PREFACE

In 2004, I intend to spend two semesters on a study abroad as a postgraduate exchange student at the University of Technology in Sydney, Australia. Each opportunity to enhance my English (both written and spoken) will take me a step closer to a successful and unforgettable year in Down Under.

As a result, all assignments will be written in English. I suppose that this decision of mine should not cause any problems since the language used in nowadays' mathematical papers *is* English. I appreciate all hints concerning an improper or wrong use of English terms. Thanks for your kind understanding.

PROBLEM 1

? *Generate 100 $B(1,\vartheta)$-distributed random variables consisting of n samples each. Determine the asymptotic confidence intervals for*

1. $\vartheta = 0.8$ *and* $n = 50$
2. $\vartheta = 0.5$ *and* $n = 50$
3. $\vartheta = 0.8$ *and* $n = 80$

Compare the widths of these intervals.

The huge amount of data to be generated forced me to use dedicated software intensively. In my previous lectures I got into contact with the excellent program Maple 8 (trial version, available for free). Therefore, it will accompany me again on my knowledge-seeking way through this (and maybe the following) problem assignment.

Basically, each $X_i \sim Bin(1,\vartheta)$ distribution – where the actual $\vartheta \in (0,1)$ is unknown – has the characteristics

$$EX_i = \vartheta \qquad \text{and}$$

$$VarX_i = \vartheta \cdot (1-\vartheta)$$

If we further define $\mu = EX_i = \vartheta$ and $\sigma^2 = VarX_i = \vartheta \cdot (1-\vartheta)$ and take into consideration that all our random variables are i.i.d. (independently and identically distributed, because they are the outcome of a close-to-perfect random number generator) then the Law of Large Numbers helps in finding the confidence intervals. It allows us to approximate

$$\hat{\vartheta} = \frac{1}{n} \cdot \sum X_i$$

by the normal distribution

$$\sqrt{n} \cdot \frac{\hat{\vartheta} - \mu}{\sigma} \sim N(0,1)$$

The consistency gives

$$S_n^2 \to \sigma^2$$

where

$$S_n = \hat{\vartheta} \cdot (1-\hat{\vartheta})$$

Therefore; we get

$$\sqrt{n} \cdot \frac{\hat{\vartheta} - \mu}{S_n} \sim N(0,1)$$

A normal distribution's confidence interval for μ is defined by

$$I = \left[\hat{\vartheta} - t_{n-1;\, 1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S_n^2}{n}},\ \ \hat{\vartheta} + t_{n-1;\, 1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S_n^2}{n}} \right]$$

which leads us to

$$I = \left[\hat{\vartheta} - t_{n-1;\, 1-\frac{\alpha}{2}} \cdot \frac{\hat{\vartheta} \cdot \left(1 - \hat{\vartheta}\right)}{\sqrt{n}},\ \ \hat{\vartheta} + t_{n-1;\, 1-\frac{\alpha}{2}} \cdot \frac{\hat{\vartheta} \cdot \left(1 - \hat{\vartheta}\right)}{\sqrt{n}} \right]$$

If we set $\alpha = 0.05$ then $t_{n-1;1-\frac{\alpha}{2}} = t_{99;0.975} \approx 1.984$.

On the next page I will explain the Maple 8 (trial) program used to generate the according random numbers, their confidence intervals and its widths. It was my intention to do quite more than requested in order to get back the "feeling" for statistical problems and Maple's programming language.

First, some packages are required in order to define all used functions:

```
> with(stats): with(random): with(plots): with(statplots):
```

A real programmer (like me ☺) frequently strives to move the main routines into separate, dedicated functions. One of the most-called functions is **observation**. Its purpose is to generate all events for a single random variable X_i, i.e. a series of 0's and 1's:

```
> observation:=(theta, size) -> [binomiald[1, theta](size)]:
```

Unfortunately, there is no deeper interest in that series. What we actually need is the frequency of 1's finally giving the estimated $\hat{\vartheta}$.

```
> estimation:=(theta, size) -> sum(observation(theta, size)[i],
                                   i=1..size)/size;
```

One hundred estimations of $\hat{\vartheta}$ form a quite solid **experiment**. Thus, we can determine the confidence intervals.

```
> experimentsize:=100:
> experiment:=(theta, size) -> sort([seq(estimation(theta, size),
                                         j=1..100)]):
```

The student-t distribution plays such an important role that it deserves its own function:

```
> studentt:=(alpha, size) -> statevalf[icdf, studentst[size]](1-alpha/2):
```

Throughout the whole assignment, the level of confidence and the number of estimations remains unchanged so I increased Maple's performance by magnitudes by defining a constant **t**:

```
> t:=studentt(0.05, experimentsize);
```

Next comes **confidenceIntervals** which is responsible for generating the requested confidence intervals:

```
> confidenceIntervals:=(samples, size) ->
                seq([samples[i]-t*sqrt(samples[i]*(1-samples[i])/size),
                   samples[i]+t*sqrt(samples[i]*(1-samples[i])/size)],
                  i=1..experimentsize);
```

All the mentioned functions perform the core work. They are instantiated by just these six lines:

```
> samples1:=experiment(0.8, 50):
> samples2:=experiment(0.5, 50):
> samples3:=experiment(0.8, 80):
> intervals1:=confidenceIntervals(samples1, 50):
> intervals2:=confidenceIntervals(samples2, 50):
> intervals3:=confidenceIntervals(samples3, 80):
```

Numbers can be quite abstract, hence, a plot often helps to gain further insights. Here is the code used for experiment 1, it does not differ much from experiment 2 or 3:

```
> intervalplot1:=seq(plot([[(intervals1[i][1]),i], [intervals1[i][2],i]],
                          color=coversTheta(intervals1[i][1],
                                            intervals1[i][2],0.8,
                                            red,blue)),
                     i=1..100) :
> display([intervalplot1, plot([[0.8,0],[0.8,100]], color=black)]);
```

Experiment 1

For experiment 1 ($\vartheta = 0.8$ and $n = 50$) we observe that 96 out of 100 intervals, i.e. 96%, actually cover ϑ which comes pretty close to our confidence level of 95%. In the following plot all intervals that overlap ϑ are colored red, ϑ itself is represented by a black vertical line:

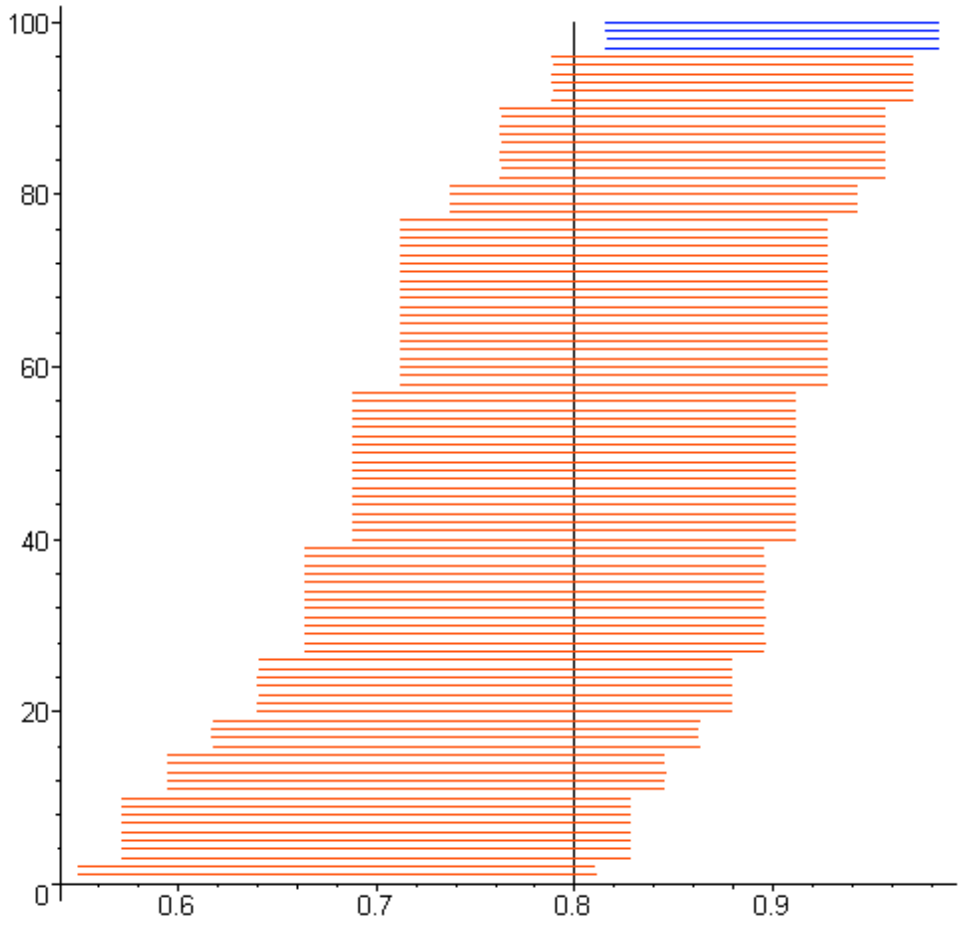

Figure 1 Confidence intervals of experiment 1

Experiment 2

Experiment 2 failed in some way. I missed the proposed confidence level of 95% by far and reached only 91%. On the other hand, five of the misses' interval borders are pretty close to ϑ. Therefore, the generated numbers are not completely unusable. If you take a look at the shape of the intervals you may notice that there are a bit more symmetrically distributed than in experiment 1.

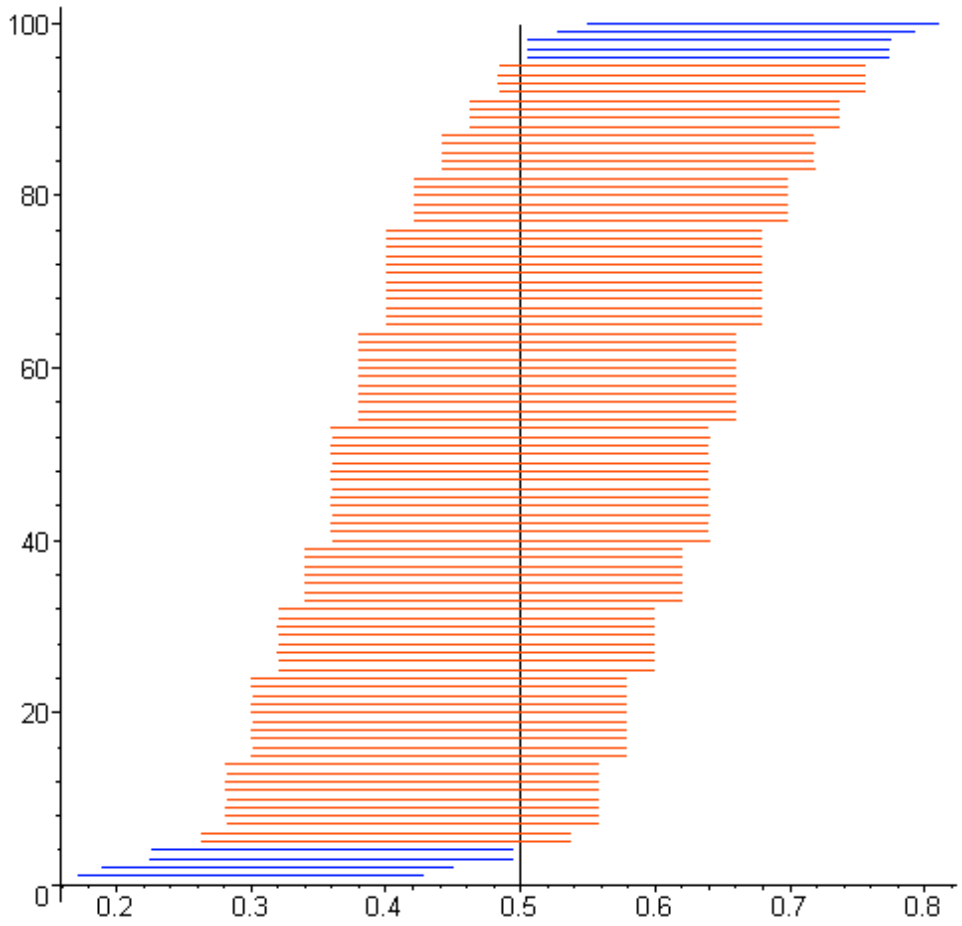

Figure 2 Confidence intervals of experiment 2

Experiment 3

The last experiment exactly matches our confidence level of 95%. The overall impression fits my expectations so there is nothing more left to say:

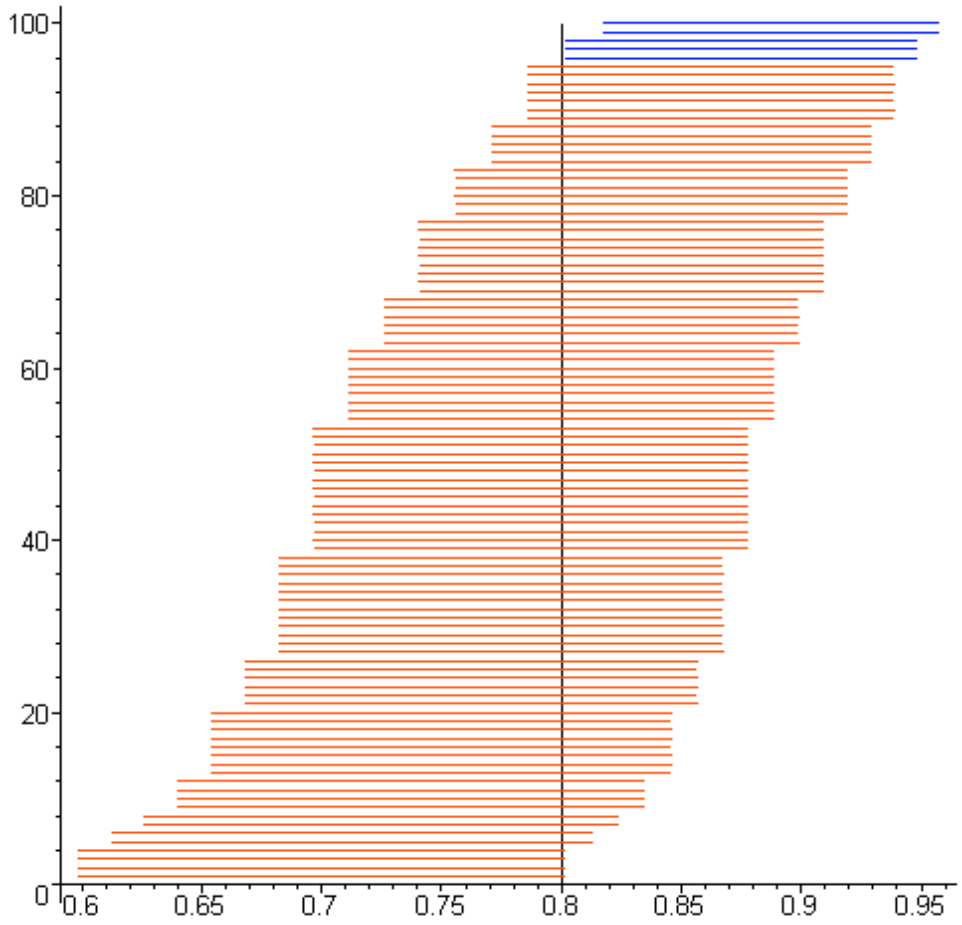

Figure 3 Confidence intervals of experiment 3

Comparing the widths

Finally, I compared the widths of all three experiments. The convincing insights gained from the plots empowered me to use a plot for the widths, too. Most of the Maple code is responsible for a properly sorted order:

```
> width1:=sort([seq(intervals1[i][2] - intervals1[i][1], i=1..100)]):
> width2:=sort([seq(intervals2[i][2] - intervals2[i][1], i=1..100)]):
> width3:=sort([seq(intervals3[i][2] - intervals3[i][1], i=1..100)]):
> widthplot1:=plot([seq([i, width1[i]], i=1..100)], i=1..100,
                   style=point, color=black):
> widthplot2:=plot([seq([i, width2[i]], i=1..100)], i=1..100,
                   style=point, color=blue):
> widthplot3:=plot([seq([i, width3[i]], i=1..100)], i=1..100,
                   style=point, color=red):
> display([widthplot1, textplot([90, 0.23, 'experiment1']),
widthplot2, textplot([90, 0.275, 'experiment2']),
widthplot3, textplot([90, 0.185, 'experiment3'])]);
```

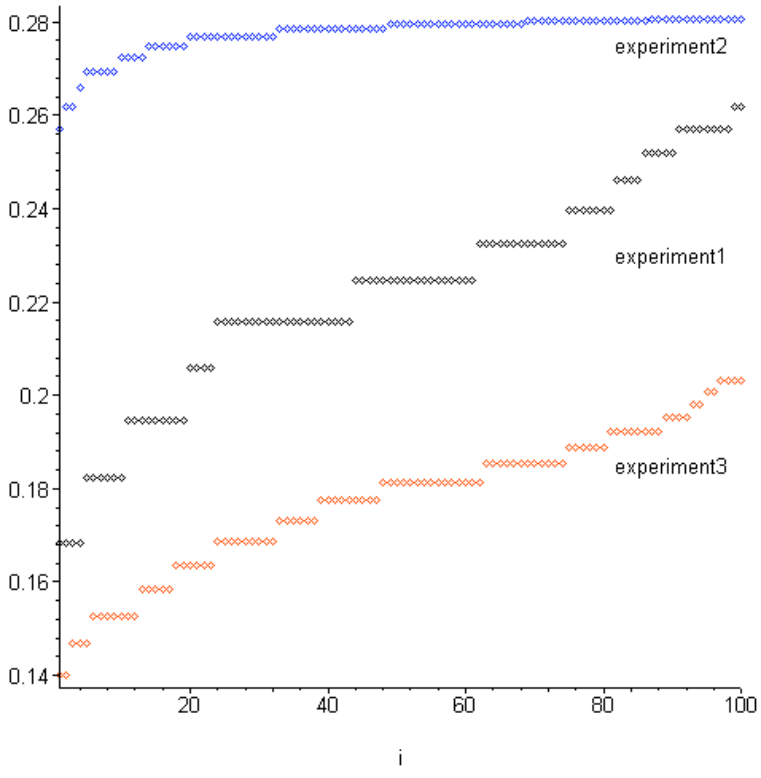

Figure 4 Widths of the confidence intervals

The plot confirms my hypothesis: experiment 2 produces the largest spread since for $\vartheta = 0.5$ the variance $\vartheta \cdot (1-\vartheta)$ is maximized. Small variations of $\hat{\vartheta}$ if $\hat{\vartheta}$ is close to 0.5 do not substantially affect the width which heavily depends on the variance as seen in Figure 5 below.

```
> plot(theta*(1-theta), theta=0..1);
```

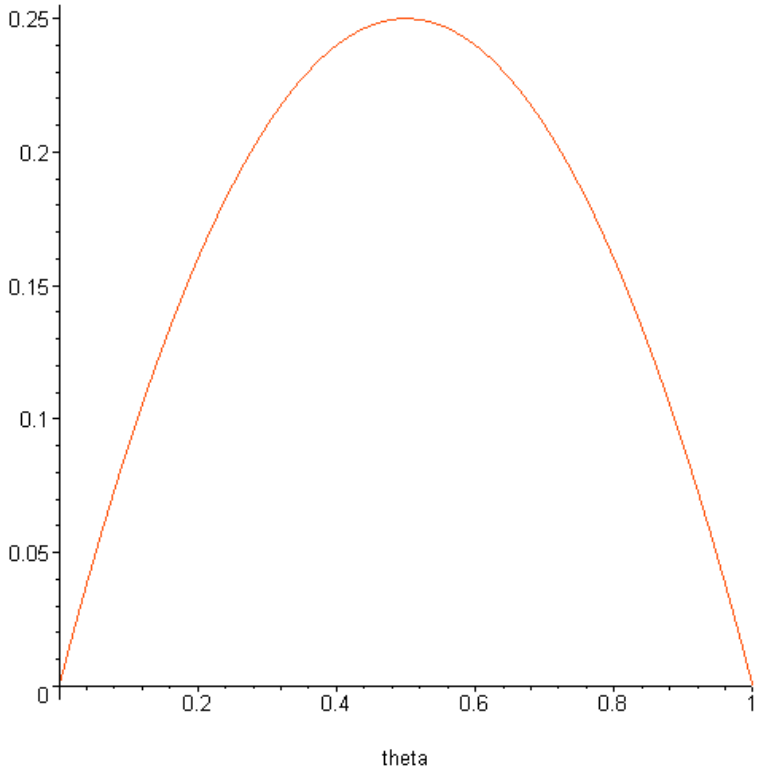

Figure 5 Variance depending on ϑ

Experiment 1 consists of 100 random variables each generated by 50 $B(1,0.8)$ observations. In contrast, experiment 3 was built of 80 $B(1,0.8)$ observation, about 50% more. This difference leads to a significantly smaller average width of experiment 3 confidence intervals'. I conclude from the experiments that the accuracy of a $B(1,\vartheta)$-based experiment can be seriously increased by either shifting ϑ away from 0.5 or by rising the number of repetitions. In the most cases, the latter is the only possible solution when aiming for a more precise estimation.