Advanced Quantitative Methods
Statistical estimators

Johan A. Elkink
jos.elkink@ucd.ie

December 18, 2010

If you encounter any mistakes in this text, please inform the author.

Contents

1 Proof of bias for sample variance

Let us say that σ^2 refers to the variance of x in the population, while s^2 refers to the sample variance of x. Let's say that we have two estimators $\hat{\sigma}^2$:

$$\hat{\sigma}_1^2 = s^2$$
$$\hat{\sigma}_2^2 = \frac{n}{n-1}s^2,$$

whereby s^2 is calculated as $s^2 = \frac{1}{n}\sum_i^n (x_i - \bar{x})^2$. How do we evaluate the bias of the two? If the estimator is unbiased, then the expected value of the estimator has to be equal to the sample parameter to be estimated, so in this context, the estimator $\hat{\sigma}_1^2$ is unbiased iff $E(\hat{\sigma}_1^2) = \sigma^2$.

For $\hat{\sigma}_1^2$:

$$\begin{aligned}
\hat{\sigma}_1^2 = s^2 &= \frac{1}{n}\sum_i^n (x_i - \bar{x})^2 \\
E(\hat{\sigma}_1^2) &= E(\frac{1}{n}\sum_i^n (x_i - \bar{x})^2) = \frac{1}{n}E(\sum_i^n (x_i - \bar{x})^2) \\
nE(\hat{\sigma}_1^2) &= E(\sum_i^n (x_i - \bar{x})^2) \\
&= E(\sum_i^n x_i^2 + \sum_i^n \bar{x}^2 - \sum_i^n 2\bar{x}x_i) \\
&= E(\sum_i^n x_i^2 + n\bar{x}^2 - 2\bar{x}\sum_i^n x_i) \\
&= E(\sum_i^n x_i^2 + n\bar{x}^2 - 2n\bar{x}^2) \\
&= E(\sum_i^n x_i^2) - E(n\bar{x}^2) = nE(x_i^2) - nE(\bar{x}^2) \\
E(\hat{\sigma}_1^2) &= E(x_i^2) - E(\bar{x}^2)
\end{aligned}$$

We now need to use the definition of the variance as $var(z) = E(z - E(z))^2$, so that:

$$\begin{aligned}
var(z) &= E(z - E(z))^2 \\
&= E(z^2 + E(z)^2 - 2zE(z)) \\
&= E(z^2) + E(E(z)^2) - E(2zE(z)) \\
&= E(z^2) + E(z)^2 - 2E(z)E(z) \\
&= E(z^2) - E(z)^2 \\
E(z^2) &= var(z) + E(z)^2,
\end{aligned}$$

which in the context of our analysis leads to:

$$\begin{aligned}
E(x_i^2) &= var(x_i) + E(x_i)^2 = \sigma^2 + \mu^2 \\
E(\bar{x}^2) &= var(\bar{x}) + E(\bar{x})^2 \\
&= var(\frac{1}{n}\sum_i^n x_i) + \mu^2 = \frac{1}{n^2}var(\sum_i^n x_i) + \mu^2 = \frac{1}{n^2}\sum_i^n var(x_i) + \mu^2 = \frac{1}{n^2}n\sigma^2 + \mu^2 \\
&= \frac{1}{n}\sigma^2 + \mu^2,
\end{aligned}$$

so we can continue:

$$\begin{aligned} E(\hat{\sigma}_1^2) &= E(x_i^2) - E(\bar{x}^2) \\ &= (\sigma^2 + \mu^2) - (\frac{1}{n}\sigma^2 + \mu^2) \\ &= \sigma^2 - \frac{1}{n}\sigma^2 \\ &= \frac{n-1}{n}\sigma^2, \end{aligned}$$

so this estimator is biased.

For $\hat{\sigma}_2^2$:[1]

$$\begin{aligned} \hat{\sigma}_2^2 &= \frac{n}{n-1}s^2 = \frac{1}{n-1}\sum_i^n (x_i - \bar{x})^2 \\ E(\hat{\sigma}_2^2) &= E(\frac{1}{n-1}\sum_i^n (x_i - \bar{x})^2) = \frac{1}{n-1}E(\sum_i^n (x_i - \bar{x})^2) \\ (n-1)E(\hat{\sigma}_2^2) &= E(\sum_i^n (x_i - \bar{x})^2) \\ &= E(\sum_i^n x_i^2 + \sum_i^n \bar{x}^2 - \sum_i^n 2\bar{x}x_i) \\ &= E(\sum_i^n x_i^2 + n\bar{x}^2 - 2\bar{x}\sum_i^n x_i) \\ &= E(\sum_i^n x_i^2 + n\bar{x}^2 - 2n\bar{x}^2) \\ &= E(\sum_i^n x_i^2) - E(n\bar{x}^2) = nE(x_i^2) - nE(\bar{x}^2) \\ \frac{n-1}{n}E(\hat{\sigma}_2^2) &= E(x_i^2) - E(\bar{x}^2) \\ &= (\sigma^2 + \mu^2) - (\frac{1}{n}\sigma^2 + \mu^2) = \sigma^2 - \frac{1}{n}\sigma^2 = \frac{n-1}{n}\sigma^2 \\ E(\hat{\sigma}_2^2) &= \sigma^2, \end{aligned}$$

so this estimator is unbiased.

The unbiased estimator for the population variance is thus $\frac{1}{n-1}\sum_i^n (x_i - \bar{x})^2$, which is indeed the often encountered textbook formula.

[1] Based on http://math.wikia.com/wiki/Proof_of_unbiased_estimator_for_variance (January 19, 2010) and http://biology.ucf.edu/ pascencio/classes/Methods/Proof%20that%20Sample%20Variance%20is%20Unbiased.pdf (January 19, 2010).

2 Probabilities and probability distributions

2.1 Probability

Imagine we are performing an experiment and there is a fixed set of possible outcomes. For example, we might toss a coin and the possible outcomes are head or tails, or we might throw a dice and the possible outcomes are 1, 2, 3, 4, 5, or 6. In these cases, we can say that the **sample space**s are $S = \{head, tails\}$ and $S = \{1, 2, 3, 4, 5, 6\}$, respectively. We can now imagine that we do this experiment many times and consider how often, say, we would throw a 3, or how often we would throw an odd number with the dice. We call the proportion of times, in a (hypothetically) large number of repetitions of the experiment, the **probability** of the particular **event**. When talking about infinitely many repetitions of an experiments, we are talking about what is in mathematics called the limit, so the formal **frequentist** definition of a probability would be: $P = lim(\frac{f}{n})$, the limit of the frequency divided by the number of trials as the number of trials goes to infinity (Wonnacott and Wonnacott, 1990, 71).[2] Say we look at the event of throwing a 3 ($E = 3$) and we have a fair dice in that all six sides have an equal chance of turning up on a throw, then we can say that $P(E) = \frac{1}{6}$. E is a subset of S (mathematically, $E \subset S$). The following rules are important to keep in mind with probabilities (Simon and Blume, 1994, 894):[3]

- Since we are talking about proportions, for any event E: $0 \leq P(E) \leq 1$.
- If E contains all possible outcomes ($E = S$), then $P(E) = 1$.
- If two events, E_1 and E_2 are mutually exclusive events, then the probability of either happening is the sum of the probabilities for each: $P(E_1 \cup E_2) = P(E_1) + P(E_2)$.
- The probability of event E not happening is: $P(\neg E) = 1 - P(E)$.

If you throw a 1 with a dice, you cannot simultaneously throw a 2 with the same dice - in other words, the outcomes $E = 1$ and $E = 3$ are mutually exclusive. Therefore, $P(E = odd) = P(E = 1) + P(E = 3) + P(E = 5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$.

Probabilities will recurr extensively when we discuss Maximum Likelihood estimation and limited dependent variable models.

2.2 Conditional probability

In some cases, the probability of one event, say E_1, might change given that another event, say E_2, occured. For example, we might have a dice where there is a bit more weight on the side of the 3. We can then conclude that the probability of throwing a 3 is higher if this is indeed the case than we would have otherwise thought. Let's say that E_2 is either true if there is an extra

[2] Note that looking at probabilities as a proportion of (hypothetical) events is the **frequentist** approach to probabilities. We can also speak of the belief that a particular outcome will occur - if we assume a fair dice, how likely do you think that a 3 will come up when I throw it? And your answer is likely to be $P(E) = \frac{1}{6}$ as well. This is the **Bayesian** interpretation of probabilities. All rules of probability discussed here apply to either interpretation.

[3] We use the following notation here: $A|B$ means "A, given B"; $A \cap B$ means "A and B"; $A \cup B$ means "A or B", which is an inclusive or, such that $(A \cap B) \subset (A \cup B)$; $A \subset B$ means "A is a subset of B"; $\neg A$ means "not A".

weight at the side of 3 or false if there is not. In mathematical formulation we can say that $P(E_1 = 3|E_2 = true) > P(E_1 = 3|E_2 = false)$ and also that $P(E_1|E_2 = true) > P(E_1)$. So the probability of throwing a 3 is larger if there is a weight than if there is not, and the probability is larger if there is a weight than we would have thought not knowing of any weight. We call $P(E_1|E_2)$ the **conditional probability** of event E_1 happening, given E_2. Another way of saying this is to say that the conditional probability of E_1 on E_2 is the probability of both occurring, divided by the probability of E_2 occuring (Simon and Blume, 1994, 894-895):

$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}$$

Conditional probabilities do not say anything about causes or directions, they are just probabilities of things occurring together. Therefore, we can also say:

$$P(E_2|E_1) = \frac{P(E_1 \cap E_2)}{P(E_1)}$$

Two events E_1 and E_2 are **independent** if $P(E_1|E_2) = P(E_1)$, or $P(E_1 \cap E_2) = P(E_1)P(E_2)$.

We can also manipulate the above formula above in two ways, such that:

$$P(E_1 \cap E_2) = P(E_1|E_2)P(E_2)$$
$$P(E_1 \cap E_2) = P(E_2|E_1)P(E_1)$$

Bayes' rule follows from the above:

$$\begin{aligned} P(E_1|E_2) &= \frac{P(E_1 \cap E_2)}{P(E_2)} \\ &= \frac{P(E_2|E_1)P(E_1)}{P(E_2)} \\ &= \frac{P(E_2|E_1)P(E_1)}{P(E_2|E_1)P(E_1) + P(E_2|\neg E_1)P(\neg E_1)} \end{aligned}$$

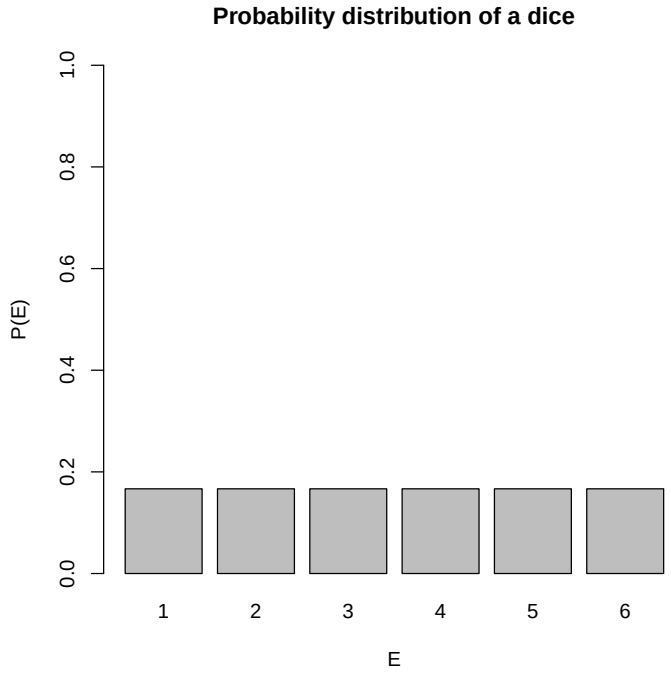
Although this formula forms the basis of Bayesian statistical analysis, it is not as such controversial in frequentist statistics - the equation simply holds for all probabilities, regardless of your interpretation of probabilities. In Bayesian estimation the above the above is used in the form:

$$P(parameters|data) = \frac{P(data|parameters)P(parameters)}{P(data)},$$

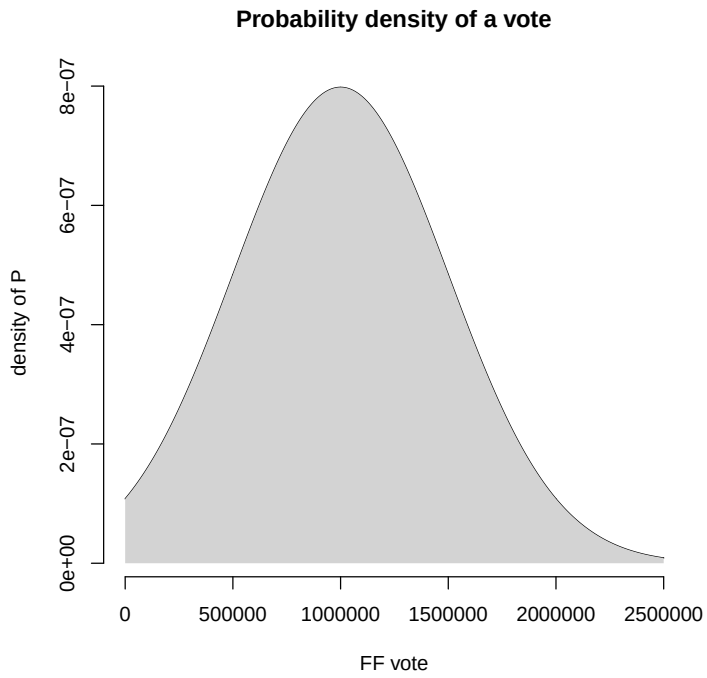
while in frequentist analysis, we estimate the parameters such that it is most likely that the data we observe occurred, in essence optimizing $P(data|parameters)$ and not $P(parameters|data)$. The Bayesian asks: "How did my belief in the parameter values change given the data I observe?", while the frequentist tries to figure out: "Under what parameters would it have been most likely that I observe the data that I did, indeed, observe?"

2.3 Probability distribution

If you look at all possible outcomes, say S, of an event, say E, we could make a simple barplot showing the probability for each outcome. For example, for a fair dice we would get:



In the case of a dice, S is a discrete set of values, with only six possible outcomes. Many probability distributions have an infinite number of possible outcomes, however, or such a large number that we normally speaking approximate them using a continuous function. For example, the number of votes for a particular party in a general election is discreet and finite, containing all integers between 0 and the size of the electorate (assuming no fraud committed), but we would normally consider this a continuous function, albeit truncated at either side. We might for example think that the probability for Fianna Fáil for a particular vote outcome to be:



We can now not really speak of the probability that the vote outcome would have a particular value, say, $E = 1032140$, since this would be infinitely small. We can, however, speak of the probability of getting a vote result between, say, one million and two million votes. Whereas for discrete events, this probability is the sum of the containing events, like we saw with throwing an odd value with a dice, here we look at the surface under the graph to get the probability.

2.4 Expected value

Let's say that x is a vector of observations from some experiments, with a discreet set of possible outcomes. The **expected values** of x is then defined by (Simon and Blume, 1994, 895):

$$E(x) = \sum_{i}^{N} P(x_i)x_i$$

Let's say that x is a collection of observations from throwing some fair dice. The expected value would then be $E(x) = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = 3\frac{1}{2}$. Another term for this is to say that this is the **mean** of random variable x, which is sometimes called the **first moment** of x.

For continuous variables, we use the surface under the graph, i.e. the integral, instead of the summation - but details are beyond the scope of this note.

Much like with matrices, there are some special rules for dealing with expectations (Gujarati, 2003, 879), taking a and b as constants and x and y as

random variables:

$$\begin{aligned} E(ax+b) &= aE(x)+b \\ E(xy) &= E(x)E(y) && \text{if } x, y \text{ independent} \\ E(y|x) &= E(y) && \text{if } x, y \text{ independent} \end{aligned}$$

Finally, the law of iterated expectations states that:

$$E(y) = E_x[E(y|x)]$$

This can be summarized as: "this law states that if we first obtain $E(y|x)$ as a function of x and take its expected value over the distribution of x values, you wind up with $E(y)$." (Gujarati, 2003, 885).

2.5 Variance

Often we do not only want to know the expected value of a random variable, but also how likely it is to get an outcome close to this mean. The **variance** of a discreet random variable is defined as (Simon and Blume, 1994, 896):

$$var(x) = E(x - E(x))^2 = \sum_i^N P(x_i)(x_i - E(x))^2,$$

which is sometimes called the **second central moment**.[4] Taking the same example of the dice, we would get:

$$\begin{aligned} var(x) &= \frac{1}{6} \cdot s(1-3\frac{1}{2})^2 + \frac{1}{6} \cdot (2-3\frac{1}{2})^2 + \frac{1}{6} \cdot (3-3\frac{1}{2})^2 \\ &\quad + \frac{1}{6} \cdot (4-3\frac{1}{2})^2 + \frac{1}{6} \cdot (5-3\frac{1}{2})^2 + \frac{1}{6} \cdot (6-3\frac{1}{2})^2 \\ &= \frac{1}{6} \cdot 6\frac{1}{4} + \frac{1}{6} \cdot 2\frac{1}{4} + \frac{1}{6} \cdot \frac{1}{4} + \frac{1}{6} \cdot \frac{1}{4} + \frac{1}{6} \cdot 2\frac{1}{4} + \frac{1}{6} \cdot 6\frac{1}{4} \\ &= 2\frac{11}{12} \approx 2.92 \end{aligned}$$

A standard deviation is simply the square root of the variance:

$$sd(x) = \sqrt{var(x)} = \sqrt{2.92} = 1.71$$

With variances we can also look at a number of specific properties (Gujarati, 2003, 881-886), again taking a, b, c, and d as constants and x and y as random

[4] To be slightly more precise, for a random variable x (Gut, 2005, 62):

- moments are $E(x)^n$, $n = 1, 2, ...$
- central moments are $E(x - E(x))^n$, $n = 1, 2, ...$
- absolute moments are $E(|x|)^n$, $n = 1, 2, ...$
- absolute central moments are $E(|x - E(x)|)^n$, $n = 1, 2, ...$

And the second central moment ($n = 2$) is called the variance (rather than vice versa).

variables:

$$
\begin{aligned}
E(x - E(x))^2 &= E(x^2) - E(x)^2 \\
var(a) &= 0 \\
var(ax + b) &= a^2 var(x) \\
cov(x, y) &= E[(x - E(x))(y - E(y))] \\
&= E(xy) - E(x)E(y) \\
cov(x, y) &= 0 \qquad \text{if } x, y \text{ independent} \\
cov(a + bx, c + dy) &= bd \cdot cov(x, y) \\
var(x + y) &= var(x) + var(y) + 2cov(x, y) \\
var(x - y) &= var(x) + var(y) - 2cov(x, y) \\
var(ax + by) &= a^2 var(x) + b^2 var(y) \qquad \text{if } x, y \text{ independent} \\
var(y|x) &= var(y) \qquad \text{if } x, y \text{ independent} \\
var(y) &= E[var(y|x)] + var(E[y|x])
\end{aligned}
$$

Note that we now sneakily also introduced, and defined, the concept of the **covariance** between two variables.

3 Limits

The **limit** of a sequence or function is the value a sequence gets close to as the n increases. More formally: "The real-valued sequence $\{x_n\}$ has the real number x^* for its **limit**, or converges to x^*, if for any positive ε, no matter how small, it is possible to find a positive integer N such that for all integers n greater than N, $|x_n - x^*| < \varepsilon$" (Davidson and MacKinnon, 1993, 102). So you could have a function of x, say, $y = \frac{1}{x}$, which as x increases, gradually gets closer and closer to 0, without ever reaching it. In this case, the limit of y is 0, or in other words:

$$\lim_{x \to \infty} \frac{1}{x} = 0.$$

Of course, many functions do not have a clear limiting value, for example:

$$\lim_{x \to \infty} 5x = \infty$$

and some functions have a limit that is something other than zero:

$$\lim_{x \to \infty} \frac{5(x + 3)}{x} = 5.$$

In §1 we show that the sample variance is a biased estimator of the population variance, because $E(s^2) = \frac{n-1}{n}\sigma^2$. Using this concept of limits, however, we can see that:

$$\lim_{n \to \infty} \frac{n - 1}{n}\sigma^2 = 1 \cdot \sigma^2 = \sigma^2,$$

therefore, although the sample variance is a biased estimator of the population variance, it is a consistent estimator - as n increases, the estimator approaches the true population variance.

For **random variables**, different types of limits exist. The first is the **probability limit**, whereby the probability that a random variable is further than some arbitrarily small distance away from some limiting variable becomes infinitely small as n increases:

$$P(\|\mathbf{x}_n - \mathbf{x}^*\| > \varepsilon) < \delta,$$

whereby ε and δ are arbitrarily small, positive real numbers. We can write:

$$\operatorname*{plim}_{n\to\infty} \mathbf{x}_n = \mathbf{x}^*$$

and $\mathbf{x}^*$ is called the **probability limit** of $\mathbf{x}$ (Davidson and MacKinnon, 1993, 103). So with the probability limit, the probability that x is not close to some value, say x^*, becomes smaller and smaller as n increases (or, if they are vectors, $\mathbf{x}$ and $\mathbf{x}^*$), and in fact becomes infinitely small as n becomes infinitely large. So you cannot say that x^* is the limit of x, because even as n is very large, there is still a very small probability that x is far from x^*. This probability gets smaller and smaller as n increases, however, and becomes eventually infinitely small.

Another type of limit for random variables is the value a sequence **almost surely** approaches:

$$P(\lim_{n\to\infty} \mathbf{x}_n = \mathbf{x}^*) = 1,$$

which we can write as:

$$\mathbf{x}_n \xrightarrow{a.s.} \mathbf{x}^*$$

(Davidson and MacKinnon, 1993, 106). So where the probability limit states that as n increases, the probability that x is more than an infinitely small distance away from x^* becomes infinitely small, the almost sure convergence states that as n increases, the probability that x equals x^* approaches one. The latter is a stronger claim and implies the former, but does not appear often in econometrics texts.

The third type of convergence is called **convergence in distribution**, which is summarized by:

$$\lim_{n\to\infty} P(\mathbf{x}_n \leq \mathbf{b}) = P(\mathbf{x}^* \leq \mathbf{b})$$

for any arbitrary $\mathbf{b}$ (Davidson and MacKinnon, 1993, 107).[5] In other words, the probability distribution of $\mathbf{x}$ approaches that of $\mathbf{x}^*$ as n increases. Note that the formulation $P(\mathbf{x}_n \leq \mathbf{b})$ represents the cumulative distribution function (c.d.f.) of vector $\mathbf{x}$, as a function of $\mathbf{b}$. The convergence in distribution can be written as:

$$\mathbf{x}_n \xrightarrow{D} \mathbf{x}^*.$$

Note that almost surely convergence or convergence in probability imply convergence in distribution (Davidson and MacKinnon, 1993, 108), so convergence in distribution is the weakest claim of the three.

Note that even if we know that something converges as n gets large, different sequences vary in the speed or **rate of convergence** (Davidson and MacKinnon, 1993, 108ff). For example, $\frac{1}{n}$ and $\frac{\log(n)}{n}$ both go to zero as $n \to \infty$, but the latter much slower than the former.

[5] The comparison operators $(<, \leq, \geq, >)$ here are assumed to operate element-by-element if $\mathbf{x}$ and $\mathbf{b}$ are vectors, such that $\mathbf{x} < \mathbf{b}$ is true iff $x_1 < b_1 \cap x_2 < b_2 \cap \ldots \cap x_n < b_n$.

It should also be noted that certainly not all limits are about "going to infinity". For example, although dividing by zero is not possible, we can say:

$$\lim_{n \to 0} \frac{1}{n} = \infty \qquad \lim_{n \to 0} \frac{\log(n)}{n} = -\infty.$$

This type of limit is much less relevant for understanding econometric methods, however, and asymptotic theory is about what happens when $n \to \infty$, hence our interest in these types of examples.

4 Monte Carlo in R

Monte Carlo simulations are a very practical way to learn about the finite sample properties of statistical estimators and tests. One can generate fake data sets on the basis of known parameter values and then attempt to estimate those parameters using some estimator or test some hypothesis regarding these parameters. Since the actual parameters are known, the difference between those and the estimated values tell us something about the quality of the estimator. This is particularly useful when the asymptotic properties of an estimator are known, but the finite sample properties are not. It is also useful to increase your intuition about statistics and sampling distributions, even when the estimators you study have well known properties.

A Monte Carlo simulation consists of the following steps:

1. Model the data-generating process (DGP)
2. Generate artificial data sets
3. Create estimates of the underlying parameters using the estimator you are testing
4. Assess the estimator's efficiency, bias and MSE relative to the (known) data
5. ... or, for tests, check proportion Type I and Type II errors

In this section we will implement a Monte Carlo experiment to demonstrate the **Central Limit Theorem**, by looking at the mean of throwing a coin. We will generate data where $P(head) = \frac{1}{2}$ and we will estimate the average value, using $head = 1, tail = 0$. We know that for this data generating process, the expected value of the ratio of number of heads by the total number of throws should be $\frac{1}{2}$.

When programming, which includes working in R, one should always use variables instead of just numbers, especially when a particular value will occur more than once in your script. The reason is that you do not want to change the same number in multiple places when you decide to change something about the script, for example the number of experiments, or the proportion of heads.

```
n <- 100
m <- 200
p <- .5
```

So here we will define three variables before we start our script: n is the number of throws per data set; m is the number of data sets; p is the proportion of heads.

We will also need a storage where we store the means from our artificial samples. We will make this storage empty (fill it with NA's, which is R's code for missing data) and the size of the number of datasets (m):

```
means <- rep(NA, m)
```

The *rep* command stands for "repeat" and thus creates a vector repeating the first value m times.

We will now create a loop for the experiments. In this particular example, a loop is actually easily avoided, and it would probably be faster not to use one, but using a loop makes the example easier to adjust for other simulations. The main command for such a loop is the *for* command, which basically counts from some number to another number and repeats whatever is between the curly brackets that many times. So to count from 1 to m and store the current value in i:

```
for (i in 1:m) {

}
```

This loop is rather useless when there is nothing in between the curly brackets, however, so this is where we now need to add the data generating process and the estimation. To create coin flips, we can use the *rbinom* function, which generates random values from the binomial distribution. The first parameter is the number of flips, the second the possible values, and the last the probability. In this case we only want ones with some probability:

```
for (i in 1:m) {
  x <- rbinom(n, 1, p)
}
```

Now we have a fake data set (m times), but we need to run the estimator, and store the result:

```
for (i in 1:m) {
  x <- rbinom(n, 1, p)
  means[i] <- mean(x)
}
```

Finally, we can look at the distribution of means - the sampling distribution of the estimator of the mean of the binomial distribution, by using various statistics and plots:

```
hist(means)
sd(means)
var(means)
mean(means)
mean(means - p)
var(means) + mean(means - p)^2
```

References

Davidson, Russell and James G. MacKinnon. 1993. *Estimation and inference in econometrics*. Oxford: Oxford University Press.

Gujarati, Damodar N. 2003. *Basic econometrics*. 4th ed. Boston: McGraw-Hill.

Gut, Allan. 2005. *Probability: a graduate course*. New York: Springer.

Simon, Carl P. and Lawrence Blume. 1994. *Mathematics for economists*. New York: W.W. Norton & Company.

Wonnacott, Thomas H. and Ronald J. Wonnacott. 1990. *Introductory statistics*. 5th ed. New York: Wiley.