Statistics - 01 Introduction

Eric Stemmler

20.01.2021

Contents

1 Personal Introduction

Contact

- email (en): rcst@posteo.de
- email (mn): byambaa3007@yahoo.com
- Room: 415 (please send an email before visiting)
- phone: +976 8868 3742

2 Learning Goals

- Formulate statistical modelling problems
- Exploratory data analysis
- Basic computations in R

3 Why is statistics important?

- Learning from data about the world
- Randomness is omnipresent
- Estimation of uncertainty vs. establishing facts
- Making decisions

4 Vocabulary

4.1 Randomness

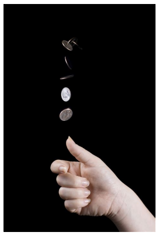

Figure 1: Uncertainty: Flipping a coin

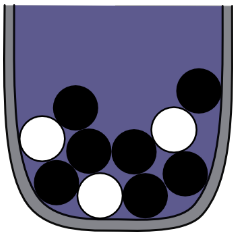

Figure 2: Variation: blindly drawing balls from an urn

Almost every real-world process has a random component. Randomness has two aspects: uncertainty and variation. Consider the flip of a coin. The laws of physics determine exactly the motion of a flipped coin and in principle, we *could* determine the exact outcome of a flip, given that we know under what conditions the flip occurred. However, the number of factors that we would need to know about the event of a flip to determine the outcome is very high and/ or impractical too measure. For instance, we would need to know the weight of the coin with high measuring precision, the starting velocity and momentum, direction, humidity and so forth. Even if we are able to measure all those factors, we would not get rid of randomness entirely, since these factors are themselves subject to random measurement error. For instance a scale can only determine the weight to a certain number of decimal places. Randomness is a general phenomenon and it appears as much as there is uncertainty about the state of things.

At the same time there is another aspect of randomness that we call *sampling variation.* Imagine an urn that contains a number of white and black balls. The balls inside the urn are what we call the *population.* Consider now that we draw a fixed number of n balls from the urn one by one, note down the fraction of white balls y and put them back in. This process is called a *trial.* Even if we know the number of balls that the urn contains, the outcome, the fraction of white balls will vary randomly across our different trials. This is called sampling variation. Statistically speaking, y varies randomly across different trials. However, y will only vary as long as n is less then the total number of balls in the urn. Since the urn only has a finite number of balls inside of it, one could just simply take out every ball and determine the fraction of white balls. But what if the number of balls (N) is very large and counting every white ball is practically impossible?

Interestingly we can still manage to estimate the *true value* of y. If we repeat this process m times, we get a sample of size m, i.e. the set of *outcomes* $y = \{y_1, y_2, \ldots, y_m\}$. As an estimate of the population mean, we can calculate the mean value of the sample, i.e. the *sample mean* or *empirical mean* $\bar{y}$ as

$$\bar{y} = \frac{1}{m} \sum_{i=1}^{m} y_i$$

$\bar{y}$ is a fair estimate of the true value. This is because the so-called *law of large numbers* guarantees that $\bar{y}$ becomes arbitrarily close to the true value as long as m is large. So we have essentially traded counting every ball against drawing a possibly large sample. In this example, it doesn't sound very exciting. But in general a statistical population can be of infinite size. In fact, the statistical population of interest in many estimation problems are of infinite size. Still, the *law of large numbers* guarantees us that we can get close to the true value. The size of a population is not always obviously infinite.

Examples for statistical estimates of infinitely large populations are

- probability for a coin to land on head
- (Human) gender ratio
- Measurement errors in physics
- Effectiveness of a vaccine
- The probability of getting cancer from smoking
- Temperature-dependent sex determination of *Crocodylus niloticus*
- ...

Let us summarize the important terms we have learned from the example of the coin flip and the urn so far:

Important terms:

- variation: the outcome of a sample varies randomly
- uncertainty: lack of knowledge of about a true value
- trial: the realization of an experiment
- population: all possible events or items
- population parameter: the true value

4.1.1 Simulation Results

Figure 3 shows a *scatter plot*. A scatter plot shows points whose x and y coordinate are equal to different parameters or variables. A scatter plot is commonly used for instance to visualize variation of data. The scatter plot in Figure 3 was printed using 2000 data points. Each data point represents 100 coin flips, which we can refer to as a trial. For each trial it was calculated

(a) length of the longest sequence of heads or tails (run)
(b) the number of runs

Resulting in 2000 values of length of longest run (y) and number of runs (x).

(Note: Since y and x can only have discrete values, i.e. 1, 2, 3, ... very many of the points in figure 3 would be drawn on top of each other, which means that many would look like a single point. To avoid this each point was randomly shifted a little to avoid the over-plotting. This is method is called adding a *jitter*.)

Figure 3 also shows one example of summarizing a data set. The actual data set consists of $2000 \times 100 = 200000$ data points. Plotting each data point individually may sometimes be of little use. For instance there could be simply too many points to plot, so that the plot is completely covered. Or when the raw data points are of little interest. In this case the data is also *nested*, which means different data points belong to different trials and it may sometimes be important to visually separate different trials, for instance because they were conducted under different circumstances. One way to solve these two problems at once is to *aggregate* the data set. To *aggregate* literally means *to consider as a whole*. In statistics aggregations mean to summarize a data set by calculating descriptive statistics as for instance the mean value or the variance. In this example however a different aggregation was chosen: We counted the number of occurrences of a certain pattern.

A common misconception about randomness is that is does not lead to patterns. By looking at figure 3 it becomes obvious that a pattern such as five times head in a row during 100 coin flips is actually very likely.

In the next section we will learn about how to calculate the probability of such a pattern.

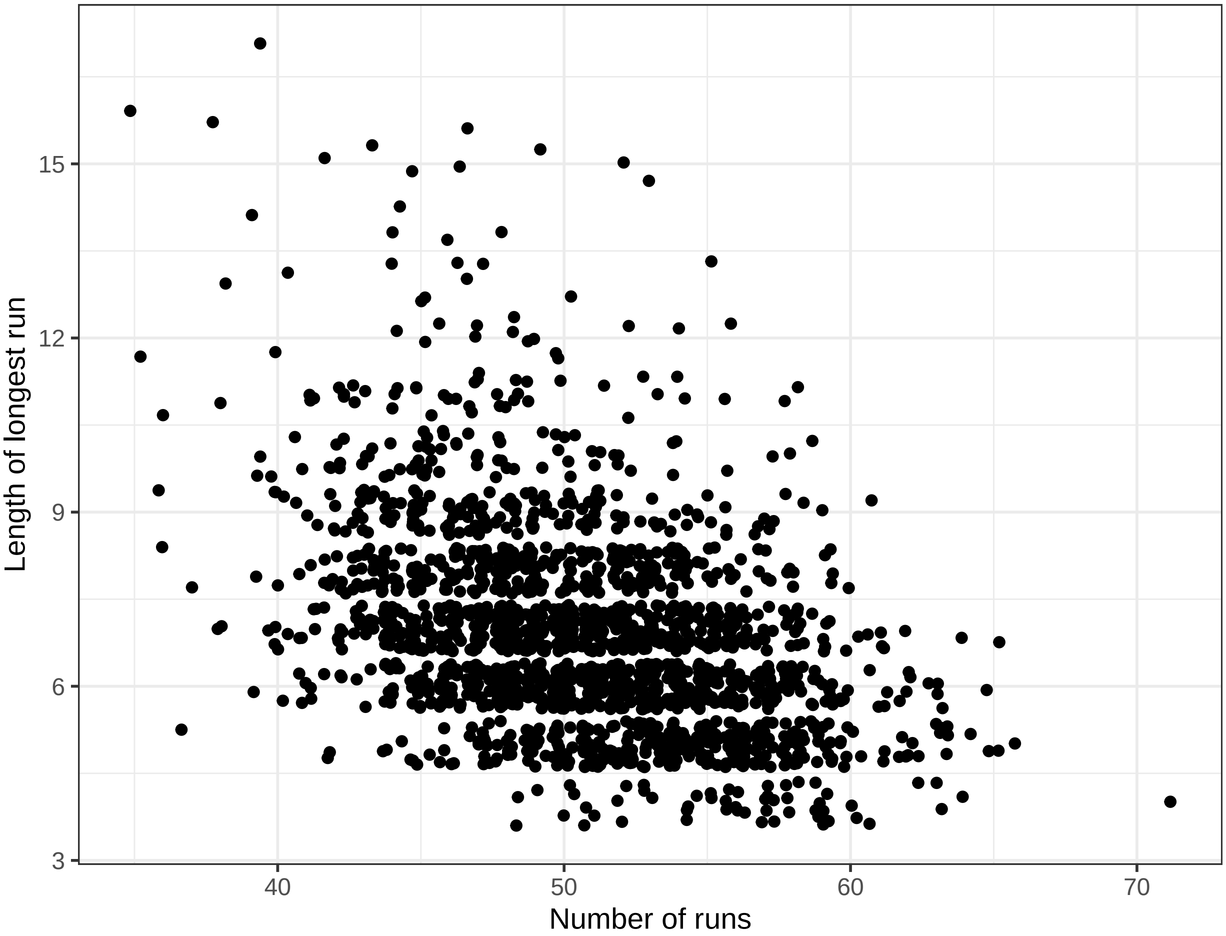

Figure 3: Length of longest run vs. number of runs from 2000 simulated experiments of 100 coin flips.

4.2 Coin flipping

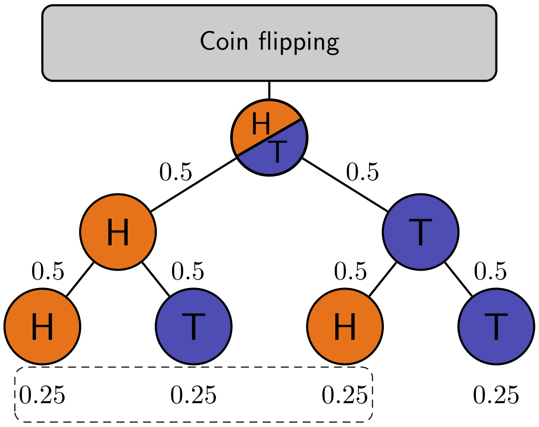

Figure 4: Probability tree for the outcomes of a coin flipping experiment

Figure 4 shows a *probability tree*. In simple cases it can be used to visualize the probabilities for specific outcomes of an experiment. Each node of the probability tree is an event. In our case, an event is either that the flipped coin shows a head ("H") or tail ("T"). If the coin is fair, then each of those events have the same probability. Since those two events are the only possible events, we know that the probability for each must be 0.5. Probabilities for distinct events must always sum up to 1 over all possible events.

If we consider flipping a coin more than once, the probabilities of those two events do not change. So the first two branches are annotated with values 0.5 denoting the probabilities for each event. If we flip the coin a second time, we again assume equal probabilities, but starting from each previous outcome we now have a total of 4 $(= 2^2)$ possible outcomes. We can continue doing so indefinitely, but the working principle remains the same.

Let us continue to consider only two coin flips, which is exactly what is shown in figure 4. We can now ask questions about the probabilities of certain outcomes of the experiment *two coin flips*. For instance we can ask: "What is the probability that we get at least one head ("H"). To determine this probability we first identify all branches that result in a least one heads, i.e. all branches with 1 or 2 times "H". Those branches have been marked in figure 4. The resulting answer is 0.75, which is the sum of the probabilities for getting

1. first flip: head, second flip: head
2. first flip: head, second flip: tail
3. first flip: tail, second flip: head

The probabilities of each of those branches alone are equally 0.25 which is the product of the probabilities along each of those branches, i.e. $0.25 = 0.5 \times 0.5$ or $0.25 = 0.5^2$, where 2 as the exponent stands for the number of flips performed.

Exercise:

- What is the probability of getting 5 heads?
- What is the probability of getting 5 heads in a row during 100 coin flips?

4.3 Binomial Distribution

There is a general solution to questions such as above: the *Binomial distribution*. The Binomial distribution is a probability function, i.e. it's value is a probability and therefore between 0 and 1. Functions like the Binomial distribution are also called *probability mass function*. The term *Binomial* tells us, that it evaluates the probability of experiments where there are only two possible outcomes. Because of this, the Binomial distribution belongs to the class of *discrete* distributions as opposed to continuous distributions.

The Binomial distribution has the following form:

$$p\left(k \mid n, \theta\right) = \binom{n}{k} \theta^{k} \left(1 - \theta\right)^{n-k} \tag{2}$$

The left side of the above equation is called a conditional probability. If one reads out this term it is said in the following way: "the probability p of k *given* (parameters) n and θ."

The parameter k is the number of *successes* in a number of n trials. θ denotes the probability of *success* in each single trial. *Success* does not have to literally be a success. A *success* can be anything, any event of interest as opposed to *fail*, which likewise doesn't literally have to mean failure (although it can be applied to experiments of success and failure). *Fail* often simply means the opposite of *success*. It is entirely up to us what we define as *success* and *fail*.

For example, let us say having 5 heads in a row is what we call a *success*. A *fail* is then simply the opposite: not having 5 heads in a row. We can use this definition together with the Binomial probability distribution to determine the probability of getting 5 heads in a row in a series of $n = 5$ coin flips.

The probability of this is

$$\begin{aligned} p\left(k = 5 \mid n = 5, \theta = 0.5\right) &= \binom{5}{5} 0.5^{5} \left(1 - 0.5\right)^{5-5} \\ &= 1 \times \frac{1}{32} \times 1 \\ &= \frac{1}{32} \approx 0.03125 \end{aligned}$$

which is the same result as we would have determined from the probability tree in figure 4.

In the next section we will apply the Binomial distribution to a real-world estimation problem, since coin flips are rarely of scientific interest.

4.4 Estimating fish population

Figure 5: Fishes in a lake

We have learned in the previous section that by using a probability distribution one can determine the probability of specific outcomes of an experiment. In general we refer to the process of defining a probability distribution in relation to phenomena as *modelling.* Of course, one can question if the probability distribution used does actually apply to the real world. This is why there are always certain assumptions connected to the use of a specific model or probability distribution. For instance, the model of the Binomial distribution that was used in the previous example assumes that the probability of getting a head in any single coin flip is independent from any previous coin flips. More commonly this is described as: The coin flips are modelled as *independent and identically distributed.* This assumption is violated for instance by the fact, that if a single coin will wears off if it is flipped 1 million times. However, it is often reasonable to ignore this violation as the change in probability might be rather small. One might always find aspects of reality that violate the theoretical assumptions of a model. This is why strictly speaking, every model must be wrong. However, this does not mean that the model is useless.

Ecologist are often interested in population dynamics in a certain geographic area. Especially in relation to environmental and climatic factors. Consider for instance the ecosystem of a lake with different specimen of fish as shown in figure 5). Different species of fish require different conditions in order to reproduce and survive, e.g. water temperature or pH value of the water. If climatic or environmental conditions of the lake change this may have an influence on the fish population. Hence monitoring of population numbers is of interest in ecology.

4.5 Modelling Fish Population - Binomial distribution

Assuming that we have no knowledge about the current number of fish in the lake, we would therefore start to take samples of fish from it. This sampling situation is in principle no different from the experiment of the urn from section 4.1 and figure 2. Again, this is only true under certain assumptions, which in this case are: The fish population of the lake is ecologically closed during the sampling period. That is, no new fish is added or removed during the sampling period. Second, the probability of observing a fish is equal for each fish and likewise each fish is chosen with equal probability.

In fact it has been shown that given that the sample of a population from an ecologically closed environment is Binomially distributed (see equation (2)), it is possible to estimate the total size of the population (Olkin et al., 1981).

We can model the sampling of fish from a lake as a Binomial distribution. The first step here is to define the meaning of the parameters of the Binomial distribution. The meaning of N is quite simple: it is the number of fish that exist in the lake, i.e. the population. This number will be of major interest for our inference, once we have collected some data and have fitted the model to this data. θ is what we can call the *capture probability.* θ can be thought of as the probability of catching any single fish. In analogy to the urn experiment, catching a fish is like drawing a white ball and not catching a fish a black ball. In this sense the previously mentioned general definition of the Binomial distribution in terms of success and failure is actually applicable: Having success means to catch a fish, failing means to not catch a fish.

Now let us assume a hypothetical lake for which we know how many fish there are. Furthermore, let us assume we know exactly the probability of catching a fish. If we start fishing in that lake, how many fish can be expect to catch? Of course this number would vary from time to time, exactly like in the urn experiment. But it would be interesting how much the sampling varies. Figure 6 shows the plotted Binomial distribution for values of y ranging from 0 to 25 for a hypothetical lake with 25 fish and a capture probability $\theta = 0.5$.

On the x axis it shows the number of fish that can be caught and on the y axis the respective probability. Interestingly there are actually two values with maximum probability: 12 and 13. Recall that the Binomial distribution is a discrete distribution, i.e. only applies to discrete data. However the expectation value of such a distribution does not need to be discrete. The expectation value, is defined as $N \times p$, which in this particular case is 12.5. Both lower and higher values than 12 or 13 are possible but less likely. So we can expect to see number of fishes captured in total ranging between 9 and 16 is perfectly possible and not unlikely.

In figure 6 we see what can be expected and also what numbers of fish caught are rather unlikely to occur.

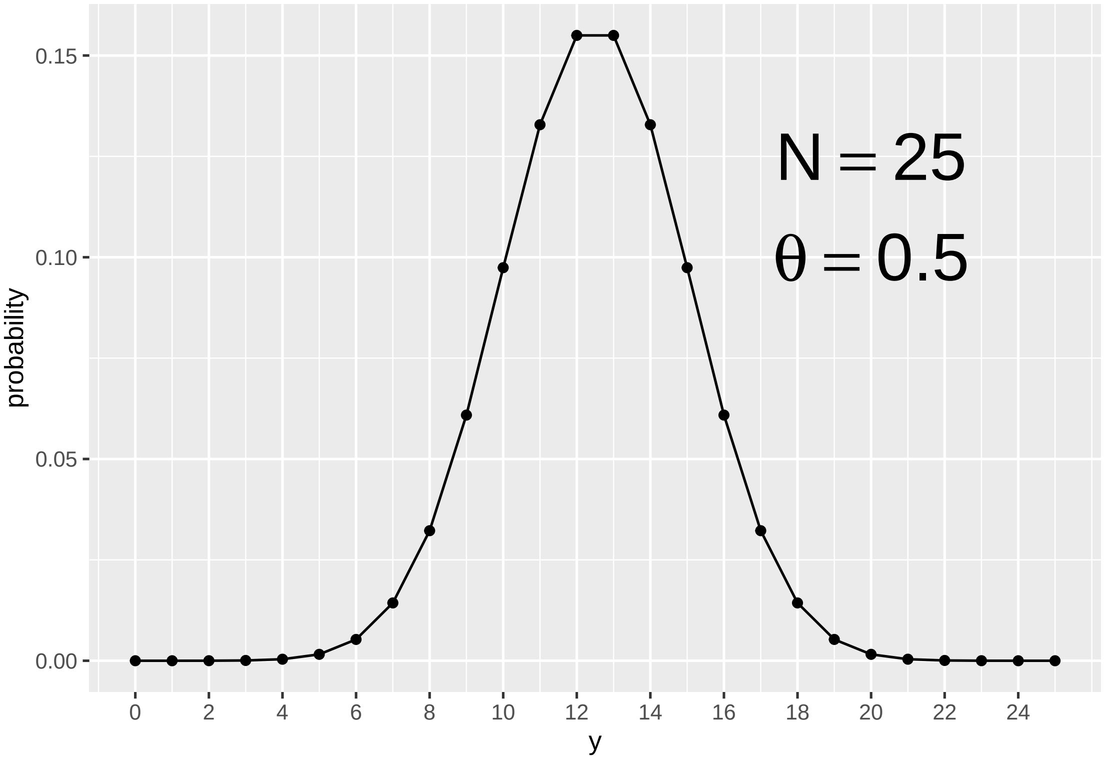

Figure 6: Graph of Binomial distribution with parameters set to $N = 25$ and $\theta = 0.5$

In this hypothetical example however, we know the values of both parameters. In practise the situation is usually the other way around: We have observed or collected some data and we want to learn about the values of the parameters. As of today there are two main strategies on how to *estimate* parameter values from data - Bayesian and Frequentist inference. However, in both cases a probability distribution is used for the estimation. Until here we have said, that a probability distribution tells us the probability of any outcome of an experiment, based on specified parameters. The same function can also be used to tell the probability of a *parameter* having a specific value. If we use a probability function for this purpose, strictly speaking, the value of the function is not called probability anymore but *likelihood*. Practically, the difference between probability and likelihood is that probabilities across all cases must sum to 1 and likelihoods not.

Let us consider the following example: Consider again you want to estimate the number of fish inside a lake. Let's assume we know the capture probability $\theta = 0.7$. We went to the lake and started to fish for 1 hour. During this time we were able to catch 3 fish. We know, that catching a fish from a closed population follows the Binomial distribution (2). How likely is it that the fish population is for instance $N = 10$? The answer is

$$\begin{aligned} p\left(y = 3 \mid \theta = 0.7, N = 5\right) &= \binom{5}{3} 0.7^3 \left(1 - 0.7\right)^{5-3} \\ &= 0.3087 \end{aligned}$$

In principle, we could do this calculation for a range of values for N and then chose the value for N with the highest likelihood as our estimate. This approach is called *maximum likelihood estimation* and in simple cases it is possible to analytically derive the so-called *maximum likelihood estimator* for a given sampling distribution to determine directly an estimate for N.

Table 1: Collected fish data: number of caught fish in 5 locations at 3 different time points.

	sampling occasions		
site	t1	t2	t3
1	2	1	2
2	3	5	5
3	0	1	1
4	2	2	1
5	3	3	3

Beginning from the following section, I will demonstrate a more practical example of estimating unknown parameters from data using the R programming language and how this looks with realistic data sets.

4.6 Data Set

Consider an ecological population monitoring study, where one of the study goals is to estimate the number of fish population in a lake. Research funds are small and many scientific staff cannot support the study, because they have to work in home office. Ecology scientists decided therefore on a simple capture counting study design as opposed to a more laborious capture-recapture study plan. Furthermore, the sampling is to be done by fishing with releasing the fish back into the water after capture. 5 different locations on the lake are chosen to have a reasonable coverage of the lake area. To capture sampling variation as good as possible and to quickly finish the study, every location is to be sampled only 3 times. See table 1 for an example of such collected data.

In order to start to get an understanding of any data set, it is always a good idea to visualize the data set at hand.If we do not yet know, how population dynamics actually work, this step would be even more important. Visualizing the data can be informative as in how to approach the modelling step and to start exploring possible models. In our example it is not as important as that, since we are using an established method of estimating populations. Nevertheless, we also want to make sure that our data looks plausible. Figure 7 shows a histogram of the data from table 1. One thing that we notice is that the distribution indeed looks as if it could have been sampled from a Binomial distribution since is vaguely resembles the typical hill shape. We also have to note, that this sample is relatively small in size given that we have relatively small capture numbers. This could mean that there is either only few fish or that fish in this lake are hard to catch, i.e. θ being small or even both. Graphically estimating, we can note that an average per site capture number of between 2 and 3 would be plausible. In total this would mean 10 to 15 fish being caught.

4.7 Fitting the model

```
## Inference for Stan model: fish.
## 4 chains, each with iter=4000; warmup=1000; thin=1;
## post-warmup draws per chain=3000, total post-warmup draws=12000.
##
##         mean se_mean    sd 2.5%   25%   50%   75% 97.5% n_eff Rhat
## p       0.75    0.00  0.15 0.30  0.70  0.80  0.86  0.93  2115    1
## lambda  3.39    0.06  2.04 1.66  2.45  3.01  3.71  7.79  1350    1
## Ntotal 16.93    0.28 10.22 8.30 12.27 15.07 18.55 38.96  1350    1
##
## Samples were drawn using NUTS(diag_e) at Mon Jan 18 18:20:21 2021.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

Once we have a data set collected and formatted into a computer readable format, we can use the R programming language to perform the next step: *model fitting*. To fit a model to data means to set the parameter y of a given model or probability distribution. In doing so, y is not consider a variable anymore and the only variables left in the equation of the distribution are the parameters that we want to estimate. We

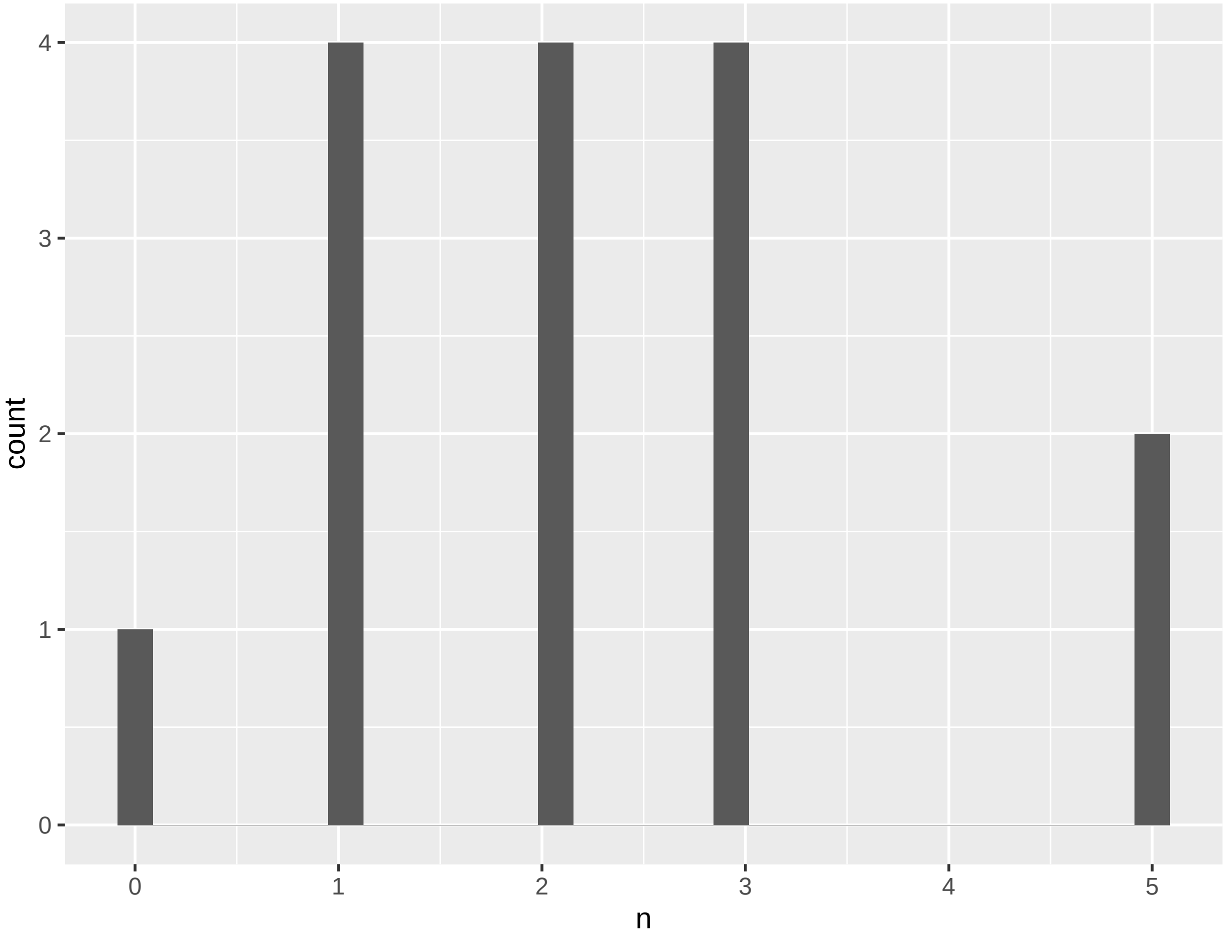

Figure 7: Histogram of the collected fish capture data.

have already seen how fitting a model to a single data point works in the above example where we estimated the probability of $N = 5$.

In practise using a single point of data will always be insufficient. We have seen from the discussion earlier on randomness and variation, that it is always possible to sample data points that are rare. Any scientific research therefore needs to reduce or better eliminate the chance that a sample is not representative. That means we need to be sure that the data we collect deliver a "typical picture". The only way to ensure this, is to consider as much data as we can.

In order to fit a model to several data points I again refer to the probability tree example as shown in figure 4. The resulting probability of a sequence of trials is the product sum of the probabilities of each single trial. Accordingly, fitting a model to several n data points $y = \{y_1, y_2, \ldots, y_n\}$ for any given value for parameter(s) θ, we calculate the individual probabilities $p_i\left(y_i \mid N, \theta\right)$ and determine $p\left(y \mid N, \theta\right) = \prod_{i=1}^{n} p_i$.

Without going into the detail of the model fitting procedure, the above print out shows a typical summary of the model fitting process. In this case the statistical modelling language STAN was used to implement the model. Parameter estimates are provided by summary statistics of the resulting parameter distributions. These distributions can be used for the next step, which is inference, in order to formulate statistical statements about the value of the parameters.

4.8 Inference - Parameters as estimates

Once we have obtained the probability distribution of the unknown parameters we can start to draw conclusions about the value of these. Figure 8 shows different percentile values. *Percentile* or *quantile* values are a possibility to summarize probability distributions. In our example they are particularly help, because we can express an degree of certainty about the upper bound of a population. In other cases estimation of parameters will focus on the mean value of distribution in order to provide an estimate. For instance, the 50th percentile for N_{total} is roughly 15. Hence we can formulate: Given the observed data of captured fish, the probability that $N_{\text{total}} \leq 15$ is 50%. More informative however, we might roughly state that we are 95% sure that $N_{\text{total}} \leq 28$, which could be especially relevant if it is about the population of an endangered species.

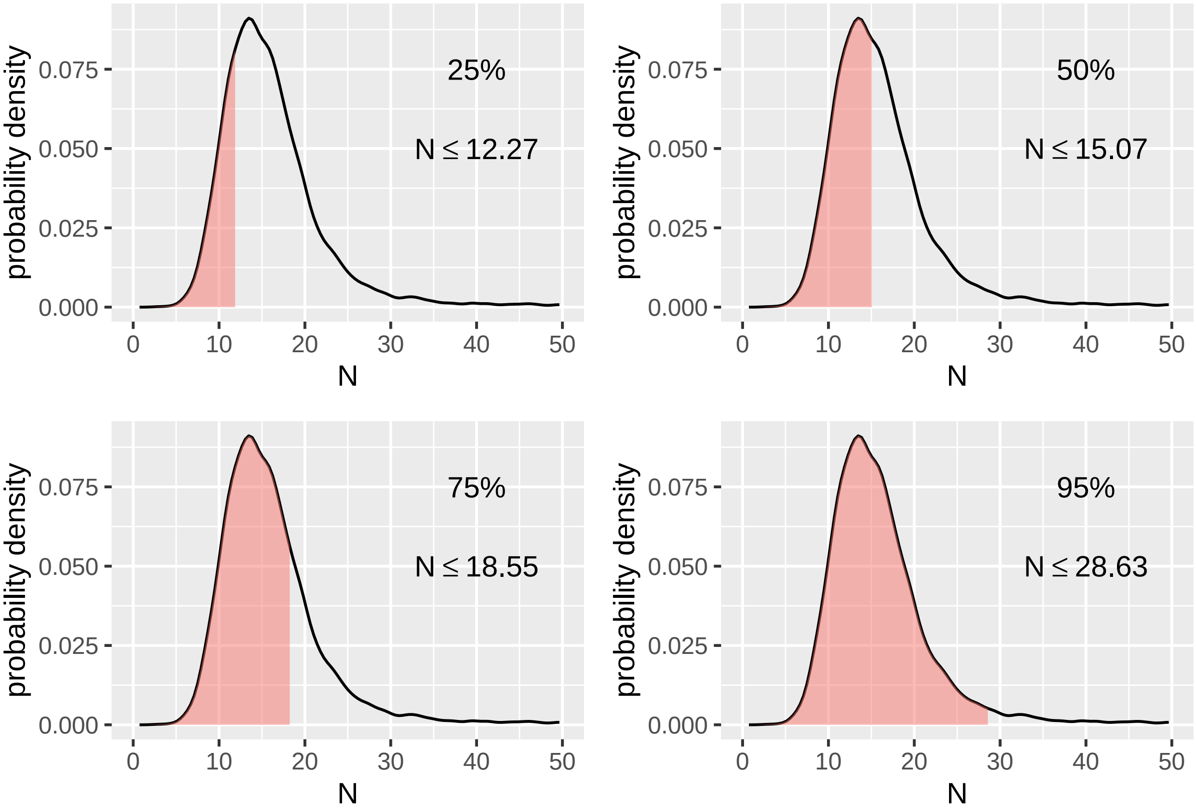

Figure 8: 25th, 50th, 75th and 95th percentile values as summary statistics of the resulting probability distribution for the value of N_{total}, estimating an upper bound for the fish population.

5 Summary

In this lesson we have started by discussing the role of statistics in science and society. We introduced basic principles and phenomena such as uncertainty, sampling variation, population, parameters and data. We clarified the difference between probability and likelihood. It was demonstrated that simple experiments and more complex and practical problems of estimation share an underlying working principle: probability distributions. The goal of many statistical estimation problems can be broken down to estimation of probability distributions of unknown parameters. Once an estimation of the probability distribution is obtained, either via Bayesian or Frequentist approach, we can formulate arbitrary statements about the value of unknown parameters. Simple distributions such as the Binomial distribution seem applicable only to "school book problems", but they can be seen as fundamental building blocks of more complex models or sometimes, as shown in the fish population example be even applicable to real world situations.

References

Ingram Olkin, A John Petkau, and James V Zidek. A comparison of n estimators for the binomial distribution. *Journal of the American Statistical Association*, 76(375):637–642, 1981.