University of Stellenbosch

Graduate School of Business

Business Statistics

Group Assignment

---

written: April 2002

from: Group 4

Mike De Kock
Org Groeneveld
Fatima Hassan
Alexander Markowski
Henrik Riedel
Karin Williams

Declaration

Hereby We, Group 4 of the Fulltime MBA Class, declare that this group work is our own original work and that all sources have been accurately reported and acknowledged, and that this document has not previously in its entirety or in part been submitted at any university in order to obtain an academic qualification.

Bellville, 2002-04-21

Contents

List of Tables

List of Figures

1 Case Study No. 1

1.1 Exercise 2.1

Type of Error	Frequency
Copy error	54
Layout	7
Omits	13
Paste-up	11
Poor reproduction	8
Ran in error	30
Rate quote	13
Space not ordered	7
Typesetting	53
Velox	28
Wrong ad	25
Wrong date	14
Wrong position	45
Wrong paste-up	5
Wrong size	6
TOTAL	319

Table 1.1: Errors for last year

1.2 Exercise 2.2

(a).

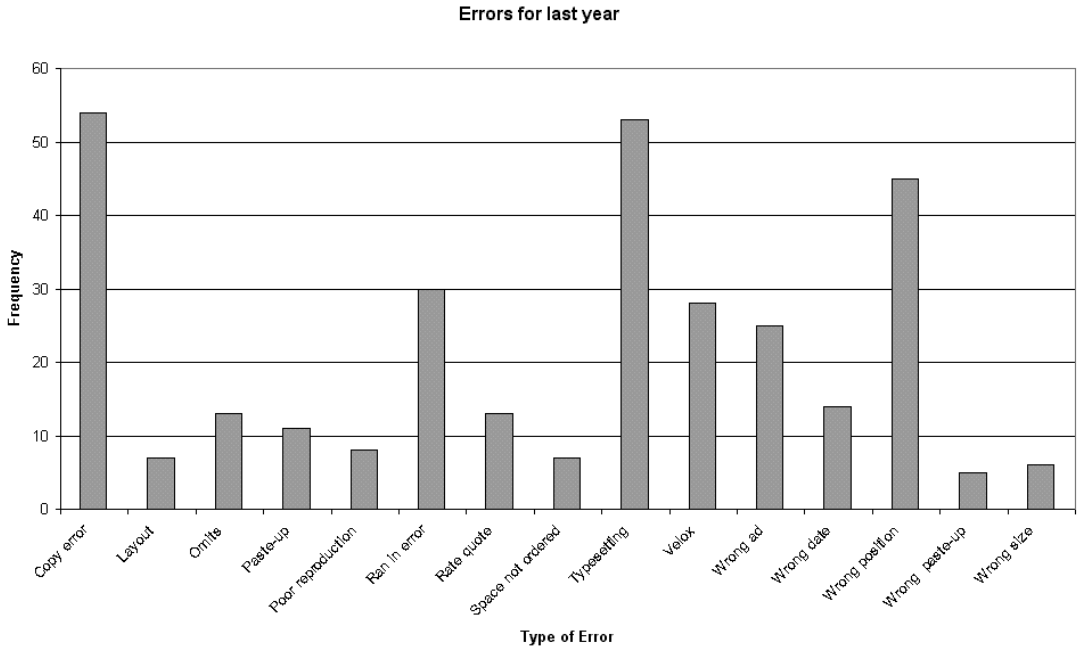

Figure 1.1: Errors for last year

(b). The graph selected is bar chart for the following reasons:

- It describes a single set of data i.e. the type of error.
- It emphasis the amount of each error within a specific period.
- It conveys the data in an easy to comprehend visual manner.

(c). Report to management:

During the last calendar year the following errors were particularly responsible for the high costs namely; copy error, typesetting, wrong position, ran in error and velox. The range in the frequency of the errors is $54 - 5 = 49$.

Recommendations:

- Reduce the errors mentioned above by finding out their causes.
- Improve efficiency.

- Improve planning.
- Improve departmental coordination

1.3 Exercise 2.3

Useful information about the errors:

- Cost of the errors.
- Components of the 5 main errors.
- Different errors in relation to the time period i.e. during the peak period which errors feature.

1.4 Exercise 2.4

Type of Error	**Frequency ($ 000)**
Copy error	32.6
Layout	3.0
Omits	36.5
Paste-up	59.4
Poor reproduction	13.0
Ran in error	108.2
Rate quote	5.3
Space not ordered	12.9
Typesetting	53.1
Velox	23.3
Wrong ad	53.6
Wrong date	35.9
Wrong position	74.9
Wrong paste-up	16.5
Wrong size	6.3
TOTAL	534.5

Table 1.2: Cost of advertising errors

1.5 Exercise 2.5

(a).

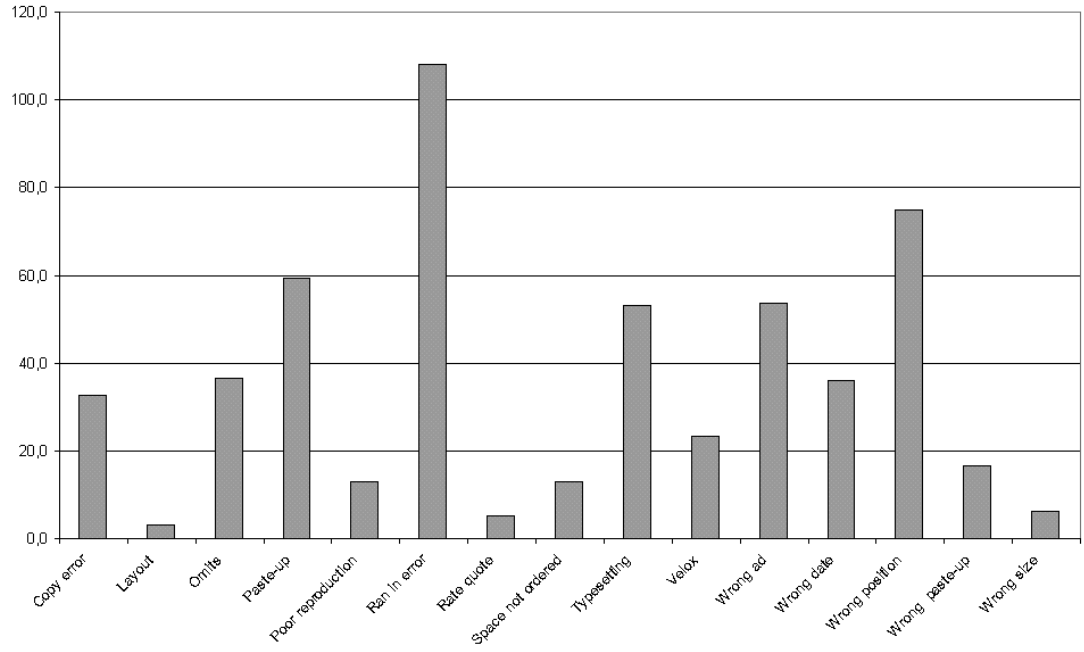

Figure 1.2: Cost of advertising errors

(b). The graph selected is bar chart for the following reasons;

- It describes a single set of data i.e. the type of error.
- It emphasis the amount of each error within a specific period.
- It conveys the data in an easy to comprehend visual manner.

(c). Report to management:

- During the last calendar year the following errors were particularly for the costs namely; ran in error, wrong position, paste up, wrong ad, typesetting.
- The range is $108.2 - 3.0 = 105.2$, which is very high.

1.6 Exercise 2.6

Recommendations:

- Find the causes of the errors in order to reduce them.
- Improved planning.
- Improve coordination.

1.7 Exercise 2.7

Type	Frequency	Amount ($ 000)
Composing room	33,33%	11,84%
Policy	53,33%	82,05%
Sales	13,33%	6,11%
Total	100,00%	100,00%

Table 1.3: Frequency and Amount of errors

1.8 Exercise 2.8

(a).

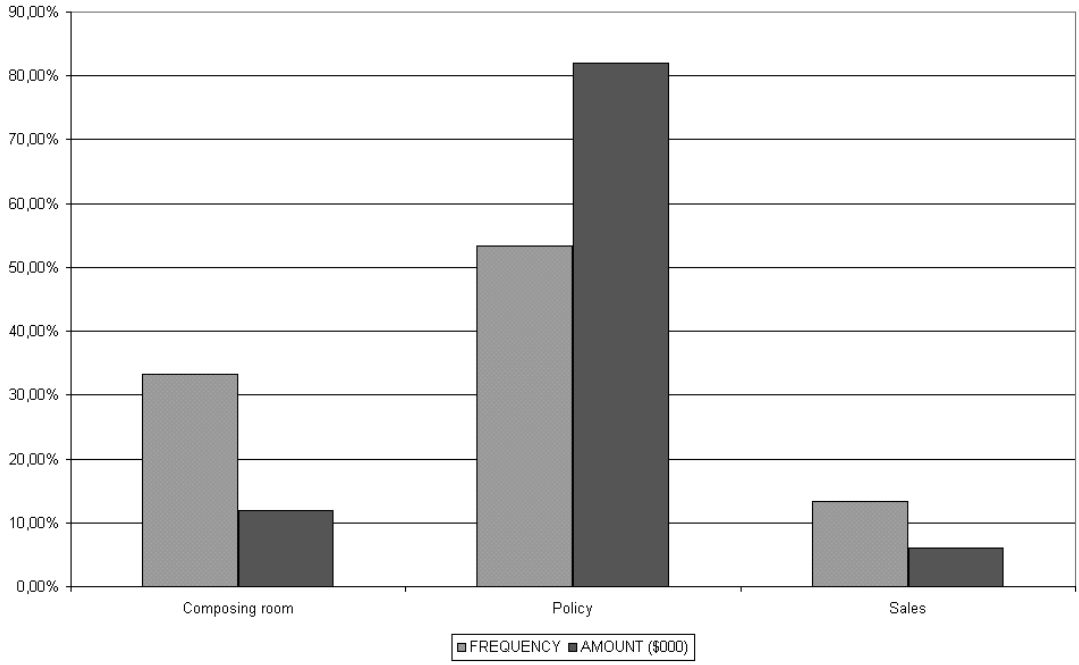

Figure 1.3: Frequency and Amount of errors

(b). The graph selected is a bar chart for the following reasons:

- The data is nominal.
- It describes two populations of data.
- Conveys information in an easy to comprehend visual manner.
- It enhances the reader's ability to grasp the substance of the data.
- It summarises the data but can't be used for in depth analysis.

(c). Bar chart was selected before, therefore no explaination.

(d). Policy is responsible for a large percentage of the ran in error followed by composing room and sales.

(e). Recommendations:

- Examine what is causing policy to have such a high percentage of errors.
- At what times is the error most occurring i.e. are there seasonal flactuations.
- Improve the planning.

1.9 Exercise 2.9

(a).

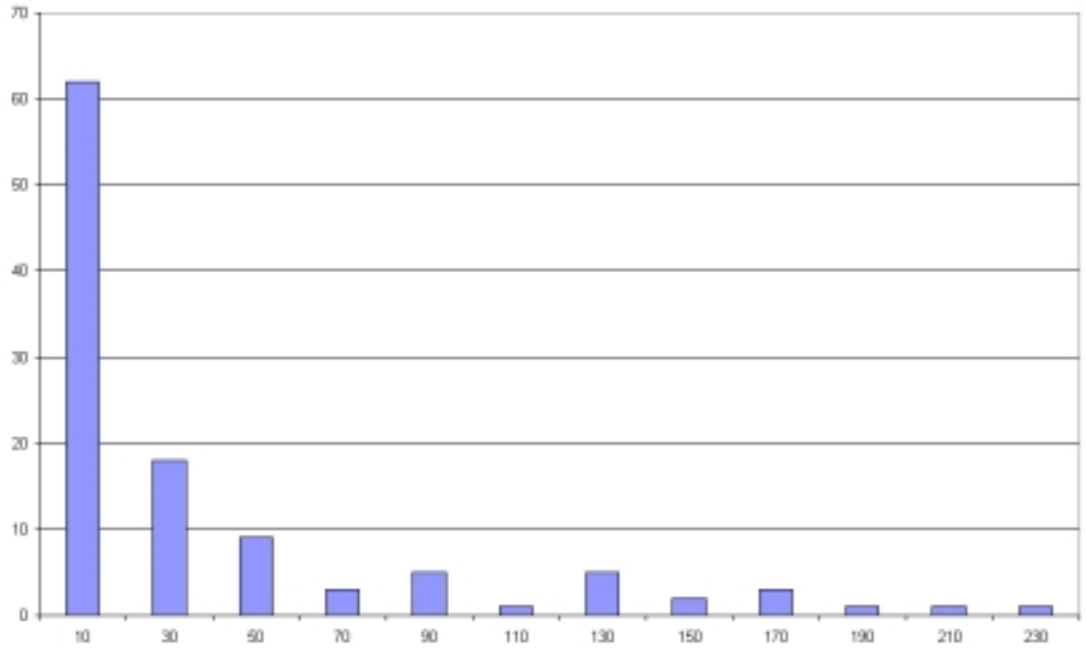

Figure 1.4: Number of Cartridges accessed per job

Class	Frequency
$0^- - 20$	62
$20^- - 40$	18
$40^- - 60$	9
$60^- - 80$	3
$80^- - 100$	5
$100^- - 120$	1
$120^- - 140$	5
$140^- - 160$	2
$160^- - 180$	3
$180^- - 200$	1
$200^- - 220$	1
$220^- - 240$	1

Table 1.4: Number of Cartridges accessed per job

(b). Report to the management:

- The distribution is skewed with large amounts of data in the first three classes.
- 20% of the data is found in the other nine classes.
- There seem to be many outliers as the frequency in the classes are ranging between 9 and 1.
- The range is very high, with the largest value being 237 and the smallest 1.

Conclusion:

- Management should investigate the use of the larger number of data cartridge and if they are not important, they can be left out.
- Management should investigate more on the inventory of smaller data cartridge as the frequency is higher.

2 Case Study No. 2

2.1 Preface

We want to compare the average length of call made early in the evening to calls made later in the evening. We used the following methode, but this depends on whether the two unknown population variances are equal.

	Early	**Late**	**Δ**
	41.3	37.1	4.2
	37.5	38.9	-1.4
	39.3	42.2	-2.9
	37.4	45.7	-8.3
	33.6	42.4	-8.8
	38.5	39	-0.5
	32.6	40.9	-8.3
	37.3	40.5	-3.2
	40.6	40.7	-0.1
	33.3	38	-4.7
	39.6	43.6	-4
	35.7	43.8	-8.1
	31.3	34.9	-3.6
	36.8	35.7	1.1
	36.3	47.4	-11.1
$\bar{x}$	36.74	40.72	-3.98
s	2.9828	3.5672	4.2862
$\sum x_i$	551.1	610.8	-59.7

Table 2.1: Call Statistic

We made use of the pooled variance estimate stated in Keller & Warrack (1999),

because in combining both samples, we produce a better estimate.

2.2 Answer 1

$$\begin{aligned} H_0 &= \mu_1 - \mu_2 = 0 \\ H_1 &= \mu_1 - \mu_2 > 0 \end{aligned}$$

$$\begin{aligned} t - stat &= \frac{(\bar{x}_1 - \bar{x}_2) - (0)}{\sqrt{S_{p^2} \cdot (\frac{1}{n_1} + \frac{1}{n_2})}} && (2.1) \\ &= \frac{(36.74 - 40.72) - (0)}{\sqrt{10.81 \cdot (\frac{1}{15} + \frac{1}{15})}} \\ &= \frac{-3.98}{1.2007} \\ &= -3.3314 && (2.2) \end{aligned}$$

$$\begin{aligned} S_{p^2} &= \frac{(n_1 - 1)s^2 + (n_2 - 1)s^2}{n_1 + n_2 - 2} && (2.3) \\ &= \frac{14 \cdot 8,8969 + 14 \cdot 12,7246}{15 + 15 - 2} \\ &= \frac{302.7}{28} \\ &= 10.8107 && (2.4) \end{aligned}$$

$$t_{\alpha,df} = 1.701 \quad (2.5)$$

In this case, we accept the null hypothesis. We conclude that there is not enough evidence to infer that the means of the length of calls early and late differ.

2.3 Answer 2

$$\begin{aligned} H_0 &= \mu_D = 0 \\ H_1 &= \mu_D > 0 \end{aligned}$$

$$\begin{aligned} t - stat &= \frac{\bar{x}_D - \mu_D}{\frac{s_D}{\sqrt{n_D}}} && (2.6) \\ &= \frac{-3.98 - 0}{\frac{4.2862}{\sqrt{15}}} \\ &= -3.61 && (2.7) \end{aligned}$$

$$t_{\alpha, n_D - 1} = 1.761 \quad (2.8)$$

We reject the null hypothesis, with proof that the difference between the two means are not null. This proofs that the means differ.

2.4 Answer 3

Maybe they need to change the time they make the calls. The used time between 5 pm and 7 pm is not very practical. It should be considered to call in the morning at work. People are busy at night.

The success of the call should be investigated instead of the length, because the length of a call does not determine the successfull closure of a sale.

3 Case Study No. 3

3.1 Phase 1

3.1.1 Case data

The marketing department already established that there is a possibility that some households will only subscribe if they receive a discount. The purpose of the experiment is to assist in deciding on the viability of an incentive scheme, and if viable, what form it should take on. The sample was taken randomly and each sample subjected to independent tests only. The results indicate the expected value of subscriptions for each incentive plan. The only probability that do not indicate a specific value is the option of a restaurant card. In order for us to compare this incentive in a fair and reasonable way, we need to first make some assumptions and then compare with other incentives. We will assume that firstly the full 15% discount will be all for the newspaper's account, or secondly that the restaurants carry loss of 5% of the discount and the newspaper foots the bill for the remaining 10%.

3.1.2 Expected value added compared to Incentives:(Calculations)

$$\text{Expected value} = \frac{\text{Revenue from Paper sales}}{100 \cdot \text{households}} = R \cdot P = E \tag{3.1}$$

$$\frac{\text{Revenue received by Herald}}{\text{paper sold}} = R \tag{3.2}$$

$$\frac{\text{Probability of sales}}{100 \cdot \text{households}} = P \tag{3.3}$$

No discount $E = 4.5 \cdot 34 = 153$

Moderate Discount	$E = 4 \cdot 37 = 148$
Substantial Discount	$E = 3 \cdot 38 = 114$
Discount Restaurant card	$E = x \cdot 53$

We compare $53 \cdot x$ to our highest expected value: $ 153(no discount) and then solve for x to calculate the maximum discount allowed to favour this option: 53x=153 Solve for x and find: x=2.89 At a 15% cost of discount to Herald, the maximum purchases per household equals $19.25 At a 10% cost of discount to Herald, the maximum purchase per household equals $28.90.

3.1.3 Summary

From the limited data available, we can infer that the households interested in subscription is around 34%. Even by slashing prices up to 25%, we only notice a rise in sales of 12%. Value created by offering incentives do not appear to be significant enough to escalate the bottom line. We would be able to build onto this data with more info on fixed/variable cost as well as profit margin. With this info we could establish how an increase in sales can dilute the fixed cost and increase the profit above the breakeven point.

The assumption is also that a lot of cost in the newspaper industry results from fixed cost per publication, such as journalists, printing press etc. Variable cost will typically include ink, labour and maintenance and a remarkable variable revenue in advertising. These factors will greatly influence the price/volume decisions.

The discount offered by the restaurant card seems very attractive. If one analyses this incentive from the prospective customer's perspective, a simple calculation tells us that the entire household needs to spend only $26.67 per week on the card before they have made up the cost of the paper per week. If we assume an average household of 4 people, that amounts to $6.67/person/week. Households can easily calculate their average cumulative restaurant expenditure per week and justify the discount card.

The recommended discount plan would be the restaurant card, but only if we can limit the expenditure on the card to less than \$19.25/week, or in the case of the Herald only compensating the restaurants for 10% of the discount, we can recommend the plan if maximum expenditure per card and week $\leq$ \$28.90. If however the discount card can not be capped, the first option of no discount is still recommended.

3.2 Phase 2

3.2.1 Case Data

The cross-tabulation results obtained, comes in the form of qualitative variables, and we would like to calculate if there is enough evidence to infer that these variables are related. We would also like to infer if differences exist among two or more populations of qualitative variables. The marketing implications of detecting possible purchase behaviour with other variables will assist us in our choice of focus in marketing campaigns. In order for us to test the relation/dependency of variables, we will use the method named the "Chi squared test of a contingency table"

3.2.2 Reading of other newspapers versus status of Home Delivery

Can we conclude that the choice to read other papers have any relation to the choice to subscribe to home delivery of the Herald? Null Hypothesis and Alternative Hypothesis:
H_0: 2 variables are independent
H_1: 2 variables are dependant.

Reads other Newspapers			
Home delivery	yes	no	total
yes	61	75	136
no	77	139	216
total	138	214	352
Test Statistic $\chi^2 = 2.9667$			
P-Value = 0.085			

Table 3.1: Contingency Table 1

Degrees of freedom = 1 and our χ^2 critical value at a confidence level of 95% indicates a value of 3.8414. Therefore we do not reject the null hypothesis. The P-value of 0.085 indicates that we stand a chance of 8.5% to be wrong, should we reject the null Hypothesis. We assume no relation.

3.2.3 Ownership of a Restaurant Card versus status of Home Delivery

Can we conclude that the choice to own a restaurant card will have any relation to the choice to subscribe to Home delivery of the Herald? Null Hypothesis and Alternative Hypothesis:

H_0: 2 Variables are independent

H_1: 2 variables are dependant

Restaurant Card			
Home delivery	yes	no	total
yes	26	110	136
no	40	176	216
total	66	286	352
Test Statistic $\chi^2 = 0.0197$ P-Value = 0.8885 DF =1 χ^2(Crit) at 5% = 3.84			

Table 3.2: Contingency Table 2

We infer that at a confidence level of 95% we cannot reject the null hypothesis. The p-value tells us that we stand a chance of 88,85% to be wrong if we reject the null hypothesis. We assume independence between variables.

3.2.4 Purchase behaviour from Monday to Saturday versus interest in trial subscription

Can we conclude that the purchase behaviour of non-subscribers from Monday to Saturday can be related to their willingness to undertake a trial subscription? Null Hypothesis

and Alternative Hypothesis:

H_0: Variables are Independent

H_1: Variables are Dependant

Monday-Saturday purchase behaviour			
Interest in trial subscription	every day	most days	total
yes	29	17	46
no	49	121	170
total	78	138	216
Test Statistic $\chi^2 = 18.3758$ P-Value = 0 DF = 1 χ^2(crit) at 5% = 3.84			

Table 3.3: Contingency Table 3

We infer at a confidence level of 95% that we can reject the null hypothesis and assume dependency between these variables. The P-value indicates a zero chance of rejecting the null hypothesis incorrectly.

3.2.5 Purchase behaviour on Sundays versus interest in trial subscription

Can we conclude that the purchase behaviour of non-subscribers on Sundays can be related to their willingness to undertake a trial subscription? We have combined the last two "Sunday purchase" columns to adhere to the "Rule of Five". Null Hypothesis and Alternative Hypothesis:

H_0: Variables are Independent

H_1: Variables are Dependant

We infer at a confidence level of 95% that we cannot reject the null hypothesis. The p-value indicates that if we reject the null hypothesis, we stand a 5.22% chance of rejecting incorrectly. We assume independence between variables.

Sunday purchase behaviour			
Interest in trial subscription	every sunday	1-3/month	total
yes	35	11	46
no	103	67	170
total	138	78	216
Test Statistic $\chi^2 = 3.7695$ P-Value = 0.0522 DF = 1 χ^2(crit) at 5% = 3.841			

Table 3.4: Contingency Table 4

3.2.6 Purchase location versus interest in trial subscription

Can we conclude that the location purchase, of non-subscribing households are related to their willingness to undertake a trial subscription? Null Hypothesis and Alternative Hypothesis:

H_0: Variables are independent

H_1: Variables are Dependant

Interest in trial subscription			
Where purchased	yes	no	total
Convenience Store/Delicatessen	12	62	74
Stationary/Candy Store	15	80	95
Vending Machine	10	11	21
Supermarket	5	8	13
Other	4	9	13
total	46	170	216
Test Statistic $\chi^2 = 14.5208$ P-Value = 0.0058 DF = 4 χ^2(crit) at 5% = 9.488			

Table 3.5: Contingency Table 5

We infer at a 95% confidence level that we can reject the null hypothesis. The p-value indicates that we stand a 0.58% chance of rejecting the null hypothesis incorrectly. We conclude dependence between variables.

3.2.7 Purchase Behaviour: Monday-Saturday versus Sunday

Can we conclude that the purchase behaviour of non-subscribing households on Sundays are related to their purchase behaviour from Monday to Saturday? For the purpose of fair comparison, we combined the column "occasionally" with "most days". "occasionally" is a lot more vague than "once/month". Null Hypothesis and Alternative Hypothesis:

H_0: Variables are independent

H_1: Variables are Dependant

Monday-Saturday purchase behaviour			
Sunday purchase behaviour	every day	most days	total
Every Sunday	55	83	138
2-3/month	19	35	54
once/month	4	20	24
total	78	138	216
Test Statistic $\chi^2 = 4.7916$ P-Value = 0.0911 DF = 2 χ^2(crit) at 5% = 5.99			

Table 3.6: Contingency Table 6

We infer at a 95% confidence level that we cannot reject the null hypothesis. The p-value indicates that we stand a 9.1% chance of rejecting the null hypothesis incorrectly. We conclude independence between variables.

3.2.8 Conclusion

From the analysis we can see that only two tables indicate a dependency between variables. These findings describe some vital relations in the purchasing behaviour. The Monday- Saturday purchasers vary between those interested in the trial subscription and those who are not interested. Households who purchase their papers more regularly during the week, tend to be more interested in the trial subscription, and this group can be targeted to switch to full-subscription.

The location of purchase, between those households interested in purchasing, and those who are not, indicates that those not interested in subscription are more likely to make their purchase at the convenience store/delicatessen and the stationary/candy store. This indicates that they probably frequent these stores more often and prefer to purchase there on choice, instead of fulltime subscription. Marketers should rather target the group who do not frequent these stores that often and persuade them to subscribe fulltime for convenience.

4 Case Study No. 4

4.1 Display of Data

Since there is only a small amount of random variation in our data, there is no need for smoothing.

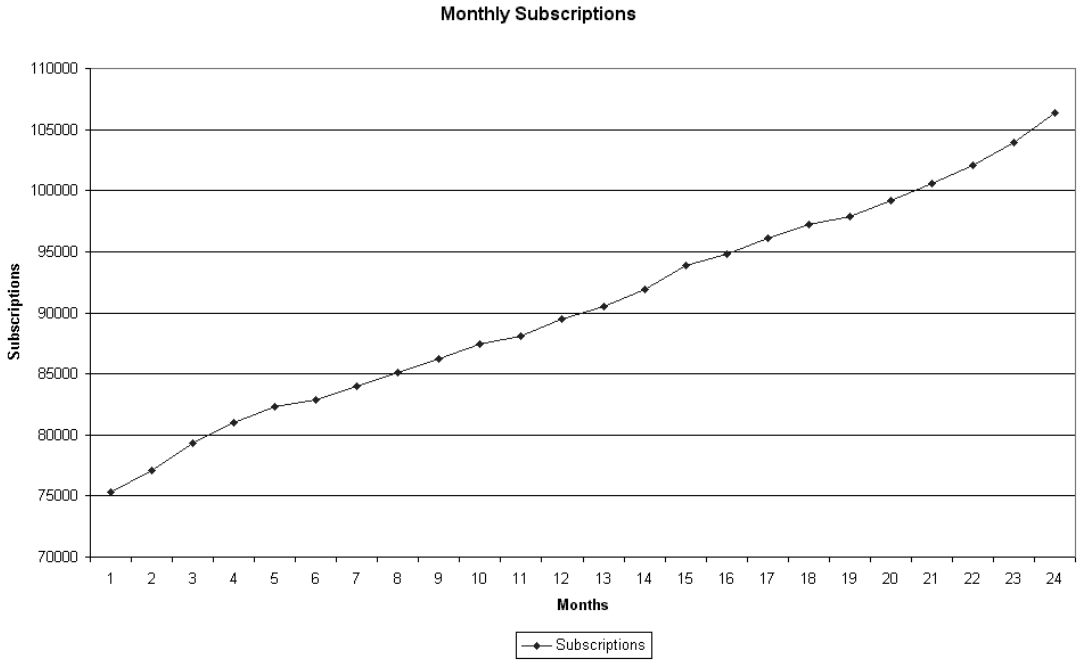

Figure 4.1: Data Plot of time series

4.2 Calculating the trend

As we can see in Figure 4.1, there seems to be a strong trend, which will be evaluated by using regression analysis. The plot seems to be very linear, what makes it unnecessary to use the quadratic model. We will use the linear model as shown in Keller & Warrack (1999) instead:

$$\hat{y} = \beta_0 + \beta_1 \cdot t + \epsilon \tag{4.1}$$

The regression statistics is done by using the data analysis from Microsoft Excel 2000 and the results are shown in table 4.2.

Regression Statistics	
Multiple R	0.9974
R^2	0.9949
Adjusted R^2	0.9946
s_ϵ	641.72
Observations	24

ANOVA

	df	*SS*	*MS*	*F*	*Sign. F*
Regression	1	$1.752 \cdot 10^9$	$1.752 \cdot 10^9$	4,255.55	$1.12 \cdot 10^{-26}$
Residual	22	$9.059 \cdot 10^6$	$0.412 \cdot 10^6$		
Total	23	$1.761 \cdot 10^9$			

	Coeff.	*Std Error*	*t-Stat*	*P-value*
Intercept	75,095	270.39	277.73	$1.70 \cdot 10^{-40}$
x-value	1,234	18.92	65.23	$1.12 \cdot 10^{-26}$

Table 4.1: Simple Regression

We can derive the regression equation directly from table 4.2 and insert the figures into equation 4.1. By doing this we can assume the trend in our time series to be:

$$\hat{y} = 75,095 + 1,234 \cdot t, where\ t = 1, 2, \ldots, 24$$

4.3 Deviations from the trend

We can now compute the $\hat{y}$ for every month as well as the percentage of trend, stated in equation 4.2.

$$\text{percentage of trend (PoT)} = \frac{y \cdot 100}{\hat{y}} \tag{4.2}$$

Month	**Subscriptions**	$\hat{\mathbf{y}}$	$\frac{\mathbf{y \cdot 100}}{\hat{\mathbf{y}}}$
1	75,327	76,329	98.69
2	77,116	77,563	99.42
3	79,341	78,797	100.69
4	80,983	80,031	101.19
5	82,326	81,265	101.31
6	82,879	82,499	100.46
7	84,006	83,733	100.33
8	85,119	84,967	100.18
9	86,182	86,201	99.98
10	87,418	87,435	99.98
11	88,063	88,669	99.32
12	89,444	89,903	99.49
13	90,507	91,137	99.31
14	91,927	92,371	99.52
15	93,878	93,605	100.29
16	94,784	94,839	99.94
17	96,109	96,073	100.04
18	97,189	97,307	99.88
19	97,899	98,541	99.35
20	99,208	99,775	99.43
21	100,537	101,009	99.53
22	102,028	102,243	99.79
23	103,977	103,477	100.48
24	106,375	104,711	101.59

Table 4.2: Computed linear trend and PoT

This percentage of trend shows the deviation of y from $\hat{y}$ as a percentage. This deviation might exist due to cyclyc or random variation, since the seasonal variation is

close to nonexistant as we can derive from figure 4.2. A seasonal variation is usually spotted within a time up to one year, all other variations ($> 1\ year$) are supposed to be cyclic.

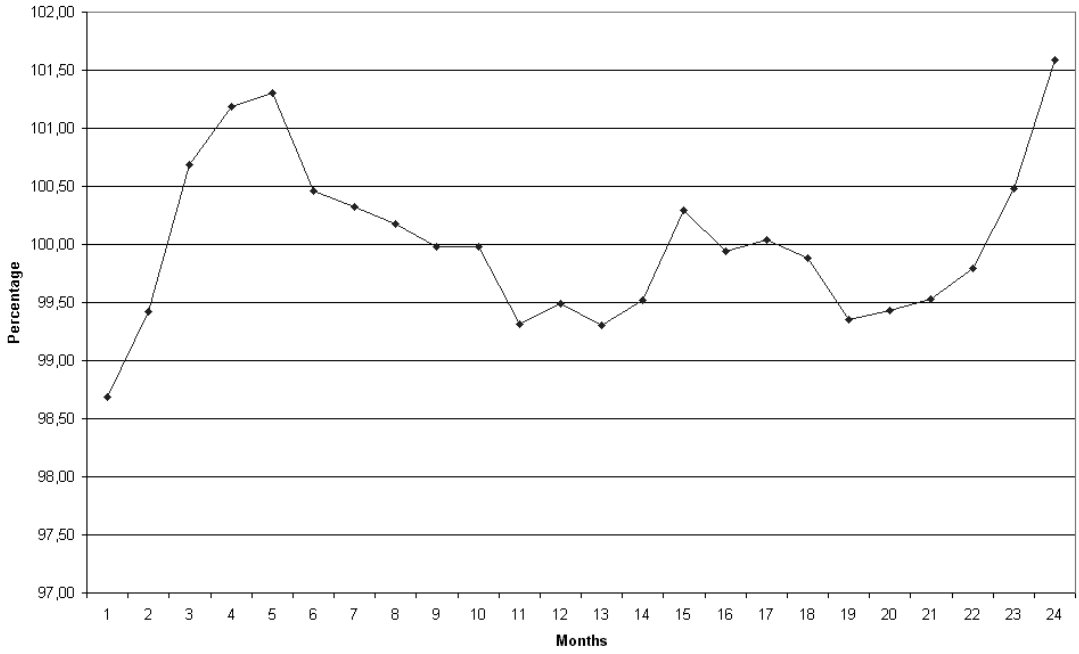

Figure 4.2: Data Plot of percentage of trend

What kind of variation is existant in our time series? Refering to figure 4.2 we can spot no cyclic pattern (not taking into account the maybe 22-month pattern, which can not be proven due to the short time series). Thererfore the variations are caused randomly, which makes it more difficult to predict values in the future. With random variations, our forecasting model stays at the level we reached with the linear regression and our model is described by the following equation:

$$\hat{y} = 75,095 + 1,234 \cdot t \tag{4.3}$$

Month	$\hat{y}$
25	105,945
26	107,179
27	108,413
28	109,647

Table 4.3: Forecasted subscriptions

4.4 Forecasting

Since our time series consists of only 2 years (24 months), there is usually not enough evidence to predice a long time into the future. Normally we would like to have approximately 4 years (48 months) of data to develop a model that is able to forecast a longer time into the future. Maybe we also discover a kind of cyclic pattern in these 48 months to improve the precision of our model.

4.5 Comparison

Due to the circumstances that we are not equipped with the table from page 851, our ability to compare the data with the results of forecasting from table 4.3 is more than poor.

Existing differences between the forecasted value and the real values may be explained by the following items:

- A cyclic pattern that was not visible within the first time series of 24 months
- The random variation we did not take into account
- Changes in the outer environment which are not observed regularly

If we take a closer look at the forecasting model and it's regression analysis, we can see a R^2 of 0.9949 indicating a very good fit for our equation. The percentage of trend never reaches the 2 % level ($\hat{y} \pm 2\%$)

List of Sources

Keller, G. & Warrack, B. (1999), *Statistics for Management and Economics*, 5th edn, Duxbury, Thomson Learning, Pacific Grove, CA.