

”We believe
Handels has 15 %
slick backs.”

”We believe
Handels has more
slick backs than
Chalmers.”

Case study: Slick backs

Miniproject 6 from Statistics for Scientists and Engineers (1999)



CHALMERS

CHALMERS TEKNISKA
HÖGSKOLA
Group 8

David Gustafsson
870528-6915

Andreas Hackl
860718-5074

Simon Standoft
850303-5613

Henrik Kjellgren
860201-2059

Kristoffer Lannö
870726-5073

TMS 136
Projekt 1
2007-09-18

Göteborg

List of contents

List of contents.....	2
1 Introduction.....	3
1.1 Background.....	3
1.2 Aim.....	3
1.2.1 Hypothesis.....	3
1.3 Sampling plan.....	3
1.3.1 Target population.....	3
1.3.2 Sampling design.....	3
1.4 Methods.....	4
1.4.1 Choosing n	4
1.4.2 Confidence intervals.....	4
1.4.3 Hypothesis test.....	4
2 Results.....	6
2.1 Data.....	6
2.2 Confidence interval.....	6
2.2.1 Analysis.....	7
2.3 Hypothesis test.....	7
2.3.1 Analysis.....	8
3 Discussion and conclusions.....	9
List of references.....	10

1 Introduction

1.1 Background

We believe that everyone sometimes makes guesses about what unknown people do for a living based on their looks. Some of these guesses are often prejudices and have various correlations to reality. We therefore thought it would be interesting to statistically investigate one of these common prejudices of what differs people at the two Gothenburg colleges Handelshögskolan and Chalmers.

1.2 Aim

The purpose of this project is to investigate the populations of men with a “slick back” - haircut on Handelshögskolan (Handels) compared to Chalmers campus Johanneberg. It will also conclude whether it is possible to say that this haircut is more likely to be seen around one of the schools than the other.

1.2.1 Hypothesis

We intentionally believe that a higher ratio of people at Handels have a “slick back” -haircut compared to people at Chalmers.

To specify our hypothesis a little we add to our presumption that 15 percent of the people recorded at Handels will have “slick back” (will be checked by the one sample binominal model for hypothesis testing).

1.3 Sampling plan

1.3.1 Target population

We have two target populations, people at Handels and people at Chalmers. The population recorded in this project will be of male gender only. We will consider all men that visit the school during lunchtime, e.g. students, teachers and personnel.

1.3.2 Sampling design

First we need to define what a “slick back” is. Our definition is stated as follows:

“A slick back is a male hair style characterised by the following properties: i) the hair must be long enough to be able to be styled with a comb but not longer than shoulder length. ii) it has to look wet in some sense, meaning that it contains either water, hair gel or some other grease substance. iii) all hair must be distinct combed backwards. The picture on the front page of Michael Douglas serve as our ideal ”slick back” -haircut.”

We will denote the haircuts we see binary as success (meaning the person has a “slick back” - haircut) or failure. To minimize problems with selection bias we will take our sample from people entering each school through the main entrance (not people leaving). Our sampling method can therefore be described as simple random sampling. A question about independence of the data has been raised in the sense that people with slick back might tend to cluster together in groups. We have investigated that possibility in a pilot study and found that it was not the case. People in groups seldom have the same hair style, regardless whether the group contain someone with a "slick back" or not. If it would have been a clustering problem, we would then have solved that by recording every fifth person passing the entrance. Our conclusions below will be valid under the assumption that slick backs do not tend to cluster in groups. There is no problem with dependence between the sample obtained at Handels and the sample obtained at Chalmers. People at Chalmers tend to stick to Chalmers and vice versa.

To obtain stationarity in the data, the ratio of "slick backs" passing by must be constant. We have achieved that by obtaining both our samples at lunchtime when most people are moving around. We will not then have the problem that certain groups might be at session, since no classes are held during lunchtime.

1.4 Methods

As we have already mentioned, our data consists of n independent Bernoulli trials. Each trial has a certain probability, p , of success (sample unit has "slick back"). Therefore the total amount of people with "slick back", Y , recorded on each school, will follow a binomial distribution model: $Y \sim b(n,p)$. In order to examine our hypothesis we will now describe the methods we used.

1.4.1 Choosing n

First we conducted a hypothesis test and constructed a confidence interval for p_{handels} and $p_{\text{handels}} - p_{\text{Chalmers}}$. We intended to approximate the binomial model with a normal model and therefore we chose n according to the specifications for that model to hold.

p is a pivotal quantity, since both Y and n are known from our sample. We know that:

$\frac{Y - np}{\sqrt{np(1-p)}}$ for $i = 1, 2$. But this approximation only holds when

$np \geq 10$ and $n(1-p) \geq 10$. Therefore we need a sample size so that this criterion is valid.

In the pilot study mentioned above, we also made an approximate judgement on p_1 and p_2 and found that $p_{1,\text{approx}} = 0.10$ (actually less than our hypothesis), $p_{s,\text{approx}} = 0.02$. For these data we get $n_1 \geq 100$ and $n_2 \geq 500$, so we took $n_1 = 150$ and $n_2 = 600$. The sample sizes are reasonable, since there are more students around Chalmers than Handels.

The fact that the sample sizes are unequal does not reject our models for deciding confidence intervals and performing hypothesis test, as we will investigate below.

Unfortunately, it turned out when the data was taken that the criterion above did not hold for p_{Chalmers} . Therefore we could not use the model approximation as we intended. However, we still used the above mentioned method for choosing n .

1.4.2 Confidence intervals

The model used to determine a level L confidence interval for the proportion of "slick backs" on Handels is the "one population large sample Confidence Intervals for the Proportion in the Binomial model" (Chen 1999):

Assumptions

- 1) First assume that the data are from a binomial distribution, $Y \sim b(n,p)$.
- 2) The sample size is large, $Y \geq 10$, $n - Y \geq 10$, so that we can approximate p to have a normal distribution.

Formulas

An approximate level L confidence interval for p is: $(\hat{p} \pm z_{\frac{1-L}{2}} \sqrt{\hat{p}(1-\hat{p})})$, where

$$\hat{p} = Y/n$$

1.4.3 Hypothesis test

One population Proportion in the Binomial model hypothesis test for $p_{\text{handels}} = 0.15$.

To decide whether our hypothesis of 15 % slick backs on Handels was correct we conducted a one population Proportion in the Binomial model hypothesis test (Chen et.al. 1999):

Assumptions

- 1) First assume that the data are from a binomial distribution, $Y \sim b(n,p)$.
- 2) The sample size is large, $Y \geq 10$, $n - Y \geq 10$, so that we can approximate p to have a normal distribution.

Formulas

Standardized Test Statistic:
$$\frac{Y - np}{\sqrt{np(1-p)}}$$

The p value is then calculated as $p = P(N(0,1) \leq z_l^*)$, for the alternative hypothesis, H_a , where

$$z_l^* = \frac{Y - np}{\sqrt{np(1-p)}}$$

$p < p_0$. If $p < \alpha = \{ \text{a pre-specified significance level, the null hypothesis can be rejected in favour for } H_a. \}$

Fisher’s exact test for slick backs at handles vs. Chalmers.

Since our result showed that the difference in population proportion did not follow the criteria for being able to be approximated by a normal distribution model, we can not use the “One population Proportion in the Binomial model hypothesis test” to conclude our second hypothesis. The normal approximation for Y is not valid. In order to solve that, MINITAB enables us to use another method for exact hypothesis testing called Fisher’s exact test, which doesn’t require (2) above.

	<u>Column</u>		
<u>Row</u>	1	2	Totals
1	Y_{11}	Y_{12}	$Y_{1.}$
2	Y_{21}	Y_{22}	$Y_{2.}$
Totals	$Y_{.1}$	$Y_{.2}$	N

Table 1

Fisher’s exact test uses a 2*2 table as table 1 above. The statistical hypothesis is stated as:

H_0 : Row and column variables are independent.

H_a : Row and column variables are not independent (in our case, the one way alternative, H_a , is that more people at handles have “slick back” hair).

To investigate if the null hypothesis can be rejected in favour for H_a , we calculate the p-value. This can be done when at least one set of marginal totals is known. Either $y_{1.}$ and $y_{2.}$ or $y_{.1}$ and $y_{.2}$. The probability for $Y_{11} = y_{11}$ is equal to:

$$\frac{\binom{Y_{1.}}{y_{11}} \binom{Y_{2.}}{y_{21}}}{\binom{N}{y_{.1}}} \quad (\text{“Fisher’s probability equation”})$$

The p-value for the test of H_0 versus H_a equals the sum of the probabilities to observe all values of Y_{11} equal to or smaller than y_{11} : $\sum_{i=0}^{y_{11}} \frac{\binom{Y_{1.}}{i} \binom{Y_{2.}}{y_{21}}}{\binom{N}{y_{.1}}}$.

Finally, to assure stationarity (which is essential for the above mentioned models to hold), we applied the method of plotting the data against order (see Results figure 1 and 2).

2 Results

2.1 Data

School	Sample size (n)	Number of slick backs (Y)	\hat{p} (%)	$\text{Var}(\hat{p}) = \hat{p}(1-\hat{p})$
Chalmers	600	7	1,167	0,0115
Handels	150	14	9,330	0,0846

Table 2. – Summarized data

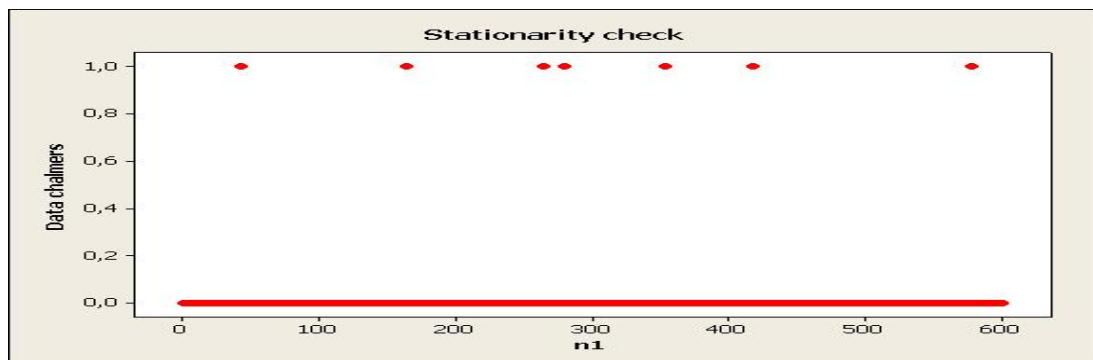


Figure 1 – Stationarity check Chalmers

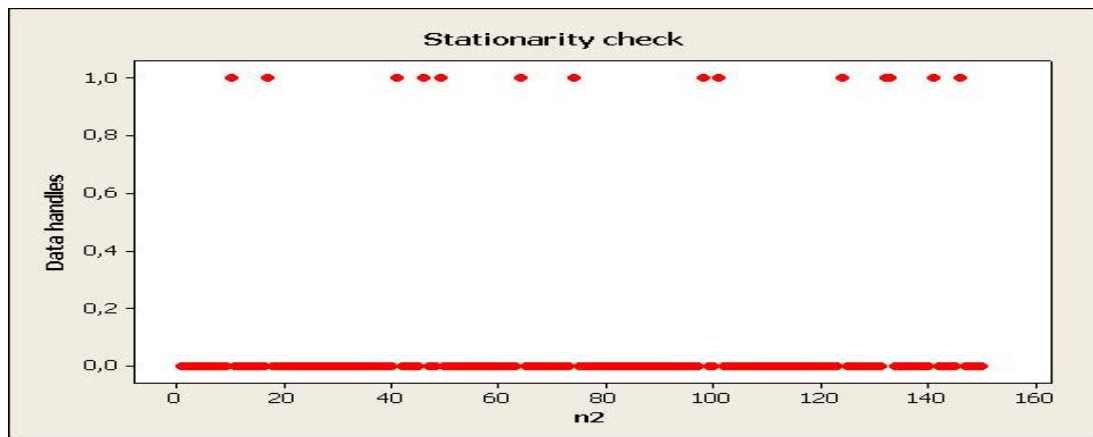


Figure 2 – Stationarity check Handels

To check stationarity we plotted data versus order. In stationary check of Chalmers and stationary check of Handels its easy to see that the data are stationary and have the same ratio of slick backs over the time. You can also see that "slick backs" do not tend to cluster.

2.2 Confidence interval

From the data as above, we create a confidence interval for the proportion, p , of "slick backs" at Handels. We use the method described earlier:

A point estimate for p is $\hat{p} = Y/n = 0.093$.

A 95 % confidence interval for p is:

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\hat{p}(1-\hat{p})}$$

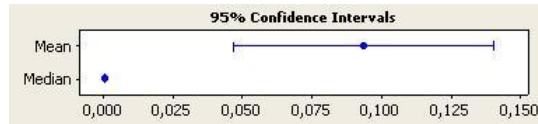


Figure 3 – Confidence interval

2.2.1 Analysis

The confidence interval can be interpreted as if we would continue to take samples from the target population, then 95 % of the calculated intervals will actually contain the true probability, p .

2.3 Hypothesis test

One population Proportion in the Binomial model hypothesis test for $p_{\text{handels}} = 0.15$.

One of our hypothesis is that 15 % of the population at Handels have a “slick back” -haircut. Hence our null hypothesis, H_0 is: $p = 0,15$ and H_a : $p \leq 0,15$. We will test this with the significance level $\alpha = 0,025$.

The variance for our data is: $\sigma^2 = np(1-p) = 4,37$. With the variance stated we can calculate the standardized test statistics for our data. With the formula provided above, $z_l^* = -1,83$. The p-value then is 0,034.

Fisher’s exact test for slick backs at handles vs. Chalmers.

Our scientific hypothesis is that more people at Handels have slick back than at Chalmers. Hence, the statistical hypothesis is:

H_0 : School and haircut are independent

H_a : School and haircut are not independent (more people at handles have slick back hair).

We want to test the p-value against the significance level, $\alpha = 0,025$, which corresponds to strong significance.

In our case we will use school for column variables and “slick back” or “not slick back“ for rows (haircut). Since the sample sizes and number of slick backs are known from our data we know the marginal totals. The data is summarized below in table 3.

		School		Total
		Handels	Chalmers	
Ha irc ut	Slick back	14	7	21
	Not slick back	136	593	739
Total		150	600	750

Table 3 – Fisher’s table for our data

To perform a Fisher’s exact test and calculating a p-value for the null versus the negative alternative hypothesis, we sum all $p(y_{11})$ for $y_{11} = 0, 1, \dots, 14$ according to the formula mentioned in methods above. This gives the **p-value = 0.0000030560401167230886**

2.3.1 Analysis

One population Proportion in the Binomial model hypothesis test for $p_{\text{handels}} = 0.15$.

With the data calculated as above, we see that the p-value $> \alpha$. We can therefore not reject the null hypothesis with significance α , which means that we can not with strongly significance say that less than 15 % of the population at Handels have slick back.

Fisher's exact test for slick backs at Handels vs. Chalmers.

Since $p < \alpha$, we can with a significance of α reject the null hypothesis in favour of H_a . We therefore conclude that the variables school and haircut are dependent with negative association, and that Handels has a higher proportion of "slick backs".

3 Discussion and conclusions

Our analysis have shown that our guess (and also null hypothesis) that 15 % of the population at Handels have “slick back”, at least was not all wrong. The null hypothesis was not possible to reject with our desired significance level ($\alpha = 0,025$) because of the larger p-value (= 0,034). So far it is stated that we are likely to be right on our hypothesis.

On the other hand the level 0.95 confidence interval for p did not contain 15 %. This means that 95 % of all confidence intervals created in repeated sampling would contain the true p. So if the true p value is 15 %, we would have got one of the 5 % intervals that do not contain the true value of p (which is not likely). If we look at our p-value again we see that it is just above our decided significance level. This, along with the confidence interval shows that we anyway should consider rejecting H_0 , but only with a p-significance level, not the desired α . But since we have already stated a significance level, we will not change it. However we can state that we are close to being able to reject H_0 . Overall, our null hypothesis is vague, and we will therefore confess that a hypothesis of 15 % slick backs on Handels is a somewhat of a prejudice. They are much more likely to be fewer, than more than 15 %.

For our second hypothesis, our result has shown that we with strong significance (α) can conclude that more people have slick back on Handels than Chalmers. The reason for this, we personally believe, is because of the difference between education programs at Chalmers and Handels. Handels is a part of the University of Gothenburg specialised in economics and law, a group of professions with a common dress code, “suit and tie”. Chalmers on the other side has a greater spread of educational programs, mostly without this clear dress code.

Can we then be confident that our results are valid with the data sampled as described? Yes, we believe so, since we have tried to think of reducing problems with sampling bias. We have also checked the data for stationarity and independence. Some factors that we could not have controlled are:

- People with habits of not passing the main entrance to enter the school are possible bias. It is hard to tell whether slick backs have a special tendency of doing so.
- Some people wore head gears and were therefore not recorded. Maybe the head gears hid a greater proportion of slick backs? Therefore our results are only valid for those of the population not wearing hats (it is fortunately very small)!
- Another possible bias is the fact that some people might have been recorded more than once if they for some reason went back inside and then came out again. Controlling this was impossible due to too much people passing by.

Problems that we encountered during the project are mainly due to wrong assumptions in taking the sample size. Our pilot study showed a different “slick back” level from what was obtained in our sample. This made using the normal approximation model a deficient alternative and we were forced to apply Fisher’s exact test instead (since all criteria for using the normal approximation was not satisfied). This made the project a lot tougher since we had to study an extensive model.

List of references

Petruccelli, Nandram, Chen, Applied Statistics for Scientists and Engineers, Pentrice Hall (1999)