Stat 107: Introduction to Business and Financial Statistics

Homework 2 Sketch Solutions

1) Suppose monthly returns of AAPL stock are normally distributed with mean 2.48% and standard deviation 12.27%. There is a 1% chance that AAPL monthly returns will be below what value?

```
qnorm(.01,2.48,12.27)
[1] -26.06429
```

2) Suppose that annual stock returns for a particular company are normally distributed with a mean of 16% and a standard deviation of 10%. You are going to invest in this stock for one year.
   a) Find that the probability that your one-year return will exceed 30%
   ```
   > 1-pnorm(30,16,10)
   [1] 0.08075666
   ```
   b) Find that probability that your one-year return will be negative.
   ```
   > pnorm(0,16,10)
   [1] 0.05479929
   ```

3) The variance of Stock A is 0.0016, the variance of the market is 0.0049 and the covariance between the two is 0.0026. What is the correlation coefficient?

   Cor(x,y)=cov(x,y)/(sd(x)(sd(y)) = .0026/(sqrt(.0016*.0049))= 0.93

4) The distribution of annual returns on common stocks is roughly symmetric, but extreme observations are more frequent than in a normal distribution. Because the distribution is not strongly nonnormal, the mean return over even a moderate number of years is close to normal.

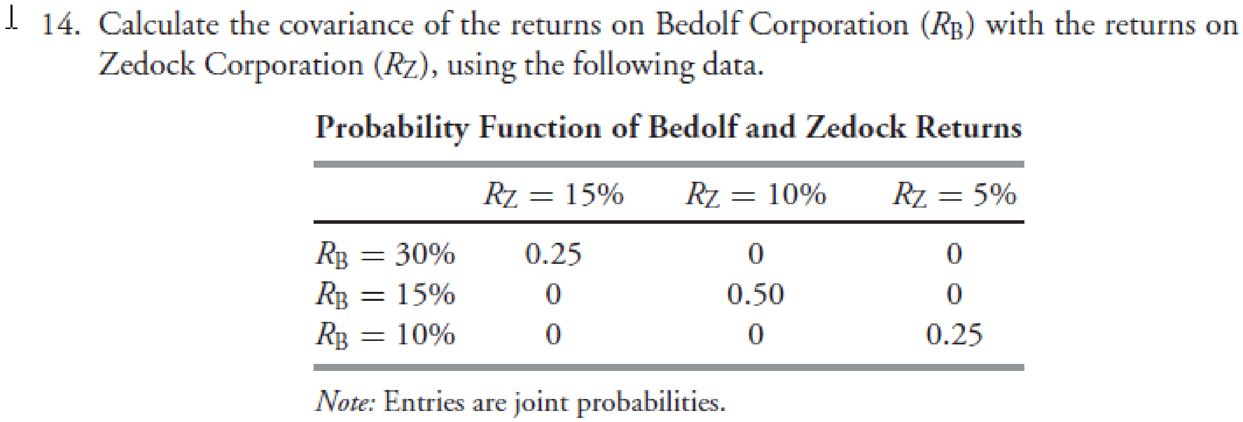
   Annual real returns on the Standard & Poor's 500-Stock Index over the period 1871 to 2004 have varied with mean 9.2% and standard deviation 20.6%. Andrew plans to retire in 45 years and is considering investing in stocks.

   a) What is the probability (assuming that the past pattern of variation continues) that the mean annual return on common stocks over the next 45 years will exceed 15%?

      By the Central Limit Theorem, $\bar{X} \sim N(\mu, \sigma^2 / n) = N(9.2, 20.6^2 / 45)$

      Then $P(\bar{X} > 15) =$
      ```
      > 1-pnorm(15,9.2,20.6/sqrt(45))
      [1] 0.02946484
      ```
   b) What is the probability that the mean return will be less than 5%?
      ```
      > pnorm(5,9.2,20.6/sqrt(45))
      [1] 0.08570424
      ```

5) QIA Workbook: Chapter 4, number 14



14. Calculate the covariance of the returns on Bedolf Corporation (R_B) with the returns on Zedock Corporation (R_Z), using the following data.

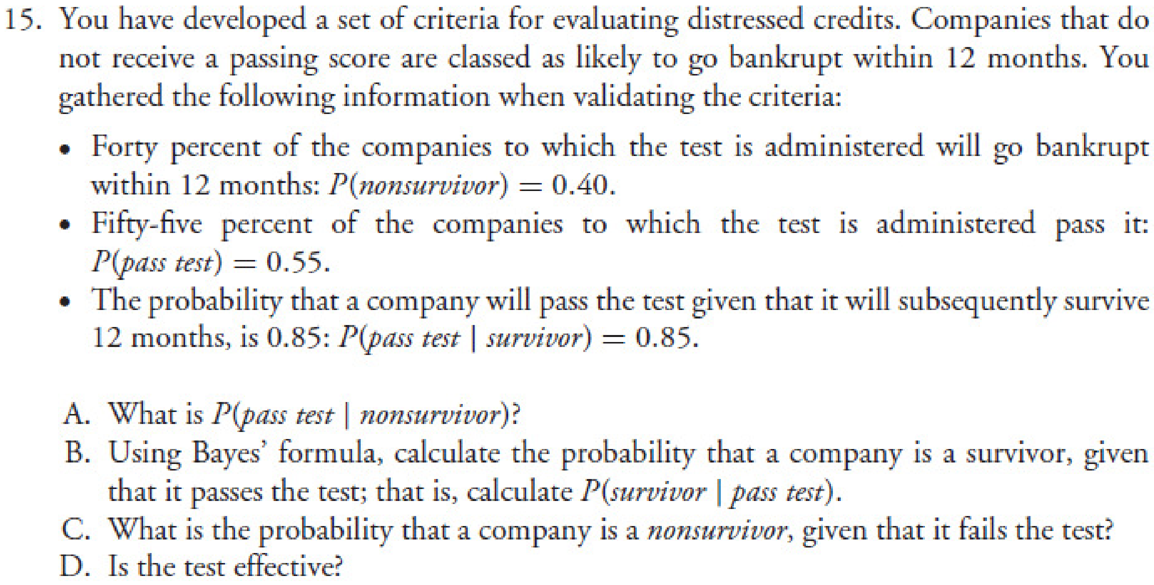
Probability Function of Bedolf and Zedock Returns

	$R_Z = 15\%$	$R_Z = 10\%$	$R_Z = 5\%$
$R_B = 30\%$	0.25	0	0
$R_B = 15\%$	0	0.50	0
$R_B = 10\%$	0	0	0.25

Note: Entries are joint probabilities.

Cov(B,Z)=E(BZ)-E(B)E(Z)
E(BZ) = (30)(15)(.25)+(10)(15)(.5)+(10)(5).25 = 200
E(B) = 30*.25+15*.5+10*.25=17.5
E(Z) =15*.25+10*.5+5*.25=10
Cov(B,Z) = 200-(17.5)(10)=25

6) QIA Workbook: Chapter 4, number 15



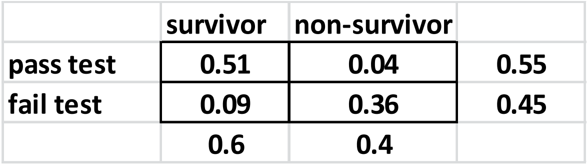
15. You have developed a set of criteria for evaluating distressed credits. Companies that do not receive a passing score are classed as likely to go bankrupt within 12 months. You gathered the following information when validating the criteria:

- Forty percent of the companies to which the test is administered will go bankrupt within 12 months: $P(\textit{nonsurvivor}) = 0.40$.
- Fifty-five percent of the companies to which the test is administered pass it: $P(\textit{pass test}) = 0.55$.
- The probability that a company will pass the test given that it will subsequently survive 12 months, is 0.85: $P(\textit{pass test} \mid \textit{survivor}) = 0.85$.

A. What is $P(\textit{pass test} \mid \textit{nonsurvivor})$?
B. Using Bayes' formula, calculate the probability that a company is a survivor, given that it passes the test; that is, calculate $P(\textit{survivor} \mid \textit{pass test})$.
C. What is the probability that a company is a *nonsurvivor*, given that it fails the test?
D. Is the test effective?

P(nonsurvivor) = 0.40
P(pass test) = 0.55
P(pass test|survivor) = 0.85 = P(pass test and survivor)/P(survivor)
So P(pass test and survivor) = (0.85)(1-0.4) = 0.51

We can now build the following 2x2 probability table

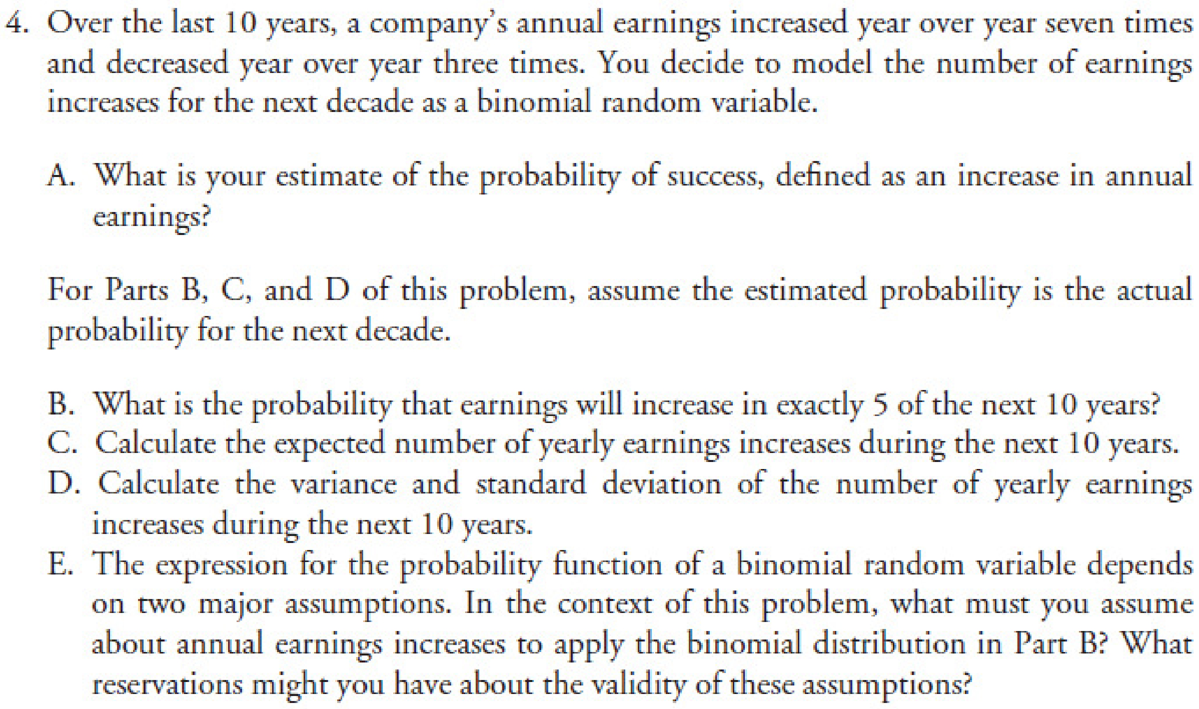


	survivor	non-survivor	
pass test	**0.51**	**0.04**	**0.55**
fail test	**0.09**	**0.36**	**0.45**
	0.6	**0.4**	

P(pass test |nonsurvivor) = 0.04/0.40 = 0.10
P(survivor|pass test) = 0.51/0.55 = 0.93
P(nonsurvivor|fail test) = 0.36/0.45 = 0.80

A company passing the test greatly increases our confidence that it is a survivor. A company failing the test doubles the probability that it is a nonsurvivor. Therefore, the test appears to be useful.

7) QIA Workbook; Chapter 5, number 4



4. Over the last 10 years, a company's annual earnings increased year over year seven times and decreased year over year three times. You decide to model the number of earnings increases for the next decade as a binomial random variable.

   A. What is your estimate of the probability of success, defined as an increase in annual earnings?

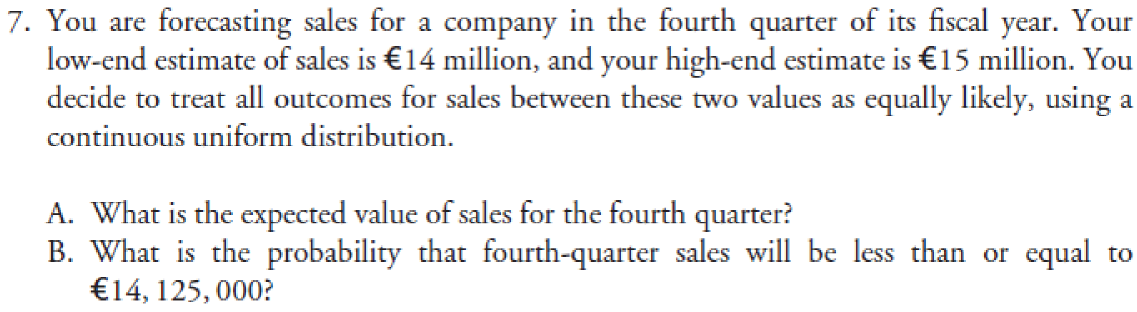
   For Parts B, C, and D of this problem, assume the estimated probability is the actual probability for the next decade.

   B. What is the probability that earnings will increase in exactly 5 of the next 10 years?
   C. Calculate the expected number of yearly earnings increases during the next 10 years.
   D. Calculate the variance and standard deviation of the number of yearly earnings increases during the next 10 years.
   E. The expression for the probability function of a binomial random variable depends on two major assumptions. In the context of this problem, what must you assume about annual earnings increases to apply the binomial distribution in Part B? What reservations might you have about the validity of these assumptions?

a) **phat = 7/10 = 0.7**
b) **`> dbinom(5,10,.7)`**
   i. **`[1] 0.1029193`**
b) **E(X)=10(.7) = 7**
c) **Var(X) = 10(.7)(.3) = 2.1 SD(X) = sqrt(2.1) = 1.45**

d) **You must assume that (1) the probability of an earnings increase (success) is constant from year to year and (2) earnings increases are independent trials. If current and past earnings help forecast next year's earnings, Assumption 2 is violated. If the company's business is subject to economic or industry cycles, neither assumption is likely to hold.**

8) QIA Workbook; Chapter 5, number 7



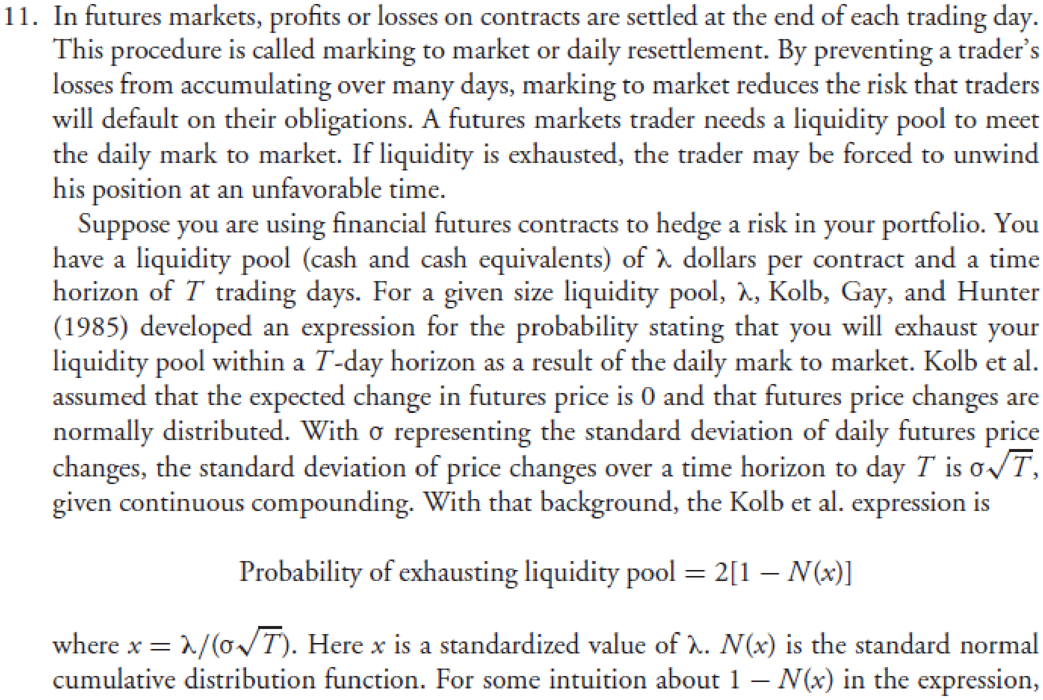
7. You are forecasting sales for a company in the fourth quarter of its fiscal year. Your low-end estimate of sales is €14 million, and your high-end estimate is €15 million. You decide to treat all outcomes for sales between these two values as equally likely, using a continuous uniform distribution.

   A. What is the expected value of sales for the fourth quarter?
   B. What is the probability that fourth-quarter sales will be less than or equal to €14, 125, 000?

a) E(X) = (15+14)/2 = 14.5million

b) P(X<4125000) = (14125000-14000000)/1000000 = 125000/1000000=0.125

9) QIA Workbook; Chapter 5, number 11

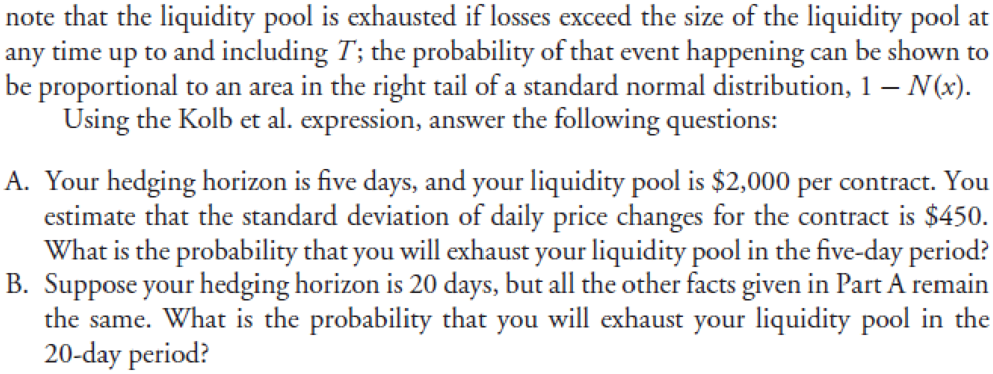


11. In futures markets, profits or losses on contracts are settled at the end of each trading day. This procedure is called marking to market or daily resettlement. By preventing a trader's losses from accumulating over many days, marking to market reduces the risk that traders will default on their obligations. A futures markets trader needs a liquidity pool to meet the daily mark to market. If liquidity is exhausted, the trader may be forced to unwind his position at an unfavorable time.

Suppose you are using financial futures contracts to hedge a risk in your portfolio. You have a liquidity pool (cash and cash equivalents) of λ dollars per contract and a time horizon of T trading days. For a given size liquidity pool, λ, Kolb, Gay, and Hunter (1985) developed an expression for the probability stating that you will exhaust your liquidity pool within a T-day horizon as a result of the daily mark to market. Kolb et al. assumed that the expected change in futures price is 0 and that futures price changes are normally distributed. With σ representing the standard deviation of daily futures price changes, the standard deviation of price changes over a time horizon to day T is $\sigma\sqrt{T}$, given continuous compounding. With that background, the Kolb et al. expression is

$$\text{Probability of exhausting liquidity pool} = 2[1 - N(x)]$$

where $x = \lambda/(\sigma\sqrt{T})$. Here x is a standardized value of λ. $N(x)$ is the standard normal cumulative distribution function. For some intuition about $1 - N(x)$ in the expression,



note that the liquidity pool is exhausted if losses exceed the size of the liquidity pool at any time up to and including T; the probability of that event happening can be shown to be proportional to an area in the right tail of a standard normal distribution, $1 - N(x)$.

Using the Kolb et al. expression, answer the following questions:

A. Your hedging horizon is five days, and your liquidity pool is $2,000 per contract. You estimate that the standard deviation of daily price changes for the contract is $450. What is the probability that you will exhaust your liquidity pool in the five-day period?

B. Suppose your hedging horizon is 20 days, but all the other facts given in Part A remain the same. What is the probability that you will exhaust your liquidity pool in the 20-day period?

a) **The probability of exhausting the liquidity pool is 4.7 percent. First calculate *x*=λ/(σ*sqrt(*T*)) = $2,000/($450*sqrt(5)) = 1.987. Thus, the probability of exhausting the liquidity pool is 2[1 – pnorm(1.99)] = 2(1 – 0.9767) = 0.0466 or about 4.7 percent.**

b) **The probability of exhausting the liquidity pool is now 32.2 percent. The calculation follows the same steps as those in Part A. We calculate *x*=λ/(σ*sqrt(*T*)) = $2,000/($450*sqrt(20)) = 0.9938. Thus, the probability of exhausting the liquidity pool is 2[1 – pnorm(0.9938)] = 32%. This is a substantial probability that you will run out of funds to meet mark to market.**

In their paper, Kolb et al. call the probability of exhausting the liquidity pool the probability of ruin, a traditional name for this type of calculation.

7) Suppose an investor can choose between two independent assets 1 and 2. The probability distributions for the rates of return for each asset are provided below.

a) Compute the expected value of each asset
Expected value of asset 1 is
0.25(5+8+12+15)=10

Expected value of asset 2 is
0.25(15+12+8+5)=10

b) Compute the standard deviation of each asset

For asset 1, E(X^2)=.25(25+64+144+225)=114.5 so Var(X)=144.5-100=44.5
SD for asset 1 = sqrt(44.5) = 6.67

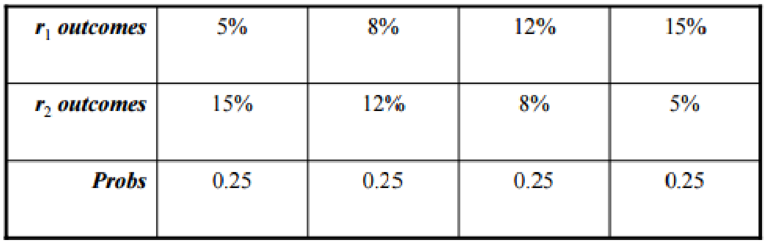
For asset 2, E(X^2)=.25(225+144+64+25)=114.5 so Var(X)=144.5-100=44.5
SD for asset 2 = sqrt(44.5) = 6.67

c) Based on expected return and the variance of return, which asset do you prefer and why? On what other basis might you prefer one asset to another?

The distributions are identical. This unfortunately is a very boring problem.

d) Assume you hold a portfolio with 50% held in asset 1 and 50% held in asset 2. What is the expected return and variance of your portfolio? Note that the portfolio return is a random variable written as R = 0.5R1+0.5R2.

Expected return = 10%
Assuming independence Var(.5R1+.5R2) = .25(44.5)+.25(44.5) = 22.25



r_1 outcomes	5%	8%	12%	15%
r_2 outcomes	15%	12%	8%	5%
Probs	0.25	0.25	0.25	0.25

8) In this problem we are going to reproduce the first table from the article https://sixfigureinvesting.com/2016/03/modeling-stock-market-returns-with-laplace-distribution-instead-of-normal/

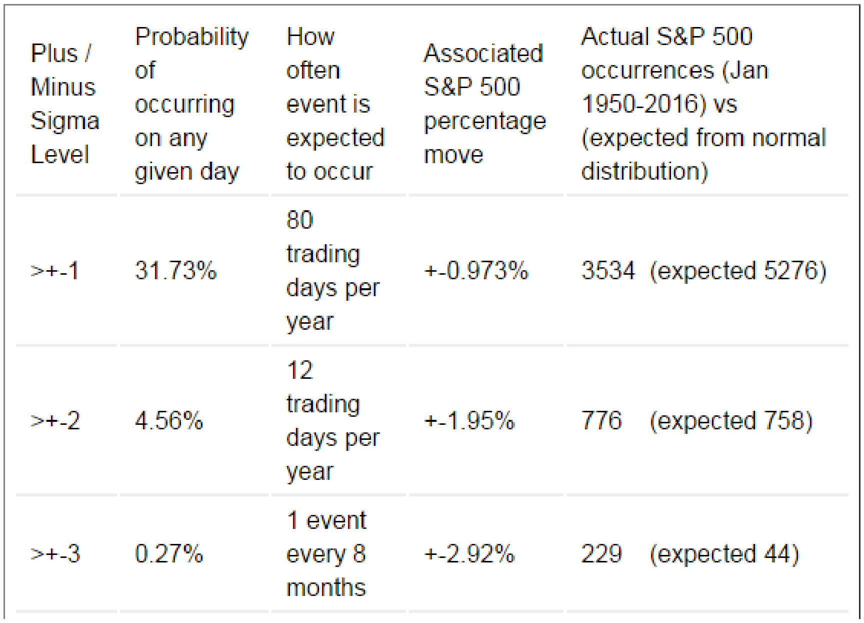
Our numbers will be slightly different but close.

To grab historical data for the S&P500 do the following

```
spydata=getSymbols("^GSPC",from="1950-01-01",auto.assign=FALSE)
spyrets=spyrets[-1]
sd(spyrets)
[1] 0.009679322
```

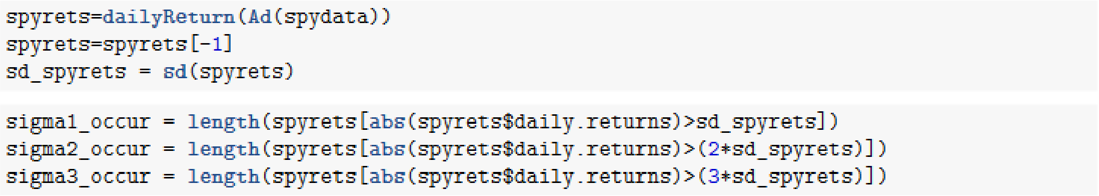
Note we find the historical standard deviation of daily returns to be 0.967% and the quoted paper uses 0.973%. We assume 250 trading days in a year.

Simply reproduce the following table using our data from R [so it will be slightly different]. The expected number is 250days*66 years/day*prob of occurrence per year.



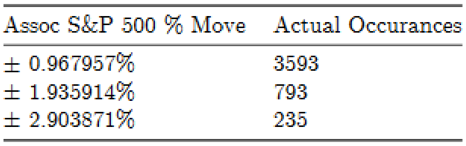
Plus / Minus Sigma Level	Probability of occurring on any given day	How often event is expected to occur	Associated S&P 500 percentage move	Actual S&P 500 occurrences (Jan 1950-2016) vs (expected from normal distribution)
>+-1	31.73%	80 trading days per year	+-0.973%	3534 (expected 5276)
>+-2	4.56%	12 trading days per year	+-1.95%	776 (expected 758)
>+-3	0.27%	1 event every 8 months	+-2.92%	229 (expected 44)

Everyone's answers will vary a little here.



```
spyrets=dailyReturn(Ad(spydata))
spyrets=spyrets[-1]
sd_spyrets = sd(spyrets)
```

```
sigma1_occur = length(spyrets[abs(spyrets$daily.returns)>sd_spyrets])
sigma2_occur = length(spyrets[abs(spyrets$daily.returns)>(2*sd_spyrets)])
sigma3_occur = length(spyrets[abs(spyrets$daily.returns)>(3*sd_spyrets)])
```



Assoc S&P 500 % Move	Actual Occurances
± 0.967957%	3593
± 1.935914%	793
± 2.903871%	235

9) Test using the S&P 500 data used previously that the daily standard deviation of returns from 1990 to 1995 equals the standard deviation of returns from 2000 to 20005. Use the R routine `var.test(x,y)`.

```
#####
spydata=getSymbols("^GSPC",from="1990-01-01",to="1996-01-
01",auto.assign=FALSE)
spyrets=dailyReturn(spydata)
spyretsearly=spyrets[-1]

spydata=getSymbols("^GSPC",from="2000-01-01",to="2006-01-
01",auto.assign=FALSE)
spyrets=dailyReturn(spydata)
spyretslater=spyrets[-1]

var.test(spyretsearly,spyretslater)
```

```
> var.test(spyretsearly,spyretslater)

        F test to compare two variances

data:  spyretsearly and spyretslater
F = 0.3638, num df = 1515, denom df = 1506, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal
to 1
95 percent confidence interval:
 0.3288791 0.4024162
sample estimates:
             daily.returns
daily.returns     0.3637971
attr(,"names")
[1] "ratio of variances"
```

This test is testing if the ratio of variances equals 1. From the small p-value we reject the null and conclude that the variances are not equal. From the confidence interval it appears the later period variance is greater.

10) Again using the historical data, note that the growth of $1 over this period can be obtained using the command **`prod((1+spyrets)).`**

a. What would the growth be if someone had missed the 10 best performing days over this period [this is the argument to not market time but instead buy and hold].
b. What would the growth be if someone had missed the 10 worst performing days over this period [this is the argument to market time and instead buy and hold].

```
myfun=function(){
spydata=getSymbols("^GSPC",from="1950-01-01",auto.assign=FALSE)
spyrets=as.numeric(dailyReturn(spydata))
spyrets=spyrets[-1]
ospyrets=sort(spyrets)
rospyrets=rev(ospyrets)

###without 10 worst days

print("Cumulative Index return from 1950")
print(prod((1+spyrets)))
print("Cumulative Index return from 1950 without 10 worst days")
spyretswolo=ospyrets[-(1:10)]
print(prod((1+spyretswolo)))
print("Cumulative Index return from 1950 without 10 best days")
spyretswohi=rospyrets[-(1:10)]
print(prod((1+spyretswohi)))
}

> myfun()
[1] "Cumulative Index return from 1950"
[1] 128.8175
[1] "Cumulative Index return from 1950 without 10 worst days"
[1] 335.153
[1] "Cumulative Index return from 1950 without 10 best days"
[1] 62.16092
```

11) Recall that the R command apply allows us to easily and quickly calculate the mean (or sd or sum or any function) of each row or column of a matrix. Using this command we can quickly write code to demonstrate the Central Limit Theorem.

The following code simulates drawing 10000 samples of size 100 from an exponential distribution with shape parameter=1, and then drawing the resulting histogram.

```
simdata=matrix(nrow=10000,ncol=100,rexp(100*10000,1))
hist(apply(simdata,1,mean))
```

Explain in easy to understand language what the above code is doing. Using the above code, show how the Central Limit Theorem works for the Gamma distribution with shape=5 and rate=5 parameters [R code is `rgamma(n,5,5)`]. Use sample sizes of 10, 50 and 100. For each simulation run, report the mean of the 10000 sample means produced and the standard deviation of the 10000 sample means produced. Comment on what you find.

This command generates 10000 samples of size 100 each. We then take the mean of each sample resulting in 10000 sample means. By the Central Limit Theorem the histogram of these sample means should look more and more normal as the sample size increases.

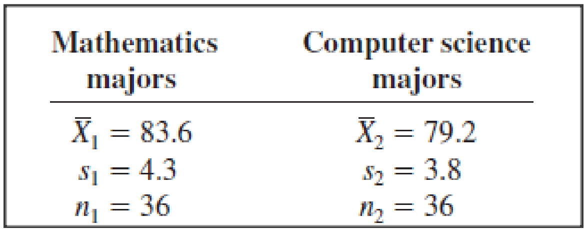
Statistical Methods Review Problems. This part of this homework is to simply reignite the brain cells involving confidence intervals and hypothesis testing. The R cookbook is the best place to go to see how to compute confidence intervals and hypothesis tests in R. For each problem below, use R to answer the question and clearly state the conclusion of your test, or clearly give the confidence interval.

12) In a study of the length of time that students require to earn bachelor's degrees, 80 students are randomly selected and they are found to have a mean of 4.8 years and standard deviation of 2.2 years. Construct a 95% confidence interval estimate of the population mean. Does the resulting confidence interval contradict the fact that 39% of students earn their bachelor's degrees in four years?

The confidence interval is $\bar{X} \pm 1.96\left(\frac{s}{\sqrt{n}}\right) = (4.31, 5.29)$

The second part of the question isn't formed that well. Could the mean of the distribution be in the interval (4.31,5.29) and the distribution still have a lot of mass at 4 years? Yes if it was skewed to the right which would be typical for graduation times.

13) Two groups of students are given a problem-solving test, and the results are compared. Find and interpret the 95% confidence interval of the true difference in means.



Mathematics majors	Computer science majors
$\bar{X}_1 = 83.6$	$\bar{X}_2 = 79.2$
$s_1 = 4.3$	$s_2 = 3.8$
$n_1 = 36$	$n_2 = 36$

(83.6-79.2)+/- 1.96sqrt(4.3*4.3/36+3.8*3.8/36)=(2.5,6.31)

From the CI we very confident that math majors score (on average) somewhere between 2.5 and 6.31 points higher than CS majors.

14) A recent Gallup poll consisted of 1012 randomly selected adults who were asked whether "cloning of humans should or should not be allowed." Results showed that 892 of those surveyed indicated that cloning should not be allowed. Construct a 95% confidence interval estimate of the proportion of adults believing that cloning of humans should not be allowed.

```
prop.test(892,1012)

        1-sample proportions test with continuity correction

data:  892 out of 1012, null probability 0.5
X-squared = 587.39, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.8595200 0.9003699
sample estimates:
        p
0.8814229
```

15) In a sample of 80 Americans, 55% wished that they were rich. In a sample of 90 Europeans, 45% wished that they were rich. Find and interpret the 95% confidence interval of the true difference in proportions.

```
> prop.test(c(.55*80,.45*90),c(80,90),correct=FALSE)

        2-sample test for equality of proportions without continuity
        correction

data:  c(0.55 * 80, 0.45 * 90) out of c(80, 90)
X-squared = 1.6942, df = 1, p-value = 0.1931
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.04982832  0.24982832
sample estimates:
prop 1 prop 2
  0.55   0.45
```

16) For this problem, download data for the last year (9/01/2014 to 9/01/2015) for AAPL and CAT. We are going to work through some items in the *R Cookbook* so look there if you get stuck. **[use adjusted closing prices for everything].**

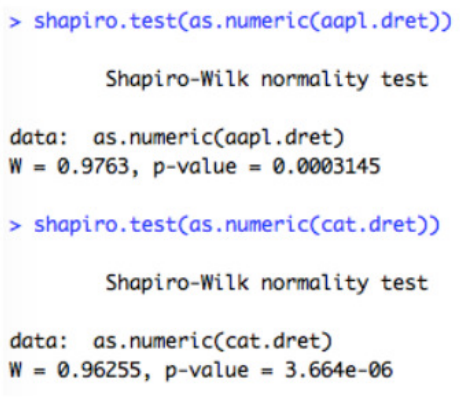
a) Run a hypothesis test to see if AAPL or CAT daily returns are normally distributed. Interpret the output, what is the conclusion of the hypothesis test?
b) Run a hypothesis test to see if AAPL or CAT monthly returns are normally distributed. Interpret the output, what is the conclusion of the hypothesis test?
c) Test if AAPL and CAT have the same average daily return (see comparing the means of two samples in the R Cookbook).
d) Test if AAPL and CAT have the same average monthly return (see comparing the means of two samples in the R Cookbook).
e) Test if AAPL and CAT have the same standard deviation of daily returns. Use the R routine `var.test(dailyReturn(Ad(AAPL)),dailyReturn(Ad(CAT)))`
f) Calculate the correlation between AAPL and CAT daily returns, and compute a confidence interval for the correlation.

Here is the code

```
# a)
library(quantmod)
getSymbols("AAPL", from="2014-09-01", to= "2015-09-01")
getSymbols("CAT", from="2014-09-01", to= "2015-09-01")
aapl.a = adjustOHLC(AAPL)
cat.a = adjustOHLC(CAT)
aapl.dret = dailyReturn(aapl.a)
cat.dret = dailyReturn(cat.a)
shapiro.test(as.numeric(aapl.dret))
shapiro.test(as.numeric(cat.dret))
# b)
aapl.mret = monthlyReturn(aapl.a)
cat.mret = monthlyReturn(cat.a)
shapiro.test(as.numeric(aapl.mret))
shapiro.test(as.numeric(cat.mret))
# c)
t.test(aapl.dret,cat.dret)
# d)
t.test(aapl.mret,cat.mret)
# e)
var.test(aapl.dret,cat.dret)
# f)
cor.test(as.numeric(aapl.dret), as.numeric(cat.dret))
```

a)

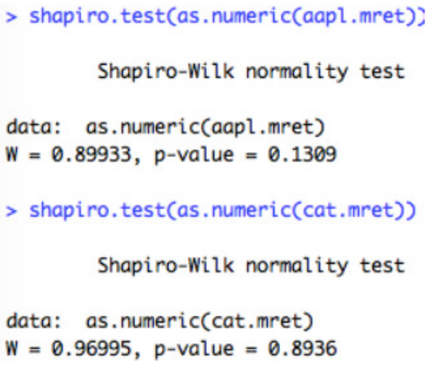
Both of our p-values are ~0, we reject the null and can claim that our data is not normally distributed:



```
> shapiro.test(as.numeric(aapl.dret))

        Shapiro-Wilk normality test

data:  as.numeric(aapl.dret)
W = 0.9763, p-value = 0.0003145

> shapiro.test(as.numeric(cat.dret))

        Shapiro-Wilk normality test

data:  as.numeric(cat.dret)
W = 0.96255, p-value = 3.664e-06
```

b)

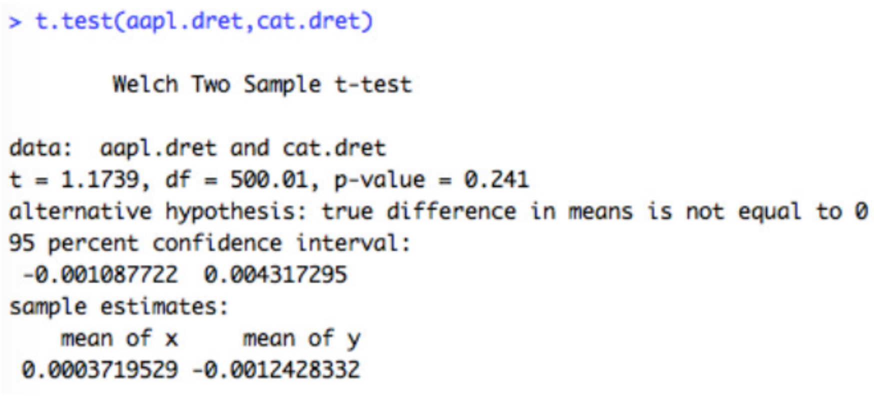
Unlike (a), we have p-values greater than .05, so we fail to reject the null hypothesis that our data is normally distributed:



```
> shapiro.test(as.numeric(aapl.mret))

        Shapiro-Wilk normality test

data:  as.numeric(aapl.mret)
W = 0.89933, p-value = 0.1309

> shapiro.test(as.numeric(cat.mret))

        Shapiro-Wilk normality test

data:  as.numeric(cat.mret)
W = 0.96995, p-value = 0.8936
```

c)

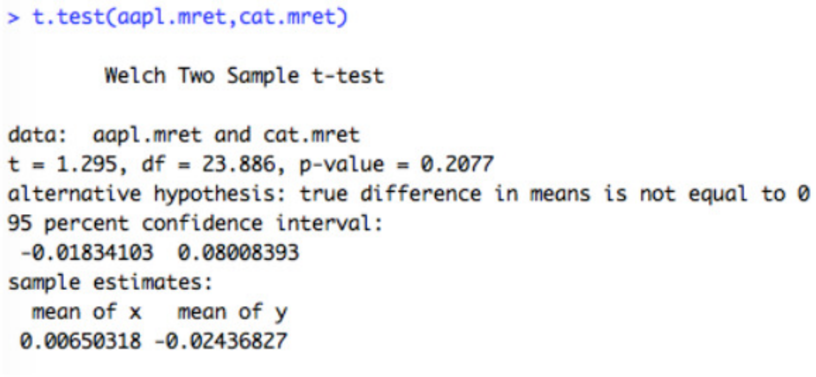
With a p-value of .241, we fail to reject the null hypothesis that the true difference in means is equal to 0:



```
> t.test(aapl.dret,cat.dret)

	Welch Two Sample t-test

data:  aapl.dret and cat.dret
t = 1.1739, df = 500.01, p-value = 0.241
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.001087722  0.004317295
sample estimates:
    mean of x     mean of y 
 0.0003719529 -0.0012428332 
```

d)

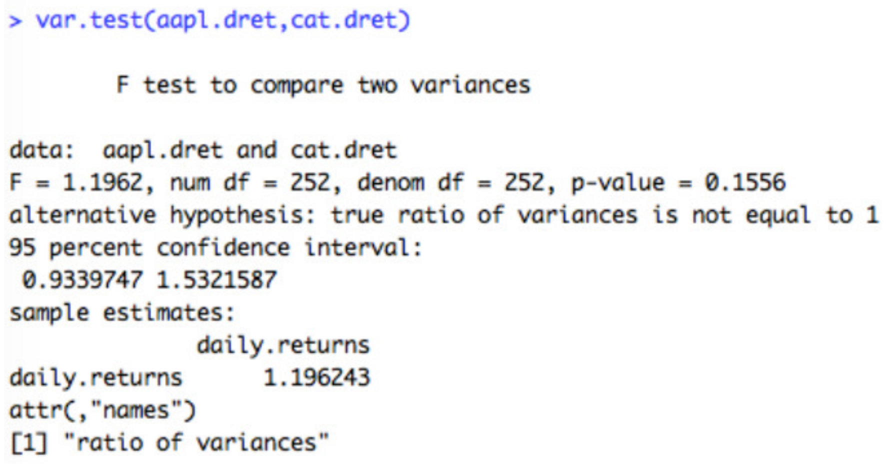
With a p-value of .2077, we again fail to reject the null hypothesis that the true difference in means is equal to 0:



```
> t.test(aapl.mret,cat.mret)

	Welch Two Sample t-test

data:  aapl.mret and cat.mret
t = 1.295, df = 23.886, p-value = 0.2077
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01834103  0.08008393
sample estimates:
  mean of x   mean of y 
 0.00650318 -0.02436827 
```

e)

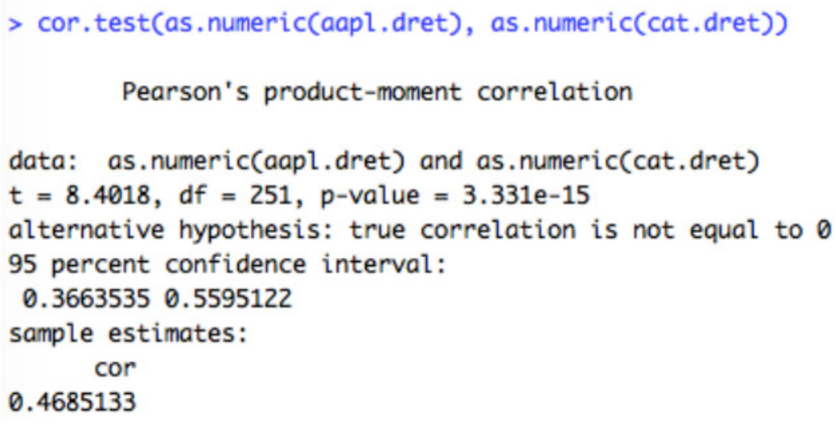
With a p-value of .1556, we fail to reject the null hypothesis that the true ratio of variances is equal to 1. Note that our confidence interval does not span 0; this is because we're testing a *ratio*, so having 1 in the CI affirms our conclusion we drew with the p-value.



```
> var.test(aapl.dret,cat.dret)

	F test to compare two variances

data:  aapl.dret and cat.dret
F = 1.1962, num df = 252, denom df = 252, p-value = 0.1556
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9339747 1.5321587
sample estimates:
              daily.returns
daily.returns      1.196243
attr(,"names")
[1] "ratio of variances"
```

f)

Our correlation was .468, with a CI of (.366, .560). Spearman correlation could also have been calculated for the purposes of this question.



```
> cor.test(as.numeric(aapl.dret), as.numeric(cat.dret))

	Pearson's product-moment correlation

data:  as.numeric(aapl.dret) and as.numeric(cat.dret)
t = 8.4018, df = 251, p-value = 3.331e-15
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3663535 0.5595122
sample estimates:
      cor
0.4685133
```