2 Statistics

Statistics can be defined as the collection, analysis, interpretation, and presentation of numerical data. We often distinguish between *descriptive statistics* which is concerned with presenting and summarizing data and *inferential statistics* which is aimed at drawing conclusions under uncertainty and random variation.

2.1 Descriptive statistics

The term *descriptive statistics* denotes methods that describe, summarize, and present data in a useful way, such that patterns, trends, and other characteristics are clearly visible. While presenting the raw data in its entirety can be useful, it might be difficult when the data set is large and complex. Often it is more effective to present and visualize summaries of the data, which focus on their most important aspects. This makes it easier to interpret and understand the data.

When we talk about a descriptive statistic, we simply mean some quantity that we compute from the observed data.

> Definition 2.1 Statistic
>
> A *statistic* is any number or quantity that is computed from data.

Two of the most important descriptive statistics are the *central tendency* and the *dispersion*.

The central tendency is a central or typical value which characterizes the data. The most common measure of central tendency is the *mean value*.

The dispersion is a measure of spread or variability. The most common measure of dispersion is the *standard deviation*.

Definition 2.2 Mean and standard deviation

Mean

The average or mean value of a set of numbers is often used to represent the central tendency of the numbers. It can be computed as:

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

Standard deviation

The standard deviation of a set of numbers is a measure of how much they are spread out. A low standard deviation (close to zero) means that all the numbers are close to the average, whereas a high standard deviation means that the numbers are dispersed on a large range. It can be computed as:

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2}.$$

The squared standard deviation, σ_x^2, is called the *variance*.

Example 2.1 Mean and standard deviation

Consider a data set which consists of five numbers:

$$x = \{1, 8, 4, 10, 2\}.$$

Let us compute the mean and standard deviation of this data:

$$\mu_x = \frac{1}{5}(1 + 8 + 4 + 10 + 2) = 5$$

$$\sigma_x = \sqrt{\frac{1}{5}\left((1-5)^2 + (8-5)^2 + (4-5)^2 + (10-5)^2 + (2-5)^2\right)}$$
$$\approx 3.46$$

Now, consider another data set which also consists of five numbers:

$$y = \{7, 5, 4, 3, 6\}$$

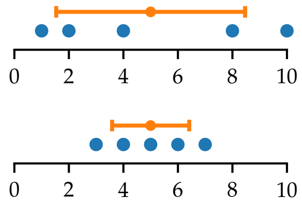

Figure 2.1: Two data sets from Example 2.1 with each observation shown as a dot. The indicated range is the mean values plus/minus one standard deviation.

The mean and standard deviation of this data is:

$$\mu_y = \frac{1}{5}(7+5+4+3+6) = 5$$

$$\sigma_y = \sqrt{\frac{1}{5}\left((7{-}5)^2 + (5{-}5)^2 + (4{-}5)^2 + (3{-}5)^2 + (6{-}5)^2\right)}$$

$$\approx 1.41$$

We can see that the two data sets have the same mean value, but the first data x have a much higher standard deviation than the second data y. In other words, the two data sets have the same central tendency but x is more dispersed (spread out) than y. The data is visualized in Figure 2.1.

2.2 Inferential statistics

The entire set of data that we want to study is called the statistical *population*. Most often we think of the population as a set of similar items (people, things, events, etc.) under study. A population can be a real and finite set (such as all all students in a class), a hypothetical set (such as all possible hands in a game of poker), or even an infinite, hypothetical set (such as all possible computer programs).

If the population is accessible and not too large, we might be able to gather all the relevant data from the population, and we can then compute any desriptive statistic we are interested in. However, often it is not practical to acquire data from the entire population. As an alternative, we can take a smaller sample from the population and use the sample statistics of the small sample to estimate the likely values of the parameters of the population.

Before we continue, let us define some of the most important terms we have just used.

Definition 2.3 Terms used in inferential statistics

Population The entire set of items of interest.

Sample A subset of the population.

Statistic A numerical value computed from the sample.

Parameter A characteristic of the population under study.

Estimator A statistic used to estimate a parameter.

Example 2.2 Exit poll

Consider the following statement that talks about an exit poll conducted after an election has taken place, but before all the votes have been counted. The aim of the exit poll is to give a quick estimate of the results of the entire election:

> "An early exit poll taken among 1000 voters shows that the incumbent president has a small lead of 51 percent against her oponent."

The population is all voters who have voted at the election.

The sample is 1000 voters who were asked in the exit poll.

The statistic from the sample is that 51 percent voted for the incumbent president.

The parameter of the population we are interested in is the percentage of all voters who voted for the incumbent president.

The estimator we use is the 51 percent from the sample, which is used to estimate the population parameter.

There are two reasons why we might not trust that the incumbent president will be re-elected with 51 percent of the votes based on the exit poll:

1. Since only 1000 voters were asked, the estimate of 51 percent will be subject to random variation, since we would probably have got a different result if we had taken a different sample.

2. Maybe the 1000 voters are not even representative of the entire population—it depends on exactly how the sample was chosen.

Clearly, when we use a statistic from a small sample to estimate a parameter of a large population, we cannot expect that our estimate will exactly match the population parameter. While the population parameter is a fixed but unknown value, the estimator will depend on exactly how we have chosen our sample. If we choose a different sample, the estimator will be a different number.

Also, it might not be clear exactly which sample statistic is the best estimator for some population parameter. It turns out that the sample mean is a good estimator of the population mean, but to estimate the population standard deviation, a small correction is made in the formula.

Definition 2.4 Estimators of mean and standard deviation

Sample estimator of mean

The population mean is estimated by the sample mean:

$$m_x = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Sample estimator of standard deviation

The population standard deviation is estimated by the *corrected* sample standard deviation:

$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - m_x)^2}.$$

Note that here we divide by $n - 1$.

When choosing a sample, it is important that the sample is representative of the population. If we are biased in the way we choose the sample, such that some items in the population are more likely than others to be included, the sample will not be representative of the population. One way to avoid any such biases is to choose the sample by a random procedure, such that all items in the population have an equal chance of being included in the sample. This is called a *simple random sample*.

Definition 2.5 Simple random sample

A set of n items chosen from the population such that each item is included with equal probability.

When we use a random sample, the estimator will be a random variable. By *random* we simply mean that the sample statistic used as our estimator will depend on exactly which sample was chosen, and that it would be different had we chosen a different sample. The good news is that we can say something about how much the sample statistic will vary across the different possible samples we can imagine.

2.2.1 *Standard error of the mean*

Let us say that we are interested in estimating the mean value in a population. Let μ and σ denote the mean and standard deviation of the population. Now, we take a simple random sample of size n, $\{x_1, x_2, \ldots, x_n\}$ and compute the sample mean

$$m_x = \frac{1}{n}(x_1 + x_2 + \cdots + x_n).$$

Since each item in the sample is independently chosen from the population, the variance of m_x can be computed as[1]

$$\sigma^2_{m_x} = \frac{\sigma^2}{n}.$$

Since we do not know the population variance σ^2, we can substitute in the sample variance s_x^2. Taking the square root gives us the final formula for the standard deviation of the sample mean, also known as the *standard error of the mean*:

Definition 2.6 Standard error of the mean

The standard deviation of the sample mean can be estimated as

$$\sigma_{m_x} = \frac{s_x}{\sqrt{n}}.$$

The formula tells us that when we estimate the mean of a population using the mean of a sample of size n, the standard deviation of our estimate is proportional to the standard deviation of the sample. We can see that if we increase the size of the sample n, the standar error will decrease. If we want a low

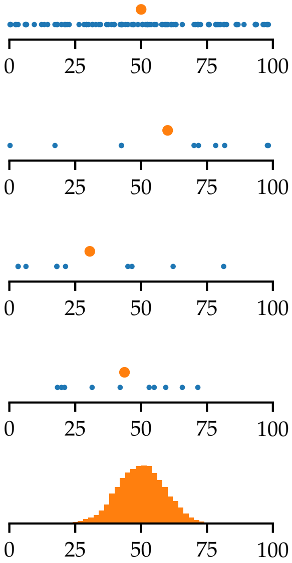

Figure 2.2: Top: A population of 100 numbers between 0 and 100 with population mean $\mu = 50$ and. Three rows in the middle: Three different simple random samples of size $n = 10$ from the population with the mean of in each sample indicated. Bottom: Histogram of sample means from 10 000 simple random samples of size $n = 10$ from the population.

[1] This can be proven mathematically, but we will not do this here. It can be a good idea to check the formula empirically by simulating several random samples from a population, computing the sample mean for each sample, and finally computing the standard deviation of the sample means.

error on our estimate, we need a large sample, but of course a large sample is typically more expensive or difficult to gather, so there is a trade-off to make.

Example 2.3 An AI system that estimates age

Let us imagine an AI system that can estimate the age of a person based on a photo. While the system is not expected to estimate the age correctly for everybody, it is supposed to be correct on average: This means that the error in the estimate should be zero on average. We decide to examine this by letting the system estimate the age of randomly chosen 10 people. We then compute the estimation error by subtracting the true age from the estimate. We get the following errors

$$x = \{7.1, -5.2, -2.6, -2.2, -6.4, -6.7, 1.5, -5., 6.8, 1.8\}.$$

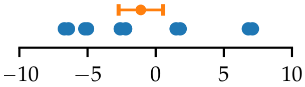

Figure 2.3: Data from Example 2.3 with each observation shown as a dot. The indicated range is the mean value plus/minus one standard error of the mean.

Based on the measurements, we compute the mean error

$$m_x = -1.09.$$

We see that the mean error is negative, so it appears that the system tends to underestimate the age a bit. But it could also just be due to random variation that occured because we only examined 10 randomly chosen people. Maybe, if we had chosen 10 other people, the result would be very different.

To look further into this, we compute the sample estimate of the standard deviation to be $s_x = 5.16$, and use this to compute the standard error of the mean

$$\sigma_{m_x} = \frac{s_x}{\sqrt{n}} = \frac{5.16}{\sqrt{10}} \approx 1.63$$

This standard error tells us something about the variation we expect in our computation of the mean, when we examine a sample of 10 people. The variation in our estimate of the mean due to the random choice of sample is so large ($\sigma_{m_x} \approx 1.63$) that we know that the estimate $m_x = -1.09$ is subject to a lot of random variation, so we should be careful with our conclusions.

2.2.2 *Confidence intervals*

We have seen how we can use a statistic computed on a sample to estimate a population parameter, and we have discussed how the standard deviation of the statistic tells us something about how much the sample statistic would vary across different possible random samples. We can use these ideas to create a *confidence interval* which roughly speaking is a range of potential values in which the population parameter is estimated to lie.

A confidence interval is also a sample statistic, which means that it is an interval that is computed from a sample that we use to infer something about the population parameter. Just like any other sample statistic, the confidence will also be subject to random variation caused by the selection of the random sample. In other words, if we had chosen another sample, we would also get another confidence interval.

A procedure for computing a confidence interval has an associated *confidence level* or *coverage*. The confidence level specifies the proportion of confidence intervals (across all possible samples we can imagine) that actually contain the true population parameter. If a procedure for computing a confidence interval has a confidence level of 95%, it means that if we imagine taking 100 random samples from the population, each of which has the same sample size, and we compute the confidence interval for each of the samples, we would expect that 95 of the computed confidence intervals would include the true population parameter.

Usually we only have a single sample from our population, and so we can only compute a single confidence interval. Then we have no way to determine whether or not the population parameter actually is in our interval or not. With a confidence level of 95%, all we can say is that 95% of the times we use our procedure to compute a confidence interval, it will include the true value, and for a single confidence interval estimate there is only a 5% chance that we were so unlucky with our random sample that the population parameter is outside our interval.

Why not set the confidence level to 100% then? It is actually very easy to make a procedure for computing a confidence interval with 100% confidence level: We can simply take the interval from minus infinity to infinity, which of course is guaranteed to include the population parameter. But such an interval is obviously not very useful. It is more interesting to find a narrow interval, but in order to do this we need to make a compromise by accepting a lower confidence level.

There are many different methods for constructing confidence intervals that are based on different assumptions. Here we will consider the most simple confidence interval which is based on the assumption that the sampling distribution of the mean is approximately a normal distribution (see Figure 2.4). In a normal distribution with zero mean and standard deviation equal to one, the probability of getting a value less than -1.96 or greater than 1.96 is 5%. So to construct a 95% confidence interval, we can simply take an interval around the sample estimate of the mean that extends to ± 1.96 times the standard error.

Definition 2.7 95% confidence interval for population mean

A 95% confidence interval for the population mean can be computed as:

$$m_x \pm 1.96 \cdot \sigma_{m_x} = 1.96 \cdot \frac{s_x}{\sqrt{n}}$$

Note that this interval is only precise when the sample size is sufficiently large ($n \geq 30$) and constitutes a small fraction of the total population ($n \leq N \cdot 10\%$).

2.2.3 *Estimating a proportion*

In some cases, the population parameter we are interested in is a *proportion*, i.e., how many items out of the total population that have some property of interest. If we had access to the entire population, we could compute the proportion π as[2]

$$\pi = \frac{\eta_x}{N},$$

where η_x is the number of items in the population that has the property, and N is the total number of items in the population. We can think of a proportion as a special kind of mean value by defining an *indicator variable*

$$z_i = \begin{cases} 1 & \text{if object } i \text{ has the desired property} \\ 0 & \text{otherwise.} \end{cases}$$

This indicator takes the value 1 if and only if the ith item has the property we are looking for. We can now compute the

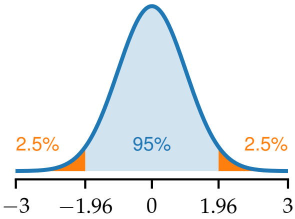

Figure 2.4: Normal distribution with zero mean $\mu = 0$ and unit standard deviation $\sigma = 1$. 5% of the probability mass lies beyond the critical value ±1.96.

[2] If the population were infinite, we would need to define the proportion as an appropriate limit.

proportion as the mean value of the indicator

$$\pi = \frac{1}{N}\sum_{i=1}^{N} z_i.$$

To estimate a population proportion from a sample, we can use the observed sample proportion.

$$p = \frac{n_x}{n},$$

where n_x is the number of items in the sample that has the property, and n is the sample size. Since the proportion is a mean value, we can compute the standard error of the proportion with the formula in Definition 2.6. The sample estimate of the standard deviation of the proportion can be computed as

$$s_p = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(z_i - p)^2} \approx \sqrt{p(1-p)}.$$

Plugging this into the definition of the standard error gives us the following:

Definition 2.8 Standard error of a proportion

The standard deviation of the sample estimate of a proportion can be estimated as

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}}.$$

Similar to the confidence interval for the population mean, we can create a confidence interval for the population proportion:

Definition 2.9 95% confidence interval for a proportion

A 95% confidence interval for the population proportion can be computed as:

$$p \pm 1.96 \cdot \sigma_p = 1.96 \cdot \sqrt{\frac{p(1-p)}{n}}$$

Note that this interval is only precise when the sample size is sufficiently large ($n \geq 30$) and constitutes a small fraction of the total population ($n \leq N \cdot 10\%$), and the population proportion is not too extreme ($10\% < p < 90\%$).

When the population proportion is either very large or very small, say less than 10% or greater than 90%, the confidence interval above is not very good. As a better alternative, we can use the Agresti-Coull interval, which is very similar but uses a correction in the estimation of the population proportion. When estimating the proportion from the sample, the correction consists of adding two to the numerator and four to the denominator,

$$\tilde{p} = \frac{n_x + 2}{n + 4}.$$

This corresponds to having a sample size of $\tilde{n} = n + 4$ where we artificially add four extra items of which two have the property that the proportion measures. With this correction, the confidence interval is given by:

Definition 2.10 Agresti-Coull interval

$$\tilde{p} \pm 1.96 \cdot \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{\tilde{n}}}$$

$$\tilde{p} = \frac{p \cdot n + 2}{n + 4}$$
$$\tilde{n} = n + 4$$

2.3 Sample size calculation

Now that we know how to compute a confidence interval both for a population mean and proportion, we can use the same formulas in reverse to calculate how large a sample size is needed to estimate a parameter within a certain desired margin or error. By isolating the sample size in Definition 2.7 we arrive at the following formula for the sample size n:

Definition 2.11 Sample size calculation for a mean

At a 95% confidence level, the required sample size to estimate a mean within a given margin of error is:

$$n = 1.96^2 \frac{s_x^2}{e^2}$$

n Required sample size

s_x^2 The expected sample variance. This is an unkown quantity, but we can substitute in our best guess, perhaps based on a small pilot study.

e Desired margin of error (half width of the confidence interval).

Prior to gathering data from a sample of the population, we can use this to give us an indication of how large a sample is needed in order to estimate a population parameter within a certain margin of error. Since we don't know the variance in advance, we can substitute in our best guess, or estimate it using a small pilot study.

Similarly, we can isolate the sample size in Definition 2.9 to arrive at a formula for computing a sample size for a proportion.

Definition 2.12 Sample size calculation for a proportion

At a 95% confidence level, the required sample size to estimate a proportion within a given margin of error is:

$$n = 1.96^2 \frac{p(1-p)}{e^2}$$

n Required sample size

p The expected sample proportion. This is an unkown quantity, but we can substitute in our best guess or use $p = 50\%$ as a worst case estimate.

e Desired margin of error (half width of the confidence interval).

The required sample size is greater when estimating proportions close to 50%, so to be most conservative we can adjust our guess a bit away from the extremes or simply plug in 50% as a worst case guess.

Example 2.4 Sample size for image recognition accuracy

Let's say we have an image recognition system that has been tested thoroughly and has a recognition accuracy of 90%. We have now developed a new system, which we think is better than the old system. To examine this, we decide to make an experiment, where we classify a bunch of images with the new system to verify its accuracy. How many images should we test the new system on?

We can calculate the required sample size using the formula in Definition 2.12. We think that the new system has an accuracy of around 95%, so we decide that we need to estimate the accuracy within a margin of error of at least ±5%. Based on this, we get

$$n = 1.96^2 \frac{0.95(1-0.95)}{0.05^2} \approx 73.$$

We then classify 73 images and end up with, say, 67 correct classifications. The sample estimate of the accuracy is then

$$p = \frac{67}{73} \approx 91.8\%$$

and we can compute the confidence interval as

$$p \pm 1.96 \cdot \sqrt{\frac{p(1-p)}{n}} \approx 91.8\% \pm 6.3\%$$

In other words, our estimated range for the accuracy is 85.5%–98.1%. Based on this we can not even tell if the new system is better than the old one or not.

On second thought, we are not so sure the new system will actually be 5 percentage points better than the old system. To be on the safe size, we decide to reduce the margin of error to ±1% and assume that the accuracy is only around 90%. With these more conservative assumptions the required sample size is

$$n = 1.96^2 \frac{0.9(1-0.9)}{0.01^2} \approx 3457$$

We new classify 3457 images and end up with, say, 3170 correct classifications. The sample estimate of the accuracy is then

$$p = \frac{3170}{3457} \approx 91.7\%$$

and we can compute the confidence interval as

$$p \pm 1.96 \cdot \sqrt{\frac{p(1-p)}{n}} \approx 91.7\% \pm 0.9\%$$

With an estimated range for the accuracy of 90.8%–92.6% we are now confident that the new system is indeed a few percentage points better than the old system.

Problems

1. Consider a data set that consists of six numbers:

$$x = \{1, 5, 2, 7, 3, 8\}$$

What is the *mean* and *standard deviation* and *variance* of the data?

2. John plans on taking a taxi home from work every day next year. He would like to know how long the commute will be on average, and to find out he takes a taxi home five times during.

   (a) In this scenario, what would be the population, sample, statistic, parameter, and estimator?

   (b) The five taxi rides take $t = \{42, 34, 29, 42, 48\}$ minutes. Compute a confidence interval for the average time.

3. A face recognition system has been designed to identify members of the public as potential criminals. When testing the system on 100 000 random people, the developers found that it had identified 1 500 people as potential criminals. Out of these 1 500 people, further analysis showed that 1 350 were actually innocent, whereas the remaining 150 were indeed criminals. Now, considering if we want to employ the system, we are interested in finding out how large a proportion of the people identified as criminals are actually criminals.

   (a) In this scenario, what would be the population, sample, statistic, parameter, and estimator?

   (b) Compute a confidence interval for the proportion.

Solutions

1. $\mu \approx 4.33$, $\sigma \approx 2.56$, $\sigma^2 \approx 6.56$.

2.(a) The population is all (hypothetical) taxi rides from work to home that John will take next year; the sample is the five rides; the statistic is the mean duration of the ride; the parameter is the mean duration of the rides next year; and the estimator is that we use the mean of the five rides to estimate the mean of next year's rides.

(b) $m_t = 39 \pm 6.56$

3.(a) The population is all (hypothetical) matches (people identified as criminals) when the system is employed; the sample are the 1 500 matches; the statistic is the sample proportion $p = \frac{150}{1\,500}$; the parameter is the proportion of correctly identified criminals when the system is employed; and we use the sample proportion as estimator.

(b) $p = 10\% \pm 1.5\%$