**Boston University Questrom School of Business**

**MF 793 – Fall 2021**

**If you do not also enter all names in GradeScope, the omitted names will get a ZERO.**

Eric Jacquier

**Problem Set 2**

**Due Monday October 11th at 11pm Boston Time**

Problems turned in after the deadline are not graded

* Do the Problem Set in groups of four at the most – students can be in different sections.
* Turn in one single copy for the group on the Gradescope site.
* No email, no paper submission, will be accepted.
* Write solutions in this word file, insert figures from R and hand-written material as pdf graphics. Then save the file as PDF.
* **A properly formatted and spaced file will be posted in a couple days. Do not start filling this file. Another theoretical problem may be added.**
* **To get a check, you need to answer all the questions.**

**If you do not also enter all names in GradeScope, the omitted names will get a ZERO.**

**If you do not also enter all names in GradeScope, the omitted names will get a ZERO.**

**If you do not do this, you can not get a check plus**

* ALL discussion and math questions answered.
* All math questions hand-written
* All Figures professionally made with X and Y axes labels and title and fig. numbers
* Tables must have row and column names, title and table number.
* Numbers in the tables must **not** contain too many useless or irrelevant digit, use your common sense as to how many digits to report in a Table. Otherwise it looks like you have no idea what matters.
* All R code as an appendix must be at the back of the homework.

Type the (up to) four team member names below. **AND**

**If you do not also enter all names in GradeScope, the omitted names will get a ZERO.**

**Last Name First Name Section (D1 or D2)**

**1**

**2**

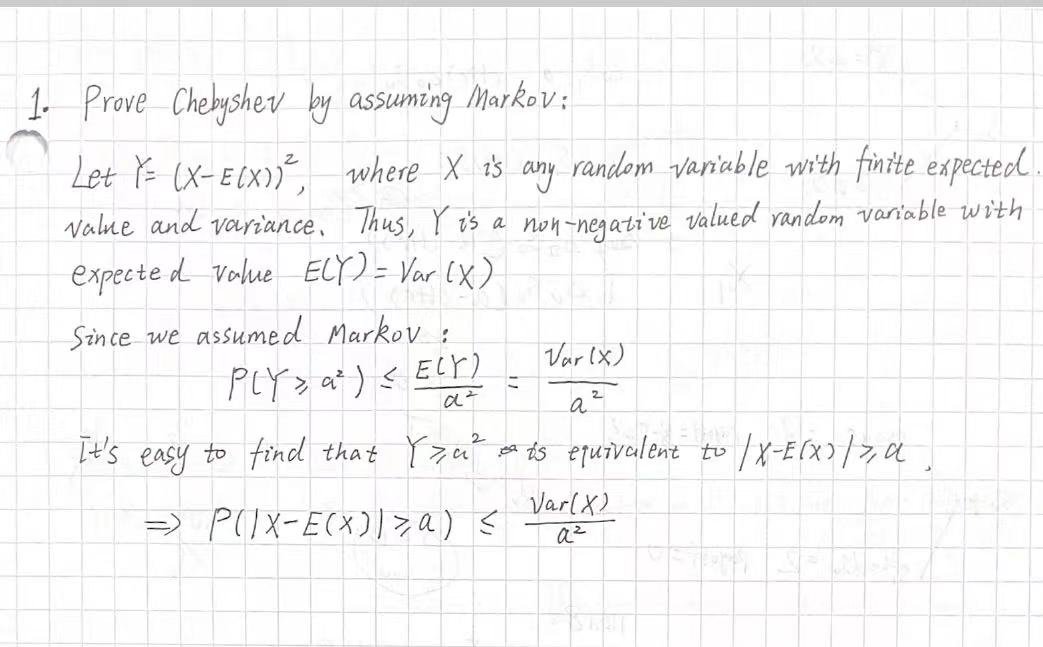
**3**

**4**

**If you do not also enter all names in GradeScope, the omitted names will get a ZERO.**

**Problem 1: Prove Chebyshev by assuming Markov**

As the title says! Assume Markov and prove Chebyshev. Answer must be hand written or the question will get zero.

****

**Problem 2: Conditional, marginal, iterated expectation, simulating distributions**

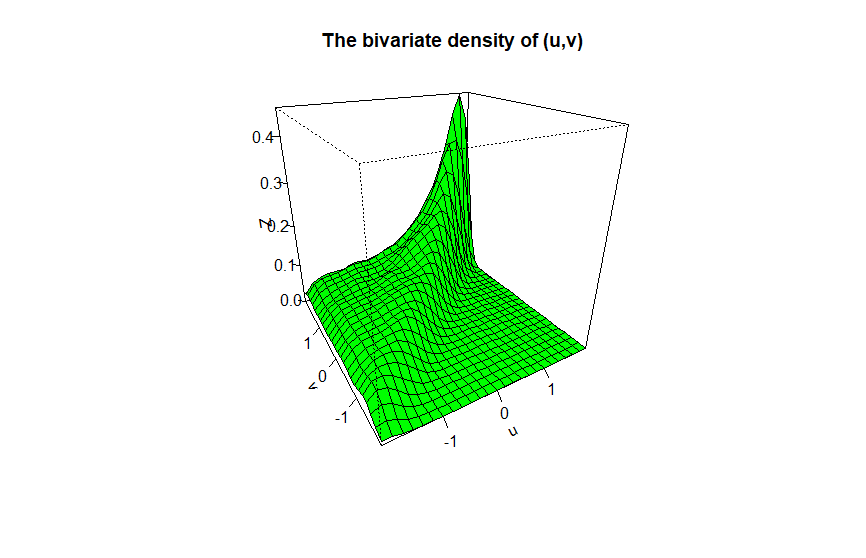
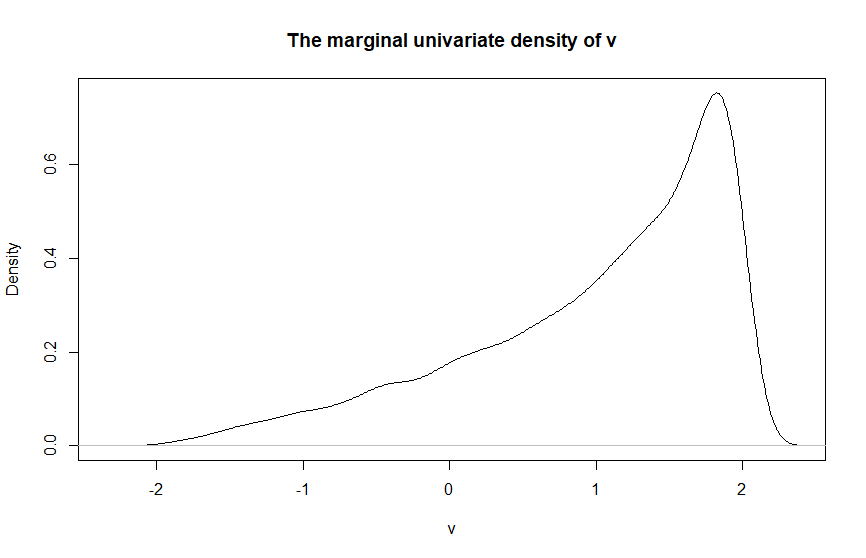
u has marginal density: p(u) ~ Uniform (-2,2). v has conditional density p(v|u) ~ Uniform(u,2).

1. Simulate 10,000 draws of u, and then a draw of v|u for each draw of u. Then use the *density* and the *persp* commands to plot, specifically:

The marginal univariate density of v: p(v) **density** command – Figure 1

The bivariate density of (u,v): p(u,v) **persp** command – Figure 2.

Here’re two plots below:

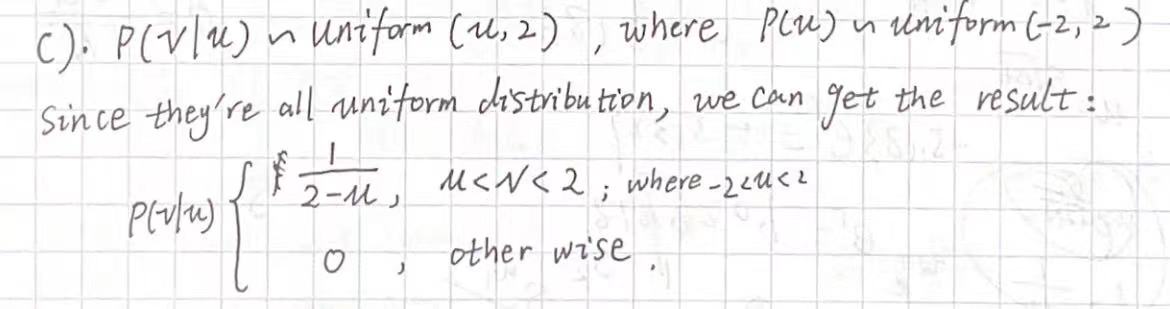


1. From your simulated data, what is your *Monte Carlo estimate* of E(v). Give a 95% confidence interval for that estimate (See moment estimation lecture note).

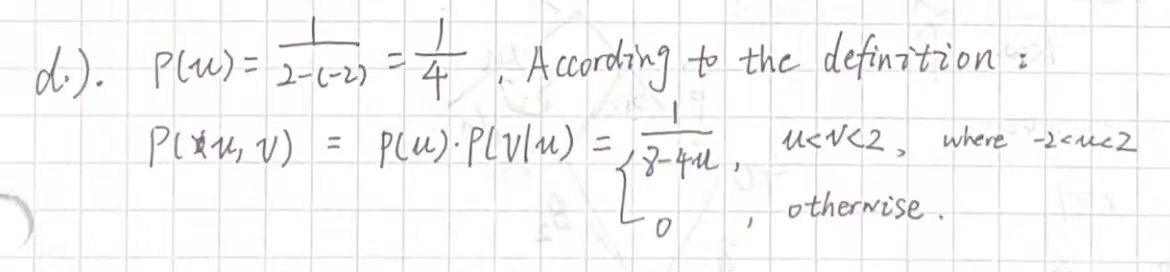
The Monte Carlo estimate of E(v) is about 0.992 ≈ **1**

And the 95 confidence interval for the estimate is about **[0.9745756 , 1.0093265]**

1. Write the theoretical conditional density p(v|u). (proof by hand or zero points)

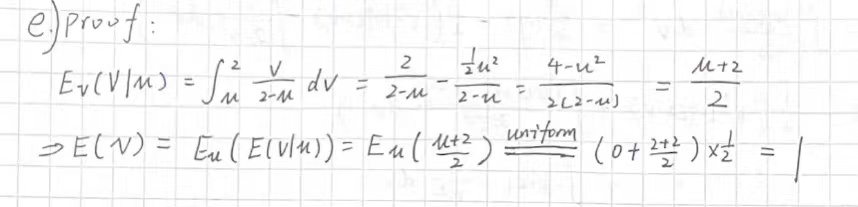


1. Write the theoretical joint density p(u,v). (proof by hand or zero points)



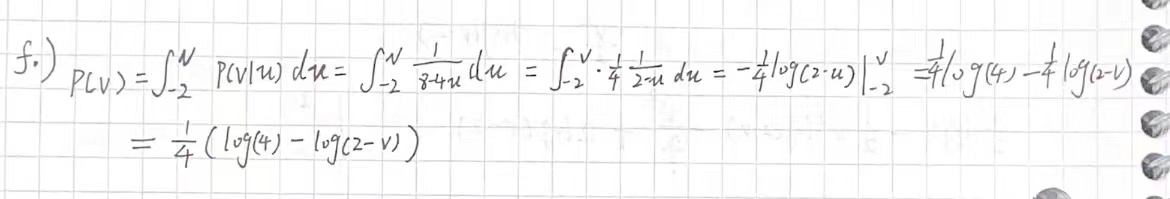
1. Compute Ev(v|u). Then use the iterated expectation rule to compute E(v) = Eu(Ev(v|u)). Proof by hand or zero points

Ev(v|u) = **(u+2)/2**



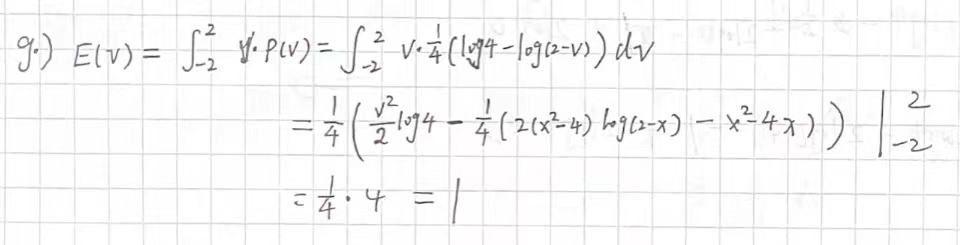
E(v) =**1**

1. Compute the marginal density p(v) by integration of p(v,u). proof by hand or zero.



\

1. Use p(v) to compute E(v) by integration. Hope you find the same result as in e)! proof by hand or zero

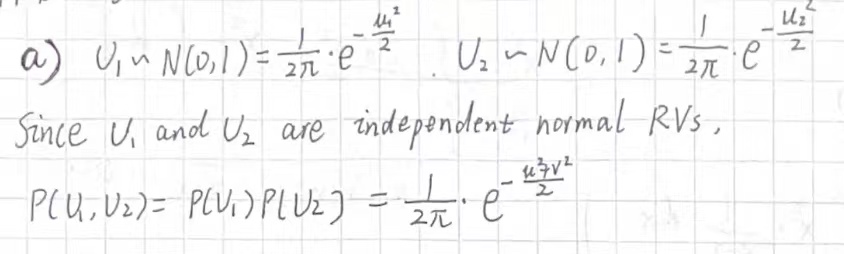


The result of g) equals to the result of e), so we got the right answer.

**Problem 3: Multivariate change of variable, the bivariate normal distribution**

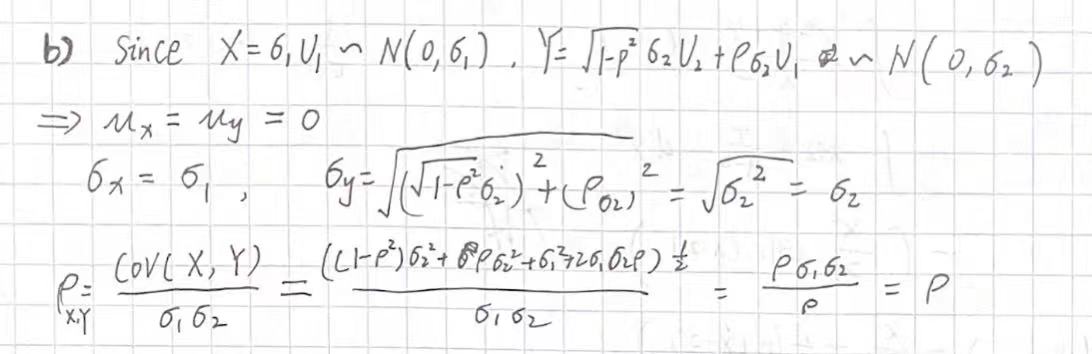
**For the entire problem 3, zero points if not by hand**

a) U1~N(0,1) and U2~N(0,1) are independent normal RVs. Write the joint density of U1 and U2, p(U1,U2).

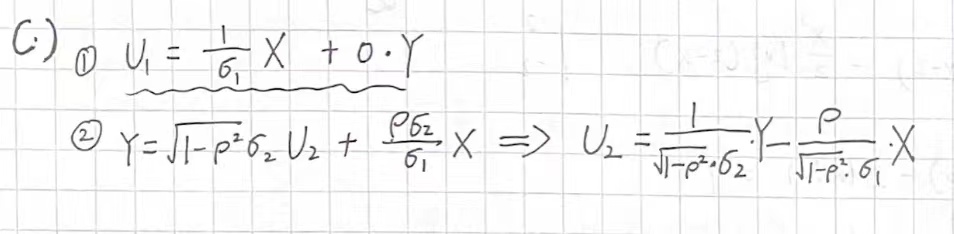


b) The bivariate function g(U1,U2) = (X,Y) is: X = σ1 U1, and Y = σ2 U2 + ρ 1.

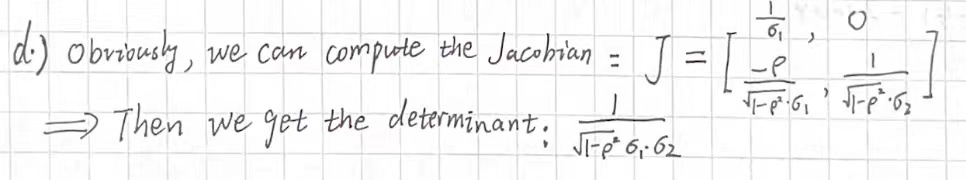
Compute μX, μY, σX, σY, and ρX,Y,



c) Write the bivariate inverse transformation: (U1,U2) = g-1(X,Y)

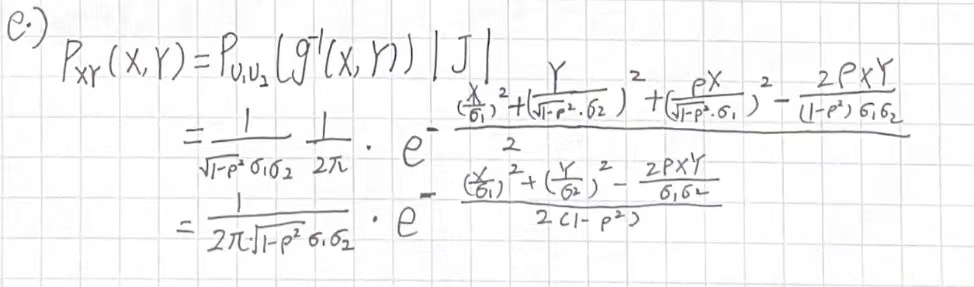


d) Use the change of variable method for multivariate densities to write the joint density p(X,Y) of two correlated zero-mean normal RVs. Same formula as in class but the absolute value of the derivative of g-1(y) is replaced by the absolute value of the determinant of the 2x2 matrix of partial derivatives, . It is called the**Jacobian** of the inverse transform.



(没怎么看懂他的意思，是让我算J还是直接算Joint density，直接算的话好像和e重复了)

e) Write pXY(X,Y) = pU1 U2(g-1(X,Y)) |J|. You just proved the joint bivariate normal density

****

**Problem 4: Data mining for persistent winners**

Your boss won’t give up this funds performance story. She goes “ Maybe we can get better results. Why did you use 15% as winning cut off, why not top 20% or 10% or 5%, to find the really persistent good funds. Why don’t you search for the winning cutoff that will show the strongest persistence? Then we could invest in these funds. Same for the market, why don’t you look at the funds that don’t just beat the market but those that beat it by some margin to search for persistence. Maybe the funds that beat the market with some margin have better persistence”.

a) Just thinking about this gives you a headache, but you remember a mandatory reading in your basic Financial Econometrics course. It was about exactly what is wrong with your boss’s thinking. Time to reread this in details to refresh your mind. Write a paragraph answering your boss, politely of course, why her approach is tricky at best.

Quote and mention relevant passages of the article.

Give her an example of why it would not be a good idea to do what she suggests.

b) This did not work, your wants you to “prove it” with a simulation. You illustrate data mining in a simple case:

A quant. picks 100 random funds, computes the t-stat of their mean return against the null H0 that the funds have zero expected return. She uses the 95th percentile to pick the funds which performance “rejects the null”, and report the top two funds to her boss. Since the simulated funds returns will all come from the null distribution, we know there is no “performance”, it’s all randomness. This will illustrate whether the quant’s procedure is biased against the null! Let’s go. Specifically:

Simulate 4 years of normal daily returns (250 days per year) for 100 funds, with all the same **annualized** mean and standard deviation μ=0, σ = 0.2/sqrt(250). To keep it simple, fund returns are uncorrelated across funds and time. Compute the 100 t-statistics against H0: μ=0.0. Save the top t-statistics, the 5th best t-statistic, and 1 t-stat randomly chosen out of the 100. Do this 10,000 times. You have 10,000 draws of the 3 t-stats under the null. Fill Table 1 with

row 1) the average of the 10,000 t-stats for each of the 3 t-stat.

row 2) the fraction of 10000 times you rejected H0 using the 5% top cutoff value – the so-called **type I error** also known as the **size of the test**.

row 3) the 95% empirical quantile of each t-statistics. (95% because we use a 1-sided test.) That is the actual 95th percentile of your 10,000 draws.

Question: What cutoff value do you use to decide whether to reject the null at the 5% level for this ***one-sided*** test, the 95th percentile of the t density?

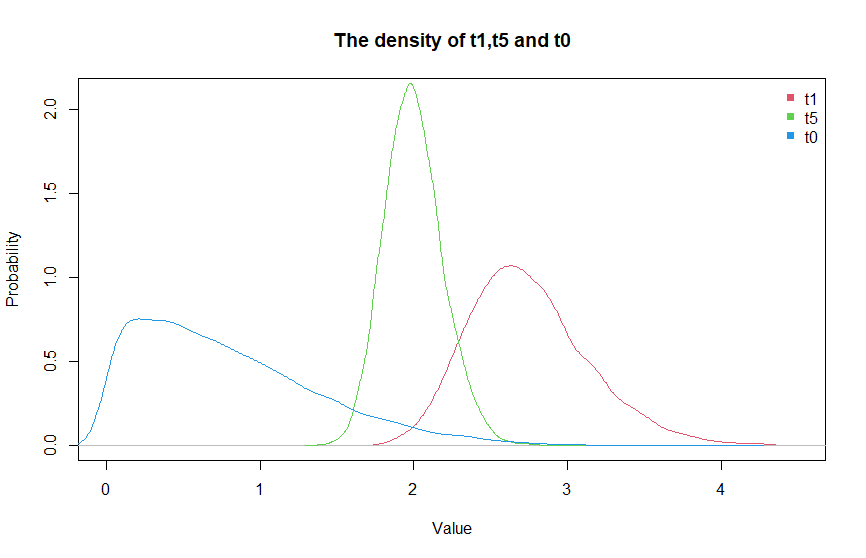
*Answer*: According to the T-test form, I decided to use 1.96 as the cutoff value.

c) Fill in Table 1

Table 1: Simulated distribution of t-tests.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Highest and 5th highest of 100 t’s | | Randomly chosen t |
|  | t1 | t5 | t0 |
| Average t | 2.753 | 2.000 | 0.798 |
| Pr.(reject H0|H0 true) | 0.993 | 0.566 | 0.049 |
| Empirical top 95th % of ti | 3.661 | 2.407 | 2.259 |

d) In figure 1, plot the four empirical densities (t1, t5, t0 the randomly chosen t) using the density command. Use a different color for each density



e) explain your results in a few sentences:

If you do 100 unrelated “5% level” tests, how often do you reject the null if the null is correct?

5%

What is the mean of a Student-t?

E(t) = 0

What is the mean of the highest of 100 Student-t’s, of the 5th highest of 100 t’s?

Mean(t1) =

Mean(t5) =

Are the highest, the 5th highest, of 100 Student-t’s distributed as Student-t’s? Why?

e) Use the numbers in Table 1 to propose an adjustment to the data mining bias.

**Problem 5: Properties of stock returns**

You will use the relevant files on our data folder.

a) Fill the following table. The daily estimate uses the appropriate data frequency to estimate μ and σ. The next 2 columns transform the direct estimate into an annual estimate by aggregation.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Direct Estimate | | By aggregation | |  |  |  |
|  | μ | σ | μann | σann | Sk | Ku | ρ(1) |
| Log(1+R) daily | 0.00063 | 0.00892 | 0.159 | 0.142 | -2.41 | 52 | 0.12 |
| Log(1+R) weekly | 0.00306 | 0.01966 | 0.160 | 0.140 | -0.69 | 8 | 0.05 |
| Log(1+R) monthly | 0.01329 | 0.04477 | 0.160 | 0.154 | -1.29 | 9 | 0.04 |
| Log(1+R) annual | 0.15952 | 0.11758 | 0.160 | 0.118 | -0.53 | 2 | -0.39 |

b) Compare the annual direct estimate to the daily, weekly, monthly “aggregated” estimates”. What do you expect and is it what you see?

For the mean: Explain

For the variance: Explain