Statistical Independence Repair PhL 8.1.13

I had to go through all of FT and clean away this wrong idea, it got done OK.

I made the common mistake of equating these to statements

X and Y are statistically independent
X and Y are uncorrelated.

The web gives a nice example of how uncorrelated is possible for X and Y dependent. (I stole it!)

Hi Henna.

Thanks for writing to the Math Forum.

Here is a simple example where the two random variables have zero correlation, yet are not independent.

Suppose X is a normally-distributed random variable with zero mean. Let Y = X^2. Clearly X and Y are not independent: if you know X, you also know Y. And if you know Y, you know the absolute value of X.

The covariance of X and Y is

Cov(X,Y) = E(XY) - E(X)E(Y) = E(X^3) - 0*E(Y) = E(X^3)

= 0,

because the distribution of X is symmetric around zero. Thus the correlation r(X,Y) = Cov(X,Y)/Sqrt[Var(X)Var(Y)] = 0, and we have a situation where the variables are not independent, yet have (linear) correlation r(X,Y) = 0.

This example shows how a linear correlation coefficient does not encapsulate anything about the quadratic dependence of Y upon X.

- Doctor Douglas, The Math Forum
http://mathforum.org/dr.math/

In this example, X and X^2 obviously are not independent, but for the symmetric pdf p(x) and zero mean X he gets

$$E(X^3) = \int dx\, x^3 p(x) = 0$$

So then

$$\text{cov}(X,X^2) = E(X^3) - \mu_x\mu_{x2} = E(X^3) = 0.$$

********** try to edit in a repair **********

(c) Basic Probability Theory

In the above discussion, if we have random variables A,B,C.... , we can write a **joint probability** density function in this manner

pdf(A=a, B=b, C=c.....) (G.2)

where now we don't distinguish whether the various sample spaces are continuous or discrete, we just write anything as a pdf (with the understanding of (G.1) above). The meaning here is that pdf(...) is the probability that random variable A has value a, while at the same time (the same experiment) random variable B has value b, and so on. For the continuous case, pdf(...) da db dc... is the same probability but for the range da of a and db of b, etc.

If all these random variables are **statistically independent** (such as A = number of dust particles on your pillow and B = temperature at some location on Pluto), this joint probability distribution factors,

pdf(A=a,B=b,C=c.....) = pdf(A=a) pdf(B=b) pdf(C=c) .... (G.3a)

For two random variables A and B, one would have

pdf(X=x,Y=y) = pdf(X=x) pdf(Y=y) . (G.4a)

We now adopt some **shorthand notations**:

p(a,b,c....) ≡ pdf(A=a,B=b,C=c.....) = $\text{pdf}_{\mathbf{ABC}}$...(a,b,c...) = $p_{\mathbf{ABC}}$..(a,b,c..) (G.5)

When one sees p(2, -4, 0. 4...) *one must remember* that the arguments correspond to values of specific random variables, and if things become unclear, one must revert to the fuller notation. One trick is to use a parameter name that reminds the reader of the random variable name, such as a for A: p(a) = P(A=a). In the shorthand notation we have

p(a,b,c...) = p(a)p(b)p(c)..... // N statistically independent random variables (G.3b)

p(x,y) = p(x)p(y) // 2 statistically independent random variables (G.4b)

Digression: One way to understand the concept of statistical independence is by the use of conditional probabilities. Consider:

$p_{\mathbf{X}}$(x | y) = probability *just for X* where X = x *given that* Y = y

which we compare to

$p_{\mathbf{XY}}$(x,y) = probability *for X and Y* where X = x and Y = y .

The connection is given by,

$$p_{\mathbf{XY}}(x,y) = p_{\mathbf{X}}(\, x \mid y \,)\, p_{\mathbf{Y}}(y).$$

If X and Y are independent, then $p_{\mathbf{X}}(\, x \mid y \,)$ has no dependence on y, X knows nothing about Y, and in this case we have $p_{\mathbf{X}}(\, x \mid y \,) = p_{\mathbf{X}}(x)$. This then yields the factored form

$$p_{\mathbf{XY}}(x,y) = p_{\mathbf{X}}(x)\, p_{\mathbf{Y}}(y) .$$

For three variables we can define

$p_{\mathbf{X}}(\, x \mid y, z \,)$ = probability *just for X* where X = x *given that* Y = y and Z = z

and then (start on the right end when reading this)

$$p_{\mathbf{XYZ}}(x,y,z) = p_{\mathbf{X}}(\, x \mid y, z \,)\, p_{\mathbf{Y}}(y \mid z)\, p_{\mathbf{Z}}(z) .$$

If X,Y,Z are statistically independent, then certainly X knows nothing about Y and Z, and Y knows nothing about Z, so then $p_{\mathbf{X}}(\, x \mid y, z \,) = p_{\mathbf{X}}(x)$ and $p_{\mathbf{Y}}(y \mid z) = p_{\mathbf{Y}}(y)$ and then

$$p_{\mathbf{XYZ}}(x,y,z) = p_{\mathbf{X}}(x)\, p_{\mathbf{Y}}(y)\, p_{\mathbf{Z}}(z)$$

giving the factored form. Often this discussion appears with the $\cap$ symbol replace our commas, and one can verify the various claims with Venn diagrams. We shall not digress more on this subject and shall take the factored form as our definition of statistical independence.

We return now to

$$p_{\mathbf{XY}}(x,y) = p_{\mathbf{X}}(x)p_{\mathbf{Y}}(y) \qquad \text{// 2 statistically independent random variables} \qquad \textit{(G.4b)}$$

If one regards $p_{\mathbf{XY}}(x,y)$ as a function $f_{\mathbf{y}}(x)$ for various fixed values of y, (G.4b) says that the *shape* of this function is not influenced by the values of y, only the *overall scale* of $f_{\mathbf{y}}(x)$ is affected by y.

After we define the correlation measure corr(X,Y) below, we will see that (G.4b) being true implies that corr(X,Y) = 0 which means X and Y are **uncorrelated**. This arrangement works only one way:

X,Y statistically independent $\quad\Rightarrow\quad$ X,Y uncorrelated.

It is easy to find examples where X,Y are uncorrelated but are not independent, we shall look at an example later.

Any pdf is **normalized** to 1 since the probability of all possible outcomes (mapped from sample spaces to the random variables) is 1. Thus,

$\int\int \ldots.\ p(x,y\ldots.)\ dx\ dy\ldots\ = 1$ or $\Sigma_{x,y\ldots.}\ p(x,y\ldots.)\ = 1$. (G.6)

********** repair has been installed **********

But I need to continue through G and make other repairs, I think all is OK at least through the start of the covariance definition. We get finally to G.26 and there are two things I want to add. I added (1) the range -1 to 1 with a simple argument, (2) the counterexample of dependent but uncorrelated.

I am now repaired up to (G.31) and there more is needed. I could just cut it down to

Fact: If X and Y are statistically independent random variables, then (G.31)
(a) p(x,y) = p(x)p(y)
(b) E(XY) = E(X)E(Y)
(c) cov(X,Y) = 0
(d) corr(X,Y) = 0
(e) X and Y are uncorrelated

(a) follows from our definition of statistical independence (G.3a) or (G.3b).

(b) $E(XY) = \int\int x\ y\ p(x,y)\ dx\ dy = \int\int x\ y\ p(x)p(y)\ dx\ dy = [\int x\ p(x)\ dx]\ [\int y\ p(y)\ dy] = E(X)E(Y)$.

(c) $cov(X,Y) \equiv E(XY) - \mu_x\mu_y = E(X)E(Y) - \mu_x\mu_y = \mu_x\mu_y - \mu_x\mu_y = 0$

(d) $corr(X,Y) \equiv cov(X,Y) / [\sigma(X)\ \sigma(Y)\ ] = 0 / [\sigma(X)\ \sigma(Y)\ ] = 0$

(e) This is the same as (d), just stated in words **QED**

********** the above has been installed **********8

Now, let's scan through the rest of the doc and see what other repairs are needed.

Pause to worry about independent versus correlation.
near (35.5) needs work
near D.3
near G.3a and G.4a
G.31

Lets try to fix G.31.

Repair of (G.31)

independent

OK, after a lot of effort I think I completely purged this error from FT. It may exist somewhere in Scrambler, but that is for later.