Mathematical Modeling of Infectious Diseases, UCSF

Lecture 3

Version 1.1—modified 23 Jan 2015

One of the things that makes infectious disease data interesting is the fact that cases give rise to more cases. Much of mathematical epidemiology is designed to represent this sort of positive feedback in different ways. Before we move on to these processes, it's time to look at some simpler, but very important, processes first.

Correlation

We've talked about variance; now we need *covariance*. The *covariance* of X and Y is defined to be

$$\text{cov}(X, Y) = E[(X - EX)(Y - EY)].$$

Let's multiply it all out:

$$\text{cov}(X, Y) = E[(X - EX)(Y - EY)] = E[XY - YE[X] - XE[Y] + E[X]E[Y]].$$

Expectation of a sum is the sum of the expectations:

$$\text{cov}(X, Y) = E[XY] - E[YE[X]] - E[XE[Y]] + E[E[X]E[Y]].$$

Expectation of a constant times a variable is the constant times the expectation:

$$\text{cov}(X, Y) = E[XY] - E[X]E[Y] - E[Y]E[X] + E[E[X]E[Y]].$$

Expectation of a constant is just the constant:

$$\text{cov}(X, Y) = E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y].$$

Now cancel:

$$\text{cov}(X, Y) = E[XY] - E[X]E[Y].$$

If $\text{cov}(X, Y) = 0$, the variables are *uncorrelated.* We also saw that independent random variables are uncorrelated.

Worked Exercise. Let X and Y be jointly distributed with $P(X = 0, Y = 1) = P(X = 0, Y = -1) = P(X = -1, Y = 0) = P(X = 1, Y = 0) = 1/4$. Prove X and Y are uncorrelated. Prove that X and Y are *not* independent. So independence implies uncorrelated, but uncorrelated does not imply independent.
Solution. Marginally, $P(X = 0) = 0.5$ and $P(Y = 0 = 0.5)$, but $P(X = 0, Y = 0) = 0$, and not 0.5.

Exercise 1. Prove $\text{cov}(X, X) = \text{var}(X)$. Prove $\text{cov}(X, aY) = a\text{cov}(X, Y)$ for a any constant. Prove $\text{cov}(X, Y) = \text{cov}(Y, X)$ for any random variables X and Y (as long as $E[X]$, $E[Y]$ exist!) ∎

Exercise 2. Prove for any X and Y independent or not, $\text{var}(X+Y) = \text{var}(X) + 2\text{cov}(X, Y) + \text{var}(Y)$. ∎

We're going to need something else. Let $U = (aX - Y)^2$ where X and Y are random variables, and a is some constant. Since $U \geq 0$, $E[U] \geq 0$ because all values of U are nonnegative. So:

$$E[U] = E[(aX - Y)^2] = E[a^2X^2 - 2aXY + Y^2] = a^2E[X^2] - 2aE[XY] + E[Y^2].$$

But since $E[U] \geq 0$,

$$E[U] = a^2E[X^2] - 2aE[XY] + E[Y^2] \geq 0.$$

This is true for any a. So let $a = E[XY]/E[X^2]$. This gives us

$$E[U] = (E[XY])^2E[X^2]/(E[X^2])^2 - 2E[XY]E[XY]/E[X^2] + E[Y^2] \geq 0.$$

$$(E[XY])^2/E[X^2] - 2E[XY]E[XY]/E[X^2] + E[Y^2] \geq 0.$$

$$(E[XY])^2 - 2E[XY]E[XY] + E[X^2]E[Y^2] \geq 0.$$

$$-E[XY]E[XY] + E[X^2]E[Y^2] \geq 0.$$

$$E[X^2]E[Y^2] \geq (E[XY])^2 \tag{1}$$

$$(E[XY])^2 \leq E[X^2]E[Y^2]$$

$$\frac{(E[XY])^2}{E[X^2]E[Y^2]} \leq 1.$$

and

$$\frac{|E[XY]|}{\sqrt{(E[X^2]E[Y^2])}} \leq 1.$$

The quantity on the left is called the population correlation coefficient ρ of course, so now we know that $|\rho| \leq 1$, or $-1 \leq \rho \leq 1$. Equation 1 goes by the name Cauchy-Schwartz inequality.

The geometric distribution

We've seen the binomial distribution, which is the number of successes in N independent identical Bernoulli trials each with success probability p. What if we start tossing Bernoulli trials and just ask how long till the first success? More specifically, on the what-th try does the first success occur? Let this be Y.

Let's do it. What is the chance the first try is a success, so no failures at all? This is p. So $P(Y = 1) = p$. What is the one and only way to get $Y = 2$? Fail the first time, succeed the second: $P(Y = 2) = p(1 - p)$. In general, $P(Y = y) = p(1 - p)^{y-1}$. This is called the *geometric distribution*.

Here, Y could take any value at all from 1 to infinity. Let's add up all the probabilities:

$$\sum_y P(Y = y) = \sum_{y=1}^{\infty} p(1-p)^y = p\sum_{y=1}^{\infty}(1-p)^y.$$

It turns out that

$$\sum_{k=0}^{\infty} a^k = \frac{1}{1-a}.$$

for $0 < a < 1$. This is called a *geometric series*.

Things in tiny print are optional extra reading.

What this really means is

$$\lim_{N\to\infty} \sum_{k=0}^{N} a^k = \frac{1}{1-a}$$

If we think of $\sum_{k=0}^{N} = f(N)$, we would have

$$\lim_{N\to\infty} f(N) = L.$$

We're not going to review all that stuff from calculus about root test and ratio test and so on. But I do want you to know what a limit is. In this case, we are looking at the limit of a sequence.

$$\lim_{N\to\infty} f(N) = L$$

means for any $\epsilon > 0$, there exists some M such that $|f(N) - L| < \epsilon$ whenever $N > M$.

For example, $\lim_{N\to\infty} \frac{1}{N} = 0$. Why? Pick any ϵ. Then we want to look at $|1/N - 0| = 1/N$. Is $1/N < \epsilon$? Sure, as long as $N > 1/\epsilon$. If we make $M = \lceil \frac{1}{\epsilon} \rceil$, then $1/N$ is within ϵ of the limit 0 whenever $N > M$. This sort of thing can seem very elusive when you first see it.

Exercise 3. Pick $a = 1/3$, and try adding $(1/3)^0 + (1/3)^1 + (1/3)^2 + \cdots = 1 + 1/3 + 1/9 + 1/27 + \cdots$. In other words, add successive terms and note the behavior. What does the formula say the limit ought to be? ∎

Exercise 4. Use the geometric series formula to show that the sum of the geometric probabilities is 1. Pay attention to the fact that the geometric probabilities start at 1, not 0. ∎

Exercise 5. Suppose that each time a case of disease is introduced into a region, $a < 1$ secondary cases are produced from each case on average. Show that the size of the n generation is a^n. Don't worry about what fractions of a person mean; interpret these in some average sense. Show that the total size of the outbreak—adding all cases together including the index case—is $1/(1 - a)$. ∎

Exercise 6. Suppose that on average a case of disease causes $R > 1$ secondary cases when no one is vaccinated or immune. Suppose now that a fraction f is vaccinated with complete protection. Discuss the epidemiological credibility of the assumption that now each case would cause (on average) $R(1-f)$ secondary cases (i.e. when the coverage fraction is f). (Do you see the expectation of a binomial possibly concealed in here?) Assume that with coverage f, each case causes $R(1-f)$ secondary cases, and that $R(1-f) < 1$; compute the total expected size of an outbreak, counting all generations. What is the critical coverage level needed for $R(1-f) < 1$? ∎

We're now going to do an important trick. Let Y be geometric. We want to write

$$g(u) = E[u^Y].$$

So.

$$g(u) = \sum_{y=1}^{\infty} u^y p(1-p)^{y-1} = up \sum_{i=0} (u(1-p))^i = \frac{up}{1-u(1-p))}$$

Take this as a function of u, for $0 \leq u \leq 1$. This quantity, for an integer count variable, has a name because it's important enough to talk about, often. It is called a *probability generating function*, and we just calculated it for the geometric distribution.

It's not obvious that this is useful or worthwhile; it's not obvious why anyone would have ever thought of this in the first place. But it works, and we will just take it for granted as useful.

Exercise 7. Use Newton's formula

$$(a+b)^N = \sum_{i=0}^{N} \binom{N}{i} a^i b^{N-i}$$

to prove that the probability generating function for a binomial is $(up + (1-p))^N$. ∎

Here is one of the reasons generating functions are so helpful. Watch what happens:

$$g'(u) = \frac{d}{du} \sum_{y=0}^{\infty} u^y P(Y=y) = \sum_{y=0}^{\infty} y u^{y-1} P(Y=y) = \sum_{y=1}^{\infty} y u^{y-1} P(Y=y).$$

Fixing $u = 1$ gives

$$g'(1) = \sum_{y=1}^{\infty} y P(Y=y),$$

which, *mirabile dictu*, is the formula for the mean. Let's try it on the geometric. Let Y be geometric with parameter p:

$$E[Y] = g'(1) = \frac{d}{du} \frac{up}{1-u(1-p)}|_{u=1}$$

$$E[Y] = \frac{(1-u(1-p))p - up(-(1-p))}{(1-u(1-p))^2}$$

Substitute in $u = 1$ and get $((1-(1-p))p + p(1-p))/(1-(1-p))^2 = (p^2 + p - p^2)/p^2 = p/p^2 = 1/p$.

Exercise 8. Suppose a needle is discarded with probability 0.2 after every usage; each disposal decision is independent of the past. What is the expected number of usages? ∎

Exercise 9. Let Y be geometric with parameter p. Prove that the conditional distribution of $Y - Y_0$ given $Y > Y_0$ for any Y_0 is geometric with parameter p also. ∎

One interesting thing we need to think about is what happens when we take various binomial distributions with $\lambda = Np$ fixed, but let N get bigger and bigger. Certainly p has to get smaller and smaller. Let's look at the generating function:

$$\lim_{N\to\infty}(up + (1-p))^N.$$

Let $p = \lambda/N$.

$$\lim_{N\to\infty}(u\lambda/N + (1-\lambda/N))^N.$$

Rearrange a bit:

$$\lim_{N\to\infty}(1 + u\lambda/N - \lambda/N)^N = \lim_{N\to\infty}(1 + (u-1)\frac{\lambda}{N})^N.$$

It turns out this is a famous limit; this comes up when you analyze instantaneous compound interest for example. And there is a standard trick for doing it, which we'll do now. Basically, let L be the limit we're looking for.

$$L = \lim_{N\to\infty}(1 + (u-1)\frac{\lambda}{N})^N$$

Then use the log of both sides:

$$\log(L) = \log\left(\lim_{N\to\infty}(1 + (u-1)\lambda/N)^N\right).$$

$$\log(L) = \lim_{N\to\infty} N\log(1 + (u-1)\lambda/N).$$

Here, we had to just use a fact from calculus—the continuity of the log—to let us switch the limit and the log. We're not proving that here; I just wanted you to notice it.

From calculus, you might remember this as the indeterminate form $0 \cdot \infty$. Let's change to $k = 1/N$; now the limit is $k \to 0$.

$$\log(L) = \lim_{k\to 0+}\frac{\log(1 + (u-1)k\lambda}{k}$$

It's now the $0/0$ indeterminate form, so it's time for L'Hopital's Rule:

$$\log(L) = \lim_{k\to 0+}\frac{1/(1 + (u-1)k\lambda)(u-1)\lambda}{1}$$

Now, we're getting somewhere:

$$\log(L) = \lim_{k\to 0+}\lambda(u-1).$$

So we just showed that

$$L = e^{\lambda(u-1)}.$$

In the limit as $N \to \infty$, the nice friendly binomial distribution has this limiting form; only λ appears, not N and p separately. We have the probability generating function of something.

The last bit of calculus for today is to look at what this is, as a Taylor series. (In this course you won't be required to compute any Taylor series, just recognise that you can expand various functions in a Taylor series.)

$$e^x = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \ldots + \frac{1}{i!}x^i + \ldots$$

But here $x = \lambda(u-1)$; $e^{\lambda}(u-1) = e^{-\lambda}e^{\lambda u}$, so

$$e^{-\lambda}e^{\lambda u} = e^{-\lambda}\big(1 + \lambda u + \frac{1}{2!}\lambda^2 u^2 + \ldots + \frac{1}{i!}\lambda^i u^i + \ldots\big)$$

Using the formula for the generating function, we can just pick off the probabilities:

$$P(Y = i) = \frac{e^{-\lambda}\lambda^i}{i!}.$$

This, of course, is the *Poisson distribution.*

Exercise 10. Intuitively we know the mean of the Poisson should be λ, from the derivation. But what if something untoward happened when we took the limit? Show that if Y is Poisson with parameter λ, then $E[Y] = \lambda$. ∎

One more trick:

$$g''(u) = \frac{d}{du}\sum_{y=1} y u^{y-1} P(Y = y) = \sum_{y=1} y(y-1)u^{y-2}P(Y = y) = \sum_{y=2} y(y-1)u^{y-2}P(Y = y).$$

It was nice putting $u = 1$ before, so what happens now?

$$g''(1) = \sum_{y=2} y(y-1)P(Y = y) = E[Y(Y-1)].$$

For the Poisson, let's get the variance then.

$$g''(u) = \frac{d}{du}\frac{d}{du}e^{\lambda(u-1)} = \frac{d}{du}\lambda e^{\lambda(u-1)} = \lambda^2 e^{\lambda(u-1)}.$$

$$g''(1) = \lambda^2 = EY^2 - EY.$$

But $EY = \lambda$, so $EY^2 = \lambda^2 + \lambda = \lambda(\lambda + 1)$.

The variance is the second moment minus the square of the mean. So $\mathrm{var}(Y) = \lambda^2 + \lambda - \lambda^2 = \lambda$. You read that right. The mean is λ, and the variance is also λ.

Worked exercise. Let X and Y be Poisson random variables with parameters λ and μ respectively. Assume these are independent. Prove $X + Y$ is Poisson, with mean $\lambda + \mu$. *Solution.* Look at the generating function: $E[u^{X+Y}] = E[u^X u^Y]$. By independence, the expectation of a product of independent variables is the product of the expectations, so $E[u^{X+Y}] = E[u^X]E[u^Y] = e^{\lambda(u-1)}e^{\mu(u-1)} = e^{(\lambda+\mu)(u-1)}$ which is the generating function of a Poisson with mean $\lambda + \mu$.

Worked problem. Let Y be Poisson with mean λ, and then conditional on this, let X be binomial with Y trials and p success probability per trial. (a) Find the mean and variance of X. (b) Find the distribution of X. *Solution.* (a) $E[X] = E[E[X|Y]]$. Since X given Y is binomial(Y,p), $E[X|Y] = pY$. Then $E[E[X|Y]] = E[py] = p\lambda$. Also, $\mathrm{var}(X) = E[\mathrm{var}(X|Y)] + \mathrm{var}(E[X|Y])$. This gives $E[Yp(1-p)] + \mathrm{var}(pY) = p(1-p)\lambda + p^2\lambda = p\lambda$.

Worked problem. Let Y be Poisson with mean λ, and then conditional on this, let X be binomial with Y trials and p success probability per trial, as in the previous worked problem. Note: this is sometimes denoted as $X = p \circ Y$. Show X is Poisson with mean $p\lambda$. *Solution.* Writing this out in terms of the law of total probability is tedious. Let's try

$$E[u^X] = E[E[u^X|Y]]$$

by iterated expectation. The conditional expectation of $u^X|Y$ is the generating function of a binomial with Y trials:

$$E[u^X|Y] = (up + (1-p))^Y.$$

Then

$$E[u^X] = E[(up + (1-p))^Y] = \sum_{y=0}^{\infty} (up + (1-p))^Y P(Y = y).$$

We could try to do this irritating sum, but better to recognize that it's already been done. It's the expectation of something to the Y for a Poisson. We know that $E[u^Y] = e^{\lambda(u-1)}$, and that's what this is, just with $up + (1-p)$ taking the place of u. So:

$$E[u^X] = e^{\lambda(up+1-p-1)} = e^{\lambda(up-p)} = e^{\lambda p(u-1)}.$$

Sure enough, this is the generating function of a Poisson variable with mean λp.

Exercise 11. Suppose that accidents occur according to a Poisson distribution with mean 1. What is the probability of zero accidents? Suppose the mean is 0.1 or 10; what is the probability of zero accidents? ∎

Exercise 12. Suppose that HIV-negative participants in a trial acquire new partners at random according to a Poisson distribution with mean λ. Supposing that the prevalence of infection is p and the probability of infection given an infected partner is β, what is the distribution of the number of *sufficient exposures*? What is the probability of zero sufficient exposures? What is the probability of the person becoming infected (i.e. of having one or more sufficient exposures)? ∎

Worked problem. Consider the following process:

$$X_{\tau+1} = p \circ X_\tau + Y_\tau,$$

where Y_τ is a sequence of independent identical Poisson variables with mean λ, and $p \circ X$ means a Binomial with X trials at p success probability per trial (assumed independent of the Y's and of each other for all time). Let $X_0 = 0$ (start with nobody). We are thinking of Y_τ as the number of incident cases, assumed to be happening randomly in time. Each case recovers or dies with probability $1-p$ (remains with probability

p) each time. So cases show up randomly, but on average λ each period, and they last on average a time $1/(1-p)$.

This is a simple example of a large class of processes called "immigration-death" processes in the mathematical demography literature.

$$E[X_{\tau+1}] = E[p \circ X_\tau] + E[Y_\tau]$$

$$E[X_{\tau+1}] = E[E[p \circ X_\tau | X_\tau]] + \lambda$$

$$E[X_{\tau+1}] = E[pX_\tau] + \lambda = pE[X_\tau] + \lambda$$

Just as before, let's iterate this to equilibrium, calling $E[X_\tau] = \mu_\tau$ and the equilibrium $\bar{\mu}$:

$$\bar{\mu} = p\bar{\mu} + \lambda$$

$$\bar{\mu}(1-p) = \lambda$$

so

$$\bar{\mu} = \frac{\lambda}{1-p} = \lambda\frac{1}{1-p}$$

In other words, the mean number you have at equilibrium—the prevalence—is the incidence rate times the expected duration. It's more or less typical for an immigration death process to have some result interpretable as prevalence equaling an incidence times a duration.

But there's more. Starting at $X_0 = 0$, X_1 is Poisson; then we binomially thin it and it stays Poisson. Then we add an independent Poisson to that and it still stays Poisson. Repeating this we can see it will stay Poisson for all time. In fact it will converge to the Poisson for any starting distribution, though we won't talk about it more today.

Finally, note that this process is also an example of a so-called *integer autoregression process* of order 1, IAR(1). These are sometimes used in the analysis of time series of counts. Let's compute the asymptotic autocorrelation of the series, which is the correlation between two neighboring counts in the far future, when we are at equilibrium.

$$\gamma = \frac{E[X_\tau X_{\tau+1}] - E[X_\tau]E[X_{\tau+1}]}{\sqrt{\mathrm{var}(X_\tau)\mathrm{var}(X_{\tau+1})}} = \frac{E[X_\tau X_{\tau+1}] - \bar{\mu}^2}{\bar{\mu}}.$$

The work is

$$E[X_\tau X_{\tau+1}] = E[(p \circ X_\tau)X_\tau] + E[X_\tau Y_\tau].$$

$$E[X_\tau X_{\tau+1}] = E[(p \circ X_\tau)X_\tau] + E[X_\tau]E[Y_\tau] = E[(p \circ X_\tau)X_\tau] + \bar{\mu}\lambda.$$

$$E[(p \circ X_\tau)X_\tau] = E[E[(p \circ X_\tau)X_\tau | X_\tau]] = E[pX_\tau X_\tau] = p\bar{\mu}(1+\bar{\mu}).$$

So

$$E[X_\tau X_{\tau+1}] = p\bar{\mu}(1+\bar{\mu}) + \bar{\mu}\lambda.$$

And

$$\gamma = (p\bar{\mu}(1+\bar{\mu}) + \bar{\mu}\lambda - \bar{\mu}^2)/\bar{\mu}$$

$$\gamma = (p(1+\bar{\mu}) + \lambda - \bar{\mu})$$

Let's unravel this: $\bar{\mu} = \lambda/(1-p)$:

$$\gamma = p\big(1 + \frac{\lambda}{1-p}\big) + \lambda - \frac{\lambda}{1-p}$$

$$\gamma = p\big(\frac{1-p}{1-p} + \frac{\lambda}{1-p}\big) + \frac{\lambda(1-p)}{1-p} - \frac{\lambda}{1-p}$$

$$\gamma = p\big(\frac{1-p+\lambda}{1-p}\big) + \frac{\lambda(1-p)-\lambda}{1-p}$$

$$\gamma = \frac{p(1-p)+p\lambda}{1-p} + \frac{-\lambda p}{1-p}$$

$$\gamma = \frac{p(1-p)+p\lambda-\lambda p}{1-p}$$

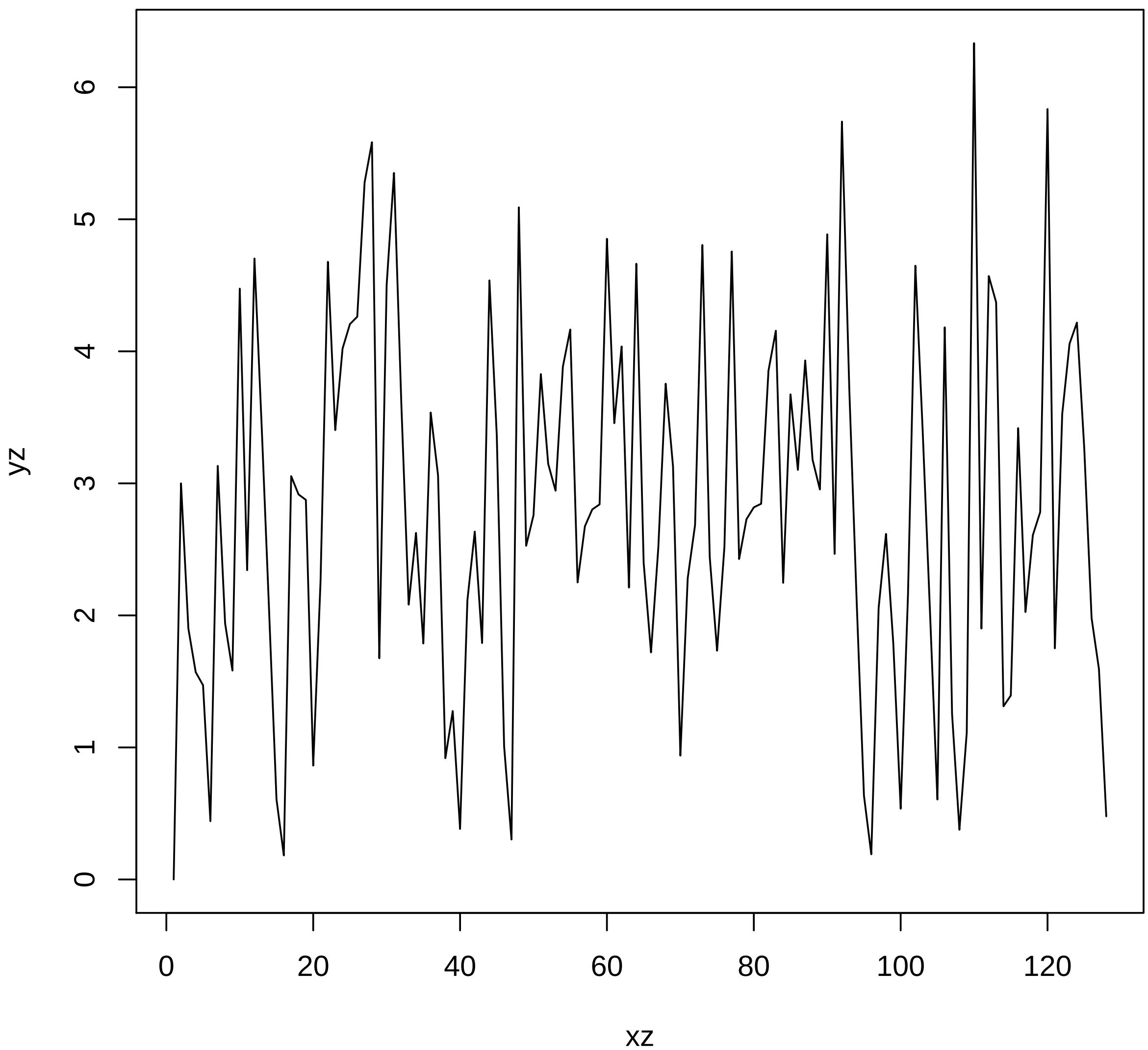
$$\gamma = \frac{p(1-p)}{1-p} = p.$$

Since p is a probability (and $p < 1$ anyway to allow equilibrium), this particular series can never have a negative autocorrelation. You can't use IAR(1) to model a process that tends to have a low count follow a high count, for example.

Let's look at some simulations of the process. Here I plot with $\lambda = 2$ (cases per day, say) and p=pp=0.3 (30% survival per day, say).

```
iar1sim <- function(len,lambda,pp,starting=0) {
  if (len <= 0 || len > 65536) {
    stop("length bad")
  }
  if (pp<0 || pp>1.0) {
    stop("p bad")
```

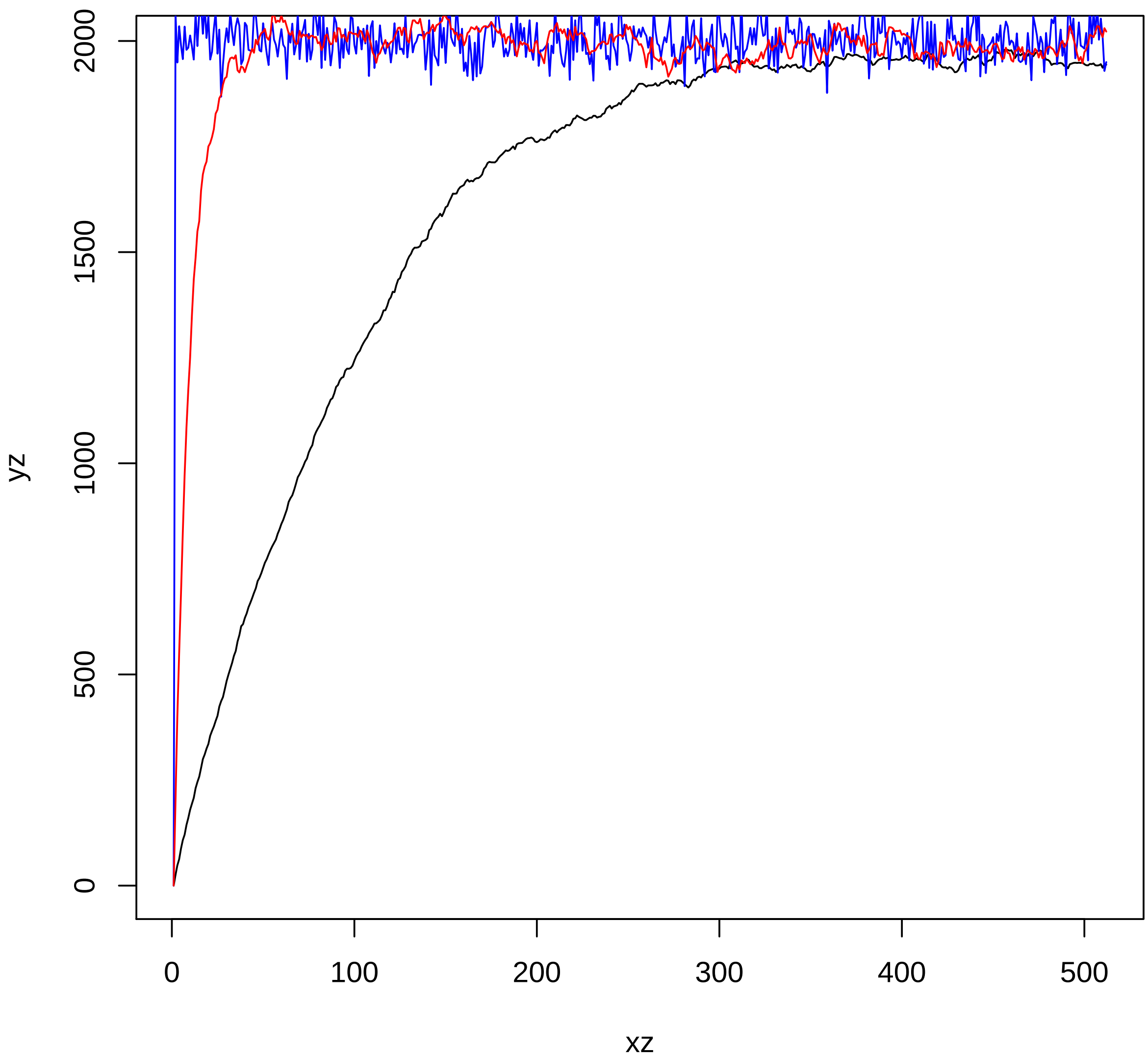
```
  }
  if (lambda < 0) {
    stop("lambda bad")
  }
  answer <- rep(NA,len)
  answer[1] <- starting
  cur <- starting
  for (ii in 2:len) {
    cur <- pp*cur + rpois(1,lambda=lambda)
    answer[ii] <- cur
  }
  answer
}
xz <- 1:128
#                       lambda=2 and pp=0.3 below
yz <- iar1sim(length(xz), 2, 0.3)
plot(xz,yz,type="l")
```



Let us compare IAR(1) sequences with the same long run mean. Consider $\lambda = 20$ and $p = 0.99$; the long run mean is 20/0.01=2000. If we had $\lambda = 200$ and $p = 0.9$ we would have the same mean, as would $\lambda = 2000$ and $p = 0$.

```
xz <- 1:512
yz <- iar1sim(length(xz), 20, 0.99)
```

```
yz2 <- iar1sim(length(xz), 200, 0.9)
yz3 <- iar1sim(length(xz), 2000, 0.0)
plot(xz,yz,type="l")
points(xz,yz3,type="l",col="blue")
points(xz,yz2,type="l",col="red")
```



Computer Exercise 13. Please use the above R code to simulate the IAR(1) process with (a) $\lambda = 10$ and $p = 0$ (no survival), $p = 0.3$, and $p = 0.8$. Part (b): Also, use $\lambda = 200$, with $p = 0.1$ in one run and $p = 0.99$ in another; please comment on the differences. Please try, in an exploratory way, different values of λ and p; feel free to set `xz <- 1:500` to go out to time 500 or even longer. Plot your results; comment on the appearance of the graph (smoothness, initial rise). Only turn in 3–4 graphs. ∎

Exercise 14. How long did this assignment take you? ∎