Boston University Questrom School of Business
MF 793 – Fall 2021

Eric Jacquier

Problem Set 1
Due Sunday September 26th at 11pm Boston Time
Solutions

- 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.**
- **To get a check, you need to answer all the questions.**

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, starting at the top of a new page.

Type the (up to) four team member names below.
Make sure to also enter the names when you upload on Gradescope.

	Last Name	First Name	Section (D1 or D2)
1			
2			
3			
4			

Problem 1: Counting

*The new MBA class has arrived, all 171 of them. The social events director wants to know the probability that exactly 5 students have the same birthday, and the rest **all** have different birthdays. Use a 365 day year.*

*a) Write the theoretical formula **by hand**.*

$$\frac{\binom{171}{5} * 365 * 364 * \ldots * (365 - 166 + 1)}{365^{171}}$$

b) Explain in words how you came up with the numerator, and the denominator.

Numerator:

Which 5 out of the 171 have the same birthday, times on what day it is (365).
The remaining 166 students can have their separate birthday on 364 days for the first, 363 days for the second, etc…

Denominator

All possible combinations of birthdays. Each student can have her birthday (or not) on any of 360 days.

c) Compute it in R and give the result. Show the code you used here below.

```
logprob<-log(choose(171,5))+lfactorial(365)-lfactorial(365-165)
 -171*log(365)
exp(logprob)
```

p = 1.32 10^{-26}.

d) Did you encounter a problem computing the R solution ?

Must use logs, then exponentiate. Factorial overflows at 171.

- The command **lfactorial** in R directly computes the log of a factorial, avoiding the overflow.
- We can log everything else.
- Ratios of overflow numbers are computed by going through the log of each part and then exponentiating.

e) What is the probability that at least (possibly more) 2 students share a birthday? Give the formula and the final result

```
The easiest of all birthday questions: 1 minus probability no
students share a birthday

1 – [ 365 * … * (365 – 171 + 1) ] / 365^171


1-exp(lfactorial(365)-lfactorial(365-171+1)-171*log(365))
```

= 1 to 12 decimals.

f) Write the definition of the Γ (gamma) function. What is its relation with the factorial function.

We often need to compute density ordinates which include the gamma function in the normalization constant.

Γ(x) = $\int_0^\infty t^{x-1} e^{-t}\, dt$

Recall (you need to know this result) that: Γ (n) = Factorial(n-1)

.. even though the gamma is a function on the real line and factorial is for integers.

In that setup, n is often a sample size which can be large, think of 100s or 1000s. Then the Gamma function overflows.

R has the log of the gamma function: **lgamma(x)**

Problem 2: Conditional probabilities, total probability theorem, forecasting stock returns

The folks at TAN Inc. (TisAllNoiz) don't know much finance. They do know that people really care whether the market return is positive. They want to start a weekly financial letter which (seems to) predict the direction of the weekly S&P500 return. They consider three strategies:
1) always predict >0 return,
2) always predict <0 return,
3) randomly predicts >0 return 60% of the time.
Your published research shows that the weekly market return is positive with probability p(U)=60%, and is unpredictable. They must have found this interesting, so they hired you as a consultant for this project.

a) Give them the weekly probability of success, and the expected number of correct predictions after 52 weeks for each rule..

Rule 1:

p(C) = p(C|U)p(U)+p(C|D)p(D) = 1*0.60+0*0.40 = **0.60**

Expect 52*0.6 = **31**.2 correct predictions

Rule 2:

p(C) = 0 * 0.60 + 1 * 0.4 = **0.4**

Expect 52*4 = **21**.2 correct predictions

Rule 3:

p(C) = p(C|up) p(up)+p(C|down)p(down) = 06 * 0.6 + 0.4 * 0.4 = **0.52** **<---- !!**

Expect **27** correct predictions

b) Explain in words, what is wrong with strategy 3) aka the "randomizing strategy".

You lose information with the randomization strategy.

Up has the higher probability of occurrence, **ANY** random (non conditional) strategy that is down sometimes has a lower performance than "up all the time".

Why do people even think like this ?? This is the RAF life jacket vs parachute story!
This bears to the difference between **rational (expectation)** behavior and **adaptive behavior.** It may be a **population survival trait**.

Under some circumstances, a population where individuals exhibits rational behavior has fewer chances of surviving extinction than one with adaptive behavior.

Say the village is in the valley and can all drown in a flash flood or be crushed by a buffalo stampede (1% chance) or be on top of the mesa and die in a freak snow storm (1.5% chance).
Where does a rational go to live? What does that do to the risk of elimination ?

*c) TAN is getting serious; they want to do **conditional** prediction. The rules in question a) were unconditional rules. They ask you to investigate conditional rules, maybe markets are not efficient ! Even though you already know the answer, you will be able to charge them for some data analysis. You hop on to Ken French's web site and download the **weekly** US stock market excess return over the risk free rate.[1] You use only data from Jan. 2012 to Dec. 2020. Of course, you need to assume that future returns will behave consistently with these data. You will write a nice report for Tan; three versions of the two-way table:*

1) simple counts 2) joint probabilities 3) conditional probabilities.

- **470** weeks over 9 years, about 52 weeks per year as expected.
- Should we use total returns or excess returns?
  Excess returns is more interesting but it makes no difference here.
- We find $\widehat{p(U)}$ = 0.62 ! You could remove the week with Rm = 0 from the count.

Simple Count Table

	$R_t < 0$	$R_t > 0$
$R_{t-1} < 0$	64	113
$R_{t-1} > 0$	114	176

Joint Probability Table

	$R_t < 0$	$R_t > 0$
$R_{t-1} < 0$	0.14	0.24
$R_{t-1} > 0$	0.24	0.38

Conditional Probability Table ($R_t \mid R_{t-1}$)

	$R_t < 0$	$R_t > 0$
$R_{t-1} < 0$	0.36	0.64
$R_{t-1} > 0$	0.39	0.61

Make sure you can compute these probabilities cold.
$P(R_t < 0 \mid R_{t-1} < 0) = P(R_t < 0 , R_{t-1} < 0) / P(R_{t-1} < 0) = (64 / 467) / (177 / 467) = 0.36$
$P(R_t > 0 \mid R_{t-1} < 0) = 1 - 0.36 = 0.64$
$P(R_t > 0 \mid R_{t-1} > 0) = P(R_t > 0 , R_{t-1} > 0) / P(R_{t-1} > 0) = (176 / 467) / (290 / 467) = 0.61$

*d) Use the numbers in the conditional table to give the **unconditional probability** of success of the rule that predicts that every week the market is up (down), it will go up (down).*

Total Probability Theorem again – every time, to compute a marginal probability!

$$\mathbf{P(C)} = p(C|R_{t-1}>0)\ p(R_{t-1}>0) + p(C|R_{t-1}<0)\ p(R_{t-1}<0)$$
$$= p(R_t>0|R_{t-1}>0)\ p(R_{t-1}>0) + p(R_t<0 \mid R_{t-1}<0)\ p(R_{t-1}<0)$$
$$= \mathbf{0.513}$$

e) Given your tables, can a conditional rule improve on this rule?

The conditional rule does the same as the random rule, and is soundly beaten by the "always up" rule!

f) Given these results, how efficient do you think the US market is?

The stock market is brutally efficient, we can't predict its weekly direction.

[1] It is a good time to go on Ken French's web page and look at all the data available there. You can't live in the data based portfolio management space without knowing this data source. There you will find explanations on the data and collection process. Always read the details and description of the data

Problem 3: Bayes rule and conditioning properly

The Wales Cargo company company is in trouble for account manipulations. The North East region manager, Mrs Head Bump, has been instructed to shut down two of three branches in Metro Boston. She communicated to the branch managers of Burlington (Mr Bean), Natick (Mrs Natty), Needham (Mr Veegan), that two of them will be fired and their banks closed, only one will keep his/her job. They all have the same probability of being fired: p(B) = p(N) = p(V) = 2/3.
Mrs. Head Bump knows which branch will remain open but obviously she must not tell. At a virtual not-so happy hour (and maybe over one too many Summer Ales), she confided to the Burlington manager that the Needham branch would close.

- *Mr Bean tells himself: Good news! I now know that Needham will close, so either my branch or Natick will close. My probability of being fired is only ½, not 2/3.*
- *Mrs. Head Bump realizes that she broke the rule of silence, but she tells herself: "Either Natick or Needham must close anyway, so I have given Mr Bean no information on **his** branch, so he should still think his probability of being fired is 2/3".*

Who is wrong, Mrs Head Bump or Mr Bean? It is a question of proper conditioning!
Denote "HB" the event: Mrs Head Bump tells Mr Bean that Needham will close (V will be fired).

a) Compute p(B|HB). Was Mrs Head Bump or Mr Bean correct?

Bayes: P(B|HB) = P(HB|B) P(B) / P(HB) So we need to compute P(C)

Total Probability Theorem for the denominator as usual but **potential mistake 1**:
Can we write: P(HB) = p(HB|B) p(B) + p(HB|V) p(V) + p(HB|N) p(N) ?
No, because B, N, and V overlapping events since 2 branches will be close.

To use the Total Probability Theorem we must use "keeping the job" as an event. These are non-overlapping since only one can keep her job.
So, "not B" means Mr Bean is not fired

P(HB) = $p(HB \mid \bar{B})$ (1-p(B))+ $p(HB \mid \bar{V})$ (1-p(V)) + $p(HB \mid \bar{N})$ (1- p(N))
= [1/2 + 0 + 1] * 1/3 = **1/2**.

$p(HB \mid \bar{N}) = 1$ clearly: if Natick does not close, Mrs Head Bump tells Mr B that V will be fired
$p(HB \mid \bar{V}) = 0$ Mrs Head Bump can't tell Mr Bean that V will be fired if he won't !

Now we can finish Bayes Theorem:
P(B|HB) = p(HB|B) p(B) / p(HB) = (1/2)* (2/3) / (1/2) = **2/3**

Mrs. Head Bump is correct, she revealed nothing to Mr. Bean, his probability of being fired is still 2/3.

b) What "wrong ?" conditioning did Mr Bean use to come up with p(B| ?) = 1/2

Whether Mr Bean knows Bayes Theorem or not, the mistake (potential mistake 2) he most likely made was to confuse two different events

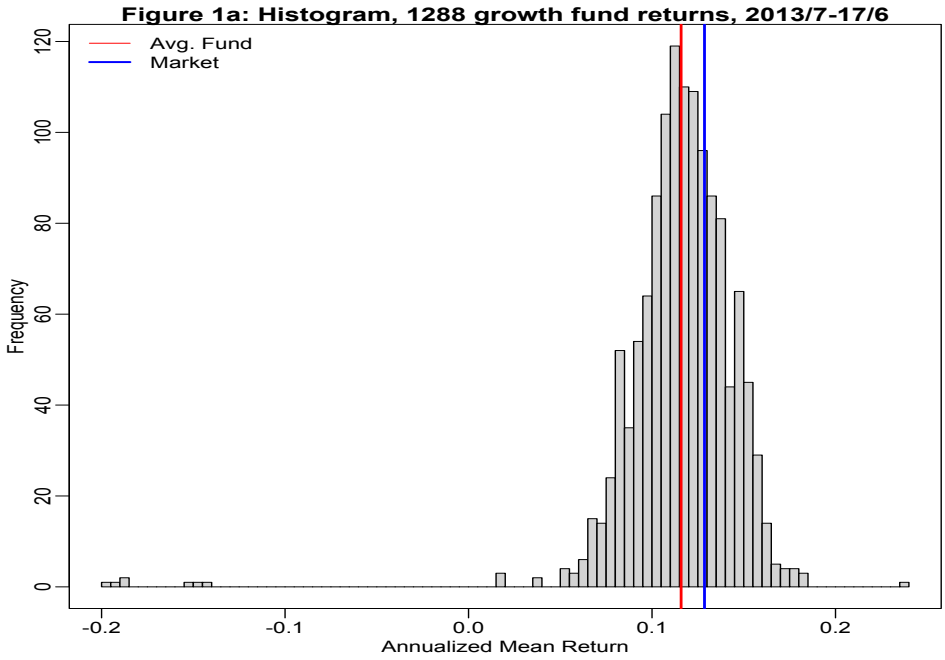
- " Needham will close" (probability 2/3)
- "Mrs Head Bumps tells him that Needham will close" (probability 1/2)

If he conditions on the event "Needham will close", we have
p(B|V) = p(V|B) p(B) / p(V) = (1/2) * (2/3) / (2/3) **= 1/2** !

Problem 4: Conditional probabilities, fund performance

Your boss wants a quick and not dirty recent performance analysis of mutual funds. She reads that most funds don't beat the market but can't get a recent review of fund performance.
You collected monthly returns from Jan. 2010 to June. 2021 for 1288 funds fully invested in the stock market, file funds-1288g-monret.csv in the Data folder.
You know you can go to Ken French's data web site and get the monthly return on the market index for the exact same period, so you do it!
a) Compute the ***average monthly return*** *for each fund for the period (2013/7-2017/6), and the market.*

- *In Fig. 1 Plot a histogram of these 1288 average fund returns. Annualize these monthly averages by multiplying them by 12 so they have kind of an annual magnitude to them. Add a vertical bar in black for the average* ***of the 1288 averages****, and a vertical bar in blue for the market average.*



- *What % of funds beat the market for that period?* 31%
- *What would you expect the result to be if the* ***fund managers were*** *randomly picking stocks ?*

  If fund managers had no special ability, for example were picking stocks randomly, we would expect 50% of them to beat the market.

- In a couple **sentences** tell your boss what you think of mutual fund performance for the first period.

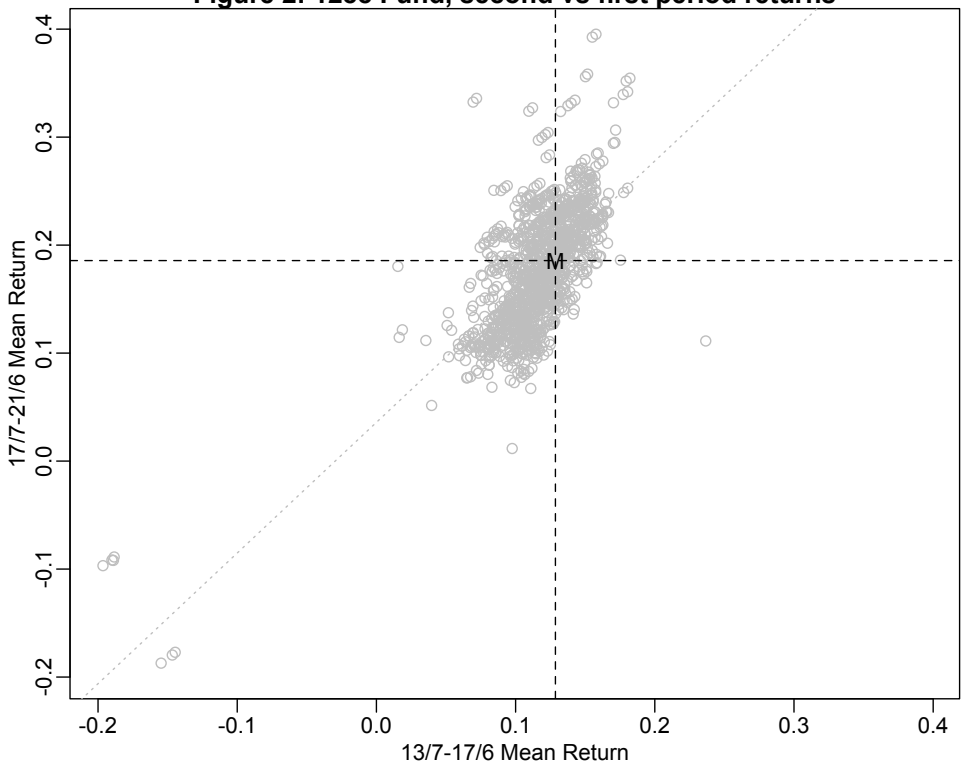
These 1288 growth funds perform worse as a group than expected if they were randomly picking stocks. Maybe by trading stocks often they incur transaction costs that further reduce their performance relative to the market. All in all, a pretty bad result.

A very small number of funds had a horrendous performance. How can one have < 0 average returns while at the same time the market return was 13% over the period? Even ignoring these funds, the group performance is below what one would expect randomly!

b) Your boss says that this is all dandy but she wants you to find out if even a small number of funds can beat the market consistently. "Ahah" you say, "this is why I saved the 2017-21 period, I will prepare a persistence analysis using basic concepts of joint and conditional probabilities". Are some funds consistently the best?

- *So you compute Period 1 (2013/7-2017/6) and Period 2 (2017/7-2021/6) mean returns for each fund. Plot R2 vs R1. This scatter plot has 1288 points.*
  Add the Market as a point to the graph; choose a symbol that makes it visible (see help("par") and help("points") in R)
  Add a vertical and a horizontal dashed lines going through the market (see abline, and lty graphic parameter in R)

Figure 2: 1288 Fund, second vs first period returns



- *Do you see any pattern in this plot?*
- *What would the plot look like if there was no persistence in performance, strong persistence?*

The boss is right. If these 31% perform persistently, they should be more likely than not to beat the market again. That would be interesting evidence

With no persistence at all, the plot should look like a circular scatter ; no evidence that high means in P1 correspond to high means in P2.

There seems to be some evidence of persistence since the plotted regression line is not horizontal.
The crazy poor performers skew the plot and may affect the impression of persistence. We should remove them.
Means are measured with a **lot** of sampling error. So we can do a "robust" version of this plot looking at the **ranks of the funds.**

- Define W (win) as a fund being top 15%, L (lose) bottom 15%, and M (middle) the middle 70% range. Fill three versions of the two-way table for period 2 vs 1:
  1A) simple counts out of 1288
  1B) joint probabilities
  1C) conditional probabilities. Keep only 2 digits for probabilities

Table 1A: Persistence in fund ranking, total counts

	L2	M2	W2
L1	94	94	5
M1	98	706	98
W1	1	102	90

.

Table 1B: Persistence in fund ranking, joint probabilities

	L2	M2	W2
L1	0.07	0.07	0.00
M1	0.08	0.55	0.08
W1	0.00	0.08	0.07

Table 1C: Persistence in fund ranking, conditional probabilities

	L2	M2	W2
L1	0.49	0.49	0.03
M1	0.11	0.78	0.11
W1	0.01	0.53	0.47

- Add a Table 1D, to show what Table 1C should be if funds had all no persistence in abilities.

Table 1D: Conditional probabilities under **no persistence** in ranking

	L2	M2	W2
L1	0.15	0.70	0.15
M1	0.15	0.70	0.15
W1	0.15	0.70	0.15

- In a few sentences summarize these results to your boss, explaining the evidence on the ability of the best funds to remain the best, the so-called "hot-hand".

Very Interesting! Well.. I think, don't you? Table 1C revealssome amount of persistence. Winners are equally likely to be **Middlers** or Winners again. That's different from a 15% chance under randomness.
Losers likely too remain losers or maybe middlers.
Middlers seem to conform to the zero persistence distribution.
Very strong performance does carry information: 47% chances of Winning is very different from 15%, so is 49% chances of losing !

c) *This was a nice effort but your boss is a bit bored and confused. "Anyway, what do I care that the best remain the best if they can't reliably beat the market! I will still advise our clients to buy DFA! Do me a table showing persistence in beating the market. Just show me the last one, the conditional thing you call it?"*

- *This time you prepare a two-by two set of three tables 2A, 2B, 2C,2D. You define Win / Lose by having a larger/ smaller return than the market.*

Table 2: Conditional Probability of Beating the Market

	L2	W2
L1	**0.73**	0.26
W1	0.18	**0.81**

- *Conclude with respect to persistence in ability to beat the market.*

  Comparing Period 2 (17-21) to Period 1(13-17), we do find **strong evidence** of persistence in ability to outperform (and underperform!) the market for both

  Question: What would be the "no persistence " table of conditional probabilities to compare Table 2.

- *As you conclude your presentation and your boss seems interested, Frankie pops his head in and quips "I know how you collected your data, performance is overblown, you have survival bias" He is really annoying, thinks he is a hot shot because he did this MSMF at BU.*
  Now you need to explain what survival bias is to your boss and how it can affect your results.

  Survivorship bias is known to spuriously increase the estimates of performance and persistence.

  We ignored funds that did so poorly that they disappeared during the period. Typically these funds are high volatility funds, they can either do very well or very poorly.

  Consider a bunch of funds trading randomly with a high level of risk. They will do very well or very badly.

  We removed the funds that did very badly and kept those that did very well, leaving us with the impression that the remaining set of funds exhibited positive performance.

  In contrast funds taking less risk do not do badly enough to cease existence.

Problem 5: Attitude to risk and return

Your total wealth is $10,000. A project can earn 30% or lose 10% with probability 0.5. Your utility of wealth has the shape U(W) = -1/W. You consider whether to invest your total wealth into it!

a) What is the $CE of your total wealth if you undertake the project?

EU = 0.5*(-1/13000)+0.5*(-1/9000) = -9.401709e-05 a meaningless number !

CE= U^{-1}(EU) = **$ 10636**

Do you do it? **Yes**

b) You can borrow at 0%. What is the maximum $$B_{max}$$ you can borrow without going bankrupt in the down case.

You must always repay your borrowing, in all scenarios. And you invest your borrowing. You are bankrupt if your wealth reaches $0. In the down scenario, your wealth is

$$(10{,}000 + B)* 0.9 - B \geq 0$$

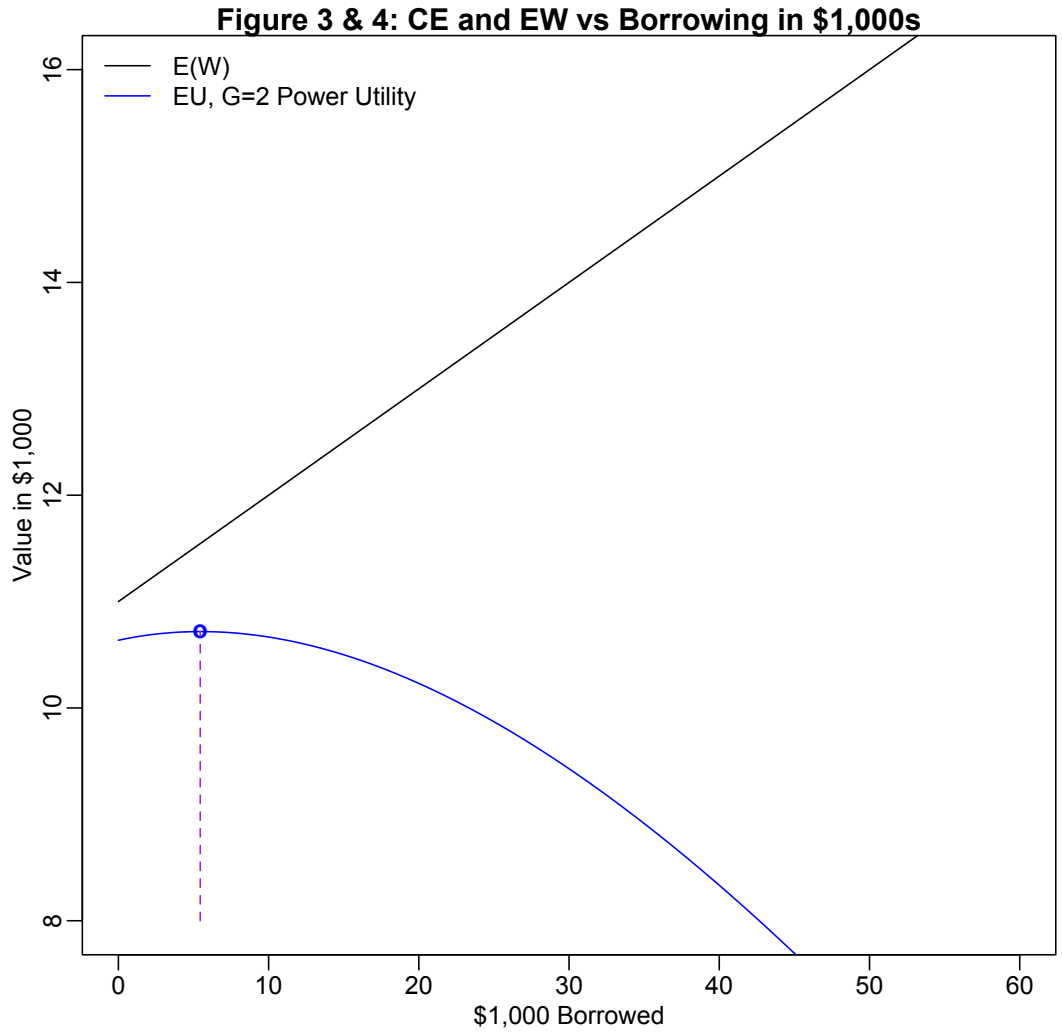
B_{max} = $ 90,000

c) In Figure 3 plot your EXPECTED wealth vs the amount borrowed $B in [0 , B_{max}]

d) Write the simple formula of your $CE as a function of $B. In Figure 4, Use R to plot $CE vs $B. Again have B in [0 , B_{max}]

CE = -1/(-0.5/((10000+B)*1.3-B)-0.5/((10000+BB)*0.8-B))

Figures 3 and 4 together going to B=$60,000 only



d) How much would you borrow optimally to increase your expected utility? Show that point on your Figure 4.

You borrow **$5468** for a Max(CE) of **$10718**, very small improvement over $10636

Analytics: Take the derivative of expected utility and set it to 0.

R CODE BELOW

The Following R commands are of **great** interest

- Do help("commandname") for details on how to use a command

- Repeat after me: **"In R I will try to avoid loops as much as possible"**

- Counting: look at the commands choose, factorial, lfactorial
  # What if you need factorial(171) .... or bigger?
  # For ratios of overflow numbers, make partial divisions and take the product after
  # or remember that: ratio = exp(log(ratio)), and log(ratio) = log(numerator)-log(denominator)
  # some functions that can overflow can have their log directly computed, lfactorial, lgamma

- Reading Data
  Easiest is simple rectangular data sets with a header row in .csv format
  arrayname <- read.csv("filename.csv")
  # it assumes a header, first row contains name variables
  also: read.table(), scan()

- Writing Data for your results
  - Think of putting table results in a R array, then write the array to a .csv file. The .csv file can then be put directly in a .xls table and into a word document
  - There are many ways to prepare latex tables in R for those using that more sophisticated quantitative paper formatting package. Not needed for our problem sets.

  write.csv(arrayname, "table.csv")
  also: write.table(). write()

- Choosing subsets of data

  Say we have a matrix with 200 rows and 5 columns. We can select any subset we want
  smallmat<- bigdatamat[101:200,3:5]
  # creates small mat as the indicated subset with columns 3 to 5, rows 101 to 200
  smallmat<- bigdatamat[101:200,c(2,4)]
  # takes columns 2 and 4 of bigmat
  smallmat<-bigdatamat[bigdatamat[,1]>20171231,]
  # takes rows with date after 12/13/2017

  Can also use multiple conditions with the & (and) , | (or), and ! (not) signs. Do help("Logic") to see more
  retmat[retmat[,1]>20171231 & retmat[,2]>0,] #
  selects all periods past 20171231 where the first return is >0.

Can directly count over a condition

```
sum(rets[,2]>0)      # gives the number of positive returns in column 2
length(rets[,2])     # total number of observations
```

- Building blocs

```
1:5                  # a sequence of numbers from 1 to 5
c(1,5)               # the numbers 1 and 5
c(vector1,vector2)   # concatenates two vectors into 1
cbind(mat1,mat2)     # joins 2 matrices next to each other,        column bind
rbind(mat1,mat2)     # joins 2 matrices with mat1 on top of mat2,  row bind
length(vector)       # gives the length
dim(matrix)          # gives the dimension
```

Basic loop

```
for (i in 4:10){
    sales[i] <- price[i]*quantity[i]
    }
```

But we can (and should) avoid loops:

```
    sales<-price*quantity          # Voila!
```

The simple * in R is not an inner product but the Hadamard (element-wise) product.. very convenient!

So.. what is the inner product of vectors or matrices? a %*% b
There we need a to be a row and b a column

- prod and sum commands give the product or sums of elements of a vector.

```
prod(1+ ret)-1       # compound returns contained in vector ret
```

- Plots

```
plot(x,y, pch="O" ,xlab="This is x", ylab="This is y",col="blue")
    # pch allows you to choose the symbol.
    # By default the graph comes out as a scatter plot with points
```

Other useful qualifiers:

```
plot(x1,y1,  xlim= c(low,high),   ylim=     ) # axes limits,
points(x2,y2, col="red",pch="*")     # adds points at (x2,y2) to the graph with red
                                     # color, and a star as symbol
title("Figure 1: This is my title")  # adds title to the plot
```

Do help("par") for graphics tweaks

```
abline(h=10)          # adds horizontal line at chosen value of 10
abline(v=   )         # adds vertical line
abline(a=… ,  b= … )  # adds straight line with intercept a, slope b
abline(lsfit(x,y))    # adds regression line to plot
```

- Histograms

```
hist(rets, nclass=40, prob=T)
    # always try to  increase the default number of bins (nclass) to make the
    # histogram  look more realistic. Otherwise it's pretty useless
    # prob=T scales the histogram to be a density.
```

- Descriptive stats

```
mean(myret)
sd(myret)          # standard deviation
var(myret)         # variance
cor(ret1,ret2)     # correlation
```

If myret is a matrix of returns, not a single vector, var(myret) computes the covariance matrix.
Use the apply command below (we don't loop) to get only the variances

- apply(matrix, 2, fun)
computes the function fun for each column (2) or row (1) of a matrix. Result is a vector.

For example:

```
Per1mean<- apply(retmatrix[retmatrix[,1]<20140000 , 2: 501],2,mean)
        #       returns a vector of 500 means using data up to Dec 2013
        #       Use & for more than one conditions, like A & B  (A and B)
        #       Ummh, this can be useful!
```