Lab Report #1: Moments & Cumulants
(Started: July 19, 2011; Revised: February 4, 2012)

Due at the start of class. You may speak to others, but whatever you hand in should be your own work.

> **Solution:** The solutions to this assignment are computed in the Matlab program listed at the end and posted here. If you run the program yourself, the output will give you most of what's asked for. The answers below go beyond this.

1. (cumulants of Bernoulli random variables) Consider a random variable x that equals δ (an arbitrary number) with probability p (a number between zero and one) and 0 with probability $1 - p$. We'll use its cumulant generating function (cgf) to find its first four cumulants, representing, respectively, its mean, variance, skewness, and kurtosis. Feel free to use Matlab — I would.

   (a) Verify that this is a legitimate probability distribution.

   (b) Derive the moment generating function. (If you're confused about this, apply the definition.) What is the cumulant generating function?

   (c) Differentiate the cgf to find the first four cumulants, labelled κ_1 through κ_4.

   (d) Derive the standard measures of skewness and excess kurtosis:

   $$\begin{aligned} \gamma_1 &= \kappa_3/(\kappa_2)^{3/2} \quad \text{(skewness)} \\ \gamma_2 &= \kappa_4/(\kappa_2)^2 \quad \text{(excess kurtosis)} \end{aligned}$$

   How do they depend on p? δ? What is excess kurtosis when $p = 1/2$?

> **Solution:** This the Bernoulli multiplied by δ, so it illustrates the impact (or lack thereof) of scaling.
>
>   (a) Since p and $1 - p$ are nonnegative and sum to one, we're ok.
>
>   (b) See program.
>
>   (c) The cumulants are
>
> $$\begin{aligned} \kappa_1 &= \delta p \\ \kappa_2 &= \delta^2 p(1 - p) \\ \kappa_3 &= \delta^3 p(1 - p)(1 - 2p) \\ \kappa_4 &= \delta^4 p(1 - p)[1 - 6p(1 - p)]. \end{aligned}$$

Note the scaling: κ_j includes δ^j. Other than scaling, we've seen the first two before. The third one tells us that skewness depends on the sign of δ and whether p is greater or less than one half (graph the probabilities against x if this isn't clear). The fourth one depends on $p(1-p)$. At $p = 1/2$, this terms reaches its max of $1/4$, so the overall term is negative, which generates negative excess kurtosis. As p moves toward zero or one, this term shrinks.

(d) You'll note that Matlab doesn't do the obvious cancelation of δ's. Once you do, you have

$$\begin{aligned} \gamma_1 &= \operatorname{sgn}(\delta)(1 - 2p)/[p(1-p)]^{1/2} \\ \gamma_2 &= 1/[p(1-p)] - 6. \end{aligned}$$

There's a subtle issue with γ_1: the magnitude of δ doesn't matter, but its sign ("sgn") does. That shows up in the ratio of $[\delta^2]^{3/2}$ to δ^3 (think about this a minute). What about excess kurtosis? At $p = 1/2$, $\gamma_2 = -2$, so there's excess kurtosis. Loosely speaking, the Bernoulli has thinner tails than the normal. As p approaches zero or one, we can reverse that. As we increase/decrease p, we get a distribution with increasing skewness and excess kurtosis.

2. (sample moments) It's often helpful to experiment with artificial test problems, so that we know when we make a mistake. Here we compute sample moments of data generated by the computer and verify what we think they are.

We'll start by generating some artificial data:

```
format compact
mu = 0;
sigma = 1;
nobs = 1000;
x = normrnd(mu,sigma,nobs,1);
```

These commands generate "pseudo-random" numbers from a normal distribution and put them in the vector x. The serve as data for what follows.

(a) Our first check is to see if the sample moments correspond, at least approximately, to our knowledge of normal random variables. For example, use the commands:

```
xbar = mean(x)
moments = mean([(x-xbar).^2 (x-xbar).^3 (x-xbar).^4])
```

What do you get? What are the first four sample moments? What are the moments of a normal random variable with this mean and variance? How do they compare?

(b) Our second check is on the Matlab commands `mean(x)`, `std(x)`, `skewness(x)`, and `kurtosis(x)`. How do they compare to sample moments you computed earlier? Are they exactly the same, almost the same, or completely different?

> **Solution:**
>
> (a) The moments are close to what the distribution implies: mean one, variance (and standard deviation) one, skewness zero, and kurtosis three. The small differences reflect sampling variability. You can see this by increasing the sample size, which typically leads the sample moments to be closer to the "population" moments. The one difference is in the variance and standard deviation, which are computed by dividing the sum of squared deviations by the number of observations minus one, rather than just the number of observations. It's a small difference, to be sure.
>
> (b) The `skewness` and `kurtosis` commands give exactly the same answers as our own calculations, which assures us that they do what we want them to do.

3. (volatility and correlation in business cycles) We're going to use macroeconomic data from FRED, the St Louis Fed's convenient online data repository, to get a summary picture of US business cycles. (Search "FRED St Louis".) Data preparation should follow these steps or the equivalent:

   • Download monthly data, 1960 to the present, for the following series: Industrial Production (series INDPRO), Employment (All Employees: Total Nonfarm, series PAYEMS), Real Personal Consumption Expenditures (PCEPI), and the S&P 500 Index (SP500). GDP is not available monthly, so the first three series serve as alternative measures of economic activity. Variation in their growth rates reflects the business cycle. The last one serves as our measure of asset prices, and its growth rate as the (approximate) return on equity. Save them all in a spreadsheet, with each series in its own column.

   • Read your data into Matlab and construct year-on-year growth rates using the formula

   $$g_t \quad = \quad \log x_t - \log x_{t-12}.$$

   [The function `log` here is the natural log. We refer to growth rates computed this way as continuously-compounded. Take that as given for now, but ask me in class if you'd like an explanation.] If the data are in an array called `data`, with each column a different variable, you can compute growth rates with

   `diffdata = log(data(13:end,:)) - log(data(1:end-12,:))`

   [The expressions `data(i,j)` have two forms here. The first index is a range (eg, `x:y`), where `end` means the last one. The second index (`:`) means use all the available columns.]

   • A short-cut for the adventurous. I don't recommend this unless you're fairly comfortable with Matlab, and maybe not even then, but my colleague Kim Ruhl has written a script that automates data input from FRED; that is, you can skip the spreadsheet step. If you'd like to try it, see

   http://www.kimjruhl.com/computing/.

Once you have the data ready to go:

(a) For each variable (that is, its growth rate), compute the standard deviation, skewness, and excess kurtosis. (The commands `std`, `skewness`, and `kurtosis` are helpful here.) Which variables have the highest "volatility" (standard deviation)? Do any of them seem "nonnormal" (non-Gaussian) to you?

(b) For the same variables, compute correlations with employment growth. (The command `corrcoef` may come in handy.) Which variable has the highest correlation?

> **Solution:** See the Matlab program. Note that there's negative skewness in most of these series, also positive excess kurtosis. The outlier here is consumption — I may have to check into that. Industrial production and employment are highly correlated (above 0.8), the S&P 500 less so (correlation about 0.3 with the former, 0.15 with the latter).
>
> I recommend, in addition, that you plot the series and look at their histograms.

4. **Optional.** (sum of normals is normal) The idea here is to use cumulant generating functions (cgfs) to show that the sum of independent normal random variables is also normal. It's helpful to break the problem into manageable pieces, like this:

(a) Consider two independent random variables x_1 and x_2, not necessarily normal. Show that the cgf of the sum $y = x_1 + x_2$ is the sum of their cgfs:

$$k(s;y) = k(s;x_1) + k(s;x_2).$$

Hint: note the form of the pdf and apply the definition of the cgf.

(b) Suppose $x_i \sim \mathcal{N}(\kappa_{i1}, \kappa_{i2})$. [This bit of notation means: x_i is normally distributed with mean κ_{i1} and variance κ_{i2}.] What is x_i's cgf? Hint: we did this in class.

(c) Use (a) to find the cgf of $y = x_1 + x_2$, with (x_1, x_2) as described in (b) (namely, normal with given means and variances). How do you know that y is also normal? What are its mean and variance?

(d) Extend this result to $y = ax_1 + bx_2$ for any real numbers (a, b).

> **Solution:**
>
> (a) If x_1 and x_2 are independent, they're pdf factors: $p_{12}(x_1, x_2) = p_1(x_1)p_2(x_2)$. That means the mgf of $x_1 + x_2$ is the product of their individual mgf's and their cgf is the sum.
>
> (b) $k(s; x_i) = s\kappa_{1i} + s^2 \kappa_{2i}/2$. (If this isn't burned into your memory, please burn it now.)
>
> (c) Sum the cgf's:
>
> $$\begin{aligned} k(s;y) &= k(s, x_1) + k(s, x_2) \\ &= (s\kappa_{11} + s^2\kappa_{21}/2) + (s\kappa_{12} + s^2\kappa_{22}/2) \\ &= s(\kappa_{11} + \kappa_{12}) + s^2(\kappa_{21} + \kappa_{22}). \end{aligned}$$

It's normal because it's mgf has the form of a normal random variable.

(d) Still normal, but with a slight change in mean and variance:

$$k(s; y) = s(a\kappa_{11} + b\kappa_{12}) + s^2(a^2\kappa_{21} + b^2\kappa_{22})/2.$$

Matlab program for questions 1 to 3:

```
%  hw1_s12
%  Matlab program for Lab Report #1, Spring 2012
%  Written by:  Dave Backus, January 2012
format compact
clear all

disp(' ')
disp('Answers Lab Report 1')
disp('*** I put pauses in the program.  If Matlab stops, hit return')

%%{
disp(' ')
disp('--------------------------------------------------------------------')
disp('Question 1')
syms p delta omega s

mgf = 1-p+p*exp(s*delta)
cgf = log(mgf)

disp(' ')
kappa1 = subs(diff(cgf,s,1),s,0)    %  mean
kappa2 = subs(diff(cgf,s,2),s,0)    % variance
kappa3 = subs(diff(cgf,s,3),s,0)
factor(kappa3)
kappa4 = subs(diff(cgf,s,4),s,0)
factor(kappa4)

disp(' ')
gamma1 = kappa3/kappa2^(3/2)
%simplify(gamma1)
gamma2 = kappa4/kappa2^2

pause
%%}

disp(' ')
disp('--------------------------------------------------------------------')
disp('Question 2')
disp('Moments of simulated normal data')
disp('Ccompare our calculations to Matlab functions')
disp(' ')
% set parameters
mu = 0;
sigma = 1;
```

```matlab
nobs = 1000;

x = normrnd(mu,sigma,nobs,1);

xbar = mean(x)
moments = mean([x (x-xbar).^2 (x-xbar).^3 (x-xbar).^4])

disp(' ')
var_x = var(x)
compare = moments(2)

disp(' ')
std_x = std(x)
compare = sqrt(moments(2))

disp(' ')
skw_x = skewness(x)
compare = moments(3)/moments(2)^(3/2)

disp(' ')
krt_x = kurtosis(x)
compare = moments(4)/moments(2)^2

pause

disp(' ')
disp('----------------------------------------------------------------')
disp('Question 3')

%{
% This reads data directly from FRED.  I got stuck converting SP500 from
% daily to monthly.  Also need to align series.  Help welcome.
[data_ip,desc_ip]=getFredData('INDPRO');
desc_ip
[data_emp,desc_emp]=getFredData('PAYEMS');
desc_emp
[data_con,desc_con]=getFredData('PCEPI');
desc_con
[data_sp500,desc_sp500]=getFredData('SP500');
desc_sp500
%}

% read data from spreadsheet:  INDPRO, PAYEMS, PCEPI, SP500
data = xlsread('FRED_hw1.xls','Monthly');
%data(1:10,:)
```

```
% year-on-year growth rates
diffdata = log(data(13:end,:)) - log(data(1:end-12,:));

disp('Basic moments')
disp('Variables:  IP, EMP, CON, SP500')
means = mean(diffdata)
stdev = std(diffdata)
skew = skewness(diffdata)
exkurt = kurtosis(diffdata) - 3
pause

% histogram to see skewness and kurtosis
hist(diffdata(:,4))
pause

disp(' ')
disp('Correlations')
corrcoef(diffdata)

return
```