$X_i^2, i=1,2,\cdots,n$ are iid χ_1^2 (the only thing that needs to be shown is that the squared variables are independent). So a χ_n^2 variable is the sum of the squares of n Gaussian variables. We'll return to this point next time.

Note that we could expand any of the families defined in this section to include the location-scale families for each of its members. We just get a bigger family with more members.

4. Infinitely-Divisible Distributions

Why don't we use the Gaussian as a base distribution and generate the family of distributions of n independent Gaussians? As you probably already know, the sum of two independent Gaussians with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 is just a Gaussian with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$. It's instructive to prove this by means of mgfs.

Derivation: Compute the mgf of the unit Gaussian (with mean 0, variance 1). Use of the result of the Derivation above to show that the mgf of the Gaussian X with mean μ and standard deviation σ is

$$M(\theta) = \mathbf{E}[e^{\theta X}] = e^{\mu\theta + \sigma^2\theta^2/2}.$$

Derivation: Write down the mgf of the sum of two unit Gaussians. Write down the mgf of a Gaussian with mean 0, variance 2. Compare.

Derivation: Prove that if X and Y are independent random variables, then the mean of $X+Y$ is the sum of the means of X and Y (Hint: the product rule. Check that $M(0) = 1$ for any mgf.

Derivation: Let X be a random variables with mgf $M_X(\theta)$. Let $Y = (X-a)/b$. What is the mgf of Y (expressed in terms of the mgf of X)?

3. Adding Independent Random Variables: Families

We can create *families* or distributions by (a) picking a base distribution, and (b) adding together iid variables from the base distribution. The general notation we will use for families will be the family name followed by the parameters of the family in parentheses. For example, Binomial(n,p). The Bernoulli(p) as base distribution leads to the Binomial (n,p) family with $n = 1,2,3,4 \cdots$. When n is 1, the resulting distribution is just the Bernoulli and so the Bernoulli (p) is just Binomial(1,p). This family isn't a location-scale family but a family created by addition.

If we use the exponential (λ) family with pdf,

$$f(x;1,\lambda) = \lambda e^{-\lambda x}, \quad x > 0,$$

as a base distribution, we get the *Gamma family* Gamma(n, λ) whose pdf's are,

$$f(x;n,\lambda) = \Gamma_n x^{n-1} e^{-\lambda x}, \quad x > 0,$$

where Γ_n is a constant that ensures the pdf integrated from $-\infty$ to ∞ is 1.

If we return to the first example of the first lecture, we find a simple model of reaction time where we assumed that overall reaction time was the sum of n iid exponential variables all with the same rate parameter λ. The distribution of the overall reaction time, by assumption, is some one of the distributions $f(x;n,\lambda)$. If we don't know λ and n, though, we don't know which one. All parametric fitting techniques begin with the assumption that the unknown distribution that described the data lies in some parameterized family and they then try to estimate the family member or members that correspond to the data. We'll consider this *estimation* problem a bit later in the course.

The last family generated by addition is the χ^2_n ('kh-eye square') family. It's base distribution is the χ^2_1 distribution whose pdf we derived last time. Recall that the χ^2_1 is the pdf of $Y=X^2$ where X is Gaussian with mean $\mu=0$ and variance $\sigma^2=1$. A χ^2_2 variable is the sum of two independent χ^2_1 variables. It's not too difficult to show that if $X_i, i=1,2,\cdots,n$ are iid Gaussian with unit means and variances 1, then

moments of X to be

$$\mu'_m = (1-p)^2 0 + 2p(1-p)1 + p^2 2^m = 2p(1-p)+2^m p^2$$

or we can compute them from the mgf,

$$M_X(\theta) = (1-p)^2 e^{0\theta} + 2p(1-p)e^{1\theta} + p^2 e^{2\theta}$$

$$= (1-p)^2 + 2p(1-p)e^{\theta} + p^2 e^{2\theta}.$$

Either way, the computation is tedious and, for a Binomial(n) variable would become rapidly more tedious as n increased.

There's a short way to do this though, based on the following,

Fact: Let X and Y be independent random variables with respective mgf's $M_X(\theta)$ and $M_Y(\theta)$. Let $Z = X + Y$ have mgf $M_Z(\theta)$. Then,

$$M_Z(\theta) = M_X(\theta)M_Y(\theta).$$

The above theorem applies for values of θ where all of the mgf's exists. We can check it immediately by seeing whether The last equation is the square of the earlier. It is, and in general, the mgf of the Binomial(n) is,

$$M_{X_n}(\theta) = (1-p + pe^{\theta})^n.$$

and it is easy to take the mth derivative with respect to θ of the right hand side. In particular, when m=1,

$$\frac{dM_X(0)}{d\theta} = np(1-p + pe^{0})^n = np.$$

So we can readily compute the mgf of a sum of independent random variables and from that, all of the moments of the sum (if the mgf's exist). We can do more. The mgf of a random variable, if it exists in some interval around 0, determines the random variable. Put more precisely,

Fact: If X and Y are random variables with the same mgf $M(\theta)$ in some open interval around 0, the X and Y have the same cdf.

The most useful application of this theorem is when we want to show that one random variable is identical to another. Previously we could show they have the same cdf's or pdf's. Now we can also show that they have the same mgf's in some interval around 0.

Derivation: Let X and Y be independent Gaussian variables with mean 0 and variance (σ^2) equal to 2. Compute the mgf's of X and Y (obviously they are the same) and the mgf of a Gaussian with mean 0 and variance 2.

2. Moment-Generating Functions

Moment generating functions were introduced when we first discussed expectation. For a continuous random variable X with pdf $f(x)$, the *moment generating function (mgf)* is,

$$M_X(\theta) = \mathbf{E}e^{\theta X} = \int_{-\infty}^{\infty} e^{\theta x} f(x)\, dx.$$

The exponential is a rapidly increasing function and the integral may be infinite or fail to converge for some values of θ. We say that the mgf 'does not exist' for these values. The moment generating function may exist for some values of θ and not others. It always exists for $\theta = 0$ and has value 1 (why?).

Suppose that it exists on some open interval surrounding 0. If we then take the mth derivative of $M_X(\theta)$ with respect to θ under the integral sign, we have

$$\frac{d^m M_X(\theta)}{d\theta^m} = \int_{-\infty}^{\infty} x^n \frac{d^m e^{\theta x}}{d\theta^m} f(x)\, dx.$$

and, if we set θ to 0,

$$\frac{d^m M_X(0)}{d\theta^m} = \int_{-\infty}^{\infty} x^m f(x)\, dx = \mu'_m,$$

the mth raw moment. That's why we call it the moment generating function.

You know how to define the mgf for a discrete random variable and you computed the mgf's of some random variables in an earlier derivation. For the Bernoulli random variable B whose outcomes are 0 with probability $1-p$ and 1 with probability p, the mgf is,

$$M_B(\theta) = (1-p)e^{\theta 0} + pe^{\theta 1} = 1-p+pe^{\theta}.$$

The mth raw moment is

$$\frac{d^m M_B(0)}{d\theta^m} = p,$$

that is, all the raw moments are equal to p.

Consider now the variable $X = B_1 + B_2$, where the B_i are independent, identically distributed Bernoulli variables with parameter p. This variable X is the Binomial(2) variable with parameter p. The Binomial(n) variable with parameter p is gotten by adding together n iid Bernoulli variables with probability p. We can work out the distribution of X from elementary probability. The possible outcomes are 0, 1, and 2. They occur with probability $(1-p)^2$, $2p(1-p)$ and p^2. We can directly compute the raw

orders of authorships. They typically define distributional families as location and scale parameters even when the distribution under discussion is rarely used in that form. The *unit exponential* will be 'officially' defined as,

$$f(x) = e^{-x} \quad x > 0; \quad 0 \text{ otherwise}$$

and it might not be obvious that the other exponentials in the exponential family form part of a location scale family. We define a location-scale exponential family as described above. If we stare at its *rate*:

$$f(x) = \lambda e^{-\lambda x} \quad x > 0; \quad 0 \text{ otherwise}$$

and let $a=0, b=1/\lambda$, then we can see the location-scale form emerge. In practice, though, the eponential distribution is commonly given one parameter, λ just as we've been writing it. It's good to remember that the exponetial is a 'scale family' and is easily generalized to a location-scale family.

Note that parameters can be dimensional: they can be defined in terms of physical units. The rate parameter of the exponential can be in units of inverse time seconds^{-1} and the variable x in the cdf or pdf should be in units of time seconds. Note that the location and scale parameters must be in the same physical units as the x variable, and that the transformed variable $z=(x-\mu)/\sigma$ is then dimensionless. In the Gaussian case, this variable is evidently a 'z-score' and it is distributed as a unit Gaussian no matter what the values of μ and $\sigma>0$ were. We'll prove this in class.

We can generalize the concept of z-score to any location-scale family. Note that the transformed variable for the exponential family has the distribution of the unit exponential.

Note also that a location-scale family is closed under linear transformation of units (e.g. changing from *feet* to *meters*). *This is a very good thing.*

The last distribution we will work with in this section is the *Weibull distribution*, introduced last time, which has two parameters α and β:

$$F(x) = 1 - e^{-(\frac{x}{\alpha})^{\beta}} \quad x > 0$$

The Weibull is not a location-scale family, but, by transforming the x axis, it can be made into one:

In class, we will calculate the pdf of the Weibull. And show that if the pdf is plotted on semilog axes, the Weibull distribution is tranformed into a location-scale family with parameters $\mu = \log_{10}(\alpha)$ and $\sigma = 1/\beta$.

Remember that, when $\beta=1$, the Weibull is just the exponential.

resulting expected value is termed the *moment generating function* of X.

We've seen the Gaussian family in previous lectures. The *unit Gaussian* has the pdf, pdf is:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

The unit Gaussian can used to generate the *Gaussian family*. Recall a family of distributions is a set of distributions that share a common form except for a small number of parameters. The Gaussian family has two parameters μ and σ, and has pdf,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}.$$

The Gaussian has been parameterized many ways in the last three hundred years, and the most useful parametrizations have gradually won out. Of course you know that μ and σ are the mean and the square root of the variance of the Gaussian (referrred to as the standard deviation). For our purposes, though, μ and σ can better be thought of as a *location* parameter and a *scale* parameter respectively. The next two results (proved in class) sum it all up:

Derivation: Suppose that $F(x)$ is a cumulative distribution function. For any real numbers a and $b > 0$, show that $G_{a,b}(x) = F((x-a)/b)$ is also a cdf.

Derivation: Let $\Phi(x)$ denote the cdf of the unit Gaussian whose pdf is given above. Then the Gaussian pdf's correspond to cdf's of the form $\Phi((x-\mu)/\sigma)$. (The tricky part here is only that the cdf of the unit Gaussian must be written as a formal integral, then transformed as in last Derivation, and then differentiated to get the pdf's of Otherwise you will be missing a factor of $\frac{1}{\sigma}$.)

Derivation: Suppose that $F(x)$ is a distribution function with pdf $f(x)$. Define $G_{a,b}(x)$ as in the first these three Derivations. What is the pdf of $G_{a,b}(x)$ in terms of the pdf of $F(x)$? What happens to the graph of the pdf as the location and the scale parameters change?

A bit of ambiguous terminology: I should refer to $G_{a,b}$ or, more generally, any distributional family as a *distributional family*. Only when a,b take on fixed values have we specified a *distribution.* Noone observes this distinction. Everyone refers to $G_{a,b}$ as either a *distribution* or a *distributional family*, depending on what they want to emphasize.

Thus, given any distribution function, we can talk about its *location-scale family*. There is a superb 'encyclopedia' of distributional families in five volumes by Johnson, Kotz, and Balakrishnan in various

Lecture 5: More Distributional Families

We ought to remark ... that the principal element of the analysis of probabilities is an exponential integral that has presented itself in very different mathematical theories.... This ... function is connected with general physics. It is required in order to characterize the motion of light through a gaseous milieu. We have discovered in recent years that it also represents the diffusion of heat in the interior of solid substances. Finally, it determines the probability of errors of the measures and mean results of numerous observations; it reappears in the questions of insurance and in all difficult applications of the the science of probabilities.

Thus, mathematical analysis unites the most diverse effects, and discovers in them common properties. Its object has nothing of the contingent, nothing of the fortuitous. Imprinted in all of nature, it is a preexistent element of universal order. This science has necessary relations with all physical causes, and with most of the combinations of spirit. – J.-B. J. Fourier

1. Location-Scale Families

The expectation of a function $h(X)$ of a continuous random variable with pdf $f(x)$ was defined earlier as,

$$\mathbf{E}(h(X)) = \int_{-\infty}^{\infty} h(x)f(x)dx,$$

This is the usual definition of expectation for a continuous r.v., but we've expanded it to the cases of discrete and mixed r.v.s. When we interpret the formula above for a discrete random variable with probabilities p_i of taking on values $x_i, i = 1,2,3...$ the pdf turns into,

$$\mathbf{E}(h(X)) = \sum_{i=1}^{\infty} h(x_i)p_i,$$

the usual definition. You will occasionally see the following notation for expectation:

$$\mathbf{E}(h(X)) = \int_{-\infty}^{\infty} h(x)dF(x),$$

where F is the cdf of the rv. You will encounter this formula in your reading and, if you think of $dF(x)$ as shorthand for $f(x)dx$ (where $f(x)$ is the pdf) then this formula is the same as the definition we began with above. We've defined pdf so that discrete and mixed rv's have pdf's and we need only work with one kind of formula. In practice, when you run across a discrete or mixed pdf, it is usually easiest to translate the discrete part into the usual summation notation you've seen for probability.

When $h(X) = X$, the resulting expected value is called the *mean* and is often denoted μ_X. When $h(X) = (X-\mu_X)^2$, the resulting expected value is termed the *variance* of X. When $h(X) = \exp(\theta X)$, the

testing 123