Stockholm University

STOCKHOLM UNIVERSITY
Department of Statistics
Spring 2019, period A-B

Andriy Andreev (examiner)
Ulf Högnäs

FINANCIAL STATISTICS

2019-03-21

Time:	09.00 - 14.00
Place:	*Ugglevikssalen*
Approved aid:	Hand-held calculator with no stored text, data or formulas
Provided aid:	*Formula Sheet and Probability Distribution Tables*, returned after the exam

- **Problems 1 – 4: MULTIPLE CHOICE QUESTIONS – max 35 points**
  - A total of four multiple choice questions with five alternative answers per question one of which is the correct answer. Mark your answers on the attached **answer form**.
  - Marking more than one alternative will result in zero points for that question.
  - Written solutions are not required to be submitted but if submitted, they might be used to evaluate the extent of the mistake in the final answer: that is done on case-by-case basis and decided by the examiner; only your answers on the answer form are guaranteed to be considered in the assessment and final grading.

- **Problems 5 – 6: COMPLETE WRITTEN SOLUTIONS – max 25 points**
  - Use only the provided **answer sheets** when submitting your solutions and answers.
  - For full marks, clear, comprehensive and well-motivated solutions are required. Unclear and un-explained solutions may result in point deductions even if the final answer is correct.
  - Check your calculations and solutions before submitting. Careless mistakes may result in unnecessary point deductions.

- The maximum number of points is stated for each question. The maximum total number of points is 36 + 24 = 60. At least 30 points is required to pass (grades A-E). The grading scale is as follows:

  A: 54 – 60 points
  B: 48 – 53 points
  C: 42 – 47 points
  D: 36 – 41 points
  E: 30 – 35 points
  Fx: 24 – 29 points
  F: 0 – 23 points

- NOTE! Fx and F are failing grades that require re-examination. Students who receive the grade Fx or F cannot supplement for a higher grade.
- Outlines of solutions will be posted on Mondo within several days after the exam.

GOOD LUCK!

1. (Multiple Choice type question) (2points + 5points + 5points = **12 points**) (normality, approximation)

A customer who visits a web shop makes a purchase with probability $p = 0.3$.

1. (2 points) $n = 12$ customers visit the web shop, independently of each other. What is the probability that at least four of the customers buy something? Choose the alternative that is closest to your answer.

   a) 0.276 b) 0.493 c) 0.507 d) 0.750 e) 0.882

2. (5 points) $n = 500$ customers visit the web shop, independently of each other. What is the probability that **fewer** than 175 of the customers buy something? Choose the alternative that is closes to your answer.

   a) 0.01 b) 0.41 c) 0.59 d) 0.88 e) 0.99

3. (5 points) An investor analyzes the daily changes in price of the Danish global bonds fund "Strategi Invest Stabil". Based on 30 daily changes, the estimated Skewness is -0.320 and the estimated Excess Kurtosis is 2.29. Which of the below listed statements is correct, assuming that we test H_0 hypothesis that the data is NOT normally distributed?

   a) At a level of significance $\alpha = 0.5\%$ we find support for this hypothesis
   b) At a level of significance $\alpha = 1\%$ we find support for this hypothesis, but not at $\alpha = 0.5\%$
   c) At a level of significance $\alpha = 2.5\%$ we find support for this hypothesis, but not at $\alpha = 1\%$
   d) At a level of significance $\alpha = 5\%$ we find support for this hypothesis, but not at $\alpha = 2.5\%$
   e) If we use a level of significance of $\alpha = 5\%$ or lower, we do not find support for this hypothesis

---

2. (Multiple Choice type question) (2 points + 6 points = **8 points**) (ARMA)

In this task, we consider two ARMA models Y_t and Z_t, both assumed to be stationary.

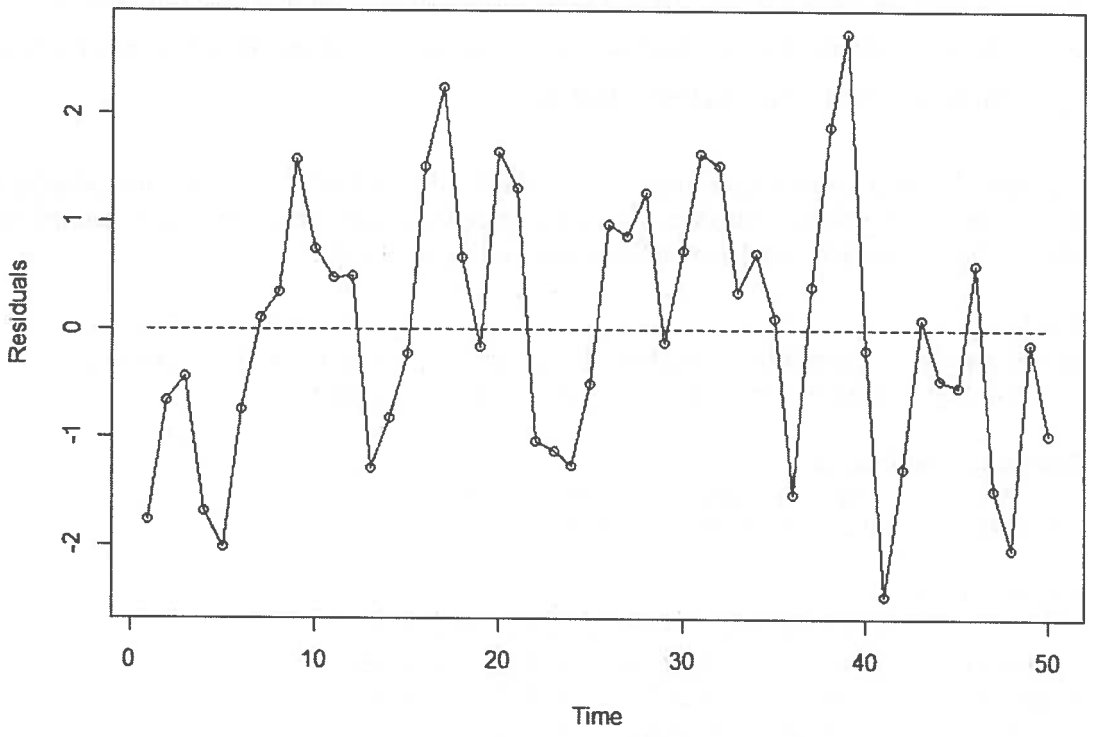
1. (2 points) Let $Y_t = 0.2Y_t + \varepsilon_t$. Calculate the correlation ρ_2 between Y_t and Y_{t-2}. Select the closest to correct answer from the options below
   a) 0.03 b) 0.04 c) 0.039 d) 0.042 e) -0.04
2. (6 points) Let $Z_t = \varepsilon_t + 0.4\varepsilon_{t-1} + 0.2\varepsilon_{t-2}$. Calculate the correlation ρ_1 between Z_t and Z_{t-1}. Select the value closest to your answer
   a) -0.4 b) -0.2 c) 0 d) 0.2 e) 0.4

---

3. (Multiple Choice type question) (4 points + 5 points = **9 points**) (Trends, Runs test)

The plot below shows 50 residuals from a regression analysis. The observations have been connected with lines to make the order more visible. Use a runs test to test whether there is evidence of **positive** autocorrelation. Use 5% level of significance.

1. (4 points) What is the value of the test variable?
   a) -2.57
   b) -1.84
   c) -0.735
   d) 0.735
   e) 2.76



2. (5 points) Given 5% level of significance, what is the critical value? What is the correct statement of the decision rule? (4 points)
   a) Critical value -1.9600. We reject the null hypothesis of "no positive autocorrelation" if $z_{obs} < -1.9600$.
   b) Critical value 1.9600. We reject the null hypothesis that "there is positive autocorrelation" if $z_{obs} < -1.9600$.
   c) Critical value -1.6449. We reject the null hypothesis of "no positive autocorrelation" if $z_{obs} < -1.6449$.

d) Critical value 1.6449. We reject the null hypothesis that "there is positive autocorrelation" if $z_{obs} < 1.6449$.

e) Critical value 1.6449. We reject the null hypothesis of "no positive autocorrelation" if $|z_{obs}| < 1.6449$.

---

4. (Multiple Choice type question) (2 points + 4 points = **6 points**) (logistic regression)

Researchers at a government agency for education wanted to examine the factors which influence the probability that a student completes sixth grade with at least satisfactory level in all subjects. Based on a sample of 25000 students, the researchers estimated the following model.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

y	the log-odds of a student reaching satisfactory level in all subjects at grade 6
X_1	Dummy variable: Parents' highest level of education is **high school**
X_2	Dummy variable: Parents' highest level of education is **up to 3 years of college**
X_3	Dummy variable: Parents' highest level of education is **more than 3 years of college**
X_4	Dummy variable: The student is **female**

Together, X_1, X_2, X_3 make out a categorical variable with multiple levels. It is not possible to belong to more than one of these categories. The base category is that neither parent has completed high school. The researcher's analysis rendered the following output:

```
Call:
glm(formula = complete ~ highsch + up.to.3 + more.than.3 + female,
    family = binomial(link = "logit"), data = grades)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1725   0.4454   0.4609   0.6982   1.3738

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.45072    0.24838  -1.815   0.0696 .
highsch      1.30856    0.24913   5.252  1.5e-07 ***
up.to.3      1.73792    0.25077   6.930  4.2e-12 ***
more.than.3  2.63928    0.24965  10.572  < 2e-16 ***
female       0.07207    0.03400   2.120   0.0340 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1) (2points) Which of these statements reflects the correct interpretation of the intercept? Note, that *log* refers to the natural *log*.

   a) There is **no reasonable interpretation** of the intercept in this model.

   b) It is the log of the **probability** that an **average** student reaches satisfactory level in all subjects.

   c) It is the log of the **odds** that an **average** student reaches satisfactory level in all subjects.

d) It is the log of the **odds** that **non-female** student reaches satisfactory level in all subjects, given that the parents' highest level of education is **no high school**.

e) It is the log of the **odds** that **female** student reaches satisfactory level in all subjects, given that the parents' highest level of education is **no high school**.

2) (4points) Find the **probability** that a randomly selected **female** student whose parents' highest level of education is **more than 3 years of college** reaches satisfactory level in all subjects at grade 6.

a) 0.0959

b) 0.906

c) 0.938

d) 0.992

e) 0.995

---

5. (Essay type question) (2points + 5 points + 5 points = **12 points**) (multiple linear regression)

Below is a regression output for a FCMG company that is in much need of investigating what affects the quantity of sold items(y). This company has a string of 30 retail stores across Sweden.

X1- Price of the immediate competitors
X2-Advertising price
X3-(1-if special offers are given at the store and 0-if no special offers)
X4-Price difference (difference between the price offered by the company's retail stores and their immediate competitor i.e competitor_price - retail_price)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.894170196
R Square	0.79954034
Adjusted R Square	0.767466794
Standard Error	74.07160966
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	547087.7161	136771.929		
Residual	25	137165.0839	5486.603358		
Total	29	684252.8			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	830.0751215	397.4632073	2.088432605	0.047101878	11.4843227	1648.666
X1	12.9633807	17.6255086	0.735489738	0.468885819	-23.33703377	49.2638
X2	0.04478463	0.037155143	1.205341333	0.239357061	-0.03173782	0.121307
X3	2.728056586	30.81557613	0.088528495	0.930162006	-60.73781049	66.19392
X4	38.5765949	13.88334659	2.77862363	0.010208439	9.983307352	67.16988

a. (2 points) Write the model explicitly and comment on the coefficient of determination R-square and the Adjusted R square, as well as on entries of the "df". Interpret at least one entry from every column in the last output matrix and make conclusions about the quality of the model. Interpret explicitly all entries for the "intercept" in the last matrix.

b. (5 points) This model includes both X1(price of the competitors) and X4(difference between the price offered by the company's retail stores and their immediate competitor). Interpret the coefficients of X1 and X4. What would be your next steps in dealing with these parameters? Describe your plan and expectations/motivation in as much detail as you can.

c. (5 points) Design and describe steps of hypothesis test for the entire Model. Set up the appropriate null and alternative hypotheses, decision rule, critical value, test statistic and conclusion.

---

6. (Essay type question) (5 points + 4 points + 4 points = **13 points**) (ARCH)

The historical behavior of an exchange rate for two currencies has been modelled with a random walk model, with a trend/drift for the average; the variance has been modelled with an ARCH(2)-model. The estimate of the trend in the random walk model was $\alpha_0 = 0.005$. The estimate in the variance model were $\alpha_0 = 0.0231, \alpha_1 = 0.600, \alpha_2 = 0.400$. The exchange rate at the three last observations were: $X_{t-2} = 6.645$, $X_{t-1} = 6.600$, and $X_t = 6.655$.

a) (5 points) Find the expected value and variance of the exchange rate at time $t + 1$.

b) (4 points) Assume that the errors are normally distributed. Find the probability that X_{t+1}(the exchange rate at time $t + 1$) is greater than 6.75.

c) (4 points) What is the advantage of using a model like ARCH- or GARCH to model the variance?

Department of Statistics

Stockholms universitet

Correction sheet

Date: 21/03/2019

Room: Uggleviksalen

Exam: Financial Statistics

Course: Financial Statistics

Anonymous code: 0028-10k

☒ I authorise the anonymous posting of my exam, in whole or in part, on the department homepage as a sample student answer.

NOTE! ALSO WRITE ON THE BACK OF THE ANSWER SHEET

Mark answered questions

1	2	3	4	5	6	7	8	9	Total number of pages
X	X	X	X	X	X				7+1
Teacher's notes									

Points	Grade	Teacher's sign.

STOCKHOLM UNIVERSITY
Department of Statistics

Spring 2019 A-B

ANSWER FORM Exam – Financial Statistics
2019-03-21

Anonymous code: 0028-NOK *(write clearly!)*

Mark your answers with a clear cross (X) in the corresponding boxes below.

NOTE! Only one cross per question. If more than one alternative has been marked, zero points will be awarded for that question.

NOTE! If, after checking your calculations properly, you are convinced that the correct answer is not included among the given alternatives, write your answer in the margin to the right and explain you reasoning on the back.

	A	B	C	D	E		
1.1	☒	☐	☐	☐	☐	0	-2
1.2	☐	☐	☐	☐	☒	5	
1.3	☒	☐	☐	☐	☐	4	-1
2.1	☐	☒	☐	☐	☐	2	
2.2	☐	☐	☐	☐	☒	6	
3.1	☒	☐	☐	☐	☐	4	
3.2	☐	☐	☒	☐	☐	5	
4.1	☐	☐	☐	☒	☐	2	
4.2	☐	☒	☐	☐	☐	4	

5) 2+5+5 = 12

6) 5+4+4 = 13

32

SU, DEPARTMENT OF STATISTICS

Room: UG Anonymous code: 0028-NOK Sheet number: 1

1. a) $P(X>4)$ $P=0.3$ $n=12$ c.c.l b.i.h

$1-(P(X\leq 4) = P(X>4) = 0.2764$ a)

b) $n=500$ $P(X<175)=$ normal ? → CLT

$E(x)=0.3\cdot 500 = 150$ $V(x)=105$ $X \sim N(\mu, \sigma^2)$

$P(X<175) = P\left(Z < \frac{175-150}{\sqrt{105}}\right) = P(Z<2.44) = 0.9927$ e)

c) H_0: Not normal H_1: Normal

$Obs = 30\left[\frac{(0.320)^2}{6} + \frac{(5.29-3)^2}{2.4}\right] = 7.067125$

Crit = $\alpha = 5\% = 5.991$ $2.5\% = 7.378$

Obs > Crit → Reject H_0

$7.067 > 5.991$ → Reject H_0

$7.067 < 7.378$ → fail to reject

$\alpha = 0.005$ → 10.597 → $7.06 < 10.597$ = fail to reject

a)

α 5% = Reject α 2.5% = Fail to reject

α → fail to reject yes 4

SU, DEPARTMENT OF STATISTICS

Room: UG Anonymous code: 0028-NOK Sheet number: 2

0028-NOK

2. AR(1)

[A] $corr(Y_t, Y_{t-2}) = a_1^2 - P_2 = a_1^2 = 0.2^2 = \underline{0.04}$

b)

[B] MA(2) $COV(Z_t, Z_{t-1}) = (\varepsilon_t + 0.4\varepsilon_{t-1} + 0.2\varepsilon_{t-2})(\varepsilon_{t-1} + 0.4\varepsilon_{t-2} + 0.2\varepsilon_{t-3})$

$= COV(0.4\varepsilon_{t-1}, \varepsilon_{t-1}) + corr(0.2\varepsilon_{t-2}, 0.4\varepsilon_{t-2})$

$= 0.4 \cdot V(\varepsilon_t) + 0.08 V(\varepsilon_t) = 0.48 V(\varepsilon_t)$

$V(Z_t) = V(\varepsilon_t) + 0.4^2 \cdot V(\varepsilon_t) + 0.2^2 \cdot V(\varepsilon_t) = 0.2 \, 1 + V(\varepsilon_t)$

$= \underline{1.2}$

$corr = 0.48 / \sqrt{1.2 \cdot 1.2} = \underline{0.4}$ e)

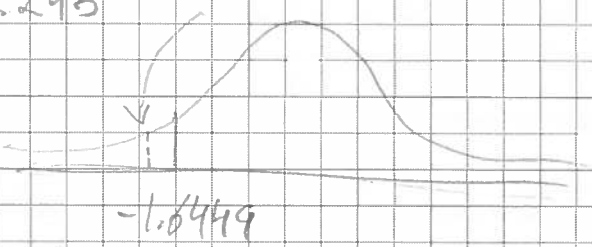
---

3. [A] $R = 17$ $H_0: P_\varepsilon = 0$ $H_1: P_\varepsilon \neq 0$

$E(R) = \frac{50}{2} + 1 = 26$ $V(R) = \frac{50(50-2)}{4(50-1)} = 12.24$

test Variabel $= \frac{17 - 26}{\sqrt{12.245}} = \underline{-2.57195}$ a)

[B] $\alpha = 5\% = 1.6449$



$-2.57195 < -1.6449 \rightarrow$ Reject H_0 c)

→

4. (A) d) $X_4 = 1$ for female

(B) $X_4 = 1$ $X_3 = 1$ Find $P(X | X_3 = 1, X_4 = 1)$

$1/1 + \exp(-(-0.45072 + 2.63928 \cdot 1 + 0.07207 \cdot 1))$

$= 0.90556 = 90.56\% \approx 90.6\%$ b)

$90.5563\% = 90.6\%$

SU, DEPARTMENT OF STATISTICS

Room: UG Anonymous code: 0028-NOK Sheet number: 3

(3 decimals used for space)

5.

A) Model: $\hat{Y}_i = 830.075 + 12.963 X_{1,i} + 0.045 X_{2,i} + 2.728 X_{3,i} + 38.577 X_{4,i} + e$ ← error term

- $R^2 \approx 80\%$ this means that 80% of the variation in the observed values of Y is explained by the model. The remaining 20% of the variation is contained in the "e" error-term.
- $R^2_{adj} \approx 77\%$ has the same interpretation as R^2 only that the R^2_{adj} takes the number of parameters in to account. Since R^2 by formula can't decrease, models of many parameters risk beeing overfitted and the R^2 will increase even if we include an non-significant parameter. The R^2_{adj} punishes models who add non-significant parameters. In our case we see that we have 3 parameters who are not significant and the R^2_{adj} has been lowered compared to the R^2.
- Under the "df" we see that we have 4 regression coefficients and that 25 dessidu-als have been created. They add up to 29 estimations. Out of 30 observations →

- Coefficients, ex: X_1 = This is the value of the regression coefficients, in this case β_1. Since β_1 is + it has a + effect on $\hat{Y}$

Standard error, ex X_1 = This is the standard errors of the estimated coefficients, in this case β_1. The standard error $\hat{\sigma}^2$ is the variance of the coefficients. The SE indicates how much the coefficient can varry. If the SE is bigger than the coefficient (which it is) it means that we have a very uncertain coefficient.

t-Stat, ex: X_1 = This is the observed value of the t-test if $t_{obs} > t_{crit}$ we reject H_0 that $\beta_1 = 0$. t-Stat for β_1 is very small we will need high α to be able to reject H_0.

P-value, ex: X_1 = The P-value is the smalest α that is required to be set in order to be able to reject H_0 $\beta_1 = 0$ in this case we need $\alpha = 46\%$ to reject H_0

lower/upper 95%, ex: X_1 = this is the interval that we with 95% confidence can say that β_1 will fall within. Since our 95% CI captures 0 we cannot at $\alpha = 5\%$ say that β_1 is $\neq 0$. A wide interval means more uncertainty in the model.

- the Intercept has the coefficient -830.075 which means that when X_1 X_2 X_3 X_4 = 0 the estimation of Y = intercept. called β_0.
the Standard error shows the variation of the intercept and that it can vary with ± SE.
The t-Stat is the observed value from the t-test (t_{obs}) if $t_{obs} > t_{crit}$ → Reject H_0 that $\beta_0 = 0$. in this case we [illegible] reject H_0 at a 5% α-level. This can also be seen from the P-value which tells us that we need an α smaller than 4.77% in order to fail to reject H_0.
The 95% CI shows us the interval of values that β_0 can take on with 95 confidence.

In general β_1, β_2, β_3 all capture 0 with their CI, have high P-values and SE which are close to the size of the coefficients. theese Parameters, one could argue should be removed from the model. β_0 and β_4 should be kept, an intercept hardly ever is removed from regression output

2

3) X_1 = For every unit of increase in the price of the competitors the estimated Q-sold increases with 12.963. Everything else held constant.

X_4 = For every unit of positive Price-diff [C-Price − R-Price] > 0. The items sold increase by 38.577. Everything else held constant.

- First of all X_1 is far from beeing significant att a 5% α-level. One alternative could then be to remove X_1 from the model. But since the difference seams to have an impact one could argue that the C-Price still plays an important role. In this case it could be wise to insert an interaction term which assumes a possitiv price diff and the looks at the C-Price. The positive price diff could then be set as a dummy variabel. The aim wold be to look att what a positive price diff, and see if we could make a significant variabel including X_1. Hoping that it plays a role.

5 →

We could also remove x_1 and add a new variable looking at the prices of closly related products or perhaps an index of the branch where we are operating. It might be factors effecting the items sold which go beyond us and our immidiate competitors. We must thus after all be aware of overfitting our model. 5

c) F-test for whole model.

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ H_1: atleast one $\beta_i \neq 0$

Reject H_0 if $F_{obs} > F_{crit}$

$$F_{obs} = \frac{MSR}{MSE} = \frac{136771.929}{5486.603358} = 24.9283$$

$F_{crit} = V_1 = 4 \quad V_2 = 30-4-1 \quad \alpha = 5\%$

$F_{crit} = V_1 = 4, V_2 = 25, \alpha = 5\% = 2.76$

$24.928 > 2.76 \rightarrow$ Reject H_0 at a 5% Significans level. Atleast one β_i is $\neq 0$ which is correct according to the P-values if we set $\alpha = 5\%$

Here we are testing the whole model seeing if it's significant what so ever. As can be seen in the H_0.

5

6. $E(X_{t+1}) = E(a_0) + E(X_t) + E(\varepsilon_{t+1})$

A) $E(a_0) = a_0$ $E(X_t) = X_t$ $E(\varepsilon_{t+1}) = 0$, ε_{t+1} is unknown

$= a_0 + X_t = 0.005 + 6.655 = \underline{6.66}$

$E(X_{t+1}) = \underline{6.66}$

$V(X_{t+1}) = V(a_0) + V(X_t) + V(\varepsilon_{t+1})$

$V(a_0) = 0$ known $\quad V(X_{t-1}) = 0$ known $\quad V(\varepsilon_{t+1}) = h_{t+1}$

$h_{t+1} = \alpha_0 + \alpha_1 \varepsilon_t^2 + \alpha_2 \cdot \varepsilon_{t-1}^2$

$\varepsilon_t = (6.655 - 6.6) - 0.005 = 0.05$

$\varepsilon_{t-1} = (6.6 - 6.645) - 0.005 = -0.05$

$h_{t+1} = 0.0231 + 0.6 \cdot 0.05^2 + 0.4 \cdot (-0.05)^2 = \underline{0.0256}$

$V(X_{t+1}) = V(\varepsilon_{t+1}) = \underline{0.0256}$

5

B) $P(X_{t+1} > 6.75)$ $\varepsilon_t \sim N(0, 0.0256)$

$E(X_{t+1}) = 6.66$

$V(X_{t+1}) = 0.0256$

Standardize $P\left(Z > \frac{6.75 - 6.66}{\sqrt{0.0256}}\right) = P(Z > 0.5625)$

$= 1 - P(Z < 0.5625) = 1 - 0.71226 = \underline{0.28774}$

$P(X_{t+1} > 6.75) \Rightarrow \underline{28.77\%}$

4

→

c) The statement that the variance is constant can often be troublesome. If we for example have a timeseries which has clusters of high variation it is resonabel to assume that the variance will be higer during these clusters. Regular models like AR and single exponential smoothing does not take this into account when estimating values. With ARCH we can take the variation of the error term into account. Under ARCH the variance of the error term depend on α_0 (the minimum variance) and the most recen ressidual between our expected change (α_0), and the actual change. Garch also takes the most recent variance of (ε_t) into account. This in tur will effect the variance of our predicted values. Using ARCH/GARCH our predicted variance for X_{t+1} will depend recent clusters of variation. For ex: If the variance has been high among recent observations the ε_t will increase and so will past h_{t-1}. We will anthe with the help of ARCH/GARCH be able to predict more accurate futhure variances of our predic-tions for example $\hat{X}_{t+1}$. This would not be possible for constan variance, and other models.

Finally, the fact that an ARCH/GARCH model has a minimum variance term α_0 means that we always assume some variation of the ε_t and therefore also in the variance of $\hat{X}_t$. Even if the series varys normally and constanly, it is still reasonable to have a minimum variance of the error terms and our forcasts/estimations.

4