Homework 1: Probability and R

36-401, Modern Regression, Fall 2015

Due at the start of class, 10 September 2015

AGENDA: Practice with probability and convergence of random variables; practice translating math into R; practice with loading, examining and plotting data.

Note: The problem set is scored out of 100 points; the problems add up to 90 points; the remaining ten points will be graded according to a writing rubric, given at the end of the assignment.

1. *Fun with the Gaussian distribution* (25) Suppose that $Y_1, Y_2, \ldots Y_n, \ldots$ are independent Gaussian random variables, all with the same mean μ and variance σ^2. As usual, $\overline{Y}_n = n^{-1} \sum_{i=1}^{n} Y_i$.
   (a) (5) Show that $\overline{Y}_n \sim N(\mu, \sigma^2/n)$. *Hint:* Use the properties of Gaussian random variables, and the rules for algebra with expectations and variances, in appendix A of the textbook.
   (b) (2) What is the distribution of $W_1 = (Y_1 - \mu)/\sigma$? Do all the W_i have the same distribution?
   (c) (2) What is the distribution of $\overline{W}_n$?
   (d) (4) Show that $\sum_{i=1}^{n} W_i^2$ has a χ^2 distribution; what is the number of degrees of freedom?
   (e) (4) The sample variance is $s^2 = (n-1)^{-1} \sum_{i=1}^{n} (\overline{Y}_n - Y_i)^2$. Find a formula for its distribution in terms of σ^2, n, and the distribution of $\sum W_i^2$.
   (f) (4) What distribution does $d = \sqrt{n}\frac{\overline{Y}_n - \mu}{s}$ follow?
   (g) An actual experiment measured the weight of a sample of 144 cats. The mean body weight was 2.72 kilograms, with a standard deviation of 480 grams. It is hypothesized that the population mean is exactly three kilograms.
      i. (2) Test the hypothesis that $\mu = 3$ against the alternative that $\mu \neq 3$ at the $\alpha = 0.05$ level of significance.
      ii. (2) Test the hypothesis that $\mu = 3$ against the alternative that $\mu < 3$ at the $\alpha = 0.01$ level of significance.

2. *Covariances, Correlations, and Sample Means* (22) Suppose $X_1, X_2, \ldots X_n$ are random variables with a common mean μ and a common variance σ^2, and that if $i \neq j$, then $\mathrm{Cov}\left[X_i, X_j\right] = \rho\sigma^2$. (That is, all the X's have the same covariance with each other.) Abbreviate $n^{-1}\sum_{i=1}^{n} X_i$ by $\overline{X}_n$.

   (a) (3) Explain why ρ must be ≤ 1 and ≥ -1.

   (b) (5) Find a formula for $\mathbb{E}\left[\overline{X}_n\right]$ in terms of μ, σ^2, ρ and n. (Some of these may not actually appear in your answer.)

   (c) (7) Find a formula for $\mathrm{Var}\left[\overline{X}_n\right]$ in terms of μ, σ^2, ρ and n, and some whole numbers. (Again, some of these may not be in the final formula.)

   (d) (7) For what values of ρ, σ^2 and μ does $\overline{X}_n$ converge on μ?

3. *Translating math into R* (12 total; 1 pt each) Give an R expression which corresponds to each of these mathematical formulas. Say whether it contains any variables. If the formula contains no variables, use R to get its value to six decimal places (but no more). If it does contain variables, note which variables you would have to define in R before evaluating the expression.

   There is generally more than one way to translate each of these formulas into a valid R expression; try to keep it simple.

   (a) $\frac{1}{\sqrt{16\pi}} e^{-(7-3)^2/16}$

   (b) $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(-7-3)^2/2\sigma^2}$

   (c) $\begin{bmatrix} 1 & 0 & -\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

   (d) $\sin(4)$

   (e) $\cos(\omega x - \phi)$

   (f) $\frac{\mathrm{pressure} \times \mathrm{volume}}{273 + \mathrm{temperature}}$

   (g) $-0.5 + 8.2 \times 42 - 0.23 \times (42)^2 + \frac{(42)^3}{6}$

   (h) $-0.5 + 8.2x - 0.23x^2 + \frac{x^3}{6}$

   (i) $[\log 1.5 \log 7 \log 0.11]$

   (j) $\sqrt{4^2 + (-8)^2 + (0.5)^2}$

   (k) $\vec{a}/\|\vec{a}\|$ (your answer must work for vectors $\vec{a}$ of any dimension)

4. *Translating R into math* (10; 2 each) Give a mathematical formula which corresponds to each R expression. Make the formulas as concise as possible. You may assume `design` is a 5×5 matrix and `rate` is a scalar numerical variable, and introduce algebraic symbols for them.

   (a) `design[1,1]+design[2,2]+design[3,3]+design[4,4]+design[5,5]`

(b) `exp(-abs(x*rate))/2*rate`

(c) `pnorm(0.95,0,1)`

(d) `sum(diag(design))`

(e) `dt(abs(x)*rate,7)`

5. *Loading and manipulating data* (21 total) The file `http://www.stat.cmu.edu/~cshalizi/mreg/15/hw/01/fha.csv` contains information about the size of American cities, and the total number of miles people drive in each city every day.

   (a) (2) Give the command you would use to load the file, and to check the number of rows and columns it has. How many rows and columns does it have? (It should have at least 400 rows and no more than ten columns.)

   (b) (2) What row of the file contains information for Pittsburgh? How do you know?

   (c) (2) How many miles are driven per person per day in Pittsburgh? Give the R commands you use to calculate this, and round the answer to the nearest mile.

   (d) (3) Calculate the number of miles driven per person per day for *every* city. Store the answer either as a vector or as a new column in the data set. Check that the new vector or column contains the right number for Pittsburgh. Give all the commands you use to do this.

   (e) (2) Make a histogram of the population of cities.

   (f) (2) Make a histogram of the number of miles driven per person per day.

   (g) (3) What is the mean number of miles driven per capita? The median number? The standard deviation?

   (h) (3) Make a scatterplot where population is on the horizontal axis and per-capita miles driven is on the vertical axis.

   (i) (4) Add a straight line with intercept 22 and slope 0 to the plot. Does it seem to fit?

RUBRIC (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Questions which ask for a plot or table are answered with both the figure itself and the command (or commands) use to make the plot. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is either properly integrated with a tool like R Markdown or knitr, or included as a separate R file. In the former case, both the knitted and the source file are included. In the latter case, the code is clearly divided into sections referring to particular problems. In either case, the code is indented, commented, and uses meaningful names. All code is relevant to the text; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output. For full credit, all code runs, and the Markdown file knits (if applicable).

EXTRA CREDIT (10): A basic result in probability is "Markov's inequality": if X is a random variable which is non-negative ($\Pr(X \geq 0) = 1$), then for any non-random number $a \geq 0$,

$$\Pr(X \geq a) \leq \mathbb{E}\left[X\right]/a \; .$$

(a, 5 pts) Use Markov's inequality to prove that for any random variable Y with mean μ and variance σ^2,

$$\Pr(|Y - \mu| \geq r) \leq \sigma^2/r^2 \; .$$

(This is called "Chebyshev's inequality".)
(b, 1 pt) Consider using the test statistic $T = |Y|/\sigma$ to test the hypothesis that $\mu = 0$ against $\mu \neq 0$ (assuming σ is known). If we set the critical level of the test statistic to 3 and assume Y is Gaussian, what is the size of the test?
(c, 4 pts) How large could the actual size be if Y is not Gaussian?