Name: ____________________

I.D. #: ____________________

University of Waterloo

STATISTICS 231
Final Examination
Wednesday April 10, 2002.
7 - 10 PM

Instructor (check one)

M. Zhu (8:30 - 10:30)	☐
W. H. Cherry (10:30 - 12:30)	☐
R.W. Oldford (12:30 - 2:30)	☐
C. C. Springer (2:30 - 4:30)	☐

Time: 3 hours

Instructions

1. Calculators and dictionaries **are** permitted (subject to inspection). Probability tables are provided separately. There are 13 pages including this one.
2. Answer:
   - in the space provided. Use the back of the preceding page if necessary. No extra paper is necessary.
   - in the language of the course. **Be precise and clear.**
   - There is one idea or point you should make for each mark. Part marks will be available.
3. Context:
   - There are two: Airbags (questions 1 to 3; pp. 2 to 9) and Stock prices (quest. 4 and 5; pp. 10 to 13).
   - Background material is distinguished by a different font.
   - Numbered questions have essentially independent contexts. Relevant context appears on the first page associated with each numbered question (you might want to turn the top corner of each relevant page). Question 4 (pages 10-11) can be answered without reference to the context. Question 5 uses the context (and mathematical results) of Question 4.

Question	Pages	Marks Available	Marks Earned
1	2 - 5	30	
2	6 - 7	11	
3	8 - 9	9	

Airbags

(Source: The Economist via The Globe & Mail, March 22, 1997.)

Some countries now require automobile manufacturers to install airbags in new model cars. Airbags rapidly inflate in front of the driver (or passenger) during an automobile crash. The objective is to minimize injury to the person by preventing them from hitting the car interior and/or exiting through the windshield during impact. This is especially important when seatbelts are not worn by the car's occupants (compliance with seatbelt legislation is relatively low in the U.S. for example).

In the **first three numbered questions,** we apply the PPDAC approach to **three separate** ***fictional*** **investigations** on airbags that might reasonably have been carried out in recent years.

1. *(30 marks)* Investigation 1

   Initial airbag legislation required new model cars to have an airbag restraint system for the driver (not necessarily for the passenger). After the airbag legislation had been in place for two years, we will suppose that the Canadian federal government set out to determine the effect of the airbag legislation upon serious injuries suffered by drivers in automobile accidents. Suppose further that each calendar year, the Ontario Ministry compiled (from several different sources) a list of traffic accidents occurring in Ontario during that year. For each automobile involved in the accident, the following items were recorded (or could be inferred): whether or not the driver suffered serious injury, make and model of the car, year of car, and whether or not an airbag restraint system was in place.

   In this investigation, we will suppose that the Canadian federal government is keen to know whether outfitting all cars with driver side airbags would reduce the proportion of accidents which result in serious injury to the driver. Of all the accidents recorded on the Ontario list, 1000 were selected for closer examination; these were the first 500 accidents on the list involving a car with a driver side airbag as well as the first 500 accidents on the list that involved cars without airbags.

   (a) *(6 marks)* **Problem** stage. For this investigation (Investigation 1), clearly describe each of the following:

      i. **(1 mark)** The *target* population.

      ii. **(1 mark)** A single *unit* of the target population.

      iii. **(1 mark)** The *response variate* and indicate clearly what *values* it could take.

      iv. **(1 mark)** The *focal* explanatory variate mentioned in the description of this study and indicate clearly what *values* it could take.

      v. **(1 mark)** One *population attribute* mentioned in the description of this study.

(b) *(11 marks)* **Plan** stage. Within the context of this investigation (Investigation 1; page 2):

i. **(1 mark)** Describe the *study population* here.

ii. **(1 mark)** Describe the *sample* here.

iii. **(2 marks)** What is meant by *study error* in general? Illustrate it by reference to the context.

iv. **(2 marks)** What is meant by *sampling bias* in general? Illustrate it by reference to the context.

v. **(1 mark)** Is this plan an *observational* or an *experimental* one?
Why?

vi. **(2 marks)** Briefly describe how one explanatory variate mentioned above could be *confounded* with the presence or absence of airbags.

vii. **(2 marks)** Does each of 500 cars selected constitute a *block* in the statistical sense?

(c) *(2 marks)* **Data** stage. Suppose that the data were gathered as described in the Plan yielding data on 500 cars with airbags and 500 cars without airbags.

Four major components comprise the Data stage. They are
1. Plan execution, 2. Data Monitoring, 3. Data Examination, and 4. Data Storage.

i. **(1 mark)** What is the purpose of the second component – "Data Monitoring"?

ii. **(1 mark)** What is the purpose of the third component – "Data Examination"?

(d) *(9 marks)* **Analysis.** Because of the potential confounding, it was decided to concentrate only on those cars with a driver side airbag. Out of all such cars involved in accidents, we wish to estimate the proportion in which the driver suffered a serious injury.

To this end, the relevant data for $i = 1, \ldots, 500$ are

$$y_i = \begin{cases} 1 & \text{if the driver of car } i \text{ was seriously injured} \\ 0 & \text{otherwise.} \end{cases}$$

Of these, only 7 y_is equalled 1, the rest were 0.

One mathematical model which might usefully describe this situation is that where each y_i is a realization of a Bernoulli random variate Y_i with

$$Pr(Y_i = y_i) = \pi^{y_i}(1-\pi)^{1-y_i} \quad ,$$

the parameter π being the probability that the i'th car selected has a driver with serious injury and $Y_1, Y_2, \ldots, Y_{500}$ being distributed independently of one another.

i. **(2 marks)** Describe *in words* what each of the following represent and how they differ from one another:

$\widehat{\pi}$

$\widetilde{\pi}$

ii. **(3 marks)** Show that the *log-likelihood function* for π is

iii. **(4 marks)** Derive the maximum likelihood estimate of π for these data, showing it to be 0.014.

Suppose that historical data (prior to the introduction of airbags) show that in all accidents involving new cars (at most two years old) the proportion of drivers to suffer serious injury is 0.042.

(e) *(2 marks)* **Conclusion.** Based on the data from the current investigation, suppose that the above historical value of 0.042 has been determined to be implausible for the value of π

i. **(1 mark)** Write a one sentence conclusion for the Government of Canada describing the main finding of our analysis.

ii. **(1 mark)** Write down one important limitation of this conclusion due to the nature of the investigation.

Further investigation of serious injuries showed that airbags are not without their pitfalls. While much better than being thrown through a windshield, having a nylon reinforced plastic balloon travelling at several hundred kilometres per hour hit you in the chest and face does not seem especially healthy.

Indeed, the U.S. National Highway and Traffic Safety Administration has determined that at least 50 drivers and passengers were killed during 1994–1997 from the impact of an airbag's inflation. Small adults, drivers close to the steering wheel, and children appear to be most vulnerable.

These "dumb" airbags inflate at the same speed regardless of the size and the position of the person; they are perhaps optimum for an average size adult male driver seated midway back from the wheel. So-called "smart" airbags are now being developed which adjust the speed of their inflation according to the person's size and distance from the airbag.

The key to safety is slow inflation. For a large adult positioned well back in the seat, the bag should inflate most quickly; for a small adult seated close to the wheel, the airbag should inflate more slowly but not so slowly as to be ineffective (time to inflation is measured in milliseconds).

Existing airbags are inflated by detonating a charge made of sodium azide which fills the bag with hot nitrogen. New systems use a dual charge inflator – upon impact, a small charge is detonated and the bag begins inflating and, a short time later, a second charge detonates to continue the inflation. For large adults well back in the seat both charges are detonated at once, for small persons close to the airbag only the first charge is detonated, and for all persons and positions between these extremes the time between charges is adjusted – the greater the time between the charges the more slowly the airbag inflates.

The simplest technology is that designed by Breed Engineering of Lakeland, Florida. A switch is mounted under the seat to determine how far back a person is sitting while a piezo-electric crystal (similar to those used in electronic bathroom scales) inside the seat cushion measures the person's weight.

2. *(11 marks)* In this question, you will play the role of a consulting statistician at Breed Engineering.

   Investigation 2

   Problem. Breed has a laboratory capable of conducting crash tests. In each test, an automobile travelling 100 kph (a standard speed) is crashed into a concrete wall. The automobile contains one or more crash test dummies (full scale models of humans) representing the driver and any passengers. Each crash test dummy is equipped with sensors that measured (in newtons) the force of impact it experienced in the crash. Crash test dummies of varying sizes are available for use.

   Breed would like to assess the quality of their product.

   (a) *(5 marks)* **Plan**. Breed has funds to purchase 12 automobiles that could then be crash tested. A driver's side airbag of their design would be installed in each. Each would be crash tested at 100 kph.

   The following model would be used to describe the possible force measurements on the driver dummy:

$$Y_j = \mu + R_j \qquad \text{for } j = 1, \ldots, 12 \tag{1}$$

   with $R_j \sim G(0, \sigma)$ independently for each j.

   i. **(2 marks)** Replication is an important concept in the design of statistical studies. Within this particular study, and model (1), describe: 1. How you would choose to replicate, and 2. What the purpose of this replication would be.

      - How?

      - Purpose of replication?

iii. **(1 mark)** In terms of the safety of the occupants during a crash, why is the value of μ important?

iv. **(1 mark)** In terms of the safety of the occupants during a crash, why is the value of σ important?

(b) **(6 marks) Data** and **Analysis**. We suppose now that Breed Engineering carries out your plan resulting in force measurements $y_1, \ldots y_{12}$. You fit the model (1) as

$$y_j = \mu + r_j \qquad \text{for } j = 1, \ldots, 12$$

to these (fictional) data producing estimates $\widehat{\mu} = 2400$ and $\widehat{\sigma} = 112$.

Based on these data, Breed Engineering would like to promote their airbags to the big auto manufacturers through advertising. The advertising is to provide a numerical range for the force of impact a *typical* driver of an automobile (one with their airbags installed, of course) would experience in a 100 kph collision.

The probability level to be associated with the interval is 0.999.

Construct the appropriate interval.

More exotic technology is the "Advanced restraint system" designed by TRW Vehicle Safety System of Washington, Michigan. They use a "camera" that would appeal to a bat – an ultrasonic transceiver mounted inside the passenger compartment that bounces ultrasonic waves off the interior to determine the size and the position of an occupant.

Safety advocates would like to see the new systems in use immediately but manufacturers will want to be certain that they do not get sued as a result of introducing technology with unexpected flaws. As Walter Kosich, a systems engineer at General Motors's Delphi division, puts it, "You have to make sure it won't be fooled if someone is sitting in the front seat reading a newspaper." An unexpected airbag in your face could really put you off the comics.

3. *(9 marks)* Investigation 3

Problem. Imagine that you are in the position of a car maker and that you are being approached by the designers of the new "smart" airbags; they would like to convince you to employ their airbag in your future models. Both Breed and TRW have had independent laboratories conduct crash tests on their airbags using cars from a variety of manufacturers and provide you with the results.

TRW's ultrasonic system is likely to be more expensive than Breed Engineering's simpler system so we would like to see for ourselves whether it is worth the extra cost.

(a) *(2 marks)* **Plan.** To this end we take 10 cars, 2 of each of the 5 different models we manufacture. For each model, one of the two cars is chosen at random to have installed a TRW airbag system – the other gets a Breed Engineering airbag system. Each car is to be crash-tested at 100 kph with the *same* size crash-test dummy in each test.

Using the index i to indicate airbag type ($i = 1$ for Breed and $i = 2$ for TRW) and the index $j = 1, 2, 3, 4, 5$ to indicate car model, a model for the repeated execution of this Plan where force is the response variate is

$$Y_{ij} = \mu_i + \alpha_j + R_j \tag{2}$$

with $R_{1,j} \sim G(0, \sigma)$ independently for $i = 1, 2$ and $j = 1, \ldots, 5$.

i. **(1 mark)** What is the principal advantage of assigning the airbag type at random?

ii. **(1 mark)** What does the parameter α_j of model (2) above represent?

Data. Suppose the results are as follows

Manufacturer	Model 1	Model 2	Model 3	Model 4	Model 5
Breed Engineering ($i = 1$)	2593	2463	2273	2413	2420
TRW ($i = 2$)	2340	2088	2176	2219	2289
Difference	253	375	97	194	131

The following summary statistics were calculated as well.

Breed Engineering ($i = 1$):

$$\overline{y}_1 = 2432.4 \quad \text{and} \quad \sqrt{\frac{\sum_{j=1}^{5}(y_j - \overline{y}_1)^2}{4}} = 114.75$$

(b) *(6 marks)* **Analysis.** Here you will address a single issue – are the two airbag types equally effective?

i. **(1 mark)** In terms of the parameters of model (2, page 8), formally express the hypothesis that the two airbag types are equally effective.

ii. **(5 marks)** Based on the above data, formally assess whether there is a difference in the effectiveness of the two airbag types.

(c) **(1 mark) Conclusion**. Suppose that our analysis strongly indicates that indeed the two airbag types were different in that the more expensive TRW system produced an average force of 100

Stock prices: rates of return

Investment firms provide estimates of the systematic risks of stocks, called *betas*. A stock's *beta* measures the relationship between the stock's rate of return (% change from previous time period) and the rate of return for the stock market as a whole.

The term *beta* derives its name from the beta coefficient for the slope in a straight line model in which a stock's rate of return is the response variate, Y, and the market's rate of return is the explanatory variate x as in the following model:

$$Y = \alpha + \beta(x - \overline{x}) + R$$

4. *(15 marks)* We'll first consider some of the mathematical features of this model for a single stock, taken at 10 different times ($i = 1, \ldots, 10$). The following mathematical model relates the response variate Y to the explanatory variate x for 10 (x_i, Y_i) data pairs.

$$Y_i = \alpha + \beta(x_i - \overline{x}) + R_i; \text{ for } \quad i = 1, \ldots, 10,$$

where, for the selected time period i, Y_i represents the rate of return for the stock, x_i the rate of return for the market as a whole, and where $R_i \sim G(0, \sigma)$, independently.

Answer the following questions with respect to this model.

(a) **(3 marks)** If the above model is written in matrix form,

$$\underline{Y} = X\underline{\theta} + \underline{R},$$

clearly write down the elements of the vectors $\underline{Y}$, $\underline{R}$, $\underline{\theta}$, and of the matrix X.

(b) **(1 mark)** Show that $\sum_{i=1}^{10}(x_i - \overline{x}) = 0$.

(c) **(5 marks)** If the observed values of the response variate are $y_1, y_2, \ldots, y_{10}$, show (by whichever method you like) that the least squares estimates of α and β are

$$\widehat{\alpha} = \overline{y} \qquad \text{and} \qquad \widehat{\beta} = \{\sum_{i=1}^{10}(x_i - \overline{x})y_i\}/\{\sum_{i=1}^{10}(x_i - \overline{x})^2\}.$$

(d) **(5 marks)** Show that $\widetilde{\beta} \sim G\left(\beta,\ \sigma \ / \sqrt{\sum_{i=1}^{10}(x_i - \overline{x})^2} \right)$. Take care to show your reasoning.

5. *(7 marks)* We will now use this context and the mathematical results of the previous question to analyze the rate of return for two different stocks – the computer manufacturer IBM, and the foodstuff company General Mills (GM).

Problem The problem is to analyze the rate of return of various stocks to determine a good investment portfolio.

Plan We are interested in choosing between two stocks, IBM and General Mills. We will select 10 years to work with and record the rate of return (% change) for each stock (IBM and General Mills) as well as the rate of return of the stock market as a whole.

The model is as in the previous question except that there are different parameters for each stock (indicated with subscripts "IBM" and "GM") and the parameter σ is the *same* for both stocks.

The following model, then, relates the rate of return each stock ($i = 1, \ldots, 10$ for IBM and $i = 11, \ldots, 20$ for General Mills) to the overall stock market rate of return, x.

$$\begin{aligned} Y_i &= \alpha_{IBM} + \beta_{IBM}(x_i - \overline{x}) + R_i, \quad \text{for} \quad i = 1, \ldots, 10, \quad \text{and} \\ Y_i &= \alpha_{GM} + \beta_{GM}(x_i - \overline{x}) + R_i, \quad \text{for} \quad i = 11, \ldots, 20, \end{aligned}$$

with $R_i \sim G(0, \sigma)$, independently for $i = 1, \ldots, 20$. Note that because $i = 1$ is the same year as $i = 11$ we have $x_1 = x_{11}, x_2 = x_{12}, \ldots, x_{10} = x_{20}$.

Data

	Rates of Return (%)									
Year	1	2	3	4	5	6	7	8	9	10
IBM	21.1	-21.9	-29.7	37.4	28.1	7.3	1.6	-9.1	10.8	4.3
General Mills	57.4	-23.0	-34.4	50.4	27.3	-11.3	9.7	11.8	33.1	1.6
Market	19.0	-14.7	-26.5	37.2	23.8	-7.2	6.6	18.4	32.4	2.1

Analysis From this data and from fitting this model via least-squares, we have

$$\sum_{i=1}^{10}(x_i - \overline{x})^2 = \sum_{i=11}^{20}(x_i - \overline{x})^2 = 3887.829$$

and parameter estimates of

$$\widehat{\alpha}_{IBM} = -2.660, \quad \widehat{\beta}_{IBM} = 0.840, \quad \widehat{\alpha}_{GM} = 0.107, \quad \widehat{\beta}_{GM} = 1.334, \quad \text{and} \quad \widehat{\sigma} = 12.77.$$

Note also that the model has been constructed so that all corresponding estimators $\widetilde{\alpha}_{IBM}$, $\widetilde{\alpha}_{GM}$, $\widetilde{\beta}_{IBM}$, $\widetilde{\beta}_{GM}$, and $\widetilde{\sigma}$ are distibuted independently from one another.

(a) **(6 marks)** Are the data consistent with the claim that $\beta_{IBM} = \beta_{GM}$?
Formally justify your answer.

(b) **(1 mark)**

If $\beta > 1$, then the stock's rate of return (% change) is greater than that of the market as whole; if $\beta < 1$, its rate of return is less than the market as a whole. A stock is called "aggressive" if $\beta > 1$, "neutral" if $\beta = 1$, and "defensive" if $\beta < 1$.

Conclusion If you were someone who did not mind taking risks (i.e. you were an aggressive investor), which stock would you invest in, IBM or General Mills? Or would it matter? Explain