

ECON 623
Fall 2011
Pablo Salinas Macario and John Rust
Solutions Final Exam

Do Question 1 and 2 out of 3 of the remaining questions below.

1. [200 points] Probability question. A casino uses what it believes to be fairly balanced dice in gambling. The casino is considering introducing a new gambling game where a pair of dice are rolled 50 times. On each roll of the pair of dice, the customer (“roller”) gets 1 point if the sum of the numbers on the dice is even, or 0 points otherwise. After the 50 draws, if the roller has more 40 points or higher, they get a prize of \$100. You are asked to calculate an entry fee E so that each time a casino customer chooses this game, the casino expects to make a \$10 expected profit on the gamble.

A 20 points Can you calculate the entry fee exactly? If so, calculate the exact entry fee that yields the casino a \$10 expected profit each time this gamble is played.

Answer Let \tilde{S} be the total score earned after $n = 50$ rolls of the dice. Clearly S takes on values (has support equal to) the set $\{0, 1, \dots, 50\}$. The probability of getting an even sum on the two dice is intuitively equal to $p = 1/2$ since an even sum can be achieved if the realized values on both of the dice are both even or both odd, and even and odd outcomes are equally likely with probability $1/2$. So the probability that the values on the two dice are both even is $1/4 = (1/2)(1/2)$ and the probability that both of the dice have odd-valued outcomes (i.e. 1, 3, or 5) is also $1/4 = (1/2)(1/2)$ so the probability of an even-valued or odd-valued outcomes for both of the dice is $1/4 + 1/4 = 1/2$. Alternatively you can simply enumerate the set of possible even outcomes a sum of 2 occurs with probability $1/36$ (since the only draw that a sum of 2 can occur is the draw $(1, 1)$ where both dice have a value of 1), a sum of 4 can occur with probability $3/36 = 1/12$ (since a sum of four can occur from a draw of $(2, 2)$, $(1, 3)$ or $(3, 1)$), and so on. Via this complete enumeration of the possible even-valued sums on the two dice by summing the corresponding probabilities, we also compute that the probability of an even sum is $p = 1/2$. Then if the successive rolls of the two dice are treated as *IID* draws, it follows that \tilde{S} is a binomial random variable with parameters $n = 50$ and $p = 1/2$. We can then compute the probability $P\{\tilde{S} \geq 40\} = 1 - P\{\tilde{S} \leq 39\} = 1.19306658 \times 10^{-5}$ using the `cdf` command in Matlab, i.e. `1-cdf('bino', 39, 50, 1/2)`. Symbolically we have

$$Pr\{\tilde{S} \geq 40\} = \sum_{j=40}^{50} \binom{50}{j} (.5)^j (.5)^{50-j} \quad (1)$$

and it is also acceptable to numerically evaluate that sum. Using this probability, the expected cost of this gamble is $E\{\tilde{C}\} = 100 \times Pr\{\tilde{S} \geq 40\} = 100 \times 1.193 \times 10^{-5} = 1.193 \times 10^{-3}$. This amounts to less than 1/10 of a cent, 0.001193. To earn an expected profit of \$10, the casino would have to charge essentially \$10 to play the game, or precisely, \$10.001193 in order to offset the small expected cost of paying out the \$100 prize in the very low probability event that $\tilde{S} \geq 40$.

B 20 points Suppose Joey Breakaleg, the casino manager, is not so good at probability theory but he vaguely recalls that there was something called the “cento limit serum” that had something to do with a “bar bell curve” from his days back at Wharton, when he was taking classes with Donald Trump. Can you explain to Joey Breakaleg that he was probably thinking of the *central limit theorem*

and show how it can be used to approximate the probability that this prize would be paid and hence the entry fee? Compare the CLT approximate probability and implied entry fee to the exact probability and entry calculated in part A above. How much more or less than \$10 in profits per game can the casino expect to get using the approximated rather than true probability of paying out the prize? (Extra credit: plot the exact CDF of the number of points entered in 50 rolls, and on the same plot, show the approximate CDF based on the Central Limit Theorem approximation to this distribution: is the CLT approximation uniformly close to the actual CDF, or are there parts of the distribution where it is a poor approximation?)

Answer We can write \tilde{S} as a sum of 50 independently and identically distributed *Bernoulli* random variables $\tilde{B}_i, i = 1, \dots, 50$. The \tilde{B}_i is defined as

$$\tilde{B}_i = \begin{cases} 1 & \text{if the sum of the two dice is even} \\ 0 & \text{if the sum of the two dice is odd} \end{cases} \quad (2)$$

clearly $E\{\tilde{B}_i\} = p = 1/2$ and $\text{var}(\tilde{B}_i) = p(1-p) = 1/4$, so σ , the standard deviation of \tilde{B}_i is $\sigma = 1/2$. So we have

$$\tilde{S} = \sum_{i=1}^{50} \tilde{B}_i \quad (3)$$

Now the Central Limit Theorem tells us that if \tilde{S}_N is given by

$$\tilde{S}_N = \frac{1}{N} \sum_{i=1}^N \tilde{B}_i \quad (4)$$

then a standardized version of \tilde{S}_N

$$\frac{\tilde{S}_N - 1/2}{\frac{\sigma}{\sqrt{N}}} = \sqrt{N}(\tilde{S}_N - 1/2)/(1/2) \implies \tilde{Z} \sim N(0, 1) \quad (5)$$

Taking a blind leap of faith that $N = 50$ is “large enough” for the Central Limit Theorem to be a good approximation to the distribution of the standardized sum of *i.i.d.* Bernoulli’s \tilde{B}_i , we can then similarly standardize \tilde{S} and approximate its CDF by the CDF of the standard normal random variable \tilde{Z} as follows

$$\frac{\tilde{S} - 25}{\sqrt{50/4}} \sim \tilde{Z} \quad (6)$$

Then we are interested in using the Central Limit Theorem approximation to evaluate $P\{\tilde{S} \geq 40\}$, we can do so as follows

$$\begin{aligned} Pr\{\tilde{S} \geq 40\} &= Pr\{\tilde{S} - 25 \geq 40 - 25\} \\ &= Pr\{(\tilde{S} - 25)/\sqrt{50/4} \geq (15/\sqrt{50/4})\} \\ &\simeq Pr\{\tilde{Z} \geq 4.24264\} \\ &= 1 - \Phi(4.24264) \\ &= 1.1045248 \times 10^{-5} \end{aligned}$$

In part A we computed the exact probability that $\tilde{S} \geq 40$ and found it equals 1.193066×10^{-5} so the approximation error in using the Central Limit Theorem is a 7.4% underestimate of the true

probability. We conclude that using the Central Limit Theorem seems quite reasonable as it has an infinitesimal impact on the price you would recommend to Joey Breakaleg to charge for this gambling game. (Note I evaluated the standard normal CDF above in Matlab using the command `1-cdf('norm',4.24264,0,1)` and I am trusting that Matlab uses a good numerical method to approximate the CDF of a standard normal distribution: this is usually done via the *error function* that is closely related to the standard normal CDF. It goes beyond what I was expecting you to answer to evaluate numerical errors in algorithms Matlab or other software use to approximate a standard normal or binomial CDF).

- C 20 points** Suppose Joey is actually a really bad student of probability theory and cannot even recall how to apply the central limit theorem here, and he distrusts probability theory (because “only da really bad guys — dem Wall Street scamsters — do dat stuff”) but he does trust computer simulations so he hires you to write a Matlab program that can conduct a monte carlo experiment to estimate the probability of the payoff and hence the entry fee. But because you fear that Joey might break your leg if you get your monte carlo calculation wrong, and you realize there will be some Monte carlo sampling error in your estimate, you set the number of Monte carlo replications used to estimate this probability to a sufficiently large number so that the probability of an error in the estimated probability greater (in absolute value) of more than 0.00001 (i.e. 1 in 100,000) is less than 0.001 (i.e. one tenth of 1%). What is the smallest number of Monte Carlo replications N that you need to achieve this degree of accuracy?

Answer Suppose we let \tilde{S}_i be a computer simulation of a random variable with a binomial(50,1/2) distribution. If we do a total of N computer draws from these random variables, we can use the *empirical CDF* to evaluate the probability that $\tilde{S} \geq 40$. That is, we can see what fraction of the N draws of $\{\tilde{S}_i\}$ are such that the $\tilde{S}_i \geq 40$. That is, we can estimate the probability via Monte Carlo as

$$\begin{aligned} Pr\{\tilde{S} \geq 40\} &= 1 - Pr\{\tilde{S} \leq 39\} \\ &\simeq 1 - \hat{F}_N(39) \end{aligned}$$

where \hat{F}_N is the empirical CDF given by

$$\hat{F}_N(x) = \frac{1}{N} \sum_{i=1}^N I\{\tilde{S}_i \leq x\}. \quad (7)$$

Define $\tilde{B}_i(x) = I\{\tilde{S}_i \leq x\}$. It follows that $\tilde{B}_i(x)$ is a Bernoulli random variable with parameter $p(x) = F(x)$, where $F(x)$ is the CDF of the random variable \tilde{S}_i (which is a binomial distribution in this case). The *Glivenko-Cantelli Theorem* tells us that with probability 1 as $N \rightarrow \infty$ we have

$$\sup_x |F_N(x) - F(x)| \rightarrow 0 \quad (8)$$

So for sufficiently large N the empirical CDF $\hat{F}(x)$ will be a uniformly good approximation to the true CDF $F(x)$. But how big does N need to be in order to guarantee that our estimate of $Pr\{\tilde{S} \geq 40\}$ via the empirical CDF, $1 - \hat{F}_N(39)$, is within 0.00001 of the true probability with probability at least $1 - .001 = .999$? We can appeal to the Central Limit Theorem to calculate the value of N since we have

$$\sqrt{N} \left(\frac{\hat{F}_N(x) - F(x)}{\sqrt{\hat{F}_N(x)(1 - \hat{F}_N(x))}} \right) \implies \tilde{Z} \sim N(0, 1) \quad (9)$$

So we calculate the requisite value of N as follows. We are interested in choosing N large enough so that the probability of the event $\{|Pr\{\tilde{S} \geq 40\} - 1 - \hat{F}_N(39)| \leq .00001\}$ is at least 0.999. Call the desired tolerance 0.00001 δ for short. Then consider the probability statement, we seek N sufficiently large so that

$$Pr\{|1 - \hat{F}_N(39) - Pr\{\tilde{S} \geq 40\}| \leq \delta\} \leq 0.999 \equiv 1 - \epsilon \quad (10)$$

We use the Central Limit Theorem to approximate the probability on the left hand side of the inequality above. Let $F(x)$ be the true CDF of a Binomial(50,1/2) random variable. We have

$$\begin{aligned} Pr\{|1 - \hat{F}_N(39) - Pr\{\tilde{S} \geq 40\}| \leq \delta\} &= Pr\{|F_N(39) - F(39)| \leq \delta\} \\ &= Pr\left\{\left|\frac{\sqrt{N}(\hat{F}_N(39) - F(39))}{\sqrt{F(39)(1-F(39))}}\right| \leq \frac{\delta\sqrt{N}}{\sqrt{F(39)(1-F(39))}}\right\} \\ &\simeq Pr\left\{|\tilde{Z}| \leq \frac{\delta\sqrt{N}}{\sqrt{F(39)(1-F(39))}}\right\} \end{aligned}$$

Now, for a standard Normal random variable we have

$$Pr\{|\tilde{Z}| \leq 3.2905\} = \Phi(3.2905) - \Phi(-3.2905) = 0.999 \quad (11)$$

(the value 3.2905 was computed in Matlab using the inverse CDF command `icdf` as `icdf('norm', 1-0.0005, 0, 1)`). Assume that N is sufficiently large that the Central Limit Theorem provides a sufficiently good approximation to the true distribution of the normalized random variable

$$\tilde{Z}_N = \frac{\sqrt{N}(\hat{F}_N(39) - F(39))}{\sqrt{F(39)(1-F(39))}} \quad (12)$$

Let $\Phi_N(x)$ denote the CDF of \tilde{Z}_N and $\Phi(x)$ be the CDF of \tilde{Z} , the $N(0,1)$ standard normal limiting distribution implied by the Central Limit Theorem. The *Berry Esseen Bound* gives us a bound on how well $\Phi(x)$ approximates the actual distribution of $\Phi_N(x)$

$$|\Phi_N(x) - \Phi(x)| \leq \frac{C\rho}{\sigma^3\sqrt{N}} \quad (13)$$

where C is an absolute constant (the best current estimate of the upper bound for C is 0.4785 by Tyurin (2010), see Wikipedia) and ρ is $\rho = E\{|\tilde{X}_1|^3\}$ and $\sigma^2 = \text{var}(\tilde{X}_1)$, where \tilde{X}_1 is first of the N IID random variables entering the standardized sum \tilde{Z}_N . In the case at hand, we have

$$\rho = F(39)(1-F(39))(1-2F(39)) \quad (14)$$

and

$$\sigma^3 = \left(\sqrt{F(39)(1-F(39))}\right)^3 \quad (15)$$

Let us initially assume that $\Phi(x)$ is a sufficiently close approximation to $\Phi_N(x)$ that we don't have to worry about the approximation error involved in using Φ in place of Φ_N in our first attempt to compute the smallest value of N such that inequality (10) holds (that is, we ignore that the last step in (11) involves an "approximate equality" \simeq instead of an actual equality $=$, so we may need to

worry about the degree of approximation error in that last step, and we will check this below using the Berry-Esseen approximation error bound). Using equations (11) and (11) we find that if we choose N large enough so that

$$3.2905 = \frac{\delta\sqrt{N}}{\sqrt{F(39)(1-F(39))}} \quad (16)$$

then the probability that using the Monte Carlo probability $1 - \hat{F}_N(39)$ to estimate $1 - F(39) = Pr\{\tilde{S} \geq 40\}$ will be within $\delta = 0.00001$ with probability at least 0.999. Solving this equation for N using the *true probability* $F(39)$ we obtain

$$N = \left[\frac{3.2905\sqrt{F(39)(1-F(39))}}{\delta} \right]^2 = 1291764. \quad (17)$$

Thus, we would need over 1 million monte carlo replications to be able to estimate the small probability $1 - F(39) = 1.93 \times 10^{-5}$ with the required degree of accuracy. Many trials are required because winning the \$100 jackpot is a “rare event” and so we need to run many Monte Carlo trials to be sure of estimating this small probability with sufficient accuracy.

We check the Berry-Esseen bounds to see what the worst-case deviation between the approximate standard normal distribution $\Phi(x)$ and the actual finite sample distribution $\Phi_N(x)$ (the latter is a standardized binomial distribution). We have

$$\frac{0.4785[F(39)(1-F(39))(1-2F(39))]}{[\sqrt{F(39)(1-F(39))}]^3\sqrt{N}} = 0.1218 \quad (18)$$

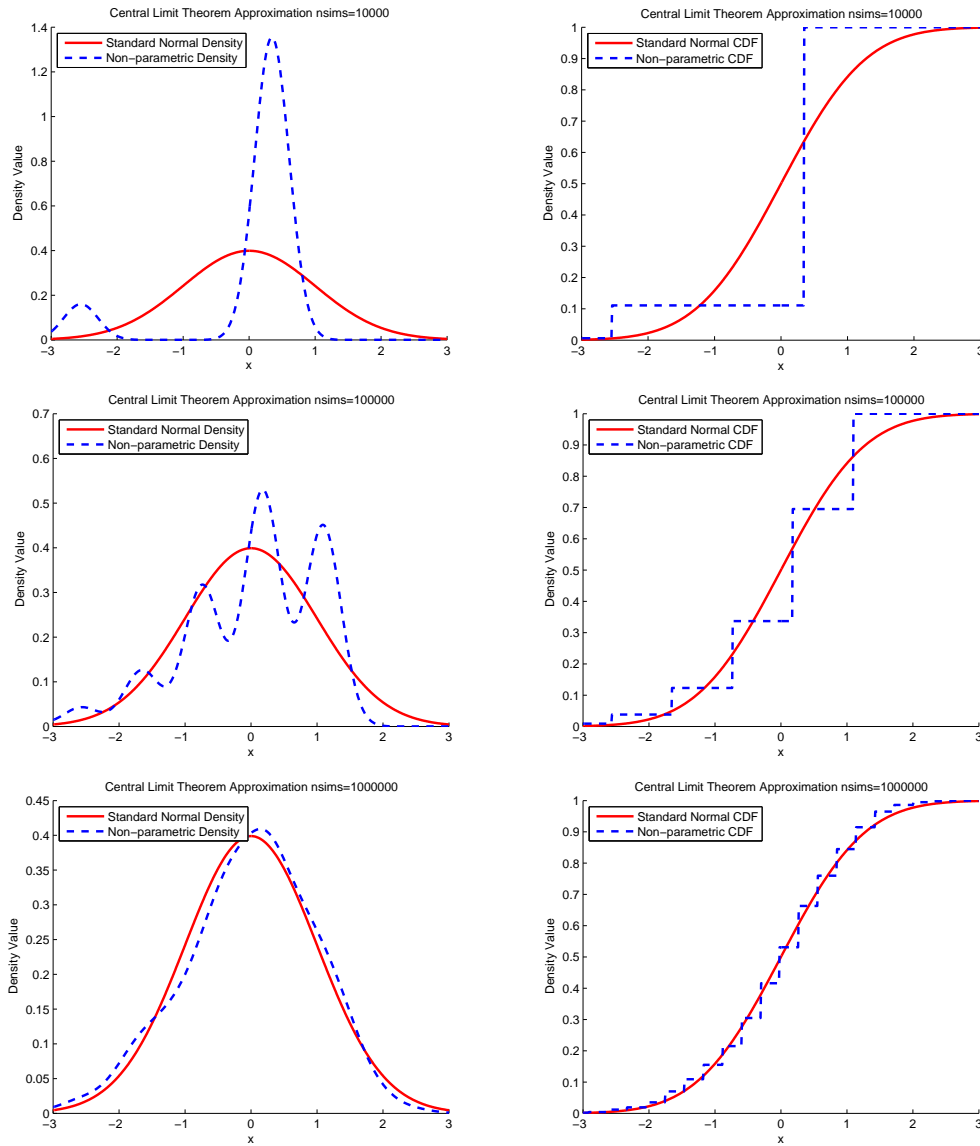
when $N = 1291764$. Thus, to ensure that the Central Limit Theorem provides a sufficiently good approximation, we need to use considerably more than 1 million Monte Carlo draws. Suppose we require that N is sufficiently large so that $\Phi_N(x)$ is also within 0.00001 of Φ . Then N would have to satisfy

$$N \geq \left[\frac{0.4785[F(39)(1-F(39))(1-2F(39))]}{0.00001[\sqrt{F(39)(1-F(39))}]^3} \right]^2 = 1.919 \times 10^{14} \quad (19)$$

This is an astronomically large number, nearly 200 trillion Monte Carlo draws, to guarantee that the actual normalized CDF Φ_N is uniformly within 0.00001 of the standard normal CDF Φ . We note that the Berry-Esseen bounds are *worst case bounds* and the actual number N needed to achieve sufficient accuracy may be far less than this number.

Figure 1 illustrates the quality of the approximation of Φ by plotting the estimated Φ_N (both as a density and as a CDF) via a Monte Carlo study involving 1000 samples of size N for $N = 10,000$, $N = 100,000$ and $N = 1,000,000$, respectively. We see that for $N = 10,000$ the Central Limit Theorem works poorly and Φ is a poor approximation to the actual finite sample distribution Φ_N . Only 11.1% of the 1000 replications resulted in an empirical distribution for which $|\hat{F}_N(39) - F(39)| < 0.00001$. The approximation improves when $N = 100,000$, and then in 69.5% of the 1000 replications $\hat{F}_N(39)$ was within 0.00001 of $F(39)$. By the time $N = 1,000,000$, the distribution Φ_N is reasonably close to Φ , though the step function nature induced by the fact that Φ_N is a standardized binomial distribution is still evident. At $N = 1,000,000$ observations, 100% of the 1000 Monte Carlo replications resulted in $\hat{F}_N(39)$ being within 0.00001 of $F(39)$.

Figure 1: Central Limit Theorem Approximations for Increasing Sample Sizes N



It is evident from Figure 1 that the quality of the approximation is better in the tails of the distribution than in the middle of the distribution, where Φ is increasing the fastest. Since we are trying to estimate a *tail probability* in this case, the necessary sample size N is far less than the maximum sample size necessary to provide a uniform approximation to Φ_N by the normal distribution Φ for all x , and in the worst case for any possible data generating mechanism. It appears from these Monte Carlo studies that approximately 1 million Monte Carlo draws is necessary to assure that with probability at least 0.999, the maximum error in estimating $F(39)$ will be no greater than 0.00001.

Note that we have calculated these bounds using the *true probability* $F(39)$ but of course this is the unknown quantity that we are trying to estimate by Monte Carlo in the first place! So there

is an additional layer of approximation to consider if we do an initial Monte Carlo experiment to estimate $F(39)$ by $F_N(39)$. As we noted in equation (9) we can use the estimated standard deviation $\sqrt{\hat{F}_N(39)(1 - \hat{F}_N(39))}$ instead of the unknown true standard deviation $\sqrt{F(39)(1 - F(39))}$ to standardize the difference between $\hat{F}_N(39)$ and $F(39)$ in equation (9). The Central Limit Theorem will still apply. However there is a serious problem in the tails since unless N is not very large, it can happen that $F_N(39) = 1$ and then the estimated standard deviation $\sqrt{\hat{F}_N(39)(1 - \hat{F}_N(39))}$ will equal 0 and it will not be possible to standardize the difference $F_N(39) - F(39)$. Thus, in practice, N must be large enough so that $\hat{F}_N(39) < 1$. In the example above, when $N = 1,000,000$, we found that $\hat{F}_N(39) < 1$ in every one of the 1000 Monte Carlo replications and the nonparametrically estimated density and CDF from these 1000 replications is very similar to the ones we obtained when we standardized using the true $F(39)$.

- D 20 points** Joey wants to know how many people you expect to select this gamble and wants you to give him a ball park estimate. Suppose you believe that people coming into the casino are expected utility maximizers and have logarithmic utility functions. Suppose the alternative to this gambling game is to play a “1 armed bandit” at a price of 50 cents per pull. The one armed bandit pays out \$100 in quarters with probability 1 in 100,000 (i.e. probability .00001). Which would a person with \$100,000 in net worth prefer: the dice game described above, or play the one armed bandit? What price does the dice game have to be adjusted to (up or down) in order for this customer to be indifferent?

Answer The expected utility of a person with logarithmic utility and wealth of \$100,000 of playing the 1-armed bandit once is

$$E\{U_{1AB}\} = p \log(100000 + 100 - .5) + (1 - p) \log(100000 - .5) = 11.51292 \quad (20)$$

using a probability of winning of $p = 0.00001$. For the proposed dice game, the expected utility for this person of playing the game is

$$E\{U_{DG}\} = p \log(100000 + 100 - 10.00193) + (1 - p) \log(100000 - 10.00193) = 11.51282 \quad (21)$$

where $p = 0.0000193$. So the expected utility of the 1 armed bandit game is higher than the dice game even though the prizes are the same and the the probability of winning the prize is nearly twice as large in the dice game, the \$10 fee of the dice game is too high relative to the 50 cent price of playing the one armed bandit, and as a result, this customer would choose to play the one armed bandit over the dice game. The fee for the dice game that makes this customer indifferent between playing the dice game and the one armed bandit is the solution F to the equation

$$E\{U_{1AB}\} = p \log(100000 + 100 - F) + (1 - p) \log(100000 - F) \quad (22)$$

where $p = 0.0000193$. Solving this equation using the `fsolve` function of Matlab, we obtain $F = .500192$. Thus, the fee for the dice game would have to be virtually the same as the fee for the one armed bandit since even though the probability of winning in the dice game is about twice as high as in the one armed bandit, the probability of winning either game is sufficiently already that the casino cannot charge hardly anything more for the dice game than the one armed bandit. In fact, given that prices must be rounded to the nearest cent, the price would have be the same for both,

50 cents. Then, since the prizes are the same but the dice game has a slightly higher probability of paying out the \$100 prize than the one armed bandit, the dice game would be less profitable than the one armed bandit game for the casino. Somehow the dice game should be modified to change either the value of the prize or the probability of winning it to provide a better “deal” to the gambler so that the gambler would be willing to pay more to play the game.

Overall, gambling is a puzzle from the standpoint of traditional economic theory of expected utility maximization, if we also believe gamblers are rational and risk-averse. Suppose we believe the casino is risk-neutral expected profit maximizer and it makes its profits on the difference between the fee F for a gamble less the expected payoff \tilde{P} , or $\Pi = F - E\{\tilde{P}\}$. A customer with wealth W and concave utility function $u(W)$ has a “reservation utility” of $u(W)$ representing their utility if they do not go to the casino at all. In order for the customer to want to come to the casino to do the gamble, then, we need

$$E\{U(W + \tilde{P} - F)\} \geq U(W) \quad (23)$$

However by Jensen’s Inequality, for a strictly concave utility function such as $U(W) = \log(W)$ and a non-degenerate gamble \tilde{P} (i.e. a gamble that has some risk of different payoffs) we have

$$U(W + E\{\tilde{P}\} - F) > E\{U(W + \tilde{P} - F)\} \geq U(W) \quad (24)$$

But if $F > E\{\tilde{P}\}$ in order to insure positive expected profits for the casino, we have

$$U(W) > U(W + E\{\tilde{P}\} - F) \quad (25)$$

Thus, the general conclusion is that an expected profit maximizing casino could never make money off of rational, risk-averse, expected utility maximizing customers. The casino is subjecting them to risk and the customers would need a “subsidy” to be willing to take on this risk voluntarily. But that subsidy would represent an expected loss of doing any gamble to the casino. So casinos must be making all of their money off of less than fully rational customers, or customers who are risk-loving or have some other reason (gambling addiction?) for coming to the casino that is inconsistent with our standard economic model of a rational, risk averse expected utility maximizer.

E 50 points You are assigned to test the fairness of dice using a “dice roller machine”. This is an automatic machine that mechanically rolls a single die 6000 times in a row. If it is fair, we expect 1000 of the draws to be 1’s, 1000 to be 2’s etc. Suppose the actual count for a particular test of a single die is 1: 991, 2: 975, 3: 997, 4: 1053, 5: 1005 and 6: 979. Can you devise a test of the hypothesis that this die is fairly balanced? If so, can you reject H_o : *this die is fair* at the 1% level of significance using your test and these data?

Answer Use a Chi-squared test to test the null hypothesis.

$$\chi^2 = \sum_{i=1}^6 \frac{(O_i - E_i)^2}{E_i} = 3.99 \quad (26)$$

where $E_i = 1000$ the expected frequency of getting a 1, 2, 3, etc in the 6000 rolls of the die, and O_i are the observed counts. It can be shown that χ^2 is distributed asymptotically as a Chi-squared random variable with 5 degrees of freedom. The value of this statistic in this case, 3.99, is at the 45th percentile of this distribution, but in order to reject the hypothesis the value of χ^2 would have

to be greater than the 99th percentile of this distribution, which can be calculated (using the Matlab `icdf` command, `icdf('chi2', .999, 5)`), to be 20.515. Since $3.99 \leq 20.515$, we fail to reject H_0 at the 1% significance level. The die appears to be fair.

The derivation of the asymptotic distribution of the χ^2 statistic is as follows. Let p be the 6×1 vector of the probabilities of the six possible outcomes on the die if it is fairly balanced. Then $p_i = 1/6$, $i = 1, \dots, 6$. Let \hat{p}_N be the sample frequencies, or estimated probabilities of these outcomes from the die-rolling machine, and let $p(i)_N$ be the i^{th} component of this vector. It can be written as a sample average of indicator functions, where the indicator function is 1 when the die roll is equal to i and 0 otherwise:

$$\hat{p}(i)_N = \frac{1}{N} \sum_{j=1}^N I\{\tilde{D}_j = i\} \quad (27)$$

where \tilde{D}_j is the outcome of the j^{th} roll of the die from the rolling machine. Let $\tilde{B}(i)_j = I\{\tilde{D}_j = i\}$. This is a bernoulli random variable with mean p_i . Assuming the draws of \tilde{D}_j are *IID* then the $\tilde{B}(i)_j$ will also be *IID* random variables as well. As a result we have with probability 1, by the Strong Law of Large Numbers,

$$\lim_{N \rightarrow \infty} p_N = p^*. \quad (28)$$

Here, p^* is the true probability vector representing the true probabilities of a roll of the die equals each of the 6 possible values, $i = 1, \dots, 6$. The Central Limit Theorem implies that

$$\sqrt{N}(\hat{p}_N - p^*) \implies \tilde{Z} \sim N(0, \Omega) \quad (29)$$

where “ \implies ” denotes convergence in distribution to a multivariate normal random vector \tilde{Z} with mean 0 and Ω is the 6×6 covariance matrix.

$$\Omega = NE\{(\hat{p}_N - p^*)(\hat{p}_N - p^*)'\} = \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_5 & -p_1p_6 \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_5 & -p_2p_6 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -p_5p_1 & -p_5p_2 & \cdots & p_5(1-p_5) & -p_5p_6 \\ -p_6p_1 & \cdots & \cdots & -p_6p_5 & p_6(1-p_6) \end{bmatrix} \quad (30)$$

Define a diagonal matrix D as

$$D = \text{diag}(p^{-1/2}) = \begin{bmatrix} p_1^{-1/2} & 0 & 0 & \cdots & 0 \\ 0 & p_2^{-1/2} & \cdots & 0 & 0 \\ 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & \cdots & p_5^{-1/2} & 0 \\ 0 & 0 & \cdots & 0 & p_6^{-1/2} \end{bmatrix} \quad (31)$$

We can write the χ^2 statistic as

$$\chi^2 = N(\hat{p}_N - p)^t D^2 (\hat{p}_N - p) \quad (32)$$

Further we have

$$\sqrt{ND}(\hat{p}_N - p) \implies D\tilde{Z} \sim N(0, D\Omega D) \quad (33)$$

We have

$$D^2\Omega = I - \Pi \quad (34)$$

where I is the 6×6 identity matrix and Π is the 6×6 matrix given by

$$\Pi = \begin{bmatrix} p_1 & p_2 & p_3 & p_4 & p_5 & p_6 \\ p_1 & p_2 & p_3 & p_4 & p_5 & p_6 \\ p_1 & p_2 & p_3 & p_4 & p_5 & p_6 \\ p_1 & p_2 & p_3 & p_4 & p_5 & p_6 \\ p_1 & p_2 & p_3 & p_4 & p_5 & p_6 \\ p_1 & p_2 & p_3 & p_4 & p_5 & p_6 \end{bmatrix} \quad (35)$$

It is not hard to show that $I - \Pi$ is an *idempotent* matrix, i.e.

$$[I - \Pi]^2 = [I - \Pi][I - \Pi] = [I - \Pi] \quad (36)$$

Also it is not difficult to show that $\Omega\Pi = 0$. This implies that $D\Omega D$ is idempotent

$$[D\Omega D][D\Omega D] = D\Omega D^2\Omega D = D\Omega[I - \Pi]D = D\Omega D. \quad (37)$$

The *Jordan decomposition theorem* implies that if Γ is an idempotent matrix, it can be represented as

$$\Gamma = O\nabla O \quad (38)$$

where O is an *orthonormal matrix* (i.e. $O'O = I$) and ∇ is an idempotent diagonal matrix, i.e. the diagonal elements of ∇ are either 1's or 0's. It is not hard to show that if $\tilde{X} \sim N(0, \Gamma)$ and Γ is idempotent, then $\tilde{X}'\tilde{X}$ is a Chi-squared random variable with $\text{trace}(\Gamma) = \text{rank}(\Gamma)$ degrees of freedom. It follows that $\sqrt{ND}(\hat{p}_N - p)$ converges in distribution to $N(0, D\Omega D)$. Since $D\Omega D$ is idempotent, we have

$$\chi^2 = [\sqrt{ND}(\hat{p}_N - p)]'[\sqrt{ND}(\hat{p}_N - p)] \implies \tilde{X}'\tilde{X} \quad (39)$$

where $\tilde{X} \sim N(0, D\Omega D)$, so it follows from the Continuous Mapping Theorem that the χ^2 statistic converges in distribution to a Chi-squared distribution with degrees of freedom equal to the trace of $D\Omega D$. It is not hard to show that

$$\text{trace}(D\Omega D) = \text{trace}(I - \Pi) = 5 \quad (40)$$

so χ^2 is approximately distributed as a Chi-squared random variable with 5 degrees of freedom under the null hypothesis.

F 70 points The file `dice-draws.txt` contains the actual results from the 6000 draws of the dice roller machine on the die tested above. You suspect that there might be a mechanical problem in the machine so that the rolls produced by this machine are not actually independently distributed draws, i.e. you suspect some sort of *dependence* in the results of successive draws from this machine. Can you think of a way of testing your suspicion using regression or some other means? Given whatever test you think up, test the hypothesis H_o : *draws from the machine are IID draws from the die*. Can you reject this hypothesis using your test at the 1% significance level? If you find dependence, how does this affect your conclusions in part E above?

Answer There are many ways to test for dependence, but an easy way (though perhaps not resulting in the most powerful test of H_o) is to do an *autogression*, that is you regress

$$\tilde{D}_t = \beta_0 + \beta_1 \tilde{D}_{t-1} + \varepsilon_t \quad (41)$$

where \tilde{D}_t is the number on the die on roll t of the dice rolling machine. If the dice rolls are independent, then $\beta_1 = 0$ and $\beta_0 = E\{\tilde{D}_t\} = 3.5$, the mean outcome. Using the data in `dice-draws.txt` the estimated regression coefficients are $\hat{\beta}_0 = 3.47$ (standard deviation 0.05) and $\hat{\beta}_1 = 0.01$ (standard deviation 0.0129). Since both of the regression parameter estimates are within one standard error of their values predicted under the null hypothesis of independence in the successive draws of the dice by the dice rolling machine, we suspect that we cannot reject the null hypothesis H_o of independence. Translating H_o to $H_o : \beta_0 = 3.5, \beta_1 = 0$ we can conduct a chi-squared test of this latter hypothesis using the statistic

$$\chi^2 = (\hat{\beta}_0 - 3.5 \quad \hat{\beta}_1 - 0) [X'X/\hat{\sigma}^2] (\hat{\beta}_0 - 3.5 \quad \hat{\beta}_1 - 0)' \quad (42)$$

It is not difficult to show that as $N \rightarrow \infty$ that the distribution of χ^2 converges to a Chi-squared distribution with 2 degrees of freedom. Calculating the statistic using the regression results, we get $\chi^2 = 0.70$. The 99th percentile if a Chi-squared distribution with 2 degrees of freedom is 9.21, so we fail to reject H_o at the 1% significance level.

Note that the program I wrote to generate dice draws and the `dice-draws.txt` file really *did* have dependence in successive draws. However the dependence in successive draws was weak and hard to detect via simple linear regression. I generated successive draws from the multinomial probabilities

$$P_t(j) = \frac{\exp\{\log(1/6) + \delta I\{\tilde{D}_{t-1} = j\}\}}{\sum_{i=1}^6 \exp\{\log(1/6) + \delta I\{\tilde{D}_{t-1} = i\}\}} \quad (43)$$

where $\delta = 0.02$. If $\delta = 0$ then $P_t(j) = 1/6$ and there is no dependence in successive draws. However if $\delta > 0$, the chance of drawing a value j is higher if the previous roll of the die was also equal to j (so that $I\{\tilde{D}_{t-1} = j\} = 1$) so the data generating mechanism did actually display a small degree of positive serial correlation in the successive draws of the die, but the regression was unable to detect it. Motivated students might consider estimating a multinomial logit model of the *transition probability* $P(\tilde{D}_t | \tilde{D}_{t-1})$ given by the multinomial logit probability above by maximum likelihood using the `dice-draws.txt` data. Then a more powerful test of H_o could be constructed by testing the hypothesis that $\delta = 0$. However I did not expect students to consider doing this, since it would require clairvoyance about the data generating mechanism that I used to generate the `dice-draws.txt` data.

2. [200 points] Suppose a consumer has an inventory of groceries at home and shopping trips occur when the consumer want to *adjust their desired inventory of groceries upward*. Let the desired *adjustment* to the level of inventories of groceries be given by the regression equation

$$y_{i,t} = X_{i,t} \beta + \varepsilon_{i,t} \quad (44)$$

where $y_{i,t}$ is household i 's desired adjustment to their stock of groceries in week t and X contains explanatory variables affecting the desired adjustment, including household income, the household's estimated current stock of groceries, number of people in the household, with breakouts for number of teenagers in the household, dummy variables for whether there are significant sales occurring during the week, and so forth.

- A. **50 points** It is natural to assume that if $y_{i,t} < 0$, i.e. the household is “overstocked” with groceries and would actually like to reduce its stock of groceries, that the household might actually throw away of older stale or spoiled food or groceries, but when $y_{i,t} > 0$ then the household would have a motive to go shopping to replenish its stock of groceries. Assume that the survey we have does not record when household throws away groceries, so we do not observe cases where $y_{i,t} < 0$ and further, due to positive transaction and hassle costs of going shopping, household i will not actually go shopping unless $y_{i,t} > K_i$ where $K_i > 0$ is some threshold that must be passed to make it worthwhile for someone in the household to go out shopping to increase the stock of groceries. If the survey then only records the $(y_{i,t}, X_{i,t})$ observations for households that have actually gone out shopping (and thus filled out a diary recording the amount spent, $y_{i,t}$, as well as any other information that is changing at the weekly level, $X_{i,t}$ such as whether there was a significant sale that week, and what the inventories of groceries were at home that week, etc), will a regression limited to just the data on the $(y_{i,t}, X_{i,t})$ observations recorded in the consumer diaries and reported in the survey result in consistent estimates of β and the K_i parameters? Please explain your answer as carefully as possible to receive full credit.
- B. **50 points** If the length of time we observe a particular household i is relatively short, say for 4 weeks, but we have a large number of households i , say N households where $N > 1000$, do you think it will be possible to consistently estimate both β and the N household-specific thresholds K_i , $i = 1, \dots, N$? If you believe is it possible sketch an estimator and an argument for the consistency of your estimator. If you think it is not possible, describe as carefully as you can why you think it is impossible to consistently estimate these parameters by regression or any other means.
- C. **50 points** Suppose we are willing to make additional *distributional assumptions*. In particular if you are willing to assume that $\{\varepsilon_{it}\}$ are *IID* $N(0, \sigma^2)$ random variables and that the K_i are also $N(\mu, \gamma^2)$ random variables and that K_i and ε_{it} are independently distributed for each i , and that for any household i $\varepsilon_{i,t}$ and $\varepsilon_{i,s}$ are independently distributed for $s \neq t$, and finally, that the random variables $(K_i, \{\varepsilon_{i,t}\}_{t=L_i}^{\bar{t}_i})$ are distributed independently of the random variables $(K_j, \{\varepsilon_{j,t}\}_{t=L_j}^{\bar{t}_j})$ for any two households $i \neq j$, can you show how you can use these extra prior assumptions to construct a consistent estimator of the parameters $\theta = (\beta, \sigma^2, \mu, \gamma^2)$?
- D. **50 points** Suppose the household diary records $y_{i,t}$ on *every* week that the household is in the survey by asking a member of the household to directly report their *subjective assessment of their desired adjustment of groceries* $y_{i,t}$ in each week t . If you had these data, would ordinary linear regression (OLS) of the grocery inventory adjustment model (44) result in consistent estimates of β and the $\{K_i\}_{i=1}^N$ assuming that we have a relatively large number of households N in our sample but we assume these households only for 4 consecutive weeks? If you think OLS might not be able to consistently estimate all of these $4N + K + 1$ parameters (where K is the number of regression parameters β and the extra 1 is for the unknown $\sigma^2 = \text{var}(\varepsilon_{i,t})$), can you think of some other estimator that would be able to at least estimate the parameters (β, σ^2) , but assuming that you are NOT willing to impose a normality assumption on the $\{K_i\}$ or the $\{\varepsilon_{i,t}\}$? If you think it is possible, do you need to impose any assumptions on the independence of the error terms $\{\varepsilon_{i,t}\}$ across households i or across time, t ?
3. **[200 points]** Suppose that you have just landed a job at a top economic consulting firm and that you are having a disagreement with your boss about an econometric model. You think that the data are generated by

$$y = X\beta_o + u \quad (45)$$

where $\beta_o \in R^k$, X is $n \times k$, $y \in R^n$ and $\{u_i\}$ are IID random variables with mean 0 and variance σ_o^2 . On the other hand, your boss says that years of experience point her to the model

$$y = X\beta_o + Z\rho_o + u \quad (46)$$

where Z is also $n \times k$ since, after all, “it can’t hurt to add more variables to the model”. You are not sure about that so you set out to investigate her claim.

- A. **[65 Points]** Suppose that your model is the correct one but that you estimate the parameters by applying OLS to the competing model. Derive an expression for $\hat{\beta}$ and show that this estimator is unbiased, stating clearly any assumptions that you make. Then derive an expression for the variance of this estimator. Apply OLS to the correct model and call the estimator $\tilde{\beta}$. Repeat the previous steps.
- B. **[30 Points]** How can we compare our two estimators? Invoke a well-known theorem to make your argument and then show explicitly the difference in efficiency between $\hat{\beta}$ and $\tilde{\beta}$. Which estimator is more efficient? When will there be no loss of efficiency? Start with the one-parameter case and then extend your argument to the k parameter case.
- C. **[85 Points]** Now, on the contrary, suppose that it is your boss that has the correct model. Repeat the calculations you performed above. Suggest some criteria by which to evaluate the performance of your estimator and derive an expression to compare to the variance of the correctly specified model. Is one estimator unambiguously better than the other? Discuss.
- D. **[20 Points]** What do you conclude about your boss’ claim? Do your conclusions depend in any way on the size of the sample to be analyzed?
4. **[200 points]** Let y be $n \times 1$, X be $n \times (k + 1)$ and suppose that $E(y|X) = X\beta$. Consider the linear programming problems

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \left\{ \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^k x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^k \beta_j^2 \right\} \quad (47)$$

and

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^k x_{ij}\beta_j)^2, \quad (48)$$

subject to

$$\sum_{j=1}^k \beta_j^2 \leq t \quad (49)$$

- A. **[20 Points]** Show that there is a one-to-one correspondence between the parameters λ and t above. Offer an interpretation of these parameters and compare the objective functions above to the one for OLS.

- B. **[15 Points]** Provide a convincing argument for why the intercept is not penalized in the problems above.
- C. **[60 Points]** Now suppose the data has been centered so that the data matrix X has k columns. Rewrite the first problem above in matrix form and show that the solution $\hat{\beta}$ is a linear function of y . Be careful to provide conditions that allow $\hat{\beta}$ to be well defined.
- D. **[10 Points]** Is the problem well defined if $X'X$ is *not* of full rank?
- E. **[60 Points]** Find $E\{\hat{\beta}|X\}$. Use the conditions you outlined in part C above to show that $\hat{\beta}$ is biased for β unless $\beta = 0$.
- F. **[15 Points]** Find the probability limit of $\hat{\beta}$ as $n \rightarrow \infty$. Is $\hat{\beta}$ consistent for β ? What happens to $\hat{\beta}$ as $\lambda \rightarrow 0$?
- G. **[20 Points]** Now let $\lambda = an$ where $a > 0$ is fixed and $n \rightarrow \infty$. Find the probability limit of $\hat{\beta}$ as $n \rightarrow \infty$.