to define $P(H|D,I)$ and evaluate it for the actual data observation that we made. In particular if we want to compare two hypotheses H_1 and H_2, we can consider the ratio[1]

$$\frac{P(H_1|D,I)}{P(H_2|D,I)} = \frac{P(D|H_1,I)}{P(D|H_2,I)}\frac{P(H_1|I)}{P(H_2|I)} \tag{1.6}$$

Now, different people with different background information I would likely assign different values to the prior odds ratio $\frac{P(H_1|I)}{P(H_2|I)}$, but much less information is used to calculate the likelihood ratio $\frac{P(D|H_1,I)}{P(D|H_2,I)}$. When we think of it as the factor by which we multiply the prior odds ratio $\frac{P(H_1|I)}{P(H_2|I)}$ to get the posterior odds ratio $\frac{P(H_1|D,I)}{P(H_2|D,I)}$, we call it the *Bayes Factor*.

In the frequentist approach, we're not allowed to assign probabilities to different hypotheses, so all we can consider is how the likelihood[2] $P(D|H,I)$ looks for different choices of the hypothesis H and evaluate it for the particular data D observed. The problem is that we can't address the question about the relative likelihood of different hypotheses and instead have to compare the data observed to different data that could have been observed for each hypothesis. In exchange for not having to say anything about $P(H|I)$ we can't actually say anything about $P(H|D,I)$.

[1] Note that the denominator $P(D|I)$ has cancelled out, which is a good thing since to calculate it we'd need to consider every possible hypothesis.

[2] In the strict frequentist interpretation, we're also not allowed to use H or I in a conditional probability so it's usually written something like $P(D;H)$ or $P_H(D)$, but let's stick with the Bayesian notation since we know what it means.

Thursday, September 14, 2017

2 Probability Distributions

In what follows, we will often suppress the explicit mention of the background information I on which all of our probabilities are conditional. The logical propositions to which we often assign probabilities involve the values of some random or otherwise unknown quantities. So for example, $N_{\text{counts}} = 37$ or $70\,\text{km/s/Mpc} < H_0 < 75\,\text{km/s/Mpc}$. Sometimes the notation gets a bit confused between a quantity and its value, and you'll see things like X for a "random variable" and x for a value it can take on. You'd like to be able to specify the probability that $X = x$, as a function of x. In practice, this is slightly complicated by whether we think of X as taking on only discrete values, or if it can take on any value in a continuous range.

If X is discrete, we can talk about its *probability mass function* $p_X(x) = P(X = x)$. This is often just written $p(x)$ or $P(x)$. For instance, if X is the number of events in a particular interval from a stationary process in which the events are independent of one another, and the average number of events expected in the interval given the long-term event rate is μ it is described by the Poisson distribution

$$p(x) = P(X = x) = \begin{cases} \frac{\mu^x}{x!}\, e^{-\mu} & x = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

However, it often happens that X is continuous, so that it is vanishingly unlikely that it takes on one specific value. For instance, the height of a randomly chosen person will not be exactly 175 cm. If you measure it to more significant figures, it will turn out to be 175.25 cm or 175.24732 cm etc. So instead we want to talk about the probability for X to be in a small interval, which we call the *probability density function*

$$f(x) = \lim_{dx \to 0} \frac{P(x < X < x + dx)}{dx} \tag{2.2}$$

so that

$$P(a < X < b) = \int_a^b f(x)\, dx \tag{2.3}$$

The pdf might be called $\text{pdf}(x)$ or even $P(x)$. A useful notation for the pdf is $\frac{dP}{dx}$, which tends to make the impact of changes of variables more obvious. In the end, it's a bit hopeless to try to stick to one letter, since you might want to talk about the joint probability distribution associated with some discrete and some continuous random variables. To give a concrete example, a common probability distribution is the Gaussian distribution with parameters μ and σ, which has pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/2\sigma^2}\,, \qquad -\infty < x < \infty \tag{2.4}$$

2.1 Expectation Values

In either case, you can define an operation known as the *expectation value*

$$E\left[g(X)\right] = \begin{cases} \sum_x g(x)\, p(x) & X \text{ discrete} \\ \int_{-\infty}^{\infty} g(x)\, f(x)\, dx & X \text{ continuous} \end{cases} \tag{2.5}$$

with the mean $\mu_X = E\left[X\right]$ as a special case, and also the variance

$$\text{Var}(X) = E\left[(X - \mu_X)^2\right] \tag{2.6}$$

Note that the linearity of the expectation value means that

$$\begin{aligned} \text{Var}(X) &= E\left[(X - \mu_X)^2\right] = E\left[X^2\right] - 2\mu_X E\left[X\right] + \mu_X^2 \\ &= E\left[X^2\right] - 2\mu_X^2 + \mu_X^2 = E\left[X^2\right] - \mu_X^2 \end{aligned} \tag{2.7}$$

To have a sensible probability distribution, we should satisfy a normalization condition $\sum_x p(x) = 1$ or $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

One particularly useful expectation value is known as the *moment generating function*, defined as

$$M(t) = E\left[e^{tX}\right] \tag{2.8}$$

If we use the Taylor series for the exponential, we can see that

$$M(t) = E\left[\sum_{k=0}^{\infty} \frac{(tX)^k}{k!}\right] = \sum_{k=0}^{\infty} \frac{t^k}{k!} E\left[X^k\right] \tag{2.9}$$

Considered as a function of t, we can think of this in terms of a McLaurin series whose kth coëfficient $M^{(k)}(0)/k!$ equals the kth moment divided by $k!$. I.e., if we take the kth derivative of the mgf, evaluated at $t = 0$, we get the kth moment:

$$M^{(k)}(0) = E\left(X^k\right) \tag{2.10}$$

Note that the moment generating function doesn't exist for every probability distribution. In general, the expectation value $E[h(X)]$ only exists if $E[|h(X)|]$ is finite. Consider, for example, the *Cauchy distribution*, which has the probability density function

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2} \qquad -\infty < x < \infty \tag{2.11}$$

It is properly normalized, since

$$\int_{-\infty}^{\infty} \frac{dx}{1+x^2} = \pi \tag{2.12}$$

(which you can show with the trigonometric substitution $x = \tan\theta$, but in any event, the integral is well-defined because the integrand goes like x^{-2} as x goes to $\pm\infty$). This distribution doesn't have a mean or in fact any moments, because

$$E\left[X^k\right] = \int_{-\infty}^{\infty} |x|^k \frac{dx}{1+x^2} \tag{2.13}$$

diverges because the integrand goes to $|x|^{k-2}$ for large $|x|$. Even if $k = 1$, we get logarithmic divergence because the integrand goes like $|x|^{-1}$, which doesn't go to zero fast enough.

A closely related function is the *characteristic function* $\Phi_X(\xi) = E\left[e^{i\xi X}\right]$. For a continuous distribution with pdf $f(x)$, this is

$$\Phi(\xi) = \int_{-\infty}^{\infty} e^{i\xi x} f(x)\,dx \tag{2.14}$$

which is, up to conventions about 2π and i vs $-i$, the Fourier transform of the pdf. The nice thing about the characteristic function is that it will always exist, because $\int_{-\infty}^{\infty} f(x)\,dx = 1 < \infty$. If it is an analytic function, we can use analytic continuation[3] to define $M(t) = \Phi(-it)$. In some cases the moment generating function won't exist, which typically means some of the moments are not defined. But if it does, it uniquely determines the probability distribution, basically because the inverse Fourier transform is unique.

2.2 Change of Variables in Probability Distributions

Imagine you have a random variable X and another random variable $Y = h(X)$ whose value is given by acting on the random value of X with the deterministic function $h()$. How can we determine the probability distribution for Y from the probability distribution for X?

Well, if they're discrete random variables, things are pretty straightforward:

$$p_Y(h(x)) = P(Y = h(x)) = P(X = x) = p_X(x) \tag{2.15}$$

[3] The characteristic function for the Cauchy distribution is $\Phi(\xi) = e^{-|\xi|}$, which is not analytic at $\xi = 0$, which is why we cannot use analytic continuation to define the mgf.

The pmf for Y has the same value as the pmf for X; you just have to evaluate it at the appropriate value.

Things get more interesting, though, for continuous random variables, since the pdf is a density, and not the probability of a specific value. It's fundamentally related to the fact that the probability for an event like $x_1 < X < x_2$ has to be the same as that for an equivalent event, say $y_1 < Y < y_2$ where $Y = h(X)$ and either $y_1 = h(x_1)$ and $y_2 = h(x_2)$ (if $h(x)$ is monotonically increasing) or $y_1 = h(x_2)$ and $y_2 = h(x_1)$ (if $h(x)$ is monotonically decreasing):

$$\int_{x_1}^{x_2} f_X(x)\,dx = \int_{y_1}^{y_2} f_Y(y)\,dy \tag{2.16}$$

You can derive the formula in detail, but the appropriate transformation is suggested by the notation:

$$\frac{dP}{dy} = \frac{\frac{dP}{dx}}{\left|\frac{dy}{dx}\right|} \tag{2.17}$$

i.e.,

$$f_Y(h(x)) = \frac{f_X(x)}{|h'(x)|} \tag{2.18}$$

Why the absolute value? Because the probability density for X and Y is defined to be positive, even if the transformation is such that Y decreases with increasing X. (Basically, it's a property of the way densities transform.)

2.2.1 Example: Inclination

As an example of the importance of a change of variables, consider the inclination ι between an arbitrary direction (say the normal to the orbital plane of a binary star system) and our line of sight. This is uniformly distributed in $\chi = \cos\iota$, from $\chi = -1$ to $\chi = 1$, so

$$f(\chi) = \frac{dP}{d\chi} = \frac{1}{2} \qquad -1 \le \chi \le 1 \tag{2.19}$$

If we change variables to ι, we have to use

$$\frac{d\chi}{d\iota} = \frac{d}{d\iota}\cos\iota = -\sin\iota \tag{2.20}$$

to find

$$f(\iota) = \frac{dP}{d\iota} = \frac{dP}{d\chi}\left|\frac{d\chi}{d\iota}\right| = \frac{1}{2}\sin\iota \qquad 0 \le \iota \le \pi \tag{2.21}$$

which is not uniform in ι.

Of course, this can also be extended to a probability density in the inclination and an azimuthal angle ψ, and leads to a direction uniformly distributed over the sphere:

$$\frac{d^2P}{d^2\Omega} = \frac{d^2P}{\sin\iota\,d\iota\,d\psi} = \frac{d^2P}{d\chi\,d\psi} = \frac{1}{4\pi} \tag{2.22}$$

Tuesday, September 19, 2017

2.3 Multivariate Distributions

Of course, you need not be dealing with only one unknown or random quantity at a time, and it's often useful to consider the joint probability distribution for multiple quantities at once. If they are discrete, this is pretty simple, for example

$$p(x_1, x_2) = P(X_1 = x_1, X_2 = x_2) \tag{2.23}$$

If they're continuous, you can again define a probability density

$$\begin{aligned} f(x_1, x_2) &= \frac{d^2P}{dx_1\,dx_2} \\ &= \lim_{\substack{dx_1\to 0\\ dx_2\to 0}} \frac{P(x_1 < X_1 < x_1 + dx_1, x_2 < X_2 < x_2 + dx_2)}{dx_1\,dx_2} \end{aligned} \tag{2.24}$$

We can use a version of the probability sum rule which says that if $\{B_1, B_2, \ldots, B_n\}$ are a set of mutually exclusive exhaustive alternatives, so that $A = A, B_1 + A, B_2 + \cdots + A, B_n$,

$$P(A|I) = \sum_{k=1}^{n} P(A, B_k|I) \tag{2.25}$$

This process is known as *marginalization*, and is useful if we really don't care which of the alternatives B_k is true. Likewise, if we want to focus on the probability distribution for X_1 and ignore the value of X_2, we can integrate out the joint probability density

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2 \tag{2.26}$$

This can be combined with the product rule $f(x_1, x_2) = f_{X_1|X_2}(x_1|x_2) f_{X_2}(x_2)$ to write

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2 = \int_{-\infty}^{\infty} f_{X_1|X_2}(x_1|x_2)\, f_{X_2}(x_2)\, dx_2 \tag{2.27}$$

where $f_{X_1|X_2}(x_1|x_2)$ is the conditional probability distribution for X_1, assuming that X_2 takes on the value x_2.

2.3.1 Expectation Values, Variance, Covariance and Correlation

If we have a multivariate probability distribution, we can define the expectation value of a function of the variables with the appropriate sum or integral, e.g., for a continuous distribution with density $f(x, y)$, the expectation value of a function $g(X, Y)$ is

$$E\left[g(X,Y)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\, f(x,y)\, dx\, dy \tag{2.28}$$

In addition to the usual means and variances:

$$\mu_X = E\left[X\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\, f(x,y)\, dx\, dy = \int_{-\infty}^{\infty} x\, f_X(x)\, dx \tag{2.29}$$

and

$$\mathrm{Var}(X) = \sigma_X^2 = E\left[(X-\mu_X)^2\right] = E\left[X^2\right] - \mu_X^2 \tag{2.30}$$

we can define the *covariance*

$$\mathrm{Cov}(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right] = E\left[XY\right] - \mu_X\mu_Y \tag{2.31}$$

and the dimensionless *correlation*

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \tag{2.32}$$

We can also define the moment generating function for a series of variables, e.g.,

$$M(t_1, t_2) = E\left[\exp(t_1 X_1 + t_2 X_2)\right] \tag{2.33}$$

where the moments are written in terms of partial derivatives

$$E[{X_1}^{m_1} {X_2}^{m_2}] = \left.\frac{\partial^{m_1+m_2}}{\partial^{m_1} t_1 \partial^{m_2} t_2} M(t_1, t_2)\right|_{(t_1,t_2)=(0,0)} \tag{2.34}$$

2.3.2 Change of Variables

For the case of the transformation of continuous random variables we have to deal with the fact that $f_{X_1,X_2}(x_1, x_2)$ and $f_{Y_1,Y_2}(y_1, y_2)$ are probability *densities* and the volume (area) element has to be transformed from one set of variables to the other. If we write $f_{X_1,X_2}(x_1, x_2) \sim \frac{d^2P}{dx_1 dx_2}$ and $f_{Y_1,Y_2}(y_1, y_2) \sim \frac{d^2P}{dy_1 dy_2}$, the transformation we'll need is

$$\frac{d^2P}{dy_1 dy_2} \sim \left|\det \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)}\right| \frac{d^2P}{dx_1 dx_2} \tag{2.35}$$

where we use the determinant of the Jacobian matrix

$$\frac{\partial(y_1, y_2)}{\partial(x_1, x_2)} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{pmatrix} \tag{2.36}$$

which may be familiar from the transformation of the volume element

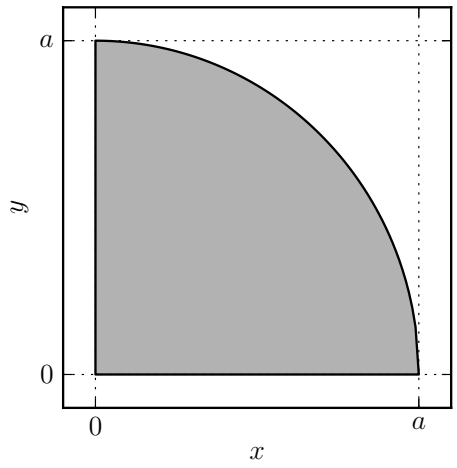
$$dy_1\, dy_2 = \left| \det \frac{\partial(y_1, y_2)}{\partial(x_1, x_2)} \right| dx_1\, dx_2 \tag{2.37}$$

if we change variables in a double integral.

To get a concrete handle on this, consider an example. Let X and Y be continuous random variables with a joint pdf

$$f_{X,Y}(x, y) = \begin{cases} \frac{4}{\pi} e^{-x^2-y^2} & 0 < x < \infty;\ 0 < y < \infty \\ 0 & \text{otherwise} \end{cases} \tag{2.38}$$

If we want to calculate the probability that $X^2 + Y^2 < a^2$ we have to integrate over the part of this disc which lies in the first quadrant $x > 0$, $y > 0$ (where the pdf is non-zero):



The limits of the x integral are determined by $0 < x$ and $x^2 + y^2 < a$, i.e., $x < \sqrt{a^2 - y^2}$; the range of y values represented can be seen from the figure to be $0 < y < a$, so we can write the probability as

$$P(X^2 + Y^2 < a^2) = \int_0^a \int_0^{\sqrt{a^2-y^2}} \frac{4}{\pi} e^{-x^2-y^2} dx\, dy \tag{2.39}$$

but we can't really do the integral in this form. However, if we define random variables $R = \sqrt{X^2 + Y^2}$ and[4] $\Phi = \tan^{-1}(Y/X)$, so that $X = R\cos\Phi$ and $Y = R\sin\Phi$, we can write the probability as

$$P(X^2+Y^2<a^2) = P(R<a) = \int_0^{\pi/2} \int_0^a f_{R,\Phi}(r, \phi)\, dr\, d\phi \tag{2.41}$$

if we have the transformed pdf $f_{R,\Phi}(r, \phi)$. On the other hand, we know that we can write the volume element $dx\, dy = r\, dr\, d\phi$. We can get this either from geometry in this case, or more generally by differentiating the transformation

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r\cos\phi \\ r\sin\phi \end{pmatrix} \tag{2.42}$$

[4]Note that we can only get away with using the arctangent $\tan^{-1}(y/x)$ as an expression for ϕ because x and y are both positive. In general, we need to be careful; $(x, y) = (-1, -1)$ corresponds to $\phi = -3\pi/4$ even though $\tan^{-1}([-1]/[-1]) = \tan^{-1}(1) = \pi/4$ if we use the principal branch of the arctangent. For a general point in the (x, y) plane, we'd need to use the function

$$\texttt{atan2}(y, x) = \begin{cases} \tan^{-1}(y/x) - \pi & x < 0 \text{ and } y < 0 \\ -\pi/2 & x = 0 \text{ and } y < 0 \\ \tan^{-1}(y/x) & x > 0 \\ \pi/2 & x = 0 \text{ and } y > 0 \\ \tan^{-1}(y/x) + \pi & x < 0 \text{ and } y \geq 0 \end{cases} \tag{2.40}$$

$\phi = \texttt{atan2}(y, x)$ to get the correct $\phi \in [-\pi, \pi)$.

to get

$$\begin{pmatrix} dx \\ dy \end{pmatrix} = \begin{pmatrix} \cos\phi\, dr - r\sin\phi\, d\phi \\ \sin\phi\, dr + r\cos\phi\, d\phi \end{pmatrix} = \begin{pmatrix} \cos\phi & -r\sin\phi \\ \sin\phi & r\cos\phi \end{pmatrix} \begin{pmatrix} dr \\ d\phi \end{pmatrix} \tag{2.43}$$

and taking the determinant of the Jacobian matrix:

$$\det \frac{\partial(x,y)}{\partial(r,\phi)} = \begin{vmatrix} \cos\phi & -r\sin\phi \\ \sin\phi & r\cos\phi \end{vmatrix} = r\cos^2\phi + r\sin^2\phi = r \tag{2.44}$$

so the volume element transforms like

$$dx\, dy = \left| \det \frac{\partial(x,y)}{\partial(r,\phi)} \right| dr\, d\phi = r\, dr\, d\phi \tag{2.45}$$

Even if we knew nothing about the transformation of random variables, we could use this to change variables in the integral (2.39) to get

$$\int_0^a \int_0^{\sqrt{a^2-y^2}} \frac{4}{\pi} e^{-x^2-y^2} dx\, dy = \int_0^{\pi/2} \int_0^a \frac{4}{\pi} e^{-r^2} r\, dr\, d\phi \tag{2.46}$$

If we compare the integrands of (2.46) and (2.46) we can see that the transformed pdf must be

$$f_{R,\Phi}(r,\phi) = \begin{cases} r\, e^{-r^2} & 0 < r < \infty;\ 0 < \phi < \pi/2 \\ 0 & \text{otherwise} \end{cases} \tag{2.47}$$

Incidentally, we can calculate the probability as

$$\begin{aligned} P(R < a) &= \int_0^{\pi/2} \int_0^a \frac{4}{\pi} e^{-r^2} r\, dr\, d\phi = \int_0^a e^{-r^2}\, 2r\, dr = -\, e^{-r^2} \Big|_0^a \\ &= 1 - e^{-a^2} \end{aligned} \tag{2.48}$$

To return to the general case, we see there are basically two things to worry about: one is the Jacobian determinant relating the volume elements in the two sets of variables, and the other is transforming the ranges of variables used to describe the event, as well as the allowed range of variables. In general terms, if $\mathcal{S}$ is the *support* of the random variables X_1 and X_2, i.e., the smallest region of $\mathbb{R}^2$ such that $P[(X_1, X_2) \in \mathcal{S}] = 1$ and $\mathcal{T}$ is the support of Y_1 and Y_2, we need a transformation of the pdf $f_{X_1,X_2}(x_1,x_2)$ defined on $\mathcal{S}$ such that

$$\begin{aligned} P[(X_1,X_2) \in A] &= \iint_A f_{X_1,X_2}(x_1,x_2)\, dx_1\, dx_2 \\ &= \iint_B f_{Y_1,Y_2}(y_1,y_2)\, dy_1\, dy_2 = P[(Y_1,Y_2) \in B] \end{aligned} \tag{2.49}$$

where B is the image of A under the transformation, i.e., $(x_1,x_2) \in A$ is equivalent to $\{u_1(x_1,x_2), u_2(x_1,x_2)\} \in B$. Since a change of variables in the integral gives us

$$\begin{aligned} &\iint_A f_{X_1,X_2}(x_1,x_2)\, dx_1\, dx_2 \\ &\quad = \iint_B f_{X_1,X_2}(w_1(y_1,y_2), w_2(y_1,y_2)) \left| \det \frac{\partial(x_1,x_2)}{\partial(y_1,y_2)} \right| dy_1\, dy_2 \end{aligned} \tag{2.50}$$

we must have, in general,

$$f_{Y_1,Y_2}(y_1,y_2) = \left| \det \frac{\partial(x_1,x_2)}{\partial(y_1,y_2)} \right| f_{X_1,X_2}(w_1(y_1,y_2), w_2(y_1,y_2)) \qquad (y_1,y_2) \in \mathcal{T} \tag{2.51}$$

which is the more careful way of writing the easier-to-remember formula we started with:

$$\frac{d^2P}{dy_1 dy_2} \sim \left| \det \frac{\partial(x_1,x_2)}{\partial(y_1,y_2)} \right| \frac{d^2P}{dx_1 dx_2} \tag{2.52}$$

2.3.3 General Formula

You can also perform this a change of variables on a joint probability density to write it in terms of another set of variables. Suppose you have N random variables $\{X_i\}$ from which you can determine the values of N random variables $\{Y_i\}$. Then the transformation uses the Jacobian determinant:

$$\frac{d^N P}{d^N y} = \left(\frac{d^N P}{d^N x}\right) \Big/ \left|\det\left\{\frac{\partial y_i}{\partial x_j}\right\}\right| \tag{2.53}$$

You can see this is the right thing to do because the Jacobian determinant is used to transform the measure of a multiple integral:

$$d^N y = \left|\det\left\{\frac{\partial y_i}{\partial x_j}\right\}\right| d^N x \tag{2.54}$$

and these probability densities are meant to be put under multiple integrals. Written in more standard notation, if we define

$$\mathbf{x} \equiv \{x_i\}, \qquad \mathbf{y} \equiv \{y_i\} \tag{2.55}$$

and

$$J_{\mathbf{yx}}(\mathbf{x}) = \det\left\{\frac{\partial y_i}{\partial x_j}\right\} \tag{2.56}$$

then

$$f_{\mathbf{Y}}(\mathbf{h}(\mathbf{x})) = \frac{f_{\mathbf{X}}(\mathbf{x})}{|J_{\mathbf{yx}}(\mathbf{x})|} \tag{2.57}$$

Thursday, September 21, 2017

3 Probability Distributions

3.1 Some Specific Probability Distributions

See Gregory, Chapter 5

Before we delve into frequentist and Bayesian applications of probability distributions, it's useful to consider some specific random variables, the sort of physical situations to which they're relevant, and what their probability distributions look like.

3.1.1 The Binomial Distribution (discrete)

Consider a sequence of identical independent yes/no questions, for example, the outcomes of a series of n flips of a coin, where the probability of heads on each flip is θ. Let H_i denote the event "the ith flip is heads" and $T_i = \overline{H_i}$ denote "the ith flip is tails", so that $P(H_i|\theta, I) = \theta$ and $P(T_i|\theta, I) = 1 - \theta$ If we want to get the probability of a particular sequence of heads and tails, we can use the product rule, e.g., the probability of a head on the first flip followed by a tail on the second flip is

$$P(H_1, T_2|\theta, I) = P(H_1|\theta, I)P(T_2|H_1, \theta, I) \tag{3.1}$$

If the information I includes the fact that coins have no "memory", i.e., we're not any more or less likely to flip a head on the second flip because we've got a head on the first flip, then $P(T_2|H_1, \theta, I) = P(T_2|\theta, I) = 1 - \theta$, so that

$$P(H_1, T_2|\theta, I) = P(H_1|\theta, I)P(T_2|\theta, I) = \theta(1 - \theta) \tag{3.2}$$

Note, this is a special case of independence of events. Saying that two events A and B are independent in the context of information I means that we can use a specialized product rule

$$P(A, B|I) = P(A|I)P(B|I) \qquad \text{iff } A \ \& \ B \text{ independent} \tag{3.3}$$

If we consider the event H_1, H_2, T_3, the probability is

$$\begin{aligned} P(H_1, H_2, T_3|\theta, I) &= P(H_1|\theta, I)P(H_2|\theta, I)P(T_3|\theta, I) \\ &= (\theta)(\theta)(1 - \theta) = \theta^2(1 - \theta)^1 \end{aligned} \tag{3.4a}$$

Similarly

$$\begin{aligned} P(H_1, T_2, H_3|\theta, I) &= P(H_1|\theta, I)P(T_2|\theta, I)P(H_3|\theta, I) \\ &= (\theta)(1-\theta)(\theta) = \theta^2(1-\theta)^1 \end{aligned} \tag{3.4b}$$

$$\begin{aligned} P(T_1, H_2, H_3|\theta, I) &= P(T_1|\theta, I)P(H_2|\theta, I)P(H_3|\theta, I) \\ &= (1-\theta)(\theta)(\theta) = \theta^2(1-\theta)^1 \end{aligned} \tag{3.4c}$$

So each possible sequence of two heads and a tail has a probability of $\theta^2(1-\theta)^1$. In general, the probability for each possible sequence of k heads and $n-k$ tails is $\theta^k(1-\theta)^{n-k}$. To get the overall probability of k heads and $n-k$ tails in n flips, we need to add up the probabilities for all of the relevant sequence; the result is the probability mass function for a binomial random variable K, $p(k|n,\theta)$. For example, we see there are three possible sequences of two heads and a tail, so

$$p(2|3,\theta) = 3\theta^2(1-\theta)^1 \tag{3.5}$$

In general, the number of ways to pick k of the n flips to be heads is "n choose k", $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, also known as ${}_nC_k$, or the "binomial coëfficient"[5] so the pmf for a binomial distribution is

$$p(k|n,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k} \qquad k = 0, 1, \ldots, n \tag{3.6}$$

We can show that this probability distribution is normalized using the binomial expansion formula

$$(a+b)^n = \sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k} \tag{3.7}$$

[5]For more review and practical discussion of the binomial coëfficient, see section 1.1 of `http://ccrg.rit.edu/~whelan/courses/2013_3fa_STAT_405/notes03.pdf`

so that

$$\sum_{k=0}^{n} p(k|n,\theta) = \sum_{k=0}^{n}\binom{n}{k}\theta^k(1-\theta)^{n-k} = (\theta + [1-\theta])^n = 1^n = 1 \tag{3.8}$$

The same sort of calculation can be used to find the moment generating function and use it to find the mean and variance relatively quickly:

$$\begin{aligned} M(t) = E\left[e^{tK}\right] &= \sum_{k=0}^{n} e^{tk}p(k|n,\theta) = \binom{n}{k}(\theta e^t)^k(1-\theta)^{n-k} \\ &= (\theta e^t + [1-\theta])^n \end{aligned} \tag{3.9}$$

It is often easier to work with the logarithm of the mgf, known as the *cumulant generating function*, which you studied on the most recent homework:

$$\psi(t) = \ln M(t) = n\ln(\theta e^t + [1-\theta]) \tag{3.10}$$

We use the chain rule to take its derivative:

$$\psi'(t) = \frac{n\theta e^t}{\theta e^t + [1-\theta]} = \frac{n\theta}{\theta + [1-\theta]e^{-t}} \tag{3.11}$$

from which we get the mean

$$E(K) = \psi'(0) = \frac{n\theta}{\theta + [1-\theta]} = n\theta \tag{3.12}$$

and differentiating again gives

$$\psi''(t) = -\frac{-n\theta(1-\theta)e^{-t}}{(\theta + [1-\theta]e^{-t})^2} \tag{3.13}$$

from which we get the variance:

$$\mathrm{Var}(K) = \psi''(0) = \frac{n\theta(1-\theta)}{(\theta + [1-\theta])^2} = n\theta(1-\theta) \tag{3.14}$$

3.1.2 Other Related Distributions

We mention briefly a few other distributions here, which are related to the binomial distribution, and which we will consider more closely as we need them for applications.

The hypergeometric distribution describes "sampling without replacement", where we have e.g., N balls, of which M are red, and we draw n of them without replacement, so the population of available balls changes with each draw; the number of red balls in our sample will be a hypergeometric random variable x whose pmf is

$$p(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \tag{3.15}$$

(The situation described by the binomial distribution can be thought of as "sampling with replacement", where you put the each ball back after drawing it, and mix things up again, so the probability of drawing a red ball remains M/N.

The negative binomial distribution describes a situation where instead of flipping the coin a fixed number n of times, we decide to keep flipping until we have r heads. The random variable X is the number of tails before the rth head. Using the fact that there must be a string of $r-1$ heads and x tails followed by one final failure, we can show that probability mass function is

$$p(x) = \binom{x+r-1}{r-1} p^{r-1}(1-p)^x p = \binom{x+r-1}{r-1} p^r (1-p)^x \tag{3.16}$$

The negative binomial distribution is important when considering the so-called "optional stopping problem", where questions of frequentist inference depend not just on the sequence of results seen, but on when you were planning to stop the experiment.

The multinomial distribution is a generalization of the binomial distribution, where instead of a series of coin flips, each with 2 possible outcomes, we have a series of trials which each have m possible outcomes. (E.g., we roll an m-sided die n times.) The parameters are $\{\theta_1, \theta_2, \ldots, \theta_m\}$, the probabilities for each of the m outcomes. Not only must each θ_i lie in the range $0 \le \theta_i \le 1$, they must also sum to one: $\sum_{i=1}^{m} \theta_i = 1$. We then have m random variables $K_1, K_2, \ldots, K_m$ with the joint pmf

$$\begin{aligned} p(k_1, k_2, \ldots, k_m) &= \frac{n!}{k_1! k_2! \cdots k_m!} \theta_1^{k_1} \theta_2^{k_2} \cdots \theta_m^{k_m}, \\ k_1 &= 0, 1, \ldots n; \quad k_2 = 0, 1, \ldots, n - k_1; \\ &\cdots; \quad k_m = n - k_1 - k_2 - \cdots - k_{m-1} \end{aligned} \tag{3.17}$$

3.1.3 The Poisson Distribution (discrete)

Consider a random process in which discrete events happen independently of each other with an average rate of r. If we count the number of events in an interval of duration T, that number of events is a random quantity with an expected value of $\mu = rT$. We usually think of the rate r as having units of inverse time and the duration T as having units of time. Examples of processes with rates in time include popcorn kernels popping, clicks on a Geiger counter, or gamma-ray bursts observed. But the interval could also be in space, e.g., we could be counting the number of cosmic rays collected in a detector of a certain area, or the number of galaxies within a certain redshift range found in a patch of sky of a given solid angle. The number K of events is a random variable with a probability mass function

$$P(K = k) = p(k|\mu) \tag{3.18}$$

Such a process is called a *Poisson process* if we can sub-divide the interval into smaller intervals (in time, space, sky position,

or whatever) and then the number of events in each sub-interval is an independent random variable with the same properties (but a smaller rate, obviously).

One way to wrap our head around a Poisson process is to think about collecting a large number M of identical intervals, in which the expected total number of events is $M\mu$, and randomly assigning each of those $M\mu$ events to one of the M intervals, without regard to where the other events have been placed. The number of events landing in any interval will be approximately[6] a Poisson random variable with Poisson parameter μ. See the ipython notebook `http://ccrg.rit.edu/~whelan/courses/2017_3fa_ASTP_611/data/notes_probability_poisson.ipynb` for an exploration of this. (As a further investigation, try removing the artificial factor of 200 scale-up and see how the histogram generated only approximately tracks the Poisson pmf.)

We can calculate the pmf for the *Poisson random variable* K as follows: subdivide the interval into some large number N of identically-sized sub-intervals. Each one has an expected number of events of μ/N. If we choose N large enough, we can make this number very very small. This means that in any one sub-interval, the odds are pretty good that there will be no events. There is some small chance (of order μ/N) of seeing one event, and a vanishingly small chance (of order $[\mu/N]^2$) of seeing two or more events in this sub-interval:

$$p\left(0\,|\mu/N\right) = 1 - \mu/N + \mathcal{O}\left([\mu/N]^2\right) \tag{3.19a}$$

$$p\left(1\,|\mu/N\right) = \mu/N + \mathcal{O}\left([\mu/N]^2\right) \tag{3.19b}$$

$$\sum_{k=2}^{\infty} p\left(k\,|\mu/N\right) = \mathcal{O}\left([\mu/N]^2\right) \tag{3.19c}$$

but this is basically a single trial which can have a yes (there is an event) or no (there is not an event) result, and the total number K of events in the larger interval can be approximated by a binomial random variable with N trials and a probability for success of μ/N for each trial. That means

$$\begin{aligned} p(k|\mu) &= \lim_{N\to\infty} b(k|\mu/N, N) = \lim_{N\to\infty} \frac{N!}{(N-k)!k!}\left(\frac{\mu}{N}\right)^k\left(1-\frac{\mu}{N}\right)^{N-k} \\ &= \frac{(\mu)^k}{k!}\lim_{N\to\infty}\left(1-\frac{\mu}{N}\right)^N \frac{N!}{(N-k)!}\left(N-\mu\right)^{-k} \end{aligned} \tag{3.20}$$

Now,

$$\frac{N!}{(N-k)!} = N(N-1)\ldots(N-k+1) = \prod_{\ell=0}^{k-1}(N-\ell) \tag{3.21}$$

and of course

$$\left(N-\mu\right)^{-k} = \prod_{\ell=0}^{k-1}\frac{1}{N-\mu} \tag{3.22}$$

so

$$\frac{N!}{(N-k)!}\left(N-\mu\right)^{-k} = \prod_{\ell=0}^{k-1}\frac{N-\ell}{N-\mu} = \prod_{\ell=0}^{k-1}\frac{1-\ell/N}{1-\mu/N} \tag{3.23}$$

but for finite k this is the product of a finite number of things, each of which goes to 1 as $N\to\infty$, so

$$p(k|r,T) = \frac{(\mu)^k}{k!}\lim_{N\to\infty}\left(1-\frac{\mu}{N}\right)^N = \frac{(\mu)^k}{k!}e^{-\mu}\ . \tag{3.24}$$

[6] It's only approximate because the total number of events is fixed to be $M\mu$. To make the construction exact, we would need to generate one Poisson random value using a Poisson distribution with mean $M\mu$, and then use *that* number

This is the *Poisson distribution*. It's easy to check that it's normalized, i.e.,

$$\sum_{k=0}^{\infty} p(k|\mu) = e^{-\mu} \sum_{k=0}^{\infty} \frac{(\mu)^k}{k!} = e^{-\mu} e^{\mu} = 1 \ . \tag{3.25}$$

Note that as a consistency check we can go back and verify our assumptions (3.19):

$$p(0|\mu/N) = e^{-\mu/N} = 1 - \mu/N + \mathcal{O}([\mu/N]^2) \tag{3.26a}$$

$$p(1|\mu/N) = \frac{\mu}{N} e^{-\mu/N} = \frac{\mu}{N} + \mathcal{O}([\mu/N]^2) \tag{3.26b}$$

$$\sum_{k=2}^{\infty} p(k|\mu/N) = 1 - p(0|\mu/N) - p(1|\mu/N) = \mathcal{O}([\mu/N]^2) \tag{3.26c}$$

We can get the mean and variance of the Poisson distribution either by taking the limiting forms of those for the binomial distribution, or by calculating the moment generating function:

$$M(t) = \sum_{k=0}^{\infty} e^{tk} \frac{\mu^k}{k!} e^{-\mu} = \frac{(\mu e^t)^k}{k!} e^{-\mu} = e^{\mu e^t} e^{-\mu} = e^{\mu(e^t - 1)} \tag{3.27}$$

To find the mean and variance, it's useful once again to use the cumulant generating function $\psi(t) = \ln M(t)$:

$$\psi(t) = \mu(e^t - 1) \tag{3.28}$$

Differentiating gives us

$$\psi'(t) = \mu e^t \tag{3.29}$$

and

$$\psi''(t) = \mu e^t \tag{3.30}$$

so the mean is

$$E(K) = \psi'(0) = \mu \tag{3.31}$$

(which was the definition we started with) and the variance is

$$\mathrm{Var}(K) = \psi''(0) = \mu \tag{3.32}$$

Tuesday, September 26, 2017

3.1.4 The Exponential Distribution (continuous)

Let's consider further a Poisson process. The Poisson distribution gives the pmf for the total number of events in an interval, which is a discrete random variable. Now consider another question. Suppose we are observing a Poisson process with an event rate r. Let's assume the intervals are in time, so that r has units of inverse time, and the number of events in a time Δt will be a Poisson random variable with parameter $r\,\Delta t$. Now suppose we start watching at a given time and see how long we have to wait for the next event. This waiting time T will itself be a random variable, with a probability density function $f_T(t|r)$ which depends on the rate r. Note that this is a *continuous* random variable. We can actually work out the pdf from our knowledge of the Poisson process. Consider the probability that T is longer than some value t:

$$P(T > t) = \int_t^{\infty} f_T(t'|r)\, dt' \tag{3.33}$$

This is the probability that in the interval of length t, beginning when we start watching, there are no events. But we know how to write the probability that there are no events from a Poisson process in an interval of a given length. It is the probability that the corresponding Poisson random variable (which has parameter rt) will take on the value 0:

$$P(K = 0) = p(0|rt) = \frac{(rt)^0}{0!} e^{-rt} = e^{-rt} \tag{3.34}$$

Equating the two expressions for this probability gives

$$\int_t^{\infty} f_T(t'|r)\, dt' = e^{-rt} \tag{3.35}$$

We can differentiate both sides with respect to t (*not* t') and find

$$-f_T(t|r) = -r\,e^{-rt} \tag{3.36}$$

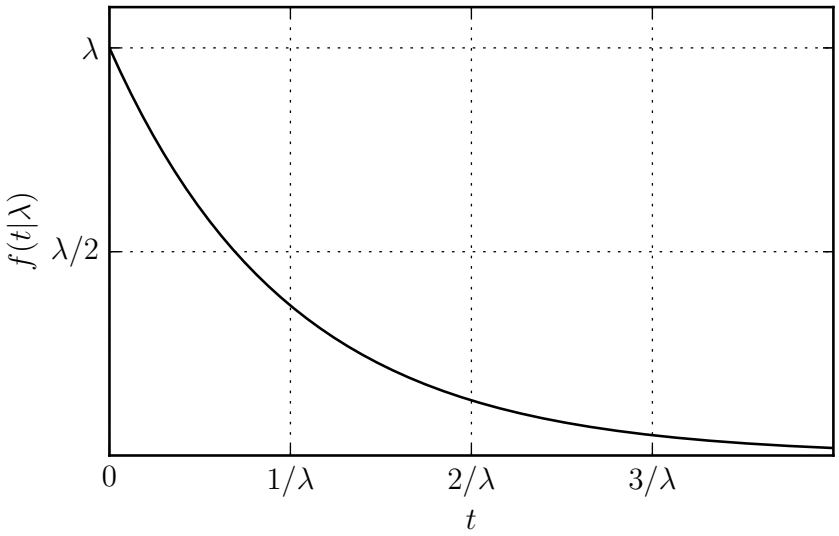
which gives us the pdf for T, the *exponential distribution*:

$$f(t|r) = r\,e^{-rt} \qquad t \geq 0 \tag{3.37}$$

It is often conventional to refer to the rate parameter as λ rather than r, so the pdf is

$$f(t|\lambda) = \lambda\,e^{-\lambda t} \qquad t \geq 0 \tag{3.38}$$

The pdf looks like this:



Note that this derivation made use of an integral of the pdf. For a general continuous random variable, we define the *cumulative distribution function*

$$F_X(x) := P(X \leq x) = \int_{-\infty}^{x} f_X(x')\,dx' \tag{3.39}$$

The derivative of the cdf is the pdf:

$$\frac{dF_X}{dx}(x) = f_X(x) \tag{3.40}$$

3.1.5 The Gamma Distribution (continuous)

The plot of the exponential distribution above has the same shape regardless of the value of the parameter λ; changing its value just changes the scales of the axes. It is, however, one member of a family of distributions known as the *Gamma distribution*, with two parameters $\alpha > 0$ and $\beta > 0$, and the pdf

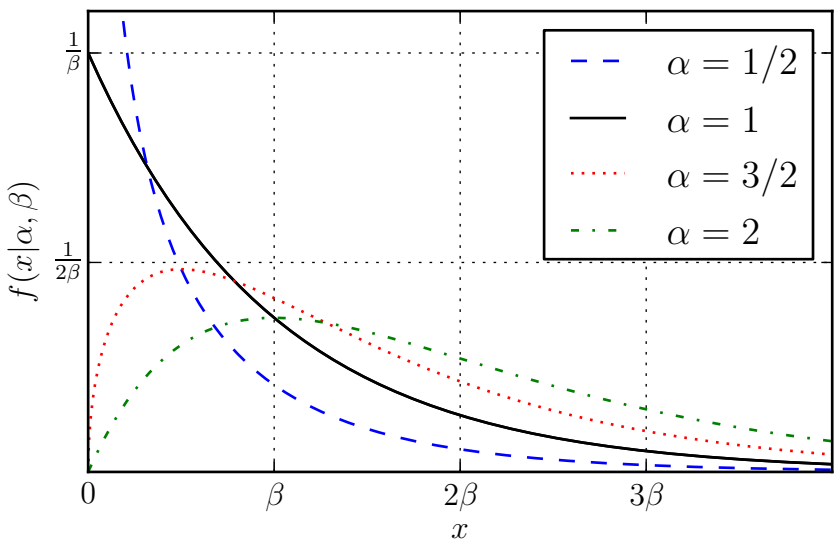
$$f(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} \qquad 0 < x < \infty \tag{3.41}$$

where

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u}\,du \tag{3.42}$$

is the Gamma function. If n is a non-negative integer, $\Gamma(n+1) = n!$, and in general $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$. Comparing the pdfs, we see that an exponential distribution with rate λ is a Gamma distribution with *shape parameter* $\alpha = 1$ and *scale parameter* $\beta = 1/\lambda$.[7] Here's the shape of the pdf for different choices of α.

[7] Note that some sources define the Gamma distribution in terms of a rate parameter which they call β, but is one over our scale parameter β. (Sources using this other convention include Wikipedia and the notes for previous versions of this class!) When in doubt, you can fall back on dimensional analysis. If β is a scale parameter, it will have the same units as x, and x/β will appear in the exponential. If it were a rate parameter, it would have units of $1/x$. Note that Gregory uses the scale parameter definition, but calls the parameter θ.



We can check that the pdf is normalized by taking the integral

$$\frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^\infty x^{\alpha-1}e^{-x/\beta}\,dx = \frac{1}{\Gamma(\alpha)}\int_0^\infty \left(\frac{x}{\beta}\right)^{\alpha-1} e^{-x/\beta}\,\frac{dx}{\beta}$$
$$= \frac{1}{\Gamma(\alpha)}\int_0^\infty u^{\alpha-1}e^{-u}\,du = \frac{1}{\Gamma(\alpha)}\Gamma(\alpha) = 1 \quad (3.43)$$

where we have made the change of variables $u = x/\beta$. The same calculation can be done to find the mgf:

$$M(t) = E(e^{tX}) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^\infty x^{\alpha-1}e^{-x/\beta}e^{tx}\,dx$$
$$= \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^\infty x^{\alpha-1}\exp\left(-\left[\frac{1}{\beta}-t\right]x\right)dx \quad (3.44)$$

If we require $t < \frac{1}{\beta}$ and make the substitution $u = (\frac{1}{\beta} - t)x$ so $x = \frac{\beta}{1-\beta t}\,u$, this becomes

$$M(t) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\left(\frac{\beta}{1-\beta t}\right)^\alpha \int_0^\infty u^{\alpha-1}e^{-u}\,du = (1-\beta t)^{-\alpha} \quad (3.45)$$

If we again construct the cumulant generating function

$$\psi(t) = \ln M(t) = -\alpha\ln(1-\beta t) \quad (3.46)$$

we find the derivative

$$\psi'(t) = \alpha\beta(1-\beta t)^{-1} \quad (3.47)$$

so the mean is

$$E(X) = \psi'(0) = \alpha\beta \quad (3.48)$$

and the second derivative

$$\psi''(t) = \alpha\beta^2(1-\beta t)^{-2} \quad (3.49)$$

so the variance is

$$\mathrm{Var}(X) = \psi''(0) = \alpha\beta^2 \quad (3.50)$$

This means that in particular, and exponential distribution (for which $\alpha = 1$ and $\beta = \lambda^{-1}$. has a mean of λ^{-1}, a variance of λ^{-2}, and a standard deviation of λ^{-1}.

Even more useful than the mean and variance, the mgf can be used to show what happens when we add two independent Gamma random variables with the same scale parameter, say X_1 with parameters (α_1, β) and X_2 with parameters (α_2, β).

First a few words about independence, as it applies to random variables. Just as events A and B are independent if and only if $P(A, B) = P(A)P(B)$, two continuous random variables are independent if and only if their joint pdf is $f(x_1, x_2) =$

$f_{X_1}(x_1)f_{X_2}(x_2)$. This then means that if we take the expectation value of some function of X_1 times some function of X_2, we have

$$\begin{aligned} E\left[h_1(X_1)h_2(X_2)\right] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x_1)h(x_2)f(x_1,x_2)\,dx_1\,dx_2 \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x_1)h(x_2)f_{X_1}(x_1)f_{X_2}(x_2)\,dx_1\,dx_2 \\ &= \int_{-\infty}^{\infty} h(x_1)f_{X_1}(x_1)\,dx_1 \int_{-\infty}^{\infty} h(x_2)f_{X_2}(x_2)\,dx_2 \\ &= E\left[h_1(X_1)\right]E\left[h_2(X_2)\right] \qquad \text{if } X_1 \text{ \& } X_2 \text{ independent} \end{aligned} \tag{3.51}$$

In particular, if $Y = X_1+X_2$, where X_1 and X_2 are independent, the mgf of Y is

$$\begin{aligned} M_Y(t) &= E\left[e^{tY}\right] = E\left[e^{t(X_1+X_2)}\right] = E\left[e^{tX_1}e^{tX_2}\right] = E\left[e^{tX_1}\right]E\left[e^{tX_2}\right] \\ &= M_{X_1}(t)M_{X_2}(t) \qquad \text{if } X_1 \text{ \& } X_2 \text{ independent} \end{aligned} \tag{3.52}$$

Specializing to the case described above where X_1 and X_2 are two Gamma random variables with the same scale parameter β, the mgf of their sum is

$$M(t) = M_1(t)M_2(t) = (1-\beta t)^{-\alpha_1}(1-\beta t)^{-\alpha_2} = (1-\beta t)^{-(\alpha_1+\alpha_2)} \tag{3.53}$$

which we see is the mgf of a Gamma random variable with parameters $\alpha_1 + \alpha_2$ and β. So, **if you add Gamma rvs with the same scale parameter, their sum is another Gamma rv whose scale parameter is the sum of the individual scale parameters.**

For example, if we add n independent exponential random variables with the same rate parameter λ, each of which is a Gamma$(1, 1/\lambda)$ random variable, their sum is a Gamma$(n, 1/\lambda)$ random variable. This special case of the Gamma distribution (where α is an integer) is called an *Erlang distribution.*

Thursday, September 28, 2017
Review for First Prelim Exam

Tuesday, October 3, 2017
First Prelim Exam

Thursday, October 5, 2017

3.1.6 The Gaussian (aka Normal) Distribution (continuous)

The Gaussian distribution, also known as the normal distribution, is a distribution with location parameter μ and scale parameter $\sigma > 0$. We refer to this as a $N(\mu, \sigma^2)$ distribution, which has pdf

$$f(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad -\infty < x < \infty \tag{3.54}$$

This turns out to be a good approximation in many situations, for reasons we'll delve into as the course goes on, but in brief

1. It is the limiting form of many distributions, as you'll investigate on the homework, which is mostly due to a result we'll see later known as the Central Limit Theorem
2. It is the natural result of a truncated Taylor expansion either of the logarithm of a pdf about a local maximum, or of an mgf about $t = 0$.
3. It is the Maximum Entropy distribution appropriate when you know nothing but the mean and variance of a distribution defined for all x.

To show that this is normalized, we take the integral

$$\int_{-\infty}^{\infty} f(x)\,dx = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-z^2/2}\,dz \tag{3.55}$$

where we have made the substitution $z = (x-\mu)/\sigma$. The integral

$$I = \int_{-\infty}^{\infty} e^{-z^2/2}\, dz \tag{3.56}$$

is a bit tricky; there's no ordinary function whose derivative is $e^{-z^2/2}$, so we can't just do an indefinite integral and evaluate at the endpoints. But we can do the definite integral by writing

$$\begin{aligned} I^2 &= \left(\int_{-\infty}^{\infty} e^{-x^2/2}\, dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2/2}\, dy\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\, dx\, dy \end{aligned} \tag{3.57}$$

If we interpret this as a double integral in Cartesian coördinates, we can change to polar coördinates r and ϕ, and write

$$\begin{aligned} I^2 &= \int_0^{2\pi}\int_0^{\infty} e^{-r^2/2}\, r\, dr\, d\phi = 2\pi \int_0^{\infty} e^{-r^2/2}\, r\, dr \\ &= -2\pi\, e^{-r^2/2}\Big|_0^{\infty} = 2\pi \end{aligned} \tag{3.58}$$

so

$$I = \int_{-\infty}^{\infty} e^{-z^2/2}\, dz = \sqrt{2\pi} \tag{3.59}$$

and the pdf does integrate to one.

To get the mgf, we have to take the integral

$$M(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} + tx\right)\, dx \tag{3.60}$$

If we complete the square in the exponent, we get

$$\begin{aligned} -\frac{(x-\mu)^2}{2\sigma^2} + tx &= -\frac{1}{2\sigma^2}\left(x - [\mu + t\sigma^2]\right)^2 + \frac{1}{2\sigma^2}\left(2\mu t\sigma^2 + t^2\sigma^4\right) \\ &= -\frac{1}{2\sigma^2}\left(x - [\mu + t\sigma^2]\right)^2 + \mu t + \frac{t^2\sigma^2}{2} \end{aligned} \tag{3.61}$$

so

$$\begin{aligned} M(t) &= \frac{1}{\sigma\sqrt{2\pi}} e^{\mu t + \frac{t^2\sigma^2}{2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x - [\mu + t\sigma^2])^2}{2\sigma^2}\right)\, dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} e^{\mu t + \frac{t^2\sigma^2}{2}} \int_{-\infty}^{\infty} e^{-z^2/2}\, dz = e^{\mu t + \frac{t^2\sigma^2}{2}} \end{aligned} \tag{3.62}$$

This means that the cumulant generating function is

$$\psi(t) = \ln M(t) = \mu t + \frac{\sigma^2 t^2}{2} \tag{3.63}$$

taking derivatives gives

$$\psi'(t) = \mu + \sigma^2 t \tag{3.64}$$

so the mean is

$$E(X) = \psi'(0) = \mu \tag{3.65}$$

and

$$\psi''(t) = \sigma^2 t \tag{3.66}$$

so the variance is

$$\mathrm{Var}(X) = \psi''(0) = \sigma^2 \tag{3.67}$$

which means that the parameters μ and σ are the mean and standard deviation of the distribution, as their names suggest. Note that this is in some sense the "simplest" possible distribution with a given mean and variance. For a general random variable, since $\psi(0) = 0$, $\psi'(0) = E(X)$ and $\psi''(0) = \mathrm{Var}(X)$, the first few terms of the Maclaurin series for $\psi(t)$ must be

$$\psi(t) = t\, E(X) + \frac{t^2}{2}\, \mathrm{Var}(X) + \mathcal{O}(t^3) \tag{3.68}$$

Given a random variable X which follows a $N(\mu, \sigma^2)$ distribution, we can define $Z = \frac{X-\mu}{\sigma}$. Its pdf will be

$$f_Z(z) = \sigma f_X(\mu + z\sigma) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \tag{3.69}$$

which is a $N(1, 0)$ distribution, also known as a *standard normal distribution.*

The cdf of a $N(\mu, \sigma)$ random variable will be

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(u-\mu)^2/(2\sigma^2)}\, du = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-\mu)/\sigma} e^{-t^2/2}\, dt \tag{3.70}$$

again, $e^{-t^2/2}$ is not the derivative of any known function, but it's useful enough that we define a function

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt \tag{3.71}$$

which is tabulated in lots of places. In terms of this, the cdf for a $N(\mu, \sigma^2)$ rv is

$$P(X \le x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \tag{3.72}$$

If we add two independent Gaussian random variables, X_1 and X_2, following $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ distributions, respectively, their sum has the mgf

$$\begin{aligned} M(t) = M_1(t)M_2(t) &= \exp\left(t\mu_1 + \frac{t^2\sigma_1^2}{2}\right) \exp\left(t\mu_2 + \frac{t^2\sigma_2^2}{2}\right) \\ &= \exp\left(t(\mu_1+\mu_2) + \frac{t^2(\sigma_1^2+\sigma_2^2)}{2}\right) \end{aligned} \tag{3.73}$$

which is the mgf of a $N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$ distribution. (In general, if we add independent random variables, their sum has a mean which is the sum of the means and a variance which is the sum of the variances, but what's notable here is the sum obeys a normal distribution, so it's characterized only by its mean and variance.

3.1.7 The Chi-Square Distribution (continuous)

Finally, suppose we have r independent Gaussian random variables $\{X_i\}$ with means μ_i and variances σ_i^2. Consider the combination

$$Y = \sum_{i=1}^{r} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2 \tag{3.74}$$

We can show that this obeys a $\chi^2(r)$ distribution

$$f_Y(y) = \frac{1}{\Gamma(r/2)2^{r/2}} y^{\frac{r}{2}-1} e^{-y/2} \tag{3.75}$$

which is a special case of a Gamma distribution with $\alpha = r$ and $\beta = 2$.

First note that the sum of r independent $\chi^2(1)$ random variables is a $\chi^2(r)$ random variable. This is because a $\chi^2(1)$ is the same as a Gamma$(\frac{1}{2}, 2)$, and if we add r independent Gamma$(\frac{1}{2}, 2)$ random variables, we get a Gamma$(\frac{r}{2}, 2)$ random variable, which is a $\chi^2(r)$ random variable.

Thus all we need to do to prove that $\sum_{i=1}^{r} \left(\frac{X_i-\mu_i}{\sigma_i}\right)^2$ is a $\chi^2(r)$ is to show that if X is $N(\mu, \sigma^2)$, so that $Z = \frac{X-\mu}{\sigma}$ is $N(0, 1)$, $Y = \frac{(X-\mu)^2}{\sigma} = Z^2$ is $\chi^2(1)$. Now, since

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad -\infty < z < \infty \tag{3.76}$$

we can't quite use the usual formalism for transformation of pdfs, since the transformation $Y = Z^2$ is not invertible. But

since $f_Z(-z) = f_Z(z)$ it's not hard to see that if we define a rv $W = |Z|$, it must have a pdf

$$f_W(w) = f_Z(-w) + f_Z(w) = 2f_Z(z) = \frac{2}{\sqrt{2\pi}} e^{-w^2/2} \quad 0 < w < \infty \tag{3.77}$$

and then we can use the transformation $y = w^2$, $w = y^{1/2}$ to work out

$$\begin{aligned} f_Y(y) = \frac{dP}{dy} = \frac{dP}{dw}\frac{dw}{dy} &= \frac{1}{2} y^{-1/2} f_W(y^{1/2}) \\ &= \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2} \qquad 0 < y < \infty \end{aligned} \tag{3.78}$$

If we recall the $\chi^2(1)$ pdf, it was

$$f(y) = \frac{1}{\Gamma(1/2) 2^{1/2}} y^{-1/2} e^{-y/2} \quad 0 < y < \infty \tag{3.79}$$

so the pdf of $Y = Z^2$ is the $\chi^2(1)$ pdf, if the value of $\Gamma(1/2)$ is $\sqrt{\pi}$. Now, if we think about it, that has to be the case, in order for the two pdfs to be normalized, but we can work out the value directly. Recall the Gamma function

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt \tag{3.80}$$

which is a finite positive number for any $\alpha > 0$. (For positive integer n, we know that $\Gamma(n) = (n-1)!$.) Thus

$$\Gamma(1/2) = \int_0^\infty t^{-1/2} e^{-t}\, dt \tag{3.81}$$

We can show that the integral is well-behaved at the lower limit of $t = 0$, and evaluate it, by changing variables to $u = \sqrt{2t}$ so that $t = u^2/2$ and $du = 2^{1/2} t^{-1/2}\, dt$; thus

$$\Gamma(1/2) = \frac{1}{\sqrt{2}} \int_0^\infty e^{-u^2}\, du = \frac{1}{\sqrt{2}} \frac{\sqrt{2\pi}}{2} = \sqrt{\pi} \tag{3.82}$$

as expected. We've used the symmetry of the integrand to say that

$$\int_0^\infty e^{-u^2}\, du = \frac{1}{2} \int_{-\infty}^\infty e^{-u^2}\, du = \frac{\sqrt{2\pi}}{2} \tag{3.83}$$

3.1.8 Summary of Properties of Gamma Distribution

- If $X \sim \text{Gamma}(\alpha, \beta)$, then $E[X] = \alpha\beta$ and $\text{Var}(X) = \alpha\beta^2$. The shape parameter α is dimensionless and the scale parameter β has the same units as X.
- If we add independent random variables, a $\text{Gamma}(\alpha_1, \beta)$ and a $\text{Gamma}(\alpha_2, \beta)$, their sum is a $\text{Gamma}(\alpha_1 + \alpha_2, \beta)$.
- If X is a $\text{Gamma}(\alpha_1, \beta)$, aX is a $\text{Gamma}(\alpha_1, a\beta)$.[8]
- The sum of the squares of r standard normal random variables is a $\chi^2(r)$, which is the same as a $\text{Gamma}(\frac{r}{2}, 2)$.
- The waiting time for the next event in a Poisson process with rate λ is an $\text{Exp}(\lambda)$ (a random variable obeying an exponential distribution), which is a $\Gamma(1, \frac{1}{\lambda})$.
- The waiting time for the nth event in a Poisson process with rate λ is a $\Gamma(n, \frac{1}{\lambda})$. (This special case of a Gamma distribution with integer α is also called an Erlang distribution.)

4 Sums of Random Variables

4.1 Mean and Variance

Consider a situation where there are two random variables X_1 and X_2, and we construct a new random variable which is their

[8] We haven't demonstrated this one, but it is easily done either by applying a change of variables to the pdf, or by using the mgf to show that $M_{aX}(t) = E\left[e^{t(aX)}\right] = E\left[e^{(ta)X}\right] = M_X(at) = (1 - \beta at)^{-\alpha}$ which is the mgf of a $\text{Gamma}(\alpha, a\beta)$

sum,

$$T = X_1 + X_2 \ . \tag{4.1}$$

If the expectation values of the random variables are

$$\mu_1 = E[X_1] \qquad \text{and} \qquad \mu_2 = E[X_2] \tag{4.2}$$

then the linearity of the expectation value operation means that the expectation value of their sum is

$$\mu_T = E[T] = E[X_1] + E[X_2] = \mu_1 + \mu_2 \tag{4.3}$$

If the random variables ave standard deviations σ_1 and σ_2 and covariance $\mathrm{Cov}(X_1, X_2)$, so that

$$E\left[(X_1 - \mu_1)^2\right] = \sigma_1^2 \tag{4.4a}$$
$$E\left[(X_2 - \mu_2)^2\right] = \sigma_2^2 \tag{4.4b}$$
$$E\left[(X_1 - \mu_1)(X_2 - \mu_2)\right] = \mathrm{Cov}(X_1, X_2) \tag{4.4c}$$

then the variance of their sum is

$$\begin{aligned}\sigma_T^2 &= E\left[(T - \mu_T)^2\right] = E\left[(X_1 + X_2 - \mu_1 - \mu_2)^2\right] \\ &= E\left[([X_1 - \mu_1] + [X_2 - \mu_2])^2\right] \\ &= E\left[(X_1 - \mu_1)^2\right] + 2E\left[(X_1 - \mu_1)(X_2 - \mu_2)\right] + E\left[(X_2 - \mu_2)^2\right] \\ &= \sigma_1^2 + \sigma_2^2 + 2\,\mathrm{Cov}(X_1, X_2)\end{aligned} \tag{4.5}$$

In particular, if X_1 and X_2 are independent or otherwise uncorrelated, then the variance of their sum is equal to the sum of their variances:

$$\sigma_T^2 = \sigma_1^2 + \sigma_2^2 \qquad \text{if } X_1 \text{ and } X_2 \text{ uncorrelated} \tag{4.6}$$

Note that this means the standard deviations are "added in quadrature":

$$\sigma_T = \sqrt{\sigma_1^2 + \sigma_2^2} \qquad \text{if } X_1 \text{ and } X_2 \text{ uncorrelated} \tag{4.7}$$

This is the standard way in which uncorrelated random errors are combined.

Note that in the special case where X_1 and X_2 are independent Gaussian random variables, we can show their sum is also a Gaussian random variable. This is pretty quick to show using the moment generating functions, since in this case

$$\begin{aligned}M(t) &= E\left[e^{X_1+X_2}\right] = E\left[e^{X_1}\right] E\left[e^{X_2}\right] = M_1(t)M_2(t) \\ &= \exp\left(t\mu_1 + \frac{1}{2}t^2\sigma_1^2\right) \exp\left(t\mu_2 + \frac{1}{2}t^2\sigma_2^2\right) \\ &= \exp\left(t(\mu_1 + \mu_2) + \frac{1}{2}t^2(\sigma_1^2 + \sigma_2^2)\right)\end{aligned} \tag{4.8}$$

Note also that this is a special case of a general linear combination of random variables

$$Y = \sum_{i=1}^{n} a_i X_i \tag{4.9}$$

for which

$$E[Y] = \sum_{i=1}^{n} a_i E[X_i] \tag{4.10}$$

and

$$\mathrm{Var}(Y) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \,\mathrm{Cov}(X_i, X_j) \tag{4.11}$$

For example,

$$\mathrm{Var}(a_1X_1 + a_2X_2) = {a_1}^2 V(X_1) + 2a_1a_2\,\mathrm{Cov}(X_1, X_2) + {a_2}^2 V(X_2) \tag{4.12}$$

4.2 IID Random Variables (Random Samples)

Considered now the example of N independent, identically distributed (iid) random variables $\{X_i\}$ with expectation values

$$E[X_i] = \mu \qquad \text{and} \qquad E[(X_i - \mu)(X_j - \mu)] = \delta_{ij}\,\sigma^2 \tag{4.13}$$

This is known as a *random sample*, and it can be used to estimate the properties of the underlying distribution. If we construct the sum

$$T = \sum_{i=1}^{N} X_i \tag{4.14}$$

then an extension of the results for a pair of random variables shows that its mean is

$$\mu_T = E[T] = \sum_{i=1}^{N} \mu = N\mu \tag{4.15}$$

and its variance is

$$\sigma_T^2 = E\left[(T - \mu_T)^2\right] = \sum_{i=1}^{N} \sigma^2 = N\sigma^2 \tag{4.16}$$

so its standard deviation is

$$\sigma_T = \sqrt{N}\,\sigma\ . \tag{4.17}$$

If we take the average of the N random variables

$$\overline{X} = \frac{1}{N}\sum_{k=0}^{N-1} X_i = \frac{T}{N}\ , \tag{4.18}$$

which is itself a random variable, we can see

$$\mu_{\overline{X}} = E\left[\overline{X}\right] = \frac{E[T]}{N} = \mu \tag{4.19}$$

and

$$\sigma_{\overline{X}}^2 = \mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{T}{N}\right) = \frac{\mathrm{Var}(T)}{N^2} = \frac{\sigma_T^2}{N^2} = \frac{N\sigma^2}{N} \tag{4.20}$$

which means

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{N}} \tag{4.21}$$

I.e., if you take the average of N iid random variables, the standard deviation is $1/\sqrt{N}$ times theie individual standard deviation.

Thursday, October 12, 2017

4.3 PDF of a Sum of Random Variables

If we consider two independent random variables X_1 and X_2 with (not necessarily identical) pdfs $f_1(x_1)$ and $f_2(x_2)$, so that their joint pdf is

$$f(x_1, x_2) = f_1(x_1) f_2(x_2) \tag{4.22}$$

and write their sum as

$$T = X_1 + X_2 \tag{4.23}$$

we can ask what the pdf $f_T(t)$ is. One way to approach this[9] is to consider the joint pdf $p(t, x_2)$. We can do this by changing variables from x_1 to $t = x_1 + x_2$; since we're actually changing

[9] An alternative, slick shortcut is to write

$$f_T(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \delta(t - [x_1 - x_2])\, f(x_1, x_2)\, dx_1\, dx_2$$

from $\{x_1, x_2\}$ to $\{t, x_2\}$, we can treat x_2 as a constant[10] and since $dt = dx_1$ in that case,

$$f(t, x_2) = f_1(t - x_2) f_2(x_2) \ . \tag{4.24}$$

if we then marginalize over x_2, we find

$$f(t) = \int_{-\infty}^{\infty} f(t - x_2)\, f_2(x_2)\, dx_2 \tag{4.25}$$

and so we see that *the pdf of a sum of variables is the convolution of their individual pdfs.*

Of course, we also know that when we add independent random variables, we multiply their mgfs. These things all tie together when you recall that the mgf is closely related to the characteristic function, which is the Fourier transform of the pdf. This means we're looking at an example of the convolution theorem: when you convolve pdfs, that's the same thing as multiplying their Fourier transforms.

4.4 The Central Limit Theorem

The central limit theorem states that if $\{X_i\}$ is a sample of size n drawn from a distribution with mean $E(X_i) = \mu$ and variance $\mathrm{Var}(X_i) = \sigma^2$, then

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \longrightarrow N(0, 1) \tag{4.26}$$

[10] Alternatively, we can consider the Jacobian determinant

$$\left\| \frac{\partial(t, x_2)}{\partial(x_1, x_2)} \right\| = \left| \det \begin{pmatrix} \left(\frac{\partial t}{\partial x_1}\right)_{x_2} & \left(\frac{\partial t}{\partial x_1}\right)_{x_1} \\ \left(\frac{\partial x_1}{\partial x_1}\right)_{x_2} & \left(\frac{\partial x_1}{\partial x_1}\right)_{x_1} \end{pmatrix} \right| = \left| \det \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \right| = 1$$

.

We can prove this using the moment generating function. I'll show this using the cumulant generating function

$$\psi(t) = \ln E(\exp[tX_i]) \tag{4.27}$$

for the distribution from which the sample is drawn. We know that $\psi(0) = 0$, $\psi'(0) = E(X_i) = \mu$, and $\psi''(0) = \mathrm{Var}(X_i) = \sigma^2$, so we can write the truncated McLaurin series for $\psi(t)$ as

$$\psi(t) = \psi(0) + t\psi'(0) + \frac{t^2}{2}\psi''(0) + o(t^2) = \psi(0) + \mu t + \frac{\sigma^2 t^2}{2} + o(t^2) \tag{4.28}$$

where $o(t^2)$ is some expression such that $\lim_{t\to 0} \frac{o(t^2)}{t^2} \to 0$; we also sometimes write this as $\mathcal{O}(t^3)$. We can write that more precisely as

$$\psi(t) = \psi(0) + t\psi'(0) + \frac{t^2}{2}\psi''(\xi(t)) = \mu t + \frac{\sigma^2 t^2}{2} + \frac{\psi''(\xi(t)) - \sigma^2}{2} t^2 \tag{4.29}$$

where $\xi(t)$ is some number between 0 and t. Since

$$\lim_{t\to 0} \psi''(\xi(t)) = \psi''(0) = \sigma^2 \tag{4.30}$$

the last term is indeed $o(t^2)$. Now we consider the cumulant generating function

$$\begin{aligned} \Psi(t; n) &= \ln E(\exp[tZ_n]) = \ln E\left(\exp\left[\frac{t}{\sigma\sqrt{n}}\left\{\sum_{i=1}^{n}(X_i - \mu)\right\}\right]\right) \\ &= \ln \prod_{i=1}^{n} E\left(\exp\left[\frac{t}{\sigma\sqrt{n}}\{X_i - \mu\}\right]\right) \\ &= \sum_{i=1}^{n} \ln E\left(\exp\left[\frac{t}{\sigma\sqrt{n}}\{X_i - \mu\}\right]\right) \\ &= n\left[\psi\left(\frac{t}{\sigma\sqrt{n}}\right) - \frac{t}{\sigma\sqrt{n}}\mu\right] \end{aligned} \tag{4.31}$$

If we use the Taylor expansion (4.29) we have

$$\Psi(t;n) = n\left[\frac{1}{2}\left(\frac{t}{\sigma\sqrt{n}}\right)^2 \psi''\left(\xi\left(\frac{t}{\sigma\sqrt{n}}\right)\right)\right] = \frac{t^2}{2\sigma^2}\psi''\left(\xi\left(\frac{t}{\sigma\sqrt{n}}\right)\right) \tag{4.32}$$

When we take the limit $n \to \infty$, the argument of $\psi''()$ goes to zero, and thus

$$\lim_{n\to\infty} \Psi(t;n) = \frac{t^2}{2\sigma^2}\psi''(0) = \frac{t^2}{2} \tag{4.33}$$

which is the natural log of the mgf of a $N(0,1)$ random variable, so by the mgf method we've shown $Z_n \longrightarrow N(0,1)$.

5 Multivariate Normal Distribution

5.1 Linear Algebra: Reminders and Notation

If $\mathbf{A}$ is an $m \times n$ matrix:

$$\mathbf{A} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} \tag{5.1}$$

and $\mathbf{B}$ is an $n \times p$ matrix,

$$\mathbf{B} = \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1p} \\ B_{21} & B_{22} & \cdots & B_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ B_{n1} & B_{n2} & \cdots & B_{np} \end{pmatrix} \tag{5.2}$$

then their product $\mathbf{C} = \mathbf{AB}$ is an $m \times p$ matrix as shown in Figure 1 so that $C_{ik} = \sum_{j=1}^{n} A_{ij}B_{jk}$.

If $\mathbf{A}$ is an $m \times n$ matrix, $\mathbf{B} = \mathbf{A}^{\mathrm{T}}$ is an $n \times m$ matrix with elements $B_{ij} = A_{ji}$:

$$\begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1m} \\ B_{21} & B_{22} & \cdots & B_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ B_{n1} & B_{n2} & \cdots & B_{nm} \end{pmatrix} = \mathbf{B} = \mathbf{A}^{\mathrm{T}} = \begin{pmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{nm} \end{pmatrix} \tag{5.4}$$

If $\mathbf{v}$ is an n-element column vector (which is an $n \times 1$ matrix) and $\mathbf{A}$ is an $m \times n$ matrix, $\mathbf{w} = \mathbf{Av}$ is an m-element column vector (i.e., an $m \times 1$ matrix):

$$\begin{aligned} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{pmatrix} = \mathbf{w} = \mathbf{Av} &= \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} \\ &= \begin{pmatrix} A_{11}v_1 + A_{12}v_2 + \cdots + A_{1n}v_n \\ A_{21}v_1 + A_{22}v_2 + \cdots + A_{2n}v_n \\ \vdots \\ A_{m1}v_1 + A_{m2}v_2 + \cdots + A_{mn}v_n \end{pmatrix} \end{aligned} \tag{5.5}$$

so that $w_i = \sum_{j=1}^{n} A_{ij}v_j$.

If $\mathbf{u}$ is an n-element column vector, then $\mathbf{u}^{\mathrm{T}}$ is an n-element row vector (a $1 \times n$ matrix):

$$\mathbf{u}^{\mathrm{T}} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \tag{5.6}$$

If $\mathbf{u}$ and $\mathbf{v}$ are n-element column vectors, $\mathbf{u}^{\mathrm{T}}\mathbf{v}$ is a number,

$$\mathbf{C} = \mathbf{AB} = \begin{pmatrix} C_{11} & C_{12} & \cdots & C_{1p} \\ C_{21} & C_{22} & \cdots & C_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ C_{m1} & C_{m2} & \cdots & C_{mp} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1p} \\ B_{21} & B_{22} & \cdots & B_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ B_{n1} & B_{n2} & \cdots & B_{np} \end{pmatrix}$$
$$= \begin{pmatrix} A_{11}B_{11} + A_{12}B_{21} + \cdots + A_{1n}B_{n1} & A_{11}B_{12} + A_{12}B_{22} + \cdots + A_{1n}B_{n2} & \cdots & A_{11}B_{1p} + A_{12}B_{2p} + \cdots + A_{1n}B_{np} \\ A_{21}B_{11} + A_{22}B_{21} + \cdots + A_{2n}B_{n1} & A_{21}B_{12} + A_{22}B_{22} + \cdots + A_{2n}B_{n2} & \cdots & A_{21}B_{1p} + A_{22}B_{2p} + \cdots + A_{2n}B_{np} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1}B_{11} + A_{m2}B_{21} + \cdots + A_{mn}B_{n1} & A_{m1}B_{12} + A_{m2}B_{22} + \cdots + A_{mn}B_{n2} & \cdots & A_{m1}B_{1p} + A_{m2}B_{2p} + \cdots + A_{mn}B_{np} \end{pmatrix} \tag{5.3}$$

Figure 1: Expansion of the product $\mathbf{C} = \mathbf{AB}$ to show $C_{ik} = \sum_{j=1}^{n} A_{ij}B_{jk}$.

known as the *inner product*:

$$\mathbf{u}^{\mathrm{T}}\mathbf{v} = \begin{pmatrix} u_1 & u_2 & \vdots & u_n \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \cdots \\ v_n \end{pmatrix} = u_1v_1 + u_2v_2 + \cdots + u_nv_n = \sum_{i=1}^{n} u_iv_i \tag{5.7}$$

If $\mathbf{v}$ is an m-element column vector, and $\mathbf{w}$ is an n-element column vector, $\mathbf{A} = \mathbf{v}\mathbf{w}^{\mathrm{T}}$ is an $m \times n$ matrix

$$\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} = \mathbf{A} = \mathbf{v}\mathbf{w}^{\mathrm{T}}$$
$$= \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix} \begin{pmatrix} w_1 & w_2 & \cdots & w_m \end{pmatrix} = \begin{pmatrix} v_1w_1 & v_1w_2 & \cdots & v_1w_n \\ v_2w_1 & v_2w_2 & \cdots & v_2w_n \\ \vdots & \vdots & \ddots & \vdots \\ v_mw_1 & v_mw_2 & \cdots & v_mw_n \end{pmatrix} \tag{5.8}$$

so that $A_{ij} = v_iw_j$.

If $\mathbf{M}$ and $\mathbf{N}$ are $n \times n$ matrices, the determinant $\det(\mathbf{MN}) = \det(\mathbf{M})\det(\mathbf{N})$.

If $\mathbf{M}$ is an $n \times n$ matrix (known as a square matrix), the inverse matrix $\mathbf{M}^{-1}$ is defined by $\mathbf{M}^{-1}\mathbf{M} = \mathbf{1}_{n\times n} = \mathbf{M}\mathbf{M}^{-1}$ where $\mathbf{1}_{n\times n}$

is the identity matrix

$$\mathbf{1}_{n\times n} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \tag{5.9}$$

If $\mathbf{M}^{-1}$ exists, we say $\mathbf{M}$ is invertible.

If $\mathbf{M}$ is a real, symmetric $n \times n$ matrix, so that $\mathbf{M}^{\mathrm{T}} = \mathbf{M}$, i.e., $M_{ji} = M_{ij}$, there is a set of n orthonormal *eigenvectors* $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ with real eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, so that $\mathbf{M}\mathbf{v}_i = \lambda_i \mathbf{v}_i$. Orthonormal means

$$\mathbf{v}_i^{\mathrm{T}} \mathbf{v}_j = \delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases} \tag{5.10}$$

where we have introduced the Kronecker delta symbol δ_{ij}. The eigenvalue decomposition means

$$\mathbf{M} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^{\mathrm{T}} \tag{5.11}$$

The determinant is $\det(\mathbf{M}) = \prod_{i=1}^{n} \lambda_i$. If none of the eigenvalues $\{\lambda_i\}$ are zero, $\mathbf{M}$ is invertible, and the inverse matrix is

$$\mathbf{M}^{-1} = \sum_{i=1}^{n} \frac{1}{\lambda_i} \mathbf{v}_i \mathbf{v}_i^{\mathrm{T}} \tag{5.12}$$

If all of the eigenvalues $\{\lambda_i\}$ are positive, we say $\mathbf{M}$ is positive definite. If none of the eigenvalues $\{\lambda_i\}$ are negative, we say $\mathbf{M}$ is positive semi-definite.

Tuesday, October 17, 2017

5.2 Special Case: Independent Gaussian Random Variables

Before considering the general multivariate normal distribution, consider the case of n independent normally-distributed random variables $\{X_i\}$ with means $\{\mu_i\}$ and variances $\{\sigma_i\}$. The pdf for X_i, the ith random variable, is

$$f_i(x_i) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2} \right) \tag{5.13}$$

and its mgf is

$$M_i(t_i) = \exp\left(t_i \mu_i + \frac{1}{2} t_i^2 \sigma_i^2 \right) \tag{5.14}$$

If we consider the random variables X_i to be the elements of a random vector

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \tag{5.15}$$

its expectation value is

$$E(\mathbf{X}) = \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} \tag{5.16}$$

and its variance-covariance matrix is

$$\mathrm{Cov}(\mathbf{X}) = \boldsymbol{\sigma}^2 = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix} \tag{5.17}$$

which is diagonal because the different X_is are independent of each other and therefore have zero covariance. We can thus write the joint mgf for these random variables as

$$\begin{aligned} M(\mathbf{t}) &= \prod_{i=1}^{n} M_i(t_i) = \exp\left(\sum_{i=1}^{n}\left[t_i\mu_i + \frac{1}{2}t_i^2\sigma_i^2\right]\right) \\ &= \exp\left(\mathbf{t}^{\mathrm{T}}\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\boldsymbol{\sigma}^2\mathbf{t}\right) \end{aligned} \tag{5.18}$$

We can also write the joint pdf

$$f(\mathbf{x}) = \prod_{i=1}^{n} f_i(x_i) = \frac{1}{\sqrt{\prod_{i=1}^{n} 2\pi\sigma_i^2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu_i)^2}{\sigma_i^2}\right) \tag{5.19}$$

in matrix form if we consider a few operations on the matrix $\boldsymbol{\sigma}^2$. First, since it's a diagonal matrix, its determinant is just the product of its diagonal entries:

$$\det\boldsymbol{\sigma}^2 = \prod_{i=1}^{n}\sigma_i^2 \tag{5.20}$$

and, for that matter,

$$\det(2\pi\boldsymbol{\sigma}^2) = \prod_{i=1}^{n} 2\pi\sigma_i^2 \tag{5.21}$$

Also, we can invert the matrix to get

$$\boldsymbol{\sigma}^{-2} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n^2} \end{pmatrix} \tag{5.22}$$

so

$$\sum_{i=1}^{n}\frac{(x_i-\mu_i)^2}{\sigma_i^2} = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\sigma}^{-2}(\mathbf{x}-\boldsymbol{\mu}) \tag{5.23}$$

which makes the pdf for the random vector X

$$f(\mathbf{x}) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\sigma}^2)}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\sigma}^{-2}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{5.24}$$

The generalization from n independent normal random variables to an n-dimensional multivariate normal distribution is to use the same matrix form for $M(\mathbf{t})$ and just to replace $\boldsymbol{\sigma}^2$, which was a diagonal matrix with positive diagonal entries, with a general symmetric positive semi-definite matrix. So one change is to allow $\boldsymbol{\sigma}^2$ to have off-diagonal entries, and another is to allow it to have zero eigenvalues. If $\boldsymbol{\sigma}^2$ is positive definite, i.e., its eigenvalues are all positive, we can use the matrix expression for the pdf $f(\mathbf{x})$ as well. As we'll see, if $\boldsymbol{\sigma}^2$ has some zero eigenvalues, we won't be able to define a pdf for the random vector $\mathbf{X}$.

5.3 Multivariate Distributions

Recall that a set of n random variables X_1, X_2, $\ldots X_n$ can be combined into a *random vector*

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \tag{5.25}$$

with expectation value

$$E(\mathbf{X}) = \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} \tag{5.26}$$

and variance-covariance matrix

$$\boldsymbol{\sigma}^2 = \mathrm{Cov}(\mathbf{X}) = E([\mathbf{X}-\boldsymbol{\mu}]^{\mathrm{T}}[\mathbf{X}-\boldsymbol{\mu}]) = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1,X_2) & \cdots & \mathrm{Cov}(X_1,X_n) \\ \mathrm{Cov}(X_1,X_2) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_1,X_n) & \mathrm{Cov}(X_2,X_n) & \cdots & \mathrm{Var}(X_n) \end{pmatrix} \tag{5.27}$$

The variance-covariance matrix must be positive semi-definite, i.e., have no negative eigenvalues. To see why that is the case, let $\{\lambda_i\}$ be the eigenvalues and $\{\mathbf{v}_i\}$ be the orthonormal eigenvectors, so that $\boldsymbol{\sigma}^2 = \sum_{i=1}^{n} \mathbf{v}_i \lambda_i \mathbf{v}^{\mathrm{T}}{}_i$. For each i define a random variable $\mathcal{X}_i = \mathbf{v}_i^{\mathrm{T}}\mathbf{X}$. It has mean $E(\mathcal{X}_i) = \mathbf{v}_i^{\mathrm{T}}\boldsymbol{\mu}$ and variance

$$\begin{aligned} \mathrm{Var}(\mathcal{X}_i) &= E([\mathbf{v}_i^{\mathrm{T}}(\mathbf{X}-\boldsymbol{\mu})]^2) = E[\mathbf{v}_i^{\mathrm{T}}(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{v}_i] \\ &= \mathbf{v}_i^{\mathrm{T}} E[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^{\mathrm{T}}]\mathbf{v}_i = \mathbf{v}_i^{\mathrm{T}}\boldsymbol{\sigma}^2\mathbf{v}_i = \lambda_i \mathbf{v}_i^{\mathrm{T}}\mathbf{v}_i = \lambda_i \end{aligned} \tag{5.28}$$

Since the variance of a random variable must be non-negative, $\boldsymbol{\sigma}^2$ cannot have any negative eigenvalues.

Incidentally, we can see that the different random variables $\{\mathcal{X}_i\}$ are uncorrelated, since

$$\begin{aligned} \mathrm{Cov}(\mathcal{X}_i,\mathcal{X}_j) &= E[\mathbf{v}_i^{\mathrm{T}}(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{v}_j] \\ &= \mathbf{v}_i^{\mathrm{T}} E[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^{\mathrm{T}}]\mathbf{v}_j = \mathbf{v}_i^{\mathrm{T}}\boldsymbol{\sigma}^2\mathbf{v}_j = \lambda_j \mathbf{v}_i^{\mathrm{T}}\mathbf{v}_j = \lambda_i \delta_{ij} \end{aligned} \tag{5.29}$$

Remember that $\mathrm{Cov}(\mathcal{X}_i,\mathcal{X}_j) = 0$ does not necessarily imply that $\mathcal{X}_i$ and $\mathcal{X}_j$ are independent. (It will, however, turn out to be the case for normally distributed random variables.)

Note that if we assemble the $\{\mathcal{X}_i\}$ into a column vector

$$\boldsymbol{\mathcal{X}} = \begin{pmatrix} \mathcal{X}_1 \\ \mathcal{X}_2 \\ \vdots \\ \mathcal{X}_n \end{pmatrix} = \begin{pmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{v}_n^{\mathrm{T}} \end{pmatrix} \mathbf{X} = \boldsymbol{\Gamma}\mathbf{X} \tag{5.30}$$

where the matrix $\boldsymbol{\Gamma}$ is made up out of the components of the orthonormal eigenvectors $\{\mathbf{v}_i\}$:

$$\boldsymbol{\Gamma} = \begin{pmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{v}_n^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} (\mathbf{v}_1)_1 & (\mathbf{v}_1)_2 & \cdots & (\mathbf{v}_1)_n \\ (\mathbf{v}_2)_1 & (\mathbf{v}_2)_2 & \cdots & (\mathbf{v}_2)_n \\ \vdots & \vdots & \ddots & \vdots \\ (\mathbf{v}_n)_1 & (\mathbf{v}_n)_2 & \cdots & (\mathbf{v}_n)_n \end{pmatrix} \tag{5.31}$$

This matrix is not symmetric, but it is orthogonal, meaning that $\boldsymbol{\Gamma}^{\mathrm{T}} = \boldsymbol{\Gamma}^{-1}$. We can see this from

$$\begin{aligned} \boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}} &= \begin{pmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{v}_n^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2^{\mathrm{T}} & \cdots & \mathbf{v}_n^{\mathrm{T}} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{v}_1^{\mathrm{T}}\mathbf{v}_1 & \mathbf{v}_1^{\mathrm{T}}\mathbf{v}_2 & \cdots & \mathbf{v}_1^{\mathrm{T}}\mathbf{v}_n \\ \mathbf{v}_2^{\mathrm{T}}\mathbf{v}_1 & \mathbf{v}_2^{\mathrm{T}}\mathbf{v}_2 & \cdots & \mathbf{v}_2^{\mathrm{T}}\mathbf{v}_n \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{v}_n^{\mathrm{T}}\mathbf{v}_1 & \mathbf{v}_n^{\mathrm{T}}\mathbf{v}_2 & \cdots & \mathbf{v}_n^{\mathrm{T}}\mathbf{v}_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = \mathbf{1} \end{aligned} \tag{5.32}$$

This matrix $\boldsymbol{\Gamma}$ can be thought of a transformation from the original basis to the eigenbasis for $\boldsymbol{\sigma}^2$. One effect of this is that it *diagonalizes* $\boldsymbol{\sigma}^2$:

$$\boldsymbol{\Gamma}\boldsymbol{\sigma}^2\boldsymbol{\Gamma}^{\mathrm{T}} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} = \boldsymbol{\Lambda} \tag{5.33}$$

Finally, recall that the moment generating function is defined as

$$M(\mathbf{t}) = E(e^{\mathbf{tX}}) \tag{5.34}$$

and that if we define $\psi(\mathbf{t}) = \ln M(\mathbf{t})$,

$$\left.\frac{\partial \psi}{\partial t_i}\right|_{\mathbf{t}=\mathbf{0}} = \mu_i \tag{5.35}$$

and

$$\left.\frac{\partial^2 \psi}{\partial t_i \partial t_j}\right|_{\mathbf{t}=\mathbf{0}} = \mathrm{Cov}(X_1, X_2) \tag{5.36}$$

This means that if we do a Maclaurin expansion of $\psi(\mathbf{t})$ we get, in general,

$$\psi(\mathbf{t}) = \mathbf{t}^{\mathrm{T}}\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\boldsymbol{\sigma}^2\mathbf{t} + \ldots \tag{5.37}$$

where the terms indicated by $\ldots$ have three or more powers of $\mathbf{t}$.

5.4 General Multivariate Normal Distribution

We define a multivariate normal random vector X as a random vector having the moment generating function

$$M(\mathbf{t}) = \exp\left(\mathbf{t}^{\mathrm{T}}\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\boldsymbol{\sigma}^2\mathbf{t}\right) \tag{5.38}$$

We refer to the distribution as $N_n(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. Note that this is equivalent to starting with the Maclaurin series for $\psi(\mathbf{t}) = \ln M(\mathbf{t})$ and cutting it off after the quadratic term.

We start with the mgf rather than the pdf because it applies whether the variance-covariance matrix $\boldsymbol{\sigma}^2$ is positive definite or only positive semi-definite, i.e., whether it has one or more zero eigenvalues. To see what happens if one or more of the eigenvalues is zero, we use the orthonormal eigenvectors $\{\mathbf{v}_i\}$ of $\boldsymbol{\sigma}^2$ to combine the random variables in $\mathbf{X}$ into n uncorrelated random variables $\{\mathcal{X}_i\}$, where $\mathcal{X}_i = \mathbf{v}_i^{\mathrm{T}}\mathbf{X}$, which have means $E(\mathcal{X}_i) = \mathbf{v}_i^{\mathrm{T}}\boldsymbol{\mu}$ and variances $\mathrm{Var}(\mathcal{X}_i) = \lambda_i$. If we combine the $\{\mathcal{X}_i\}$ into a random vector

$$\boldsymbol{\mathcal{X}} = \boldsymbol{\Gamma}\mathbf{X} \tag{5.39}$$

where

$$\boldsymbol{\Gamma} = \begin{pmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{v}_n^{\mathrm{T}} \end{pmatrix} \tag{5.40}$$

is the orthogonal matrix made up out of eigenvector components, $\boldsymbol{\mathcal{X}}$ has mean $E(\boldsymbol{\mathcal{X}}) = \boldsymbol{\Gamma}\boldsymbol{\mu}$ and variance-covariance matrix

$$\mathrm{Cov}(\boldsymbol{\mathcal{X}}) = \boldsymbol{\Lambda} = \boldsymbol{\Gamma}\boldsymbol{\sigma}^2\boldsymbol{\Gamma}^{\mathrm{T}} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} \tag{5.41}$$

The random vector $\boldsymbol{\mathcal{X}}$ also follows a multivariate normal distribution, in this case $N_n(\boldsymbol{\Gamma}\boldsymbol{\mu}, \boldsymbol{\Lambda})$. To show this, we'll show the more general result, that if $\mathbf{X}$ is a $N_n(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ multivariate normal random vector, $\mathbf{A}$ is an $m \times n$ constant matrix and $\mathbf{b}$ is an m-element column vector, the random vector $\mathbf{Y} = \mathbf{AX} + \mathbf{b}$ also obeys a multivariate normal distribution. (Note that this works whether m is equal to, less than, or greater than n!) We prove this using the mgf. The mgf for $\mathbf{Y}$ is

$$\begin{aligned} M_Y(\mathbf{t}) &= E[\exp(\mathbf{t}^{\mathrm{T}}\mathbf{Y})] = E[\exp(\mathbf{t}^{\mathrm{T}}\mathbf{AX} + \mathbf{t}^{\mathrm{T}}\mathbf{b})] \\ &= e^{\mathbf{t}^{\mathrm{T}}\mathbf{b}} E[\exp([\mathbf{A}^{\mathrm{T}}\mathbf{t}]^{\mathrm{T}}\mathbf{X})] \end{aligned} \tag{5.42}$$

Now here is the key step. $\mathbf{t}$ is an m-element column vector. $\mathbf{A}$ is an $m \times n$ matrix, so its transpose $\mathbf{A}^{\mathrm{T}}$ is an $n \times m$ matrix,

and the combination $\mathbf{A}^{\mathrm{T}}\mathbf{t}$ is an n-element column vector, whose transpose is the n-element row vector

$$[\mathbf{A}^{\mathrm{T}}\mathbf{t}]^{\mathrm{T}} = \mathbf{t}^{\mathrm{T}}\mathbf{A} \tag{5.43}$$

Therefore, the expectation value in the last line above is just the mgf for the original multivariate normal random vector $\mathbf{X}$ evaluated at the argument $\mathbf{A}^{\mathrm{T}}\mathbf{t}$:

$$\begin{aligned} E[\exp([\mathbf{A}^{\mathrm{T}}\mathbf{t}]^{\mathrm{T}}\mathbf{X})] &= M_X(\mathbf{A}^{\mathrm{T}}\mathbf{t}) \\ &= \exp\left([\mathbf{A}^{\mathrm{T}}\mathbf{t}]^{\mathrm{T}}\boldsymbol{\mu} + \frac{1}{2}[\mathbf{A}^{\mathrm{T}}\mathbf{t}]^{\mathrm{T}}\boldsymbol{\sigma}^2[\mathbf{A}^{\mathrm{T}}\mathbf{t}]\right) \\ &= \exp\left(\mathbf{t}^{\mathrm{T}}\mathbf{A}\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{A}\boldsymbol{\sigma}^2\mathbf{A}^{\mathrm{T}}\mathbf{t}\right) \end{aligned} \tag{5.44}$$

This makes the mgf for $\mathbf{Y}$ equal to $e^{\mathbf{t}^{\mathrm{T}}\mathbf{b}}$ times this, or

$$M_Y(\mathbf{t}) = \exp\left(\mathbf{t}^{\mathrm{T}}[\mathbf{A}\boldsymbol{\mu} + \mathbf{b}] + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{A}\boldsymbol{\sigma}^2\mathbf{A}^{\mathrm{T}}\mathbf{t}\right) \tag{5.45}$$

which is the mgf for a normal random vector with mean $\mathbf{A}\boldsymbol{\mu}+\mathbf{b}$ and variance-covariance matrix $\mathbf{A}\boldsymbol{\sigma}^2\mathbf{A}^{\mathrm{T}}$, i.e., one that obeys a $N_m(\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{A}\boldsymbol{\sigma}^2\mathbf{A}^{\mathrm{T}})$ distribution.

So, we return to the random vector $\boldsymbol{\mathcal{X}} = \boldsymbol{\Gamma}\mathbf{X}$, which we now see is a multivariate normal random vector with mean $\boldsymbol{\Gamma}\boldsymbol{\mu}$ and diagonal variance-covariance matrix $\boldsymbol{\Lambda}$. Its mgf is

$$\begin{aligned} M_{\boldsymbol{\mathcal{X}}}(\mathbf{t}) &= \exp\left(\mathbf{t}^{\mathrm{T}}\boldsymbol{\Gamma}\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{t}\right) = \exp\left(\sum_{i=1}^{n}[t_i\mathbf{v}_i^{\mathrm{T}}\boldsymbol{\mu} + \frac{1}{2}\lambda_i t_i^2]\right) \\ &= \prod_{i=1}^{n}\exp\left(t_i\mathbf{v}_i^{\mathrm{T}}\boldsymbol{\mu} + \frac{1}{2}\lambda_i t_i^2\right) = \prod_{i=1}^{n} M_{\mathcal{X}_i}(t_i) \end{aligned} \tag{5.46}$$

which is the mgf of n independent random variables. There are two possibilities: either $\boldsymbol{\sigma}^2$ (and thus $\boldsymbol{\Lambda}$) is positive definite, which means all of the $\{\lambda_i\}$ are positive, or one or more of the $\{\lambda_i\}$ are zero.

In the first case, we have the special case we considered before, n independent normally-distributed random variables, so the joint pdf is

$$f_{\boldsymbol{\mathcal{X}}}(\boldsymbol{\xi}) = \prod_{i=1}^{n} f_{\mathcal{X}_i}(\xi_i) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Lambda})}}\exp\left(-\frac{1}{2}(\boldsymbol{\xi}-\boldsymbol{\Gamma}\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}(\boldsymbol{\xi}-\boldsymbol{\Gamma}\boldsymbol{\mu})\right) \tag{5.47}$$

We can then do a multivariate transformation to get the pdf for $\mathbf{X} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\mathcal{X}} = \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\mathcal{X}}$. The Jacobian of the transformation is $\boldsymbol{\Gamma}^{\mathrm{T}}$, whose determinant is either 1 or -1, because

$$1 = \det\mathbf{1} = \det(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}}) = (\det\boldsymbol{\Gamma})^2 \tag{5.48}$$

(This is true for any orthogonal matrix.) This means $|\det\boldsymbol{\Gamma}| = 1$, and the pdf for $\mathbf{X}$ is simply

$$f_{\mathbf{X}}(\mathbf{x}) = f_{\boldsymbol{\mathcal{X}}}(\boldsymbol{\Gamma}\mathbf{x}) \tag{5.49}$$

If we also note that $\det\boldsymbol{\Lambda} = \prod_{i=1}^{n}\lambda_i = \det\boldsymbol{\sigma}^2$, we see that

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Lambda})}}\exp\left(-\frac{1}{2}\boldsymbol{\Gamma}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}(\mathbf{x}-\boldsymbol{\mu})\right) \\ &= \frac{1}{\sqrt{\det(2\pi\boldsymbol{\sigma}^2)}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}(\mathbf{x}-\boldsymbol{\mu})\right) \end{aligned} \tag{5.50}$$

But the combination $\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}$ is just the inverse of $\boldsymbol{\sigma}^2$, because

$$(\boldsymbol{\Lambda})^{-1} = (\boldsymbol{\Gamma}\boldsymbol{\sigma}^2\boldsymbol{\Gamma}^{\mathrm{T}})^{-1} = (\boldsymbol{\Gamma}^{\mathrm{T}})^{-1}\boldsymbol{\sigma}^{-2}\boldsymbol{\Gamma}^{-1} \tag{5.51}$$

so we find that, for arbitrary positive definite symmetric $\boldsymbol{\sigma}^2$,

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\sigma}^2)}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\sigma}^{-2}(\mathbf{x}-\boldsymbol{\mu})\right) \quad (5.52)$$

which is exactly the generalization we expected from the n-independent-random-variable case.

On the other hand, if $\boldsymbol{\sigma}^2$ has one or more negative eigenvalues, so that $\det\boldsymbol{\sigma}^2 = 0$, and $\boldsymbol{\sigma}^{-2}$ is not defined, that pdf won't make sense. In that case, consider the random variable $\mathcal{X}_i$ corresponding to the zero eigenvalue $\lambda_i = 0$. Its mgf is

$$E(e^{t_i\mathcal{X}_i}) = M_{\mathcal{X}_i}(t_i) = \exp\left(t_i\mathbf{v}_i^{\mathrm{T}}\boldsymbol{\mu} + \frac{1}{2}\lambda_i t_i^2\right) = \exp\left(t_i\mathbf{v}_i^{\mathrm{T}}\boldsymbol{\mu}\right) \quad (5.53)$$

but the only way that is possible is if $\mathcal{X}_i$ is always equal to $\mathbf{v}_i^{\mathrm{T}}\boldsymbol{\mu}$, i.e., $\mathcal{X}_i$ is actually a discrete random variable with pmf

$$P(\mathcal{X}_i = \xi_i) = \begin{cases} 1 & \xi_i = \mathbf{v}_i^{\mathrm{T}}\boldsymbol{\mu} \\ 0 & \text{otherwise} \end{cases} \quad (5.54)$$

This is the limit of a normal distribution as its variance goes to zero.

Returning to the case where $\boldsymbol{\sigma}^2$ is positive definite, so that each $\mathcal{X}_i$ is an independent $N(\mathbf{v}_i^{\mathrm{T}}\boldsymbol{\mu}, \lambda_i)$ random variable, we can construct the corresponding standard normal random variable

$$\mathcal{Z}_i = (\lambda_i)^{-1/2}(\mathcal{X}_i - \mathbf{v}_i^{\mathrm{T}}\boldsymbol{\mu}) = (\lambda_i)^{-1/2}\mathbf{v}_i^{\mathrm{T}}(X_i - \boldsymbol{\mu}) \quad (5.55)$$

which we could combine into a $N_n(\mathbf{0}, \mathbf{1})$ random vector

$$\boldsymbol{\mathcal{Z}} = \begin{pmatrix} \mathcal{Z}_1 \\ \mathcal{Z}_2 \\ \vdots \\ \mathcal{Z}_n \end{pmatrix} \quad (5.56)$$

However, it's actually more convenient to combine them into a different $N_n(\mathbf{0}, \mathbf{1})$ random vector

$$\boldsymbol{\mathcal{Z}} = \sum_{i=1}^{n} \mathbf{v}_i\mathcal{Z}_i = \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\mathcal{Z}} = \sum_{i=1}^{n} \mathbf{v}_i(\lambda_i)^{-1/2}\mathbf{v}_i^{\mathrm{T}}(X_i - \boldsymbol{\mu}) = \boldsymbol{\sigma}^{-1}(X_i - \boldsymbol{\mu}) \quad (5.57)$$

with pdf

$$f_{\mathbf{Z}}(\mathbf{z}) = \frac{1}{(2\pi)^{n/2}} e^{-\mathbf{z}^{\mathrm{T}}\mathbf{z}/2} \quad (5.58)$$

Thursday, October 19, 2017

5.5 Sample Mean and Sample Variance (Student's Theorem)

Recall the definition of a random sample $\{X_i | i = 1, \ldots, n\}$, which is n independent random variables drawn from the same distribution, and let $E[X_i] = \mu$ and $\mathrm{Cov}(X_i, X_j) = \delta_{ij}\sigma^2$. We showed in class that the sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (5.59)$$

has $E\left[\overline{X}\right] = \mu$ and $\mathrm{Var}(\overline{X}) = \sigma^2/n$. On the homework, you've shown that the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \quad (5.60)$$

has expectation value $E[S^2] = \sigma^2$. In addition,

$$\Sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \quad (5.61)$$

has expectation value $E[\Sigma^2] = \sigma^2$. These are all true for any underlying distribution.

Now, let's consider the case where the underlying distribution is Gaussian, i.e., $X_i \sim N(\mu, \sigma^2)$, which means that $\mathbf{X} \sim N_n(\mu\mathbf{e}, \sigma^2\mathbf{1}_{n\times n})$, where

$$\mathbf{1}_{n\times n} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \tag{5.62}$$

is the $n \times n$ identity matrix, and

$$\mathbf{e} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{5.63}$$

Since $\overline{X} = \frac{1}{n}\mathbf{e}^{\mathrm{T}}\mathbf{X}$ is a linear transformation of the Gaussian random vector $\mathbf{X}$ (using the $1 \times n$ matrix $\frac{1}{n}\mathbf{e}^{\mathrm{T}}$), we know that $\overline{X}$ is Gaussian-distributed, $\overline{X} \sim N(\mu, \sigma^2/n)$, and that

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \tag{5.64}$$

We also know that

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{n}{\sigma^2}\Sigma^2 \sim \chi^2(n) \tag{5.65}$$

so $\Sigma^2 \sim \text{Gamma}(\frac{n}{2}, \frac{2}{n}\sigma^2)$.

To understand the distribution of S^2 and its relationship to $\overline{X}$, we use several results which are known collectively as *Student's theorem*. ("Student" was the pseudonym of William S. Gosset, who found these results while working in the Guinness brewery in Dublin.) These results are:

1. $\overline{X}$ and S^2 are independent random variables
2. $$\sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma}\right)^2 = \frac{n-1}{\sigma^2}S^2 \sim \chi^2(n-1) \tag{5.66}$$
   which means that $S^2 \sim \text{Gamma}(\frac{n-1}{2}, \frac{2}{n-1}\sigma^2)$
3. The combination
   $$T = \frac{\overline{X} - \mu}{\sqrt{S^2/n}} \tag{5.67}$$
   obeys what's known as a Student's t-distribution with $n-1$ degrees of freedom. The t-distribution with ν degrees of freedom is the distribution obeyed by the combination
   $$\frac{Z}{\sqrt{W/\nu}} \tag{5.68}$$
   where Z and W are independent random variables with $Z \sim N(0,1)$ and $W \sim \chi^2(\nu)$. It is a straightforward but unenlightening calculation to show (by writing $f_{Z,W}(z,w) = f_Z(z)f_W(w)$, changing variables to write $f_{T,W}(t,w)$ and marginalizing to get $f_T(t) = \int_0^\infty f_{T,W}(t,w)\,dw$) that the pdf of the Student t-distribution is
   $$f_T(t;\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{5.69}$$
   If ν is very large, the t-distribution is approximately the same as the standard normal distribution. If $\nu = 1$, it's the Cauchy distribution. Note that the moment generating function for the t-distribution doesn't exist, since for $k \geq \nu$, the integral
   $$\int_{-\infty}^{\infty} t^k \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} dt \tag{5.70}$$
   diverges, as its integrand becomes, up to a constant, $t^{k-\nu-1}$ for large t.

The demonstration that Student's theorem is true is a nice application of the multivariate normal distribution. First, we can write

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\mathbf{e}^{\mathrm{T}}\mathbf{X} \tag{5.71}$$

and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}(\mathbf{X} - \mathbf{e}\overline{X})^{\mathrm{T}}(\mathbf{X} - \mathbf{e}\overline{X}) \tag{5.72}$$

Note that $\overline{X} = \frac{1}{n}\mathbf{e}^{\mathrm{T}}\mathbf{X}$ and

$$\mathbf{Y} = \mathbf{X} - \mathbf{e}\overline{X} = \left(\mathbf{1} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}}\right)\mathbf{X} \tag{5.73}$$

are both linear transformations of $\mathbf{X}$. We can combine them into an $n+1$-element random vector

$$\begin{aligned}\begin{pmatrix} \overline{X} \\ X_1 - \overline{X} \\ X_2 - \overline{X} \\ \vdots \\ X_n - \overline{X} \end{pmatrix} &= \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & 1-\frac{1}{n} \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{n}\mathbf{e}^{\mathrm{T}} \\ \mathbf{1} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}} \end{pmatrix}\mathbf{X} = \mathbf{A}\mathbf{X}\end{aligned} \tag{5.74}$$

where $\mathbf{A}$ is a $(n+1)\times n$ matrix. Since the original random vector $\mathbf{X}$ has $\boldsymbol{\mu} = \mu\mathbf{e}$ and, the transformed vector $\mathbf{A}\mathbf{X}$ has expectation value

$$E(\mathbf{A}\mathbf{X}) = \mathbf{A}\boldsymbol{\mu} = \begin{pmatrix} \frac{1}{n}\mathbf{e}^{\mathrm{T}} \\ \mathbf{1} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}} \end{pmatrix}\mu\mathbf{e} = \begin{pmatrix} \mu \\ \mathbf{0} \end{pmatrix} \tag{5.75}$$

where we've used the fact that

$$\mathbf{e}^{\mathrm{T}}\mathbf{e} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}\begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix} = n \tag{5.76}$$

and

$$(\mathbf{1} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}})\mathbf{e} = \mathbf{e} - \frac{n}{n}\mathbf{e} = \mathbf{0} \tag{5.77}$$

Since $\mathbf{X}$ has the variance-covariance matrix $\boldsymbol{\sigma}^2 = \sigma^2\mathbf{1}$, the transformed random vector $\mathbf{A}\mathbf{X}$ has variance-covariance matrix[11]

$$\begin{aligned}\mathrm{Cov}(\mathbf{A}\mathbf{X}) = \mathbf{A}\boldsymbol{\sigma}^2\mathbf{A}^{\mathrm{T}} &= \begin{pmatrix} \frac{1}{n}\mathbf{e}^{\mathrm{T}} \\ \mathbf{1} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}} \end{pmatrix}\sigma^2\mathbf{1}\begin{pmatrix} \frac{1}{n}\mathbf{e} & \mathbf{1} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}} \end{pmatrix} \\ = \sigma^2\begin{pmatrix} \frac{\mathbf{e}^{\mathrm{T}}\mathbf{e}}{n^2} & \mathbf{e}^{\mathrm{T}}\left(\mathbf{1} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}}\right) \\ \left(\mathbf{1} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}}\right)\mathbf{e} & \left(\mathbf{1} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}}\right)^2 \end{pmatrix} &= \sigma^2\begin{pmatrix} \frac{1}{n} & \mathbf{0}_{1\times n} \\ \mathbf{0}_{n\times 1} & \left(\mathbf{1} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}}\right) \end{pmatrix}\end{aligned} \tag{5.78}$$

Where we've used the fact (previously noted) that $\mathbf{1} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}}$ annihilated $\mathbf{e}$ and also that

$$(\mathbf{1} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}})^2 = \mathbf{1} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}} + \frac{1}{n^2}\mathbf{e}n\mathbf{e}^{\mathrm{T}} = \mathbf{1} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}} \tag{5.79}$$

We say that a matrix with this property is a *projection matrix*, in this case onto n-dimensional vectors perpendicular to $\mathbf{e}$.

Since the first row and column of $\mathrm{Cov}(\mathbf{A}\mathbf{X})$ are all zeros, except for the diagonal element, it means that the random variable $\overline{X}$, which is the first element of $\mathbf{A}\mathbf{X}$, and the random vector $\mathbf{Y}$ are independent, which in turn means $\overline{X}$ and $S^2 = \frac{1}{n-1}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$ are independent random variables, which is part 2 of Student's

[11] Remember that $\mathbf{A}$ is $(n+1)\times n$, $\boldsymbol{\sigma}^2$ is $n\times n$, and $\mathbf{A}^{\mathrm{T}}$ is $n\times(n-1)$, so $\mathbf{A}\boldsymbol{\sigma}^2\mathbf{A}^{\mathrm{T}}$ is $(n+1)\times(n+1)$.

theorem. We've also seen part 1, since $\overline{X}$ is a normal random variable whose mean is the first element of $E[\mathbf{AX}]$, i.e., μ and whose variance is the $(1,1)$ element of $\mathrm{Cov}(\mathbf{AX})$, which is σ^2/n.

To see that

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\mathbf{Y}^{\mathrm{T}}\mathbf{Y}}{\sigma^2} \tag{5.80}$$

is a chi-square random variable with $n-1$ degrees of freedom, consider the variance-covariance matrix

$$\mathrm{Cov}(\mathbf{Y}) = \left(\mathbf{1} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}}\right)\sigma^2 \tag{5.81}$$

This is an $n \times n$ matrix, but it is not invertable, so we can't do the usual trick to construct a $\chi^2(n)$ random variable. It's actually not too hard to work out the eigenvalue decomposition of this matrix; since $(\mathbf{1} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}})\mathbf{e} = \mathbf{0}$, we see that $\mathbf{e}$ is an eigenvector with eigenvalue zero. Because the matrix $\mathbf{1} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}}$ is a projector onto the $n-1$-dimensional subspace perpendicular to $\mathbf{e}$, we can choose any $n-1$ orthonormal vectors in that subspace, and they will be eigenvectors with eigenvalue σ^2. For example, take

$$\mathbf{v}_1 = \begin{pmatrix} 1/\sqrt{2} \\ -1\sqrt{2} \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 1/\sqrt{6} \\ 1/\sqrt{6} \\ -2\sqrt{6} \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \qquad \cdots$$
$$\mathbf{v}_{n-1} = \begin{pmatrix} 1/\sqrt{n(n-1)} \\ 1/\sqrt{n(n-1)} \\ 1/\sqrt{n(n-1)} \\ \vdots \\ 1/\sqrt{n(n-1)} \\ -(n-1)/\sqrt{n(n-1)} \end{pmatrix} \tag{5.82}$$

If we let $\mathbf{v}_n = \mathbf{e}/\sqrt{n}$, we have our complete set of orthonormal eigenvectors, with $\lambda_1 = \lambda_2 = \cdots = \lambda_{n-1} = \sigma^2$ and $\lambda_n = 0$. If we define $\mathcal{Y}_i = \mathbf{v}_i^{\mathrm{T}}\mathbf{Y}$, then $\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_{n-1}$ are $n-1$ independent normal random variables each with variance σ^2. (The last combination is trivial, since $\mathbf{v}_n^{\mathrm{T}}\mathbf{Y} = \frac{1}{\sqrt{n}}\mathbf{e}^{\mathrm{T}}\mathbf{Y} = \mathbf{0}$.) This means the following combination is a $\chi^2(n-1)$ random variable:

$$\begin{aligned}\sum_{i=1}^{n-1}\left(\frac{\mathcal{Y}_i}{\sigma}\right) &= \frac{\mathbf{Y}^{\mathrm{T}}\mathbf{v}_i\mathbf{v}_i^{\mathrm{T}}\mathbf{Y}}{\sigma^2} = \frac{1}{\sigma^2}\mathbf{Y}^{\mathrm{T}}\left(\sum_{i=1}^{n-1}\mathbf{v}_i\mathbf{v}_i^{\mathrm{T}}\right)\mathbf{Y} \\ &= \frac{1}{\sigma^2}\mathbf{Y}^{\mathrm{T}}\left(\mathbf{1} - \frac{1}{n}\mathbf{e}\mathbf{e}^{\mathrm{T}}\right)\mathbf{Y} = \frac{1}{\sigma^2}\mathbf{Y}^{\mathrm{T}}\mathbf{Y} = \frac{n-1}{\sigma^2}S^2\end{aligned} \tag{5.83}$$

This is point 2 of Student's theorem.

Finally, for point 3 we construct a t-distributed random variable.

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \tag{5.84}$$

is a standard normal random variable, and

$$\frac{n-1}{\sigma^2}S^2 \tag{5.85}$$

is a $\chi^2(n-1)$, we can take

$$T = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{n-1}{\sigma^2}S^2/(n-1)}} = \frac{\overline{X} - \mu}{\sqrt{S^2/n}} \tag{5.86}$$

and it will obey a t-distribution with $n-1$ degrees of freedom, which completes the proof of Student's theorem.

Tuesday, October 24, 2017

5.6 Reduced Chi-Squared

We know that if $\{X_i\}$ are iid Gaussian random variables, i.e., $X_i \sim N(\mu_i, \sigma_i^2)$, the combination

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2 \tag{5.87}$$

is a chi-squared random variable with n degrees of freedom, $\chi^2(n)$. This is often used as a goodness-of-fit test; if a model predicts that $X_i \sim N(\mu_i, \sigma_i^2)$, we can collect a realization of these data, call it $\{x_i\}$, and calculate

$$\sum_{i=1}^{n} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 \tag{5.88}$$

if it's much larger than we expect for a chi-square with this number of degrees of freedom, this casts doubt upon the model, since

$$P\left(\sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2 > w \right) = \int_w^\infty f_{\chi^2}(u; n)\, du \tag{5.89}$$

For example, if we have five degrees of freedom, we'd only expect to find a chi-square value greater than twenty 0.125% of the time:

```
In [1]: from scipy import stats

In [2]: stats.chi2.sf(20,5)
Out[2]: 0.0012497305630313482
```

We've seen that we can generalize this construction to the case of a multivariate Gaussian $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$; if the variance-covariance matrix is non-singular, so that $\boldsymbol{\sigma}^{-2}$ exists, we have

$$(\mathbf{X} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\sigma}^{-2} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2(n) \tag{5.90}$$

5.6.1 Minimizing χ^2 over parameters

An important generalization of the chi-squared construction comes when the model depends on some set of parameters $\boldsymbol{\lambda} \equiv \{\lambda_k | k = 1, \ldots, m\}$ where $m < n$. Typically, we assume that these parameters influence the means, not the variances, of the normal distribution, so that we have a family of models with $\boldsymbol{\mu}(\boldsymbol{\lambda})$. We can then construct

$$w(\mathbf{X}; \boldsymbol{\lambda}) = [\mathbf{X} - \boldsymbol{\mu}(\boldsymbol{\lambda})]^{\mathrm{T}} \boldsymbol{\sigma}^{-2} [\mathbf{X} - \boldsymbol{\mu}(\boldsymbol{\lambda})] \tag{5.91}$$

for a particular realization $\mathbf{x}$ of the data, the chi-square $w(\mathbf{x}; \boldsymbol{\lambda})$ will be a function of $\boldsymbol{\lambda}$, and we can find the $\boldsymbol{\lambda}$ which satisfies that by solving the set of m equations

$$0 = \frac{\partial w}{\partial \lambda_k} = -2 \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial \mu_i}{\partial \lambda_k} \left[\boldsymbol{\sigma}^{-2} \right]_{ij} [x_j - \mu_j(\boldsymbol{\lambda})] \tag{5.92}$$

for the m best-fit parameters $\{\widehat{\lambda}_k(\mathbf{x})\}$. The minimized chi-square value is then $w(\mathbf{x}; \widehat{\boldsymbol{\lambda}}(\mathbf{x}))$. In the simplifying case where the means are a linear function of the parameters, $\boldsymbol{\mu} = \mathbf{A}\boldsymbol{\lambda} + \mathbf{b}$, we can show that $w(\mathbf{X}; \widehat{\boldsymbol{\lambda}}(\mathbf{X})) \sim \chi^2(n - m)$, i.e., the constructed quantity still obeys a chi-square distribution, but the number of degrees of freedom is the number of data points minus the number of parameters that we fit. It is generally assumed that this is approximately true even if the dependence on the parameters is not linear.

To show that it works in the case where the predicted $\boldsymbol{\mu}$ is linear in the parameters, $\boldsymbol{\mu} = \mathbf{A}\boldsymbol{\lambda} + \mathbf{b}$, note that in that case $\frac{\partial \mu_i}{\partial \lambda_k} = A_{ik}$, which is independent of $\boldsymbol{\lambda}$, and we can write the condition for minimizing the chi-square as a matrix equation (dividing by -2 for convenience):

$$\mathbf{0} = \mathbf{A}^{\mathrm{T}} \boldsymbol{\sigma}^{-2} [\mathbf{x} - \boldsymbol{\mu}(\boldsymbol{\lambda})] = \mathbf{A}^{\mathrm{T}} \boldsymbol{\sigma}^{-2} [\mathbf{x} - \mathbf{A}\boldsymbol{\lambda} - \mathbf{b}] \tag{5.93}$$

or

$$\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A}\widehat{\boldsymbol{\lambda}}(\mathbf{x}) = \mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}(\mathbf{x}-\mathbf{b}) \tag{5.94}$$

Since $\mathbf{A}$ is an $n \times m$ matrix (it maps m parameters into n data values), the combination $\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A}$ is an $m \times m$ matrix. We can assume it's invertible, since if it's not we don't actually need all m parameters to describe the model. Thus the best-fit parameters are

$$\widehat{\boldsymbol{\lambda}}(\mathbf{x}) = (\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}(\mathbf{x}-\mathbf{b}) \tag{5.95}$$

and the corresponding expectation values for the data are

$$\boldsymbol{\mu}\big(\widehat{\boldsymbol{\lambda}}(\mathbf{x})\big) = \mathbf{A}(\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}(\mathbf{x}-\mathbf{b}) \tag{5.96}$$

Note that we may be tempted to simplify this with something involving the matrix inverse of $\mathbf{A}$, but that doesn't exist because $\mathbf{A}$ is *not* a square matrix. What we can do, though, is define the matrix

$$\mathbf{P} = \boldsymbol{\sigma}^{-1}\mathbf{A}\left[\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A}\right]^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-1} \tag{5.97}$$

so that

$$\boldsymbol{\mu}\big(\hat{\boldsymbol{\lambda}}(\mathbf{x})\big) = \boldsymbol{\sigma}\mathbf{P}\boldsymbol{\sigma}^{-1}\mathbf{x} \tag{5.98}$$

In fact, since

$$\mathbf{P}\boldsymbol{\sigma}^{-1}\mathbf{A} = \boldsymbol{\sigma}^{-1}\mathbf{A} \tag{5.99}$$

we see that for *any* $\boldsymbol{\lambda}$,

$$\mathbf{P}\boldsymbol{\sigma}^{-1}\boldsymbol{\mu}(\boldsymbol{\lambda}) = \mathbf{P}\boldsymbol{\sigma}^{-1}\mathbf{A}\boldsymbol{\lambda} = \boldsymbol{\sigma}^{-1}\mathbf{A}\boldsymbol{\lambda} = \boldsymbol{\sigma}^{-1}\boldsymbol{\mu}(\boldsymbol{\lambda}) \tag{5.100}$$

The matrix $\mathbf{P}$ is not only symmetric, but it's a projection matrix, since

$$\begin{aligned}\mathbf{P}\,\mathbf{P} &= \boldsymbol{\sigma}^{-1}\mathbf{A}\left[\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A}\right]^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A}\left[\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A}\right]^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-1} \\ &= \boldsymbol{\sigma}^{-1}\mathbf{A}\left[\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A}\right]^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-1} = \mathbf{P}\end{aligned} \tag{5.101}$$

This $n \times n$ matrix is a projector onto an m-dimensional subspace, since

$$\begin{aligned}\operatorname{tr}\mathbf{P} &= \operatorname{tr}\left(\boldsymbol{\sigma}^{-1}\mathbf{A}\left[\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A}\right]^{-1}\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-1}\right) \\ &= \operatorname{tr}\left(\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A}\left[\mathbf{A}^{\mathrm{T}}\boldsymbol{\sigma}^{-2}\mathbf{A}\right]^{-1}\right) = \operatorname{tr}\mathbf{1}_{m\times m} = m\end{aligned} \tag{5.102}$$

That means that it has m eigenvectors with unit eigenvalue and $n-m$ eigenvectors with zero eigenvalue. Let $\{\mathbf{u}_i | i = 1, \ldots, m\}$ be a set of m orthonormal eigenvectors with unit eigenvalue and $\{\mathbf{u}_i | i = m+1, \ldots, n\}$ be a set of $n-m$ orthonormal eigenvectors with zero eigenvalue, so that

$$\mathbf{P} = \sum_{i=1}^{m} \mathbf{u}_i\,\mathbf{u}_i^{\mathrm{T}} \tag{5.103}$$

and

$$\mathbf{1}_{n\times n} - \mathbf{P} = \sum_{i=m+1}^{n} \mathbf{u}_i\,\mathbf{u}_i^{\mathrm{T}} \tag{5.104}$$

are projection operators onto the two orthogonal subspaces.

Now, let $\boldsymbol{\lambda}_{\mathrm{true}}$ be the column vector of true, unknown parameters. Then

$$\mathbf{Z} = \boldsymbol{\sigma}^{-1}\left[\mathbf{X} - \boldsymbol{\mu}(\boldsymbol{\lambda}_{\mathrm{true}})\right] \tag{5.105}$$

is a vector of n independent standard normal random variables. Since $\{\mathbf{u}_i | i = 1, \ldots, n\}$ is an orthonormal basis, we can also construct independent standard normal random variables $\{\mathcal{Z}_i | i = 1, \ldots, n\}$ where

$$\mathcal{Z}_i = \mathbf{u}_i^{\mathrm{T}}\mathbf{Z} = \mathbf{u}_i^{\mathrm{T}}\boldsymbol{\sigma}^{-1}\left(\mathbf{X} - \boldsymbol{\mu}(\boldsymbol{\lambda}_{\mathrm{true}})\right) \tag{5.106}$$

If we take the sum of the squares of the last $n-m$ of these random variables, that will obey a chi-squared distribution with

$n-m$ degrees of freedom:

$$W_{n-m} = \sum_{i=m+1}^{n} [\mathcal{Z}_i]^2 = [\mathbf{X} - \boldsymbol{\mu}(\boldsymbol{\lambda}_{\text{true}})]^{\mathrm{T}} \boldsymbol{\sigma}^{-1} \left(\sum_{i=m+1}^{n} \mathbf{u}_i \, \mathbf{u}_i^{\mathrm{T}} \right) \boldsymbol{\sigma}^{-1} [\mathbf{X} - \boldsymbol{\mu}(\boldsymbol{\lambda}_{\text{true}})]$$
$$= [\mathbf{X} - \boldsymbol{\mu}(\boldsymbol{\lambda}_{\text{true}})]^{\mathrm{T}} \boldsymbol{\sigma}^{-1} (\mathbf{1}_{n\times n} - \mathbf{P}) \, \boldsymbol{\sigma}^{-1} [\mathbf{X} - \boldsymbol{\mu}(\boldsymbol{\lambda}_{\text{true}})] \tag{5.107}$$

But by (5.100) and (5.98),

$$(\mathbf{1}_{n\times n} - \mathbf{P}) \, \boldsymbol{\sigma}^{-1} [\mathbf{X} - \boldsymbol{\mu}(\boldsymbol{\lambda}_{\text{true}})] = (\mathbf{1}_{n\times n} - \mathbf{P}) \, \boldsymbol{\sigma}^{-1} \mathbf{X} = \boldsymbol{\sigma}^{-1} \mathbf{X} - \mathbf{P} \boldsymbol{\sigma}^{-1} \mathbf{X}$$
$$= \boldsymbol{\sigma}^{-1} \left[\mathbf{X} - \mu \left(\hat{\boldsymbol{\lambda}}(\mathbf{X}) \right) \right] \tag{5.108}$$

so

$$W_{n-m} = \left[\mathbf{X} - \mu \left(\hat{\boldsymbol{\lambda}}(\mathbf{X}) \right) \right]^{\mathrm{T}} \sigma^{-2} \left[\mathbf{X} - \mu \left(\hat{\boldsymbol{\lambda}}(\mathbf{X}) \right) \right] = W_{\text{red}} \tag{5.109}$$

which means this quantity we've constructed, which is χ^2-distributed with $n-m$ degrees of freedom (in the case where the (**??**) modelled expectation values are linear in the parameters, i.e., $\boldsymbol{\mu} = \mathbf{A}\boldsymbol{\lambda} + \mathbf{b}$), is indeed the residual chi-squared $w(\mathbf{X}; \widehat{\boldsymbol{\lambda}}(\mathbf{X}))$.