Wed, May 5, 2010.
1. The binomial and the exponential family.
2. Properties of correlograms.

1) The binomial and the exponential family.
Exponential family.
$f(y) = \exp\{[y\theta - b(\theta)]/a(\phi) + c(y,\phi)\}$
Binomial.
$f(y) = \text{choose}(m,y)\ \mu^y\ (1-\mu)^{m-y}$
$= \exp\{\log(\text{choose}(m,y)) + \log(\mu^y) + \log[(1-\mu)^{m-y}]$.
Let $\text{lch}(m,y) = \log(\text{choose}(m,y))$.
$= \exp\{y\log(\mu) + (m-y)\log(1-\mu) + c(y,\phi)\}$
$= \exp\{y\log(\mu) + m\log(1-\mu) - y\log(1-\mu) + \text{lch}(m,y)\}$
$= \exp\{y[\log(\mu)-\log(1-\mu)] + m\log(1-\mu) + \text{lch}(m,y)\}$
$\log(\mu)-\log(1-\mu) = \log(\mu/(1-\mu))$
$= \exp\{y\log(\mu/(1-\mu)) + m\log(1-\mu) + \text{lch}(m,y)\}$.
So, $\theta = \log(\mu/(1-\mu))$.

2) Properties of correlograms (and autocorrelation functions).

Sample autocov $= c_k = 1/n \sum (Y_t - \underline{Y})(Y_{t+k} - \underline{Y})$. The sum is from t=1 to n-k. $\underline{Y}$ means the sample mean of $Y = (Y_1 + \ldots + Y_n) / n$

Similarly, you can compute the
sample autocorrelation function (sample acf) $= r_k = c_k / c_0$.

A correlogram is a plot of r_k versus the lag, k.

Some properties of autocorrelations are the following.
a) Even: $\rho_k = \rho_{-k}$, and $r_k = r_{-k}$.

b) Both are between -1 and 1 ; i.e. $\leqq 1$ in absolute value.

c) $\rho_0 = r_0 = 1$.

d) If Y_t has a seasonal component with period τ (e.g. one year),

then r_k and ρ_k also have seasonal components with period τ .

Often the correlogram starts out high and gradually decays.

For iid observations Y_i with mean 0 and variance σ^2, $\rho_k = 0$, for $k \neq 0$.
$r_k \sim -1/n$, +/- about $2 / \sqrt{n}$, for $k \neq 0$.
So for confidence bounds, you can use $-1/n$ +/- $1.96/\sqrt{n}$.
People often use instead 0 +/- $2/\sqrt{n}$.

Why is $E[r_k] \sim -1/n$?
Look at r_1.
$r_1 = c_1/c_0$.

Now, the exp value of something divided by something

is not equal to the quotient of the exp values, but approximately it
is. Anyway, let's look at $E[c_1]$.

$E[c_1] = E\,[1/n \sum (Y_t - \underline{Y})(Y_{t+1} - \underline{Y})]$

$= 1/n\, E[\, (Y_1 - \underline{Y})(Y_2 - \underline{Y})$
$+ (Y_2 - \underline{Y})(Y_3 - \underline{Y})$
$+ \ldots + (Y_{n-1} - \underline{Y})(Y_n - \underline{Y})]$

$= 1/n\ \{ (n-1)\, E[\underline{Y}^2] - E[Y_1^2]/n - 2E[Y_2^2/n]$
$- \ \ldots - 2\, E[Y_{n-1}^2]/n - E[Y_n^2]/n\ \}$

(because for instance $E[Y_1\, \underline{Y}\,] = E[Y_1\, (Y_1 + \ldots + Y_n)/n] = E[Y_1^2]/n$.)

(and similarly $E[\underline{Y}^2] = E[(Y_1 + \ldots + Y_n)/n\ (Y_1 + \ldots + Y_n)/n]$
$= 1/n^2\, E[Y_1^2 + \ldots + Y_n^2] = 1/n^2\, (n)\, E[Y_1^2] = \sigma^2/n.$)

So $E[c_1] = 1/n\, [(n-1)\ \sigma^2/n + \sigma^2/n[-1 -2 -2 - \ldots -2 -1]\}$
$= \sigma^2/n^2\, [n-1 - 1 - 2(n-2) - 1]$
$= \sigma^2/n^2\, [-n + 1]$
$\sim - \sigma^2/n.$

And c_0 = an estimate of σ^2, so it's around σ^2.
(actually $E[c_0] = \sigma^2\, (n-1)/n$)
so it makes sense that $E[r_1] \sim -1/n$.

Note that if we knew μ , i.e. if μ replaced $\underline{Y}$ in these formulas, then $E[c_0]$ would be zero.

3) AR(p).

AR(1): $Y_t = \alpha_1 Y_{t-1} + \varepsilon_t$, or more generally,

AR(p): $Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots \alpha_p Y_{t-p} + \varepsilon_t$.

Motivation: Y_t depends on Y_{t-1}, Y_{t-2}, etc. (continuity, esp if α_1 is close to 1 .