Stochastic Calculus, Courant Institute, Fall 2014

http://www.math.nyu.edu/faculty/goodman/teaching/StochCalc2014/index.html

Always check the class message board before doing any work on the assignment.

Assignment 1, due September 15

Corrections (check the class message board): (The old part 5 (a recurrence relation) was removed. What was part 6 is now part 5. It is an exercise in R programming. The due date was corrected. It was September 18. Now it's September 15.)

1. (*Academic integrity*) Please read the Academic Integrity section of the class web page for this class. Please state whether you agree to abide by each of these policies. *Hint:* You will not be allowed to take this course if you do not agree to follow guidelines (2)-(4).

2. (*Central limit theorem*) Suppose Y_i are i.i.d. random variables with $E[Y_i] = 0$, $E\left[Y_i^2\right] = \sigma^2$, and $E\left[Y_i^4\right] = \mu_4 < \infty$. Suppose $X_n = \frac{1}{\sqrt{n}}\left(Y_1 + \cdots + Y_n\right)$. The CLT says that as $n \to \infty$, the distribution of X_n converges to $\mathcal{N}(0,\sigma^2)$. This means that if $V(x)$ is a reasonable function of x and $X \sim \mathcal{N}(0,\sigma^2)$, then $E[V(X_n)] \to E[V(X)]$ as $n \to \infty$. But $E[V(X)]$ depends only on σ^2. Therefore, if n is large, $E[V(X)]$ depends very little on details of the distribution of Y_i, other than the variance. This exercise checks that for $V(x) = x^4$.

   (a) Show that if $X \sim \mathcal{N}(0,\sigma^2)$, then
   $$E\big[X^4\big] = 3\sigma^4 \; . \tag{1}$$
   Hint: Write the integral formula for the expectation, use $x^4 e^{-x^2/2} = x^3\left(xe^{-x^2/2}\right)$, and
   $$xe^{-x^2/2\sigma^2} = Const \;\cdot\; \partial_x e^{-x^2/2\sigma^2} \;,$$
   and integrate by parts.

   (b) Conclude that the variance of X^2 is $\mathrm{var}\big(X^2\big) = 2\sigma^4$. We will use this formula many times this semester.

   (c) Write the formula for $E\big[X_n^4\big]$. *Hint:* Use
   $$X_n^4 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} Y_i Y_j Y_k Y_l \;,$$
   Take the expectation and figure out which of the terms are different from zero. Among those are terms that involve $E\big[Y_i^2 Y_j^2\big] = \sigma^4$ if $i \neq j$, and terms that involve $E\big[Y_i^4\big] = \mu_4$. You need to figure out how many terms of each kind there are.

(d) From your formula for part 2c, show that $E\big[X_n^4\big] \to 3\sigma^4$ as $n \to \infty$. This does not have to be a formal mathematical proof, just use the fact that $\frac{1}{n} \to 0$ as $n \to \infty$. The CLT says that the influence of μ_4 should disappear in the limit $n \to \infty$.

(e) (*not an action item*) Conclusions:

i. The CLT represents a loss of information. Details of the distribution of Y are lost in the limit process $X_n \to X$.

ii. Gaussian random variables are a simple scaling related to *thin tails* (more on this later). The variance of X^2 is proportional to the square of the variance of X.

iii. Probabilistic analysis involves knowing more than just the sizes of terms. The non-zero terms in the sum representing $E\big[X_n^4\big]$ are $\frac{\sigma^4}{n^2}$ or $\frac{\mu_4}{n^2}$. They have the same order of magnitude in a mathematical sense because they share the same power of n. You cannot tell how important terms of a certain kind are by looking at just one term or the power of n. Sometimes a lot of small terms add up to something important, even as $n \to \infty$. Sometimes they don't. The terms $\frac{\sigma^4}{n^2}$ do. The terms $\frac{\mu_4}{n^2}$ don't.

3. (*Student $t-$distribution, part 1*) The Student $t-$distribution is important in statistics because it is related to the accuracy of estimates of the mean of a Gaussian random variable. It is becoming widely used also because it is a simple probability distribution with an explicit PDF that has power-law fat tails depending on a parameter. Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$ is a one dimensional Gaussian. Suppose X_i are independent samples of X. An estimator of the distribution mean is the sample mean:

$$\widehat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \; . \tag{2}$$

An estimator of the variance is

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \; . \tag{3}$$

Both $\widehat{\mu}$ and $\widehat{\sigma^2}$ are random variables. This exercise studies their distribution.

(a) Show that if X_i is replaced by $X_i - \mu$, then the distribution of $\widehat{\sigma^2}$ is unchanged and the distribution of $\widehat{\mu}$ is only shifted by μ. This is mainly theoretical because you are unlikely to know μ in practice. The distribution of $X_i - \mu$ is $\mathcal{N}(0, \sigma^2)$. From now on, assume that $\mu = 0$.

(b) If X_i is replaced by $\frac{1}{\sigma}X_i$, how do the random variables $\widehat{\mu}$ and $\widehat{\sigma^2}$ change? From now on, assume $\sigma = 1$, so $X_i \sim \mathcal{N}(0, 1)$.

(c) Let $v_1 \in \mathbb{R}^n$ be the column vector $v_1 = \frac{1}{\sqrt{n}}(1, \ldots, 1)^t$. Find $\|v_1\|_{l^2}$. Show that there are vectors v_2, …, v_n so that the vectors v_1, v_2, …, v_n are an orthonormal basis of $\mathbb{R}^n$. You can do this by giving a formula for v_2, …, v_n, or by using abstract theorems (if you state them completely).

(d) Let $v \in \mathbb{R}^n$ be any unit vector ($\|v\| = 1$, $\|v\| = \|v\|_{l^2}$ everywhere in this exercise). Let $\mathcal{S} \subset \mathbb{R}^n$ be the plane of vectors perpendicular to v. That is, $x \in \mathcal{S}$ if and only if $x^t v = 0$. The *orthogonal projection* of x onto $\mathcal{S}$ is the $y \in \mathcal{S}$ that minimizes $\|x - y\|$. Show the basic facts about projections, not necessarily in this order:

i. $y = x - (x^t v)\, v$ $\quad$ ($(x^t v)$ is the v component of x.)

ii. y is perpendicular to $x - y$ $\quad$ (the geometry of l^2 projection)

iii. $\|x\|^2 = \|y\|^2 + (x^t v)^2$ $\quad$ (the Pythagorean theorem)

(e) From now on, $\mathcal{S}$ is the plane perpendicular to v_1 of part (3c). Let $\vec{X}$ be the column vector

$$\vec{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

Show that $\overline{X} = \frac{1}{\sqrt{n}} \vec{X}^t v_1$.

(f) Let $Y_i = \vec{X}^t v_i$ for $i = 2, \ldots, n$. Show that the Y_i are independent of each other, independent of $\overline{X}$ and have $Y_i \sim \mathcal{N}(0, 1)$.

(g) Let $\vec{Y}$ be the orthogonal projection of $\vec{X}$ onto $\mathcal{S}$. Show that

$$\left\|\vec{Y}\right\|^2 = \sum_{i=2}^{n} Y_i^2 = \sum_{i=1}^{n} \left(X_i - \left(\vec{X}^t v_1\right) v_{1,i}\right)^2 = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

(h) Show that the random variables $\widehat{\sigma}^2$ and $\widehat{\mu}$ are independent.

(i) The *chi square* distribution with k *degrees of freedom* is defined as the distribution of $Q = Z_1^2 + \cdots + Z_k^2$, where $Z_i \sim \mathcal{N}(0, 1)$ are independent. This is written $Q \sim \chi_k^2$. Show that $\widehat{\sigma^2} \sim \frac{1}{n-1} \chi_{n-1}^2$ and that $E\left[\widehat{\sigma^2}\right] = 1$.

(j) Return now to $X_i \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu \neq 0$ and $\sigma^2 \neq 1$. Then $\overline{X}$ is normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. We want to quantify $\widehat{\mu} - \mu$, which is the difference between the estimated mean and the unknown true mean. More precisely, we want to quantify how many standard deviations $\widehat{\mu}$ is from μ. But we do not know σ, only the estimated $\widehat{\sigma} = \sqrt{\widehat{\sigma^2}}$. The t–statistic is the difference between $\widehat{\mu}$ and μ, measured in estimated standard deviations of $\widehat{\mu}$:

$$\widehat{\mu} = \mu + t \widehat{\sigma}_{\widehat{\mu}} \ .$$

Show that this leads to

$$t = \frac{\sqrt{n}\,(\widehat{\mu} - \mu)}{\sqrt{\widehat{\sigma^2}}} \ . \tag{4}$$

Show that t has the same distribution as

$$t = \frac{\sqrt{n-1}Z}{\sqrt{\chi^2_{n-1}}} \ , \tag{5}$$

where Z is a one dimensional standard normal and χ^2_{n-1} is an independent chi-square random variable with $n-1$ degrees of freedom. Conclude that the t in (4) is independent of the parameters μ and σ. This independence is the basis of the *Student* $t-$*test* in statistics.

(k) (*not an action item*) This exercise illustrates the general principle that linear algebra, even abstract linear algebra, is a good way to understand multivariate Gaussians. It illustrates how important it is that you make Gaussian random variables, the Y_i in this case, independent simply by making them have zero covariance, which you calculate using linear algebra.

4. Part (3) is all you need to know about the $t-$statistic to do statistics. But the formula for the density of the t random variable is useful in many other modeling problems. In this problem, C represents a normalization constant that may be different in different places.

(a) Let $Q \sim \chi^2_n$. Let F be the distribution function $F(q) = \Pr(Q \leq q)$. Let $f(q) = F'(q)$ be the probability density. Show that

$$F(q) = C \int_{\|x\|^2 \leq q} e^{-\|x\|^2/2}\, dx = C \int_{r=0}^{\sqrt{q}} r^{n-1} e^{-r^2/2}\, dr \ .$$

Conclude that

$$f(q) = Cq^{(n-2)/2} e^{-q/2} \ . \tag{6}$$

(Note: the three C values above are all different. Their values do not matter and I hope you can do the exercise without writing formulas for them. There is a formula for the C in (6) that involves the *gamma function*, so the density (6) is sometimes called the *gamma* distribution.)

(b) The t–distribution with n *degrees of freedom* is the distribution of the random variable

$$T = \frac{\sqrt{n}Z}{\sqrt{\chi^2_n}} \ ,$$

where Z and χ^2_n are independent. We want $f(t)$, which is the probability density of T. This is "clearly" symmetric: $f(t) = f(-t)$

(why??). Since $f(-t) = f(t)$, we can find f from a modified distribution function $F(t) = \Pr(|T| \leq t)$, which satisfies $2F'(t) = f(t)$. Show that $|T| \leq t$ is equivalent to $|Z| \leq t\sqrt{Q}/\sqrt{n}$. Write the integral over z that represents

$$P(t,q) = \Pr\big(|Z| \leq t\sqrt{q}/\sqrt{n}\big)$$

for fixed t and q. Then write the double integral over z and q that represents the probability.

$$\widetilde{P}(t) = \Pr\Big(|Z| \leq t\sqrt{Q}/\sqrt{n}\Big)$$

when Q is random. The variable t appears only in the limit of integration of the inner dz integral.

(c) Differentiate this expression with respect to t to get a one dimensional integral expression of the form

$$f(t) = C\int_{q=0}^{\infty} q^p e^{-a(t)q}\,dq\ ,$$

with $a(t)$ and p explicitly given as simple functions of n and t.

(d) Use the change of integration variable $a(t)q = r$ to get an explicit formula

$$f(t) = \frac{C}{a(t)^{p+1}}\ .$$

If you did all this correctly, the answer should be

$$f(t) = \frac{C}{\left(1 + \frac{1}{n-1}t^2\right)^{n/2}}\ .$$

This is the *Student $t-$density* with $n-1$ degrees of freedom. More generally, the $t-$density with *parameters* μ and σ and n degrees of freedom is

$$f(x;\mu,\sigma,n) = \frac{C}{\left(1 + \dfrac{(x-\mu)^2}{n\sigma^2}\right)^{n/2+1}}\ . \qquad (7)$$

There is an explicit formula for C, also involving the gamma function, but it is complicated and rarely is needed. An important feature of this formula is that n does not have to be an integer. Of course, it was an integer in (5), but even that is unnecessary if we use (6) to define the chi-square distribution for non-integer n.

(e) Show that the probability density (7) has a *power law tail*, which means that $f(x) \approx Cx^{-p}$ as $x \to \pm\infty$.

(f) Show that the t–density converges to the corresponding normal as the number of degrees of freedom goes to infinity. One can do this abstractly using *Slutsky's theorem*, but here please just evaluate the limit of (7) as $n \to \infty$. Of course, the constants C depend on n, but these also have a limit, which you don't need to compute.

(g) Clearly $\mu = E[X]$ for the density (7). Show that $\text{var}(X) < \infty$ for $n > 1$ but the formula $\sigma^2 = \text{var}(X)$ is not true. *Hint #1*: It suffices to show $E\big[X^2\big] \neq 1$ when $\mu = 0$ and $\sigma = 1$, and for some value of n. *Hint #2*: This may be very time consuming. Please do not do it unless you have lots of free time and everything else is finished. The fact is more important than the proof.

(h) (*not an action item*) Why assign this big computation? One reason is to practice integration in n dimensional spaces with unknown constants. Another is to become familiar with the density (7) with an arbitrary power law tail.

5. (*Working with R*) This assignment is a rapid introduction to some aspects of scientific computing and visualization in R. This class uses R because it seems to be the scripting language that is easiest to install and use for all the platforms people are likely to have. But experience shows that following the instructions will not give the desired outcomes for all people on all platforms. Please start the R exercise as soon as possible so there is time to solve any problems related to installing R on your platform or running the posted scripts. Please post any problems you encounter on the class message board. It is likely others will have similar problems. Take the stuff about coding standards seriously, especially if you are a beginning programmer. If you spend 15 minutes making your code more readable and automatic, you will save days of pointless debugging. Follow the links on the class Resources page to read more about programming style.

(a) If you do not have the `R` package on your computer, install it from the web. There are instructions for this on the class web page. It *should* be easy. On a Linux system or a Mac (the instructions for Microsoft operating systems are similar, but I don't know what they are), you should be able to type the command `R` into a command window and get an R prompt, which looks like this;

```
>
```

Type the command `x = 2`. It creates a variable in R called `x`, gives it the value 2.3, then gives you a prompt for the next command Now the R window should look like

```
> x = 2.3
>
```

If you type an expression that evaluates to a number, R should print that number, then give you a prompt. For example, typing x*x should give $2.3^2 = 5.29$. The 5R window should look like this:

```
> x = 2
> x*x
[1] 5.29
>
```

(b) You should do almost all your R work with scripts rather than directly at the command line. An R *script* is a file that contains a sequence of R commands. Its filename ends with .R. Create a file dummy.R that contains the lines

```
x = 2
x*2
```

You run a script at the R command line by typing source("script_name.R"). You run the dummy.R script by typing source("dummy.R"). Unfortunately, you don't see the output.

```
> source( "dummy.R")
>
```

You need to add a command that types the output in the R window. One of these is cat (which stands for "catenate", it's a long story). If you change x*x to cat(x*x), you get

```
> source( "dummy.R")
5.29>
```

You see the answer, but it is on the same line as the next prompt. The reason is that cat is too literal. It sent to the terminal a sequence of characters representing x*x, but lacking the character crlf (carriage return, linefeed), that says "go to the beginning of the next line". This character is written \n. If you add the line cat("\n"), this sends to the terminal the *string* (a sequence of characters in quotes), which puts you at the beginning of the next line. If dummy.R is

```
x = 2.3
cat(x*x)
cat("\n")
```

then typing source("dummy.R") should give

```
> source( "dummy.R")
5.29
>
```

You should use the R command sprintf ("string print file") to create informative output lines with numbers and text together. The example scripts have examples.

(c) Download and save the three R scripts `IntersectingCurves.R`, and `IntersectingCurvesPlot.R`, and `PlotDensitites.R` into some directory (folder). If you have a Mac or a Linux box, go to the directory where you saved them and type `source("IntersectingCurves.R")`. This should create a file called `PlotDemo.pdf`. Compare this to the file `PlotDemoCheck.pdf`, which is posted with the assignment. The figures should be the same. Run the script `PlotDensitites.R`. The result should be

```
> source("PlotDensitites.R")
Welcome to R world
Computing with mu =    0.000, sigma =    2.000, and n =    5.000
numerical Gaussian integral =   5.0133, error = 2.117e-06
numerical Gaussian variance =   1.0000, error = -3.000e+00
t integral =    5.266
Gaussian variance =    4.000, t variance =    1.000, t var - Gaussian var = -3.000e
normalized difference is =    1.264
>
```

(d) The script `PlotDensities.R` is designed to compute the variance of X when X has the $t-$distribution with n degrees of freedom. The script does not assume that you know the normalization constant in the probability density. This is a common situation, you $X \sim f(x)$, and $f(x) = Cu(x)$, where you have a formula for u but don't know C. Of course,

$$\frac{1}{C} = \int_{-\infty}^{\infty} u(x)\,dx \;.$$

We estimate the integral numerically as

$$\int_{-\infty}^{\infty} u(x)\,dx \approx \int_{x_{min}}^{x_{max}} u(x)\,dx \approx \sum_{i=1}^{n_x} u(x_i)\Delta x \;.$$

The integration points are uniformly spaced in the integration interval, which leads to $x_i = x_{min} + (i-1)\Delta x$, and $\Delta x = (x_{max} - x_{min})/(n_x - 1)$. The script applies this first to the Gaussian, where you know the answer, then to the $t-$distribution, where you don't. You should play with the computational parameters n_x and x_{min} to see what it takes to get an accurate answer for the Gaussian. It is remarkable how large Δx can be and still get very high accuracy. The C for the $t-$distribution converges to the C for the Gaussian as $n \to \infty$. Play with the code to see this happen numerically.

(e) Modify the script `PlotDensities.R` to compute the variances of the Gaussian and the $t-$distributed variable. You know the exact answer for the Gaussian, which is a check on the numerics. For the Gaussian, you just need to modify the line

```
    g_var = g_var + ( g( x, mu, sig)/g_int)*dx
```

to estimate

$$E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{u(x)}{C}\, dx \; .$$

You should see that the variance of the $t-$distributed variable with parameter σ is larger than σ^2, but converges to σ^2 as $n \to \infty$.

(f) To start learning `R`, figure out how to change the cubic to a quadratic $y = ax^2$. The number of intersection points will be different.

(g) Add code to `PlotDensities.R` to print in one plot figure the normalized Gaussian and t densities. You can get most of the code you need from `IntersectingCurves.R` and `IntersectingCurvesPlot.R`. Make sure to put computational parameters, n, μ, σ into the plot title or elsewhere on the plot. For this, you will have to learn to use `sprintf`. Correct the legend to refer to the two curves correctly. Hand in one or two plots, but at least one with $n = 5$ to see the difference between the two curves, and the fat tails of the t.