18.465 PROBLEM SET 2, DUE FRI., FEB. 20, 2015

1. Let X be a non-constant random variable with $E(|X|^3) < \infty$ and standard deviation $\sigma > 0$. Recall that the *skewness* of X or its distribution is defined as $E((X - EX)^3)/\sigma^3$.
(a) Show that the skewness is preserved if we add a constant a to X or multiply it by a constant $b > 0$.
(b) If X has a distribution symmetric around its mean EX, in other words $X - EX$ has the same distribution as $EX - X$, show that the skewness of X is 0.
(c) In particular show that any normal distribution $N(\mu, \sigma^2)$ with $\sigma > 0$ or any uniform distribution $U[a, b]$ with $a < b$ has zero skewness.

2. (a) Find the skewness of a an exponential variable V having, for some $\lambda > 0$, density $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ and 0 for $x < 0$. *Hint*: it suffices to consider $\lambda = 1$.
(b) Generate, in R, a sample of 20 i.i.d. standard exponential random variables by the command
> x = rexp(20)
and then test it for normality by the Shapiro–Wilk test,
> shapiro.test(x)
Repeat this a few times. (You don't need to retype the commands; hit the "up arrow" key until you get back to the command you want to repeat, then "Enter" and for generating random variables you will get new ones, starting from a different seed than before. Likewise to redo the test on the new sample.) Is normality rejected, i.e. is the p-value less than 0.05, at least in most tries?
(c) A uniform $U[a, b]$ distribution is symmetric around its mean, so it has 0 skewness, although in other ways it's clearly very different from a normal distribution. We may as well consider $U[0, 1]$ as changes of location and scale won't matter. Find the kurtosis of any $U[a, b]$ distribution.
(d) Try
> x = runif(20)
which will generate $x = (X_1, ..., X_n)$ i.i.d. $U[0, 1]$ and do the Shapiro–Wilk test on x to see if normality is rejected. Try a few times. If it is not rejected, then instead of $n = 20$ try larger multiples of 10 such

Date: 18.465, Feb. 13, 2014.

as $n = 30$, 40, etc. until you find n large enough so that normality is rejected in several trials. (The Shapiro–Wilk test is "consistent against all alternatives," meaning that for any non-normal distribution, for n large enough, normality should be rejected with probability increasing toward 1.)

3. In R, the command
> z = rnorm(6)
generates 6 i.i.d. standard normal random variables. To do a simple linear regression of z on x, where x is the index one can generate by
> x = 1:6
which is the easiest way to enter the vector $x = (1, 2, 3, 4, 5, 6)$ or in R, x = c(1,2,3,4,5,6), would seem useless and uninteresting because we know that the simple linear regression hypothesis holds with $a = b = 0$, so we would not expect either the intercept $\widehat{a}$ or the coefficient $\widehat{b}$ of x in the regression to be significantly different from 0 (p-value less than $\alpha = 0.05$) except of course with probability about 0.05. We would also expect the Shapiro–Wilk test not to reject normality, even for a much larger value of n in place of x, except again with probability α.

But, suppose we take the order statistics of the z_i to get $z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(6)}$. In R, doing
> sz = sort(z)
will give sz as the vector of order statistics. It is still useless to do the Shapiro–Wilk test on sz, because it should give exactly the same output on sz as on z. But if we do the simple linear regression
> out = lm(sz ∼ x)
in R, then
> summary(out)
to see the results, it will not be surprising at all that the coefficient $\widehat{b}$ of x is positive (order statistics, by definition, have an increasing trend). It will not be too surprising if $\widehat{b}$ is significantly different from 0, the only question being perhaps whether $n = 6$ is large enough to produce a significantly non-zero slope.
(a) Do you find, in a few tries, that $\widehat{b}$ is significantly different from 0?
(b) Is $\widehat{a}$ also significantly different from 0? Explain why you think that should, or should not, happen.

4. In R, generate vectors as follows:
> x = 1:10
> y = sin(pi * x/20)
> v = y + 0.1*rnorm(10)

where "sin" does give the sine function in R, "pi" does give $\pi =$ 3.14159..., and * gives multiplication; thus 0.1*rnorm(10) is equivalent to rnorm(10,0,0.1). The sine function is increasing on the interval $[0, \pi/2]$, although nonlinearly. So when we do the simple linear regression of v on x,

> out = lm(v ~ x)

> summary(out)

(a) we expect that the coefficient of x will be positive and are not surprised if it's significantly different from 0. Are these things true?

(b) But we know that the sine function is not exactly linear. Find all the ten residuals of the simple linear regression by

> residuals(out)

and see if there is a pattern in them.

(c) Do the quadratic regression

> qout = lm(v ~ x + I(x^2))

> summary(qout)

and see if the coefficient $\widetilde{c}$ of x^2 is significantly different from 0. If so, we can reject the simple linear regression hypothesis by the test based on $\widetilde{c}$.

(d) Is the coefficient $\widetilde{c}$ positive or negative? Is the sign surprising? Why or why not?

(e) Find the residuals of the quadratic regression by

> residuals(qout)

Is there any pattern in them? Of course, we know that the sine function is not exactly quadratic, either. If we replaced 10 by some larger number n and likewise 20 by $2n$, we should eventually be able to reject quadratic regression also, by doing cubic regression and finding a significantly non-zero coefficient of x^3.