Ec240a, Fall 2015

Professor Bryan Graham

Problem Set 1

Due: October 26th, 2015

Problem sets are due at 5PM in the GSIs mailbox. You may work in groups, but each student should turn in their own write-up (including a printout of a narrated/commented and executed iPython Notebook if applicable). Please also e-mail a copy of any iPython Notebook to the GSI (if applicable).

1 Binomial distribution

Imagine it is the Fall Semester of 2012 and you are conducting a survey of the presidential voting intentions of Cal undergraduates. Let $Y = 1$ if a randomly sampled Cal undergraduate plans to vote for the incumbent, Barack Obama, and zero if they plan to vote for an alternative candidate. Among the population of Cal undergrads $\theta = \Pr(Y = 1)$ is the true population frequency of individuals who intend to vote for Obama. You take a random sample of size N from the Cal student body. Let $Z_N = \sum_{i=1}^{N} Y_i$ equal the total number of sampled students who indicate their intention to vote for Obama.

1. Derive a formula that can be used to calculate the ex ante (i.e., pre-sample) probability of the event that $Z_N < z$ for any $z \in \{1, 2, \ldots, N\}$. Provide a 3 - 4 sentence written description of your reasoning.
2. Let $\bar{Y}_N = \frac{1}{N} \sum_{i=1}^{N} Y_i$. Using your answer above, provide an expression that can be used to calculate the ex ante probability of the event $\frac{\sqrt{N}\left(\bar{Y}_N - \theta\right)}{\theta(1-\theta)} < c$.
3. Using your formula plot, in an iPython Notebook, $\Pr\left(\frac{\sqrt{N}\left(\bar{Y}_N - \theta\right)}{\theta(1-\theta)} < c\right)$ as a function of c for $N = 5, 10, 100, 1000$ and $\theta = 1/2$. Make a single figure with 4 subplots arrayed 2×2. Title each figure and label all axes.
4. Let $X \sim \mathcal{N}(0, 1)$. Plot $\Pr(X < c)$ as a function of c on *each* of the four plots created in the previous problem.
5. Repeat questions 3 and 4 with $\theta = 1/20$. Comment on your figures (4 - 6 sentences).
6. After collecting your data you form a confidence interval for θ of $\bar{Y}_N \pm \sqrt{\frac{\bar{Y}_N\left(1-\bar{Y}_N\right)}{N}} z^{1-\alpha/2}$ where $z^{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of a standard normal distribution. What is the approximate ex ante probability that this interval will contain θ? Provide a heuristic justification for this interval (3 - 4 sentences).
7. While in the Econometrics Laboratory, Clopper Pearson, an advanced graduate student, recommends the confidence interval $\left[\underline{\theta}, \bar{\theta}\right]$ where $\underline{\theta}$ and $\bar{\theta}$ respectively solve

$$\begin{aligned}
\underline{\theta} &= \max\left\{t : \sum_{i=1}^{\lfloor Z_N \rfloor} \binom{N}{i} t^i (1-t)^{N-i} > \frac{\alpha}{2}\right\} \\
\bar{\theta} &= \min\left\{t : \sum_{i=\lfloor Z_N \rfloor + 1}^{N} \binom{N}{i} t^i (1-t)^{N-i} > \frac{\alpha}{2}\right\}.
\end{aligned}$$

The lower limit, $\underline{\theta}$, is set equal to zero when $Z_N = 0$ and the upper limit, $\bar{\theta}$, equal to 1 when $Z_N = N$.

8. Write 3 - 4 sentences justifying Clopper's proposal.

9. It turns out that $\sum_{i=k}^{N} \binom{N}{i} t^i (1-t)^{N-i} = \int_0^t f_B(\theta)\, \mathrm{d}\theta$ with $f_B(\theta)$ the pdf of a Beta $(k, N-k-1)$ random variable. Let $F_B^{-1}(\tau; a, b)$ give the τ^{th} quantile of the Beta (a,b) distribution. Argue that

$$F_B^{-1}\left(\frac{\alpha}{2}; Z_N, N - Z_N + 1\right) < \theta < F_B^{-1}\left(1 - \frac{\alpha}{2}; Z_N + 1, N - Z_N\right)$$

closely approximates Clopper's interval. We continue to set the lower limit to zero when $Z_N = 0$ and the upper limit equal to 1 when $Z_N = N$.

10. Let $Y_1, \ldots, Y_N \overset{iid}{\sim}$ Bernoulli (θ). Hoeffding's Inequality states that for any $\epsilon > 0$

$$\Pr\left(\left|\bar{Y}_N - \theta\right| > \epsilon\right) \leq 2\exp\left(-2N\epsilon^2\right)$$

with $\bar{Y}_N = N^{-1}\sum_{i=1}^{N} Y_i$. Let $\mathrm{CI} = \left(\bar{Y}_N - \epsilon_N, \bar{Y}_N + \epsilon_N\right)$. Use Hoeffding's Inequality to find an expression for ϵ_N as a function of N and α such that

$$\Pr\left(\theta \in \mathrm{CI}\right) \geq 1 - \alpha.$$

11. In your iPython Notebook, generate 1,000 samples each consisting of $i = 1, \ldots, N$ independent Bernoulli random variables Y_i with $\Pr(Y_i = 1) = \theta$. For each sample compute the confidence intervals described in 6, 8 and 10 above and record whether the intervals contains θ or not. Do this for all sixteen combinations of $N = 5, 10, 100, 1000$, $\theta = 1/20, 1/2$ and $\alpha = 0.05, 0.10$. Summarize your results in a table and comment on then (7 to 10 sentences).

12. Using the Cropper's method you construct the interval $[0.48, 0.72]$; what is the probability that this interval contains θ? Comment (3 to 4 sentences).

2 Binomial-Beta learning

Let θ, as before, denote the probability than a randomly sampled Cal undergraduate intends to vote for Barack Obama. Assume that your beliefs about θ are summarized by a prior distribution. In particular the probability that you assign to different possible values of θ is given by a beta (a, b) distribution (i.e., $\theta \sim$ beta (a, b)). Let Z_N, also as before, equal the number of Cal students, out of a random sample of size N, who say they intend to vote for Barack Obama.

1. What is the conditional distribution of Z_N given θ?

2. Calculate the joint distribution of Z_N and θ.

3. Calculate the conditional distribution of θ given Z_N. What is the mean of this distribution? Why might posterior be a good name for this distribution? (5 - 6 sentences)

4. Assume that $a = b = 1/2$. Comment on this prior (2 to 3 sentences).

5. While in the Econometrics Laboratory, Harold Jeffreys, another advanced graduate student, suggests that you calculate the interval

$$F_B^{-1}\left(\frac{\alpha}{2}; Z_N + \frac{1}{2}, N - Z_N + \frac{1}{2}\right) < \theta < F_B^{-1}\left(1 - \frac{\alpha}{2}; Z_N + \frac{1}{2}, N - Z_N + \frac{1}{2}\right).$$

   Comment on Jeffreys' proposal? (4 - 6 sentences).

6. Using Z_N you evaluate Jeffreys' interval to be $[0.47, 0.73]$; what is the probability that this interval contains θ? Comment (3 to 4 sentences).

3 Multivariate normal distribution

Let $\mathbf{Y} = (Y_1, \ldots, Y_K)'$ be a $K \times 1$ random vector with density function

$$f(y_1, \ldots, y_K) = (2\pi)^{-K/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{y} - \mu)' \Sigma^{-1} (\mathbf{y} - \mu)\right),$$

for Σ a symmetric positive definite $K \times K$ matrix and μ a $K \times 1$ vector. We say that $\mathbf{Y}$ is a multivariate normal random variable with mean μ and covariance Σ or

$$\mathbf{Y} \sim \mathcal{N}(\mu, \Sigma).$$

The multivariate normal distribution arises frequently in econometrics and a mastery of its basic properties is essential for both applied and theoretical work in econometrics. This problem provides an opportunity for you to review and/or learn some of these properties. There are many useful references on the multivariate normal distribution, for example, T. W. Anderson's *An Introduction to Multivariate Statistical Analysis.*

1. Let C be a $K \times K$ nonsingular matrix. Show that $\mathbf{Z} = C\mathbf{Y}$ is distributed according to $\mathcal{N}(C\mu, C\Sigma C')$.

2. Partition $\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_2')'$ into $K_1 \times 1$ and $K_2 \times 1$ sub-vectors with $K_1 + K_2 = K$. Let $\mu = (\mu_1', \mu_2')'$ and

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

   be conformable partitions of μ and Σ (note that symmetry implies $\Sigma_{12} = \Sigma_{21}'$). Show that $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent random vectors if $\Sigma_{12} = \Sigma_{21}' = \underline{0}\underline{0}'$ (i.e., a matrix of zeros).

3. Let

$$C = \begin{pmatrix} I_{K_1} & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_{K_2} \end{pmatrix}$$

   for I_P a $P \times P$ identity matrix. Derive the distribution of $\mathbf{Z} = C\mathbf{Y}$. Are the first K_1 elements of $\mathbf{Z}$ independent from the second K_2? Interpret your result?

4. Let D be a $P \times K$ ($P \leq K$) matrix of rank P. Arrange the first P columns of D, denoted by D_{11}, such that they are non-singular. Denote the remaining $K - P$ columns by D_{12}. Find a $(K - P) \times K$ matrix E such that

$$\begin{pmatrix} \mathbf{Z} \\ \mathbf{X} \end{pmatrix} = \begin{pmatrix} D \\ E \end{pmatrix} \mathbf{Y}$$

is a non-singular transformation of $\mathbf{Y}$. Finally show that $\mathbf{Z}$ is distributed according to $\mathcal{N}(D\mu, D\Sigma D')$.

5. Consider the partition of $\mathbf{Y}$ introduced in Problem 2 above. Derive the conditional distribution of $\mathbf{Y}_1$ given $\mathbf{Y}_2 = \mathbf{y}_2$.

6. Let $\{\mathbf{Y}_i\}_{i=1}^N$ be a random sample of size N drawn from the multivariate normal population described above. Show that $\sqrt{N}\left(\overline{\mathbf{Y}} - \mu\right)$ is a $\mathcal{N}(0, \Sigma)$ random variable for $\overline{\mathbf{Y}} = \frac{1}{N}\sum_{i=1}^N \mathbf{Y}_i$, the sample mean (<u>HINT:</u> Use independence of the $i = 1, \ldots, N$ draws and your result in Problem 4 above).

7. Let $\mathbf{W} = N \cdot \left(\overline{\mathbf{Y}} - \mu\right)' \Sigma^{-1} \left(\overline{\mathbf{Y}} - \mu\right)$. Show that $\mathbf{W} \sim \chi_K^2$ (i.e., $\mathbf{W}$ is a chi-square random variable with K degrees of freedom).

8. Let $\chi_K^{2,1-\alpha}$ be the $(1-\alpha)^{th}$ quantile of the χ_K^2 distribution (i.e., the number satisfying the equality $\Pr\left(\mathbf{W} \leq \chi_K^{2,1-\alpha}\right) = 1 - \alpha$ with $\mathbf{W}$ a chi-square random variable with K degrees of freedom). Let D be a $P \times K$ $(P \leq K)$ matrix of rank P and d a $P \times 1$ vector of constants. Consider the hypothesis

$$\begin{aligned} H_0 &: D\mu = d \\ H_1 &: D\mu \neq d. \end{aligned}$$

Maintaining H_0 derive the sampling distribution of $D\overline{\mathbf{Y}}$ as well as that of

$$\mathbf{W} = N \cdot \left(D\overline{\mathbf{Y}} - d\right)' \left(D\Sigma D\right)^{-1} \left(D\overline{\mathbf{Y}} - d\right).$$

You observe that, for the sample in hand, $\mathbf{W} > \chi_P^{2,1-\alpha}$ for $\alpha = 0.05$. Assuming H_0 is true, what is the ex ante (i.e., pre-sample) probability of this event? What are you inclined to conclude after observing $\mathbf{W}$ in the sample in hand?