Standard Distributions PhL 5.30.03
updated 8.25.03

The book from Randy Sylvester is Proakis, *Digital Communications* , 3rd Ed, 1995. His Chapter 2 is a 60 page survey of fancy statistics stuff. I have two-up copy of this stuff but it is hard to read since print is very small, but the information is there and it is valuable information at that!

This section has no copy to go with it.

2-1 Probability. The sample space S; joint and conditional probabilities; independence.

2-2-1. The notion of $P(X \le x)$ defines a cdf, but we are more usually interested in the pdf which is the derivative of the cdf. For discrete stuff, he likes to still think of a continuous pdf using a delta function, as in the following:

$$p(x) = \sum_{i=1}^{n} P(x = x_i)\, \delta(x - x_i)$$

Extends to functions of two random variables X_1 and X_2. Joint, conditional, Bayes.

Here is where the copy starts:

2-1-2. Functions of a Random Variable. This is what we care about.

"Suppose you know all about the distribution of X, and you want to know about some Y = f(X). "

Example is $Y = f(X) = aX + b$. The idea is to work with the CDF first, sort of transfer over to the new variable, then take a derivative to get the PDF.

Second example is $Y = aX^3 + b$. Same idea.

Third example is $Y = aX^2 + b$, a very important example. This is 2 to 1, not 1 to 1, so a little messier, but the resultant pdf is in 2-1-47 which gets used later.

Now the key result is really (2-1-49) where he uses Y = g(X) instead of my f(X). He does not derive this general result, and the method he shows earlier for the quadratic example does not generalize very well to the general case! However, here is my own derivation of the general case:

$$Q(y)\,|dy| = P(x_1)\,|\,dx_1\,| + P(x_2)\,|\,dx_2\,| + \text{etc.}$$

Here, do NOT think of the x_i as multiple random variables, that lies in the next section. Instead, think of $x_i(y)$ as all the solutions of the equation y = g(x). Think of the case $y = x^2$ as the canonical example. There are two strips under the X distribution that map into the same strip in the Y distribution, see handwritten notes with this Word doc titled **Example #1**. The absolute values are because direction does not matter of the strip, it is the area of the strips that matter, all probabilities are positive definite. You see in the canonical case that $x_1(y) = +\sqrt{y}$ and $x_2(y) = -\sqrt{y}$ so both dx1 and dx2 will be proportional to dy. In the

above, P(x) is the distribution for X (perhaps Gaussian, perhaps not), and Q(y) is what we are looking for, the distribution for y = g(x). Divide both sides by dy and just rewrite to get:

$$Q(y) = P(x1) / \left|dy/dx\right|^{x=x1} + P(x2) / \left|dy/dx\right|^{x=x2} + \text{etc}$$

But y = g(x), so we really have this: [where x1 = x1(y), etc]

$$Q(y) = P(x1) / \left|g'(x)\right|^{x=x1} + P(x2) / \left|g'(x)\right|^{x=x2} + \text{etc}$$

and this is what 2-1-49 says. This is a far-reaching result!!!

In Example #2, I do the thing explicitly for $y = x^2$ in the special case that P(x) = Gaussian. In this case, the two P(x) are equal, and the derivs are equal, so both terms are equal and erase the 2 that comes from the derivative of x^2. If you ever get confused, come back to this example! This shows what a squarer does to a PDF if you insert Gaussian signals into it!

Clearly, once you find your new Q(y), you will need to do integrals to compute your new moments. That is not done at this point in the text.

He then goes on to consider the general case Y = Y(X) where each is multiple variables. Be careful, because this section only applies when you are basically converting from n X variables to n Y variables. You find that the PDF's are related by the Jacobian, as one would expect, so nothing too fancy here. But you could use this perhaps in going from x,y coordinates to r,θ coordinates in some noise situation. If the relation between Y and X is Y = AX where A is an invertible matrix, then J = 1/det(A).

2-1-3. Averages. This is an unbelievably clear well-written section that tells you everything you need to know about moments, central moments, correlation matrix, covariance matrix, joint moments, you name it. Everything is completely clear and concise!

Don't overlook result (2-1-63). This relates again to our issue of y = g(x) and you know about x. To get moments of y, you could on the one hand go compute PDF(y) which is our Q(y) above. But another way is to simply put g(x) inside the integral as shown with P(x).

Characteristic functions . Very very important stuff! The definition is in 2-1-71 where we see first that the $\psi(j\upsilon)$ is the Fourier Transform of the PDF(x), with a certain choice on the 2π constant location. But also notice that you can interpret $\psi(j\upsilon)$ as $< \exp(j\upsilon x) > = E(\exp(j\upsilon x))$ in the normal notation.

Derivatives bring down powers of x from the exponent and then you evaluate at parameter $\upsilon = 0$ and you get various moments of interest. Variable υ the Fourier conjugate of the PDF(x) variable x.

A very powerful result now: Suppose you know the CF for X, and you want the CF for a sum of X's. Consider y = x1 + x2. To get $\psi_y(j\upsilon)$, you need $E(\exp(j\upsilon y))$, notice the y in there. But now insert y and you get $E(\exp(j\upsilon x1)\exp(j\upsilon x2))$. If you assume the joint PDF is independent, things factorize and you find that $\psi_y(j\upsilon) = \psi_{x1}(j\upsilon)\psi_{x2}(j\upsilon)...$ a very simple result. But we know why this is. It is because in the x domain when you do y = x1 + x2, you are doing convolution (as in my paper on adding random voltages), so in the frequency domain it has to be diagonal like this. The diagonalization applies of course even if each X has a different distribution! I think I could use this idea to combine that binomial distribution stuff with the Gaussian noise in our rate discovery section, but I did not do it. The idea being that these are different distributions and you are adding the variables. I sort of assumed Gaussian for both with large enough R.

He finishes up with multidimensional CF's which I ignore for now.

2-1-4. Some useful distributions.

Binomial. Here he adds some X_i where each is 0 or 1. Does not use CF, just does the traditional derivation and we get statement of the mean, variances.

Uniform. A constant distribution, not too fancy.

Gaussian (**Normal**). This is our big friend, he has it properly normalized of course. The CDF is the error function and he gets to mention the **Q function** here as well. Of course the CF of a Gaussian being the FT is another Gaussian. He uses gaussian lower case! Then we get our famous result: if you add up a set of X_i each with a gaussian, as in white noise, you can must multiply the CF's and you get the rules that you just add means and add variances.

Chi-Square. Now getting into new stuff for me. There are two cases. In both cases, you are dealing with the situation $y = x^2$ and x is assumed to have a gaussian characteristic.

(a) central

Suppose X is a gaussian *with zero mean*. Suppose $Y = X^2$. What is the pdf for Y? We did this already in the example above and the result is quoted:

$p(y) = 1/(\sqrt{2\pi y}\ \sigma) \exp(-y/2\sigma^2)$ // See **Example #2** mentioned above

In this case we used that example 2-1-47 to get the result. This time we do the pdf first. The cdf is then an integral I have often run into which has $1/\sqrt{x}$ with $\exp(-x/2\sigma^2)$. This is not even an erf(x) thing. This is the finite integral. We can do the infinite integral to get the CF, however. as in 2-1-107.

So where are we? If you have $Y = X^2$, the pdf for Y is a Chi-square as above.

Next, add up a bunch of X_i^2 terms (all of zero mean and the same σ) , as you might get going through a multiplier. Just raise the CF to the nth power and do a reverse FT to get the 2-1-110 result which is a power of y times an expo. This is the **Chi-Square with n degrees of freedom.** Very good. If you add just two terms, you get n=2 case and the power goes away and you have just the $\exp(-y/2\sigma^2)$ part. The cdf is a mess, related to incomplete Gamma.

For the general case of adding n squared terms, author gives these results: Char Function, PDF, the CDF as an integral which can be done in some cases. The first two moments are also stated, and then the variance is given. VERY useful stuff to me to have ready to use like this.

(b) non-central

Now we start all over again where X is a gaussian with a non-zero mean! The pdf is now an expo times a cosh as in 2-1-115. Now add a series of X_i^2 where we allow each one to have a different mean, but the same σ. Compute the CF as a product of the underlying CFs, do the FT, and you end up with the PDF being a Bessel I function with parameter s^2 that is the sum of all the means squared. This result is the non-central Chi-squared. You can relate it to certain $Q_m(a,b)$ functions of Marcum. Again we get all these results stated: Char Function, PDF, CDF (no closed form), and again we get the two lowest moments and the variance.

Summary. Chi-Square is a large subject, used to see whether data fit a presumed distribution, for example. But here we just think of it as the PDF for the sum of squares of voltages. It the voltages squared have zero mean Gaussians, then you have the simpler central Chi case. Otherwise the non-central.

Rayleigh. (starts with zero mean Gaussian, related to central Chi square) This is what you get when you do the Chi-Square sum of squares, but in the end you take a square root! Proakis gets the pdf just by changing variables in 2-1-128. The CDF is trivial, so are all the moments. As before, he then moves on to the general case of n terms. The Char Function is a mess which turns out to have a ${}_1F_1$ component with a negative argument. Author first does simple case being the sum of just two terms to get all this stuff. Simple because the Chi-Square has no powers for n=2, just the expo.

Now generalize to sum of n terms, the usual next step. The result is a pdf with powers and expo.

Rice. (starts with non-zero mean Gaussian, related to non-central Chi square). The PDF is now a Bessel I function even in the general case as in 2-1-143. The square root of the sum of X_i^2 variable is called R. The CDF is state. Finally, the moments are ALL given as confluent hypergeometrics!

I think when you have a constellation point perturbed by noise in the two dimensions, you end up with a Rice situation for the magnitude of the vector, and that is why fading and multipath get associated with Rice and Rayleigh.

Application. Now I am going to apply the above to a specific situation where I am inserting statistical probes into my Xilinx RD design. I have the probe add up the squares of N numbers and then divide by N, and then take the square root. I know that in the limit that $N \to \infty$, this is the square root of the variance from zero. Since I am interested in the no-correlation case, the mean is zero, so my probe is trying to compute the square root of the variance, which is just the SD. But I don't have time to wait for $N=\infty$, so I want to know what the distribution looks like of finite SD readings about the true SD, ie, the variance of the standard deviation.

See attached Example #3 where this problem is first attacked. I have left out the divide by N part at this point. (N = n). We have a simple Rayleigh situation since means are 0 and since the data stream is assumed Gaussian for large R, even for no noise. I find these results

$$\langle r \rangle = \sqrt{2}\,\sigma\,\alpha_n \qquad \text{where } \alpha_n = \Gamma([n+1]/2) / \Gamma(n/2)$$

$$\langle r^2 \rangle = n\,\sigma^2$$

$$\text{variance} = \langle r^2 \rangle - \langle r \rangle^2 = n\,\sigma^2 \left[1 - \alpha_n^2/(n/2) \right]$$

$$\text{std} = \sqrt{n}\,\sigma \sqrt{1 - \alpha_n^2/(n/2)}$$

Now, Maple shows us that the quantity $\alpha_n^2/(n/2) \to 1$ from below, and gets quite close at n = 5 or 10, in fact. I show from an AS reference that $\alpha_n^2 \to n/2$ for large n.

Now we want to go back and add in our missing factor of $1/\sqrt{n}$ that should be present in our expression at the start of Example #3 for R. On page 2 of that example, we show what an overall constant of 2 does to things. The PDF is changed as shown [this comes from my multiplying random numbers paper, one of the examples]. The net result is that all moments get scaled by 2^k for kth moment. That means that our STD result above will get scaled in our case by $1/\sqrt{n}$, so we will then have

$$\text{STD} = \sigma \sqrt{1 - \alpha_n^2/(n/2)}$$

Again, this describes the distance of our "probe reading" from the actual standard deviation of the data. It is the standard deviation of the standard deviation. If $n \to \infty$, then STD = 0, and we should be right on the money for whatever the theory says. For example, at the R-chunk outputs, we expect for an N case to have this result:

$\sigma = \sqrt{R}$ // in theory units. We have R = 64, so expect $\sigma = 8$.

When we convert this to hardware units with our 8-bit bus, we get $\sigma = 8*127 = 1016$. This is what we would expect to see with $n \to \infty$. And now we have an explicit formula for the expected STD away from this amount for finite m. Here is an excel table,

n	alpha(n)	sqrt(n/2)	B/C	std1	std	3 sig low	3 sig high
10	2.180949	2.24	0.97535	0.220663	224	343	1689
20	3.123011	3.16	0.98758	0.157099	160	537	1495
30	3.840849	3.87	0.99170	0.128552	131	624	1408
40	4.444275	4.47	0.99377	0.111449	113	676	1356
50	4.975064	5.00	0.99501	0.099747	101	712	1320
60	5.454452	5.48	0.99584	0.091095	93	738	1294
70	5.894989	5.92	0.99644	0.084363	86	759	1273
80	6.304822	6.32	0.99688	0.078933	80	775	1257
90	**6.689596**	**6.71**	**0.99723**	**0.074431**	**76**	**789**	**1243**
100	7.053413	7.07	0.99750	0.070622	72	801	1231
200	9.987508	10.00	0.99875	0.049969	51	864	1168
300	12.23725	12.25	0.99917	0.040808	41	892	1140
400	14.1333	14.14	0.99938	0.035344	36	908	1124
500	15.80348	15.81	0.99950	0.031615	32	920	1112
600	17.31329	17.32	0.99958	0.028861	29	928	1104
700	18.70161	18.71	0.99964	0.026721	27	935	1097
800	19.99375	20.00	0.99969	0.024996	25	940	1092
900	21.20731	21.21	0.99972	0.023567	24	944	1088
1000	22.35509	22.36	0.99975	0.022358	23	948	1084
2000	31.61882	31.62	0.99988	0.01581	16	968	1064
3000	38.72661	38.73	0.99992	0.012909	13	977	1055
4000	44.71856	44.72	0.99994	0.01118	11	982	1050
5000	49.9975	50.00	0.99995	0.01	10	986	1046
6000	54.76997	54.77	0.99996	0.009129	9	988	1044
7000	59.15868	59.16	0.99996	0.008451	9	990	1042
8000	63.24358	63.25	0.99997	0.007906	8	992	1040
9000	67.08018	67.08	0.99997	0.007453	8	993	1039
10000	70.70891	70.71	0.99998	0.007071	7	994	1038
20000	99.99875	100.00	0.99999	0.005	5	1001	1031
30000	122.4735	122.47	0.99999	0.004082	4	1004	1028
40000	141.4205	141.42	0.99999	0.003536	4	1005	1027
50000	158.1131	158.11	1.00000	0.003162	3	1006	1026
60000	173.2044	173.21	1.00000	0.002887	3	1007	1025
70000	187.0822	187.08	1.00000	0.002673	3	1008	1024
80000	199.9994	200.00	1.00000	0.0025	3	1008	1024
90000	212.1314	212.13	1.00000	0.002357	2	1009	1023
100000	223.6062	223.61	1.00000	0.002236	2	1009	1023

This shows that with N=90 as I am doing, the ±3 sigma range is (789,1243) around 1016. In fact we have to go out to N = 5000 to get the STD down to 10 hardware units!

So, I think this gives me a much improved view of my statistical probes. At N = 90, σ = 76/1016 = 7.5%, so if we look at ± 3σ, we are going about 23% to each side. This is much larger than I thought it would be! This is my do5 simulation. Remember that 3σ is 99%, 2σ is 95%, and 1σ is 68%.