MAT120: Lecture 13 Handout

Introduction to Statistics

Last lecture we were doing probability, focusing on binomial experiments and probability distributions. In this lecture we begin the statistics unit by discussing what statistics is, why sampling matters, and how bias can sneak into data. We will also build the two core descriptive tools: central tendency (finding the middle) and variance (measuring spread). Finally, we will apply these ideas to our Haribo data set as a worked example. Next lecture we move into the normal distribution, and after that we will learn hypothesis testing.

1 What is Statistics?

Statistics is the mathematical study of *data*. We may think of data as merely "a set of pieces of information collected from something in the real world". The main task of statistics, therefore, is to study this information and discover the structure that it may have, so that we can use this structure to help inform us on how to make decisions about things in the real world. Statistics is ubiquitous within science and industry: statistics is used in everything from the migration patterns of birds to the fundamental theories of heat, and from the manufacturing processes of microchips to the interwoven fabric of the economy. At the introductory level, one could argue that statistics is one of the most useful and important mathematical tools to learn.

We will now spend the next three lectures introducing some basic concepts in statistics. Today, we start at the very beginning by answering questions about the nature of statistics, and the basic structures that a data set can display. In the next lectures, we will start to see how statistics can be used to make informed *tests* of systems, based on some clever mathematical machinery.

1.1 The workflow of a statistical study

A good statistical study mixes real-world context with mathematical techniques in order to discover new information about a system. One basic format for a statistical study would be the following workflow:

What each step means

- **Collection** involves measuring some system in the real world and collecting data from it.
- **Organization** involves the rearrangement of your data into a useful format.
- **Analysis** involves using informed mathematics to describe insightful features of your data set.
- **Interpretation** involves taking these findings and explaining what they mean within a real-world context.
- **Presentation** involves the concise description of your results in a format that others can easily understand.

Notice that this workflow is an educated mixture of real-world considerations and mathematical tools. This lies directly at the beating heart of statistics: it is *not* merely a mathematical theory, but instead, statistical knowledge inherently involves the real world.

1.2 The Practice of Statistics

Statistics can often feel boring and dry because it is formulaic. In fact, lots of the formulas used in statistics have a deep mathematical theory behind them, but practical statistics rarely uses that

theory directly. Evaluated from the perspective of mathematical sophistication, statistics can often feel boring because it is so simple, and many of the most common formulas use only basic arithmetic.

In fact, statistics is also very interesting for exactly the same reason: even from the most basic mathematical operations, statistics creates a bunch of clever formulas that can be used to describe overall structures in data. Statistics is interesting in that it is such a powerful and useful tool that comes from such humble beginnings. The practical use of statistics cannot be understated: the world is made up of information, and many properties about you and society can be converted into data format and analysed.

1.2.1 Descriptive and Inferential Statistics

In practice, there are two main pillars of statistics. These are: descriptive statistics and inferential statistics.

- **Descriptive statistics** deals with the description of the data that you have collected. We are inclined to think here about visual descriptions such as the graphs and charts we might see in news articles or in scientific publications. However, most of the time we can also describe data using *numbers*, such as averages and variation, and more sophisticated techniques like regression.
- **Inferential statistics** deals with the conclusions that we may reach based upon our data. Often, our statistical studies are constrained by the resources that we have, and the collection of "perfect" data may be too expensive or time-consuming, or even impossible. So, inferential statistics involves the use of clever mathematical tools that can be used to test larger systems based on several small pieces of information. Inferential statistics typically includes techniques such as hypothesis testing and estimation.

1.3 What is Data?

In reality we do not always have access to a data set that would describe a system perfectly. So, we have to settle for the data that we have. Statistics builds tools for figuring out things about the full population without perfectly knowing it.

Our data is only as good as our method of acquiring it. Unwanted effects can be accidentally built into an analysis, and it is our job to identify these effects and minimise them.

1.3.1 Types of data

Data fall into two main categories: quantitative or categorical.

- **Quantitative data** is data that is numerical.
- **Categorical data** is data that is non-numerical.

Quantitative data has a mathematical structure that can be used to help analyse it, whereas categorical data is much more similar to a list of names.

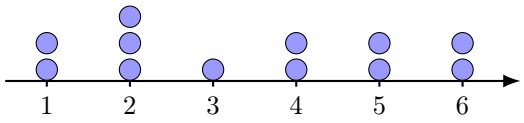
1.3.2 Representing Quantitative Data

We often write numerical data in the form $(x_1, x_2, \ldots, x_n)$. This notation is very similar to the set-theoretic curly brackets { and } we have seen before. However, there is an important difference: when writing data, we will consider the *order of the data* as well. That's why we write the data with round brackets: these brackets tell us to keep track of the order that the data is in.[1]

Example data: $(1, 2, 3, 1, 5, 4, 2, 4, 6, 2, 6, 5)$.

A helpful way to visualise repeated values is with a diagram like the one below. In this diagram, we can imagine our numerical data as small balls that are plotted on the number line. Whenever there are multiple data entries with the same numerical value, we simply stack up another ball on top of the others, so that height displays frequency. For completeness, we will always write the data set itself above its diagram.

Data: $(1, 2, 3, 1, 5, 4, 2, 4, 6, 2, 6, 5)$



Of course, the diagram above doesn't represent the order that the data is in. However, since most formulas only use basic arithmetic operations like addition and multiplication (which are commutative), the order doesn't always matter so much.

2 Sampling

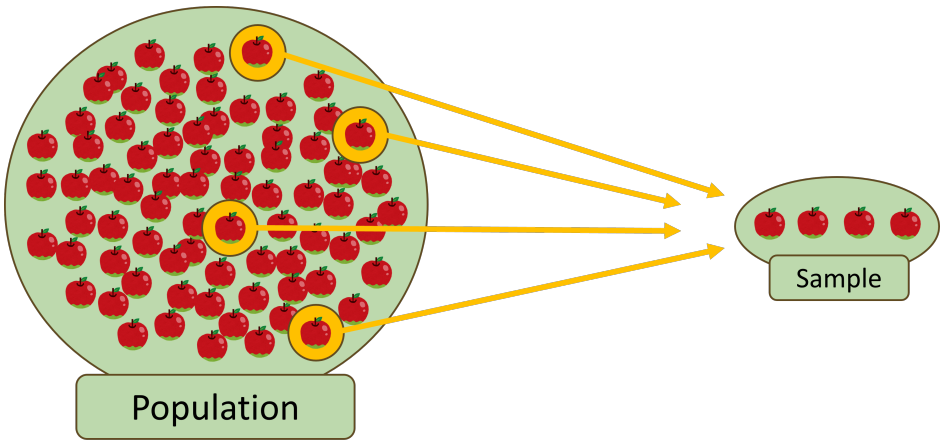
2.1 Core Definitions

Often, the collection of things we wish to study is too large to study directly. So, in practice we take a smaller subset and then study that instead. This practical consideration motivates the following terminology.

Definitions

- **Population:** the set of all the things you want to study.
- **Sample:** a chosen subset of the population.
- **Population parameter:** a fact, measurement, or quantity about the entire population.
- **Sample statistic:** a fact, measurement, or quantity computed from a sample.

[1] Actually, we saw this notation when discussing linear models in Lecture 7. There, we used "ordered pairs" to describe solutions of equations. These were pairs of numbers, written in the form (x, y). In the case of data sets, we are generalising this notation.



The process of obtaining a sample is called *sampling* from the population. Sampling is an extremely important part of statistics, since in reality we never have access to a full population. But, we should be very careful when doing so, because sometimes the way in which we sample our data may accidentally introduce a narrative that is not actually present in the population. In statistics, these errors are called *biases*.

2.2 Worked Example: Sampling in Tokyo

Suppose that Jenny wants to perform a statistical study in order to find out the average salary of everybody that lives in Tokyo. Jenny doesn't work for the municipal government, so she does not have access to census data about the population of Tokyo. Instead, she needs to take a sample. We will now discuss some possible sampling methods for Jenny. However, before we do, we will first make an important point. Consider the following two sampling methods:

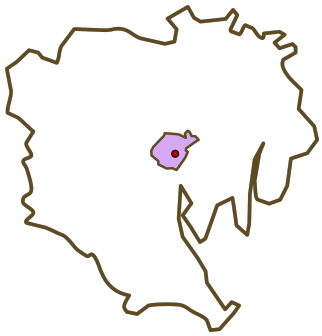
Two sampling methods

1. **Street sampling:** Jenny walks around in her neighbourhood and asks 30 people on the street how much money they make.
2. **Door-to-door sampling:** Jenny knocks on doors in her neighbourhood and asks people how much money they make.

Now it may feel as though these two methods don't actually make a difference. However, we have to keep in mind that there is a *human element* to data collection. For example: some people may feel threatened by a stranger coming to their home uninvited and asking about money. Perhaps, in those circumstances, people might be inclined to lie by undervaluing their wealth, since Jenny could be a thief scanning for targets. In the resulting sample data, perhaps the average salary would be lower than the average from the street sample.

2.3 Possible Sampling Methods

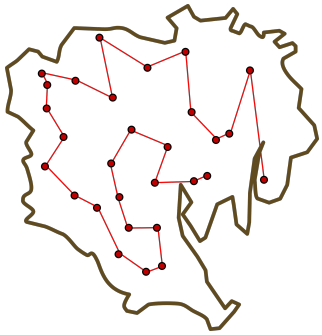
Working with the setting of Tokyo, here are four possible sampling methods that Jenny could employ.



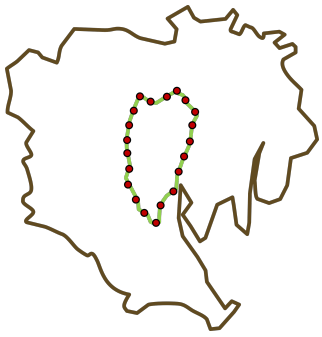
Method 1: Selecting 30 participants from her home in Chiyoda.

Lack of diversity: it is certainly convenient to take a sample of people from her local neighbourhood. However, her numbers may not represent Tokyo overall. Some neighbourhoods are richer or poorer, older or younger, or have different types of jobs. So even if she asks 30 people, she may accidentally measure the character of that neighbourhood rather than Tokyo as a whole. In the case of Chiyoda, her sample is likely to overestimate average salary.

Method 2: Randomly selecting 30 participants from anywhere in Tokyo.



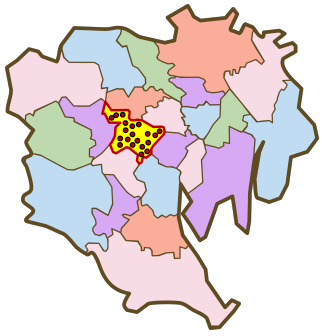
Logistical difficulty: although it may feel unbiased to sample people randomly from across the city, actually doing so would be expensive and time-consuming. Not only would it be difficult to travel to the various neighbourhoods she has randomly selected, she will also run into other potential biases depending on how she selects a representative to ask. Nevertheless, a plan that looks unbiased on paper may fail in practice because of resource constraints.



Method 3: Travelling the Yamanote line, stopping at each of the 30 stations and asking somebody from each.

Lack of accessibility: sampling "one person per station" may include tourists and outsiders who do not live in Tokyo. It can also accidentally introduce a bias towards tourist stations, commuter hubs or business districts, which can skew the sample toward certain kinds of workers. Put differently, this sample may have a bias towards a particular type of Tokyo citizen: those who work or travel around the Yamanote line.

Method 4: Randomly selecting one of the 23 wards of Tokyo, and then randomly selecting 30 participants from within that ward.



Lack of diversity: although it is more convenient than travelling throughout all of Tokyo, and less biased than Jenny's local neighbourhood of Chiyoda, there will *still* be a lack of diversity in this type of sample. Randomness does not guarantee that the diversity of a population is fully represented in small samples. Instead, it only makes the selection rule unbiased.

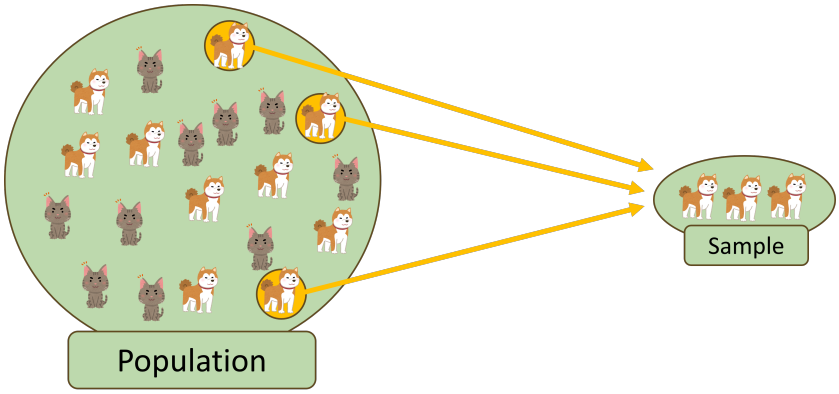
So, what is the best sampling method here? In fact, it depends on what Jenny is trying to do. In practice, there is often no single "best method". Instead, sampling is often a balancing act between resources and error, and sometimes we must compromise. Ideally, these biases ought to be understood by the researcher running the study, and this should be made as explicit as possible when communicating results.

2.4 Simple random sampling

Definition: Simple Random Sample

A simple random sample of size n is a subset of the population selected randomly in such a way that every member of the population has an equal chance of being selected. Equivalently, when sampling without replacement, every subset of size n has the same chance of being selected from the population.

Even if selection is random, it is always possible to get an unlucky sample that does not represent the population honestly. For example, if a population has 10 cats and 10 dogs, and you take a random sample of size 3, there is a chance that you get 3 dogs, therefore underrepresenting the cats in the population. As with the fourth method of the previous section, simple random samples are unbiased in the sense that there is no particular selection rule for any sample. However, that does not guarantee that any chosen sample is an honest representative of the diversity a population may have.



2.5 Examples of Sampling Biases

Three common biases

- **Survivorship bias:** "Architecture in the past was so beautiful, and modern buildings are so ugly." This can happen if we mostly remember and keep the best examples from the past.
- **Selection bias:** "I asked everybody in Kabukicho if Tokyo is noisy and chaotic. They all said yes, therefore Tokyo is noisy and chaotic." Kabukicho is not a random sample of Tokyo.
- **Funding bias (conflict of interest):** if a study is funded by a group that benefits from a particular conclusion, this can bias results.

All of the above are examples of *sampling biases* – the selection of samples that accidentally imply a conclusion that would not otherwise exist in the population. In the example of the survivorship bias, the speaker is drawing that conclusion based on the past buildings that *currently exist*. However, it is probably the case that the ugly buildings were destroyed and replaced with important new projects, whereas the historically-pretty buildings were preserved for their aesthetic value. If you were to magically travel back in time to the 1700s, it would probably not be the case that every building is as beautiful as the ones that still exist to this day.

3 Central Tendency

3.1 The meaning of central tendency

When we collect a lot of data, we often want to summarise it with a single representative number. One way to do this is to create a description of where the "middle" of the data is. This is exactly the intention of central tendency: a *measure of central tendency* is a rule that takes a whole data set and produces one number that represents a kind of "middle".

In fact, there is no single, perfect notion of the "middle of a data set". Instead, there are different, competing notions that become more or less useful in different contexts. For example, if your data contain extreme outliers (a few very large or very small values), some measures of central tendency will move a lot, and others will barely change.

3.2 Summation notation

Before getting to a discussion of central tendency, we will first need to introduce some new notation: summation notation. This is simply a compact way to write a big sum of numbers. We use the symbol Σ as a shorthand notation for a big sum of numbers.

Summation notation

$$\sum_{i=1}^{n} x_i \quad \text{means} \quad x_1 + x_2 + x_3 + \cdots + x_n.$$

The letter i is the *index* (it tells you which term you are adding). The 1 and the n tell you where the counting starts and ends. Here, n is the number of data points in the data set.

3.3 Three common "averages": mean, median, mode

In everyday language, the word "average" is ambiguous and takes many meanings. When working with basic statistics, the three common choices of average are the mean, the median and the mode. These three averages are different ways to describe the middle of a data set, and they answer slightly different questions.

- **Mean:** the "balanced middle" of the data. If you imagine each data point as a weight on a number line, the mean is the balance point. It uses *every* value, so it is sensitive to extreme outliers.

- **Median:** the middle value after sorting. It depends mainly on the *order* of the data, so it does not react strongly to outliers. It is often a good "typical value" when the data are skewed.
- **Mode:** the most frequent value. This is especially useful for *categorical* data (for example, "most common sweet type"). For numerical data, the mode can be informative, or it can be misleading, depending on how the values repeat.

Which one should I use? (rule of thumb)

- If you care about totals and balancing (for example, splitting costs fairly), the **mean** is natural.
- If you want a "typical" value that ignores extreme outliers (for example, salaries), the **median** is often better.
- If you want the *most common* category or value, use the **mode**.

3.4 Median

The median is the value that splits the data in half (half of the data lie at or below it, and half lie at or above it).

Step 1: Order all values from smallest to biggest.

Step 2: If n is odd, pick the middle entry (the $(n+1)/2$-th entry).

Step 3: If n is even, pick the number halfway between the $n/2$-th and $(n/2+1)$-th entries.

For example, for the data set $(1, 1, 3, 4, 5)$, the median is 3. For a data set with an even number of points, such as $(1, 1, 3, 4, 5, 5)$, the median is the value halfway between the two middle points (in this case, 3.5).

3.5 Mean

For a sample $x_1, x_2, \ldots, x_n$, the sample mean is written $\bar{x}$ and defined by

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The mean acts like a balance point: values above the mean pull up, values below the mean pull down, and the mean is the point where things balance. It can be seen as a kind of "geometric centre" of the data set.

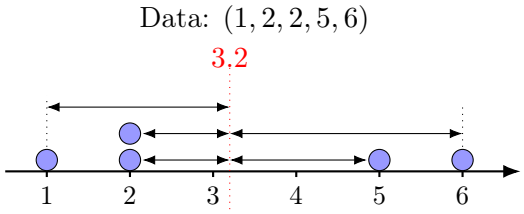
A useful fact about the mean

The sum of all the distances from the data to the mean always adds to zero:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

So, the data "above the mean" and those "below the mean" cancel each other out in total.

To see this fact in action, let's measure the distances in the above diagram:



The three arrows on the left of the mean (which is 3.2) sum to $1.2 + 1.2 + 2.2 = 4.6$, and the two arrows on the right have lengths that also sum to $1.8 + 2.8 = 4.6$.

3.6 Mode

The mode is the most frequent value. If there is a *unique* most frequent value, that value is the mode. If several values tie for most frequent, we say the data are *multi-modal*. Some textbooks say that there is simply *no* mode in this case.

Exercise

Consider the data set $(1, 1, 2, 3, 4, 5)$.

(a) What is the mean?

(b) What is the median?

(c) What is the mode?

Solution

(a) Sum: $1 + 1 + 2 + 3 + 4 + 5 = 16$ and $n = 6$, so mean $= 16/6 = 8/3 \approx 2.67$.

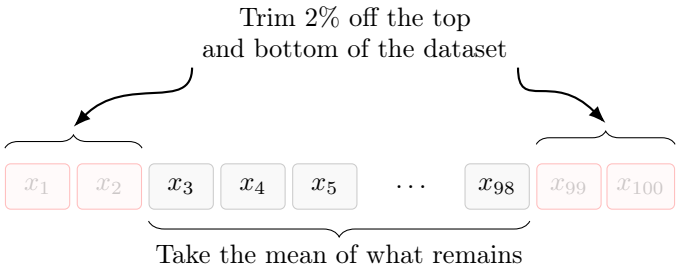
(b) The ordered list is already $(1, 1, 2, 3, 4, 5)$. Since $n = 6$ is even, the median is halfway between the 3rd and 4th entries: $(2 + 3)/2 = 2.5$.

(c) The value 1 appears twice and the others appear once, so the (unique) mode is 1.

3.7 Trimmed means

We can create an "x% trimmed mean" by cutting off the top and bottom x% of a data set, and then taking the mean of what remains. The intent here is to keep the geometric "balance" idea of the mean, whilst also reducing the influence of extreme outliers.

For example, if we were to have an ordered data set of size 100, a 2% trimmed mean would remove a pair of numbers from the two ends of the data, and then take the mean using the 96 numbers that are left. This is roughly pictured below:

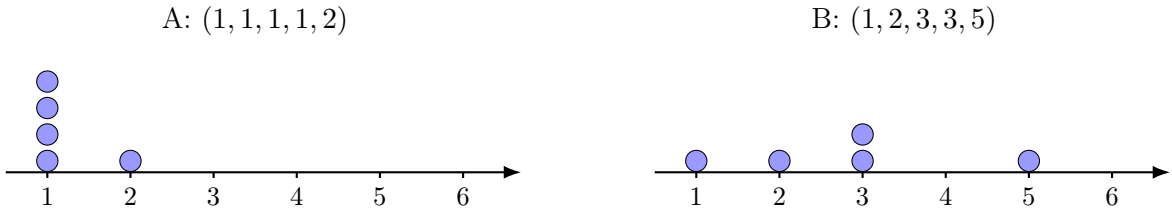


Generally speaking, trimmed means are quite useful when the data has a human component to it. For example, in sports judging, a panel of judges often use a numerical ranking system to compare athletes. Usually, an athlete's score will be calculated using a trimmed mean in which the most favourable and lowest scores are removed first. This helps the results to be more objective, and it reduces sensitivity to the preferences of individual judges who may be prone to error.

4 Variation

Central tendency roughly tells us where the data are "located", but it does not tell us how spread out it is. It is easy to construct two data sets that have the same general middle, yet wildly different visual appearance. This motivates the next idea: *variation* (also called *spread*).

For example, compare:



Intuitively, data set B is more spread out (its values wander further away from the middle).

4.1 Range

The range is the distance between the biggest and smallest value:

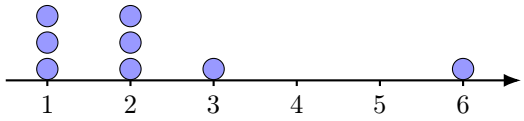
$$\text{range} = \max(\text{data}) - \min(\text{data}).$$

It is a quick first description of spread. Small range suggests less spread, and a large range suggests more spread.

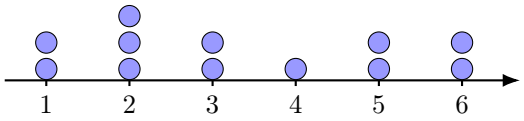
4.1.1 A problem with the range

The range only uses two numbers (the minimum and maximum data points), so in a sense the formula is blind to the majority of data in the data set. The range is also very sensitive to outliers, because one extreme data point can make the range very large. Consider the following two data sets:

A: $(1, 1, 1, 2, 2, 2, 3, 6)$



B: $(1, 1, 2, 2, 2, 3, 3, 4, 5, 5, 6, 6)$



Notice that the data set A has the value 6 as an outlier, whereas B does not. According to the range alone, these two data sets have the same amount of spread. However, we can visually see that the majority of the data points in A are clustered together, and in B this is not the case. So, at least in *some sense*, B should have a larger spread than A.

In order to capture the spread of data more carefully, we need a statistic that uses *all* the data points in its formula, not just the minimum and maximum. This naturally leads us to variance.

4.2 Variance

Variance is a more inclusive and informative measure of spread than the range. It measures spread by looking at average squared distances from the mean. The data will be more spread out whenever those squared distances tend to be larger.

A useful way to think is:

> If the data points tend to sit close to the mean, the variance is small. If the data points tend to sit far away from the mean, then the variance is large.

4.2.1 The core geometric idea

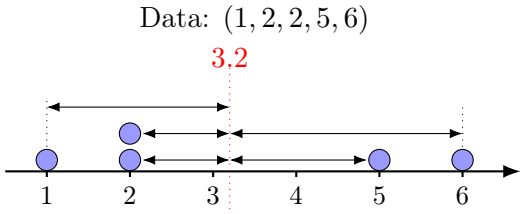
For each data point x, look at how far it is from the mean $\bar{x}$, which is simply the difference $(x - \bar{x})$. If we tried to average the raw deviations $(x - \bar{x})$, positives and negatives would cancel out, so we would get something equal to 0. To avoid this unwanted cancellation, we square the deviations.

So the basic recipe is:

Step 1: Start with the mean $\bar{x}$.
Step 2: Compute each deviation $(x_i - \bar{x})$.
Step 3: Square each deviation $(x_i - \bar{x})^2$ (so everything is positive and large deviations count more).
Step 4: Average the squared deviations.

To reiterate: Step 3 above is quite important: generally the deviation between points and the mean will be either positive or negative depending on whether the data point is to the right or left of the mean:

To see this fact in action, let's measure the distances in the above diagram:



As we saw previously, the sum of the arrows on the left of the mean will cancel out the sum of the arrows on the right of the mean, since they have the same value with opposite sign. But, squaring these values will remove this effect.

4.2.2 Formula for sample variance

For a sample $x_1, x_2, \ldots, x_n$ with sample mean $\bar{x}$, the *sample variance* is written s^2 and defined by

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

The denominator is $n-1$ (not n) because we are typically using the sample to estimate a population spread. In this course, you mainly need to know the formula and how to interpret it: bigger s^2 means more spread around the mean.

In practice, you will usually compute variance using calculators or computer software, but it is important to understand what the formula is measuring. Nonetheless, just so you can get a feel for how the computation works, we will briefly review the process.

Example. Consider the data set $(1, 1, 2, 3)$. In order to compute the variance, we first determine the mean:

$$\bar{x} = \frac{1+1+2+3}{4} = \frac{7}{4} = 1.75.$$

Now, we must compute the distances of each data point to the mean, and square the answers. It's easiest to display this as a table:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	-0.75	0.5625
1	-0.75	0.5625
2	0.25	0.0625
3	1.25	1.5625

Finally, to compute the variance we sum up the answers in the final row and divide by $n-1=3$:

$$s^2 = \frac{0.5625 + 0.5625 + 0.0625 + 1.5625}{3} = \frac{2.75}{3} \approx 0.917.$$

4.3 Standard deviation

The *standard deviation* is the square root of the variance:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

We take a square root because variance is measured in "squared units" (for example, if the data are in yen, variance is in yen^2). Standard deviation returns to the original units, so it is easier to interpret as a typical distance from the mean. This observation will become extremely important in the next lecture.

4.4 Important note on notation (sample versus population)

As mentioned previously, a central question in inferential statistics is the determination of population parameters from sample statistics. For this reason, it is quite important to keep a very clear distinction between sample statistics and population parameters. Part of this good practice is the use of different notation for the mathematics associated to samples and populations. For technical reasons, the formulas associated to a sample are slightly different from those used for a population, and importantly we use different symbols to talk about very similar concepts. The following is a summary.

Sample statistics vs population parameters

- **Sample statistics:** $\bar{x}$ (mean), s^2 (variance), s (standard deviation).
- **Population parameters:** μ (mean), σ^2 (variance), σ (standard deviation).

4.5 The Coefficient of variation

Sometimes we want to compare spread across data sets with different scales. For example, a standard deviation of 2 is large if the mean is 4, but it is small if the mean is 200. So, we can help to "normalise" the relative size of standard deviations by dividing them by the mean. Taking the ratio of standard deviation to mean gives a unitless number that is a measure of spread. This unitless number is called the *coefficient of variation*, defined as follows.

$$\text{Population: CV} = \frac{\sigma}{\mu} \qquad \text{and} \qquad \text{Sample: CV} = \frac{s}{\bar{x}}.$$

A larger CV means "more variation compared to the typical size of the values".

4.6 Central tendency and variance as a "map"

Central tendency tells you *where* the data live, and variance (or standard deviation) tells you *how widely* they are spread out. So the mean and the standard deviation are two ways to map out data:

Mean = where the centre is, Standard deviation = how far from the centre data typically are.

In the next lecture (normal distributions), this "map" idea becomes even more concrete, because bell curves let us translate "how many standard deviations away" into approximate percentages.

5 A Worked Example

A good statistical study mixes real-world context with mathematical techniques in order to discover new information about a system. We will now apply the workflow of Section 1 to the data we collected in class.

5.1 Haribo data (the raw sample)

In class we collected data from 19 medium-sized bags of Haribo sweets. For each bag we counted how many sweets of each type occurred. Here is the raw data:

Sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Rings	3	7	4	2	6	6	7	3	7	4	2	6	6	4	7	5	5	2	6
Hearts	5	4	6	4	5	2	5	7	1	3	6	3	4	6	4	3	2	4	5
Bottles	4	1	4	2	4	2	3	3	4	6	4	0	2	4	4	4	8	1	5
Eggs	4	5	3	6	0	6	1	2	6	2	4	4	1	2	1	4	3	7	1
Gummy bears	7	8	6	12	11	10	10	9	7	11	6	14	15	8	10	9	6	12	7
Total per bag	23	25	23	26	26	26	26	24	25	26	22	27	28	24	26	25	24	26	24

5.2 Totals and average counts

Total number of sweets counted across all 19 bags is 476. Average total sweets per bag is

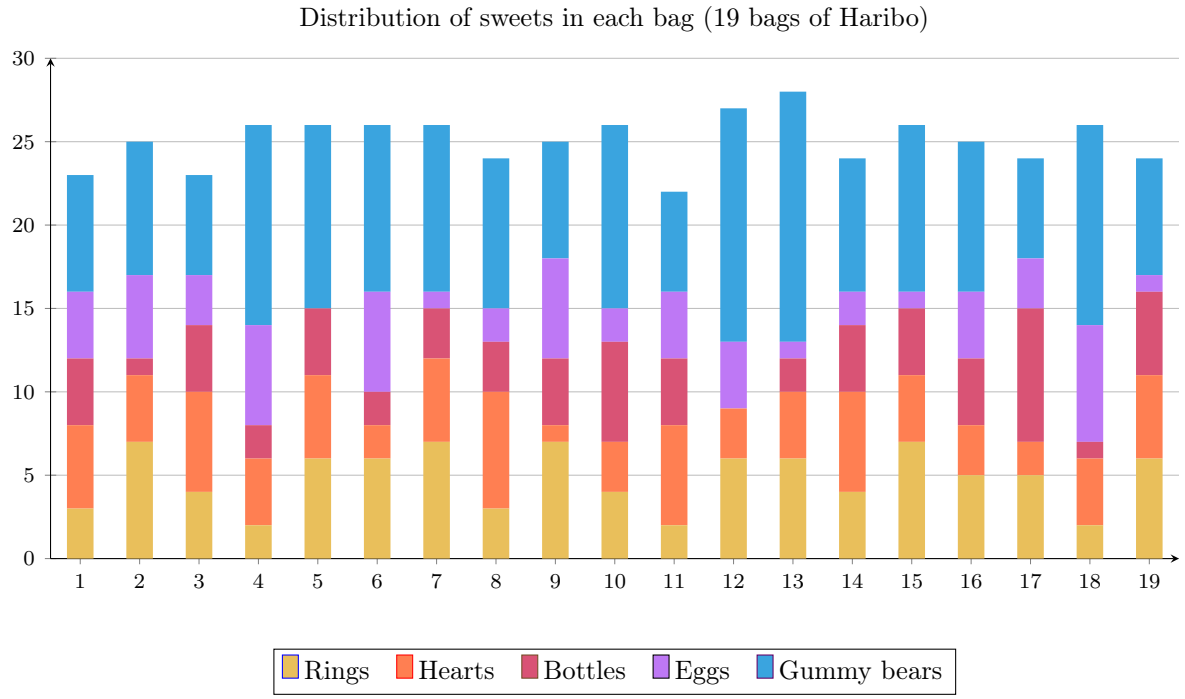
$$\frac{476}{19} \approx 25.05.$$

Totals by type, and averages per bag can be calculated with a similar formula:

Type	**Total**	**Average per bag**	**Average (approx.)**
Rings	92	92/19	4.84
Hearts	79	79/19	4.16
Bottles	65	65/19	3.42
Eggs	62	62/19	3.26
Gummy bears	178	178/19	9.36

5.3 Distribution of sweets

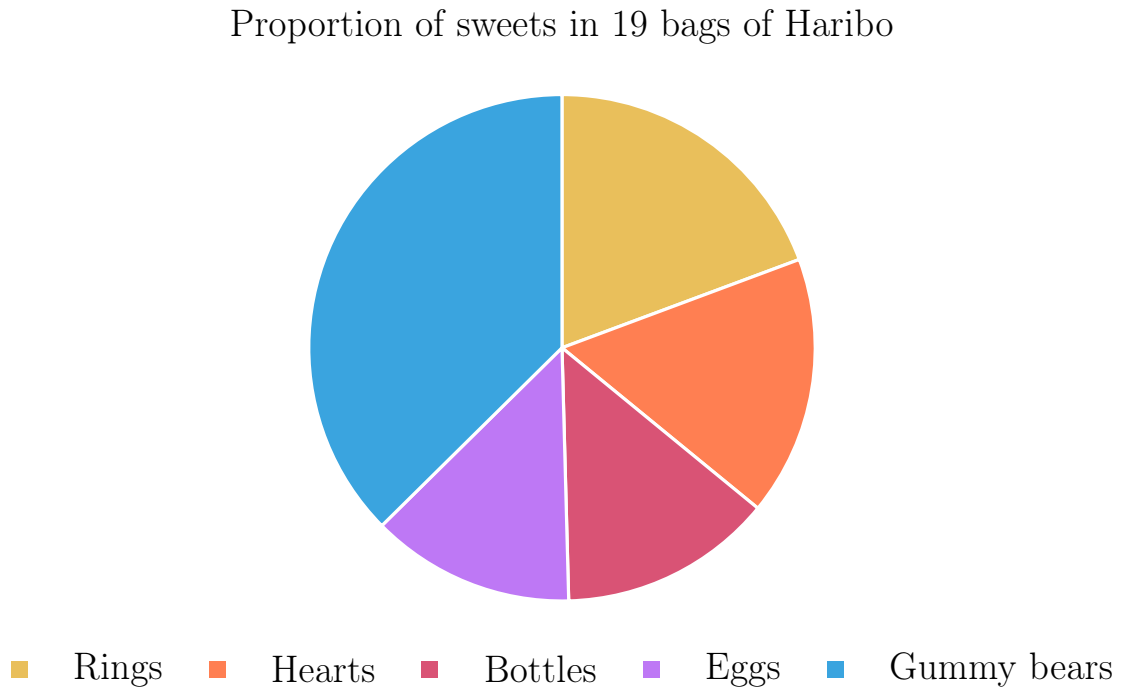
Using the raw data, we can plot the distribution of sweets in each bag in a large graph:

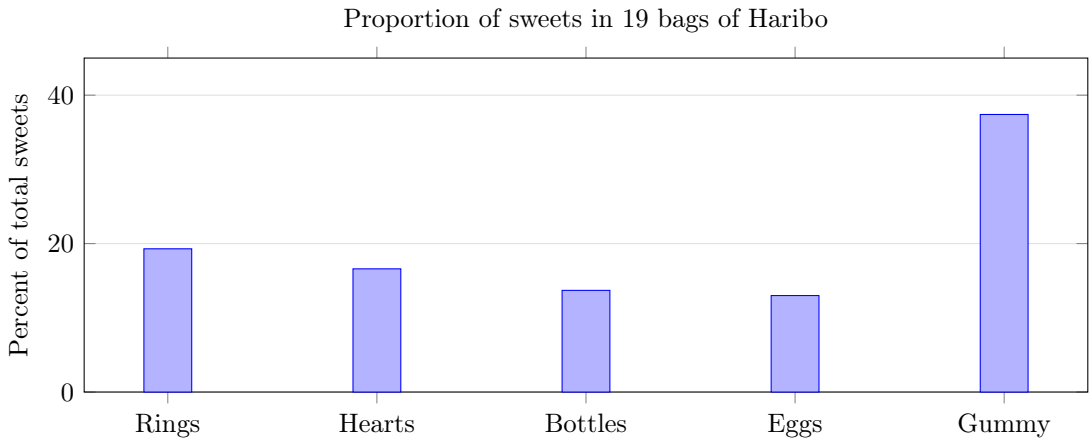


It appears as though the bag sizes are quite constant at around 25. However, the distribution of sweets in each bag seems to vary quite a lot.

5.4 Proportions of sweets (relative frequencies)

Type	Proportion	Percent (approx.)
Rings	92/476	19.3%
Hearts	79/476	16.6%
Bottles	65/476	13.7%
Eggs	62/476	13.0%
Gummy bears	178/476	37.4%





5.5 Spread and variation in the Haribo counts

Category	Mean	Variance	Std. dev.	CV
Hearts	4.16	2.47	1.57	0.38
Eggs	3.26	4.32	2.08	0.64
Bottles	3.42	3.48	1.87	0.55
Gummy bears	9.36	7.13	2.67	0.29
Rings	4.84	3.25	1.80	0.37
Total sweets per bag	25.05	2.27	1.51	0.06

The total sweets per bag have a small variation, but the coefficient of variation for each type of sweet

is quite high in comparison. This supports the earlier observation: bag size seems fairly uniform, but the internal composition varies.

5.6 How to calculate these quickly in Excel

Concept	Excel command
Adding cells	`=SUM(...)`
Sample mean	`=AVERAGE(...)`
Sample standard deviation	`=STDEV.S(...)`
Median	`=MEDIAN(...)`
Mode	`=MODE.SNGL(...)`
Coefficient of variation (CV)	`=STDEV.S(...)/AVERAGE(...)`

5.7 Interpretation of results (three claims)

Our study of a sample of 19 Haribo bags suggests three main findings:

1. In this sample, the proportions are not close to 20% each. In particular, gummy bears appear more common than the other types.
2. Bag sizes are typically uniform, at around 25 sweets per bag.
3. The mix of sweets varies noticeably from bag to bag.

5.8 Possible sampling biases

Any good statistical study ought to at least *consider* the possible biases that may have occurred as part of the study. In our case, a reasonable population to keep in mind is all medium-sized Haribo bags sold locally (for example, in Japan) during the period of our sampling, and we took a sample of 19 bags. Below are some possible biases that we should be aware of.

- Our 19 bags may have come from the same production batch.
- We do not know if everybody recorded their data correctly.
- Data entry mistakes are possible when typing values into Excel.
- We do not know if the local manufacturing plant that made these sweets is the same as other factories.
- We do not know if there was a local manufacturing error in this particular batch of Haribo.
- We do not know if the manufacturing standards of Haribo vary from country to country, and even if we did, we did not record which country manufactured this particular sample.

5.9 Summary of the study in workflow language

- **Collection:** students' data written on paper.
- **Organization:** Excel.
- **Analysis:** average counts, variation, and charts.

- **Interpretation:** what the sample suggests about typical bag size and sweet composition.
- **Presentation:** charts and summary statements.

5.10 What's next

We will now develop techniques that we can use to test these claims and figure out information about the population of all Haribo bags.