coggle
made for free at coggle.it

Statistics

how knowing your sample helps you accurately predict what the population will be like and coming up with a way of saying how reliable your predictions are

11 estimating populations and samples making predictions

estimate the mean of the population

what is it

The mean of the sample is called a point estimator for the population mean.

A point estimator of a population parameter is some function or calculation that can be used to estimate the value of the population parameter.

As an example, the point estimator of the population mean is the mean of the sample, so we can use the sample mean to estimate the population mean.

How

estimate The mean of the sample

$$\bar{x} = \frac{\Sigma x}{n}$$

μ - estimate the mean of the population, a point estimator for μ
μ the population mean
$\bar{x}$ bar - the mean of the sample
To represent the sample mean, we use $\bar{x}$ bar (pronounced "x bar")
n is the sample size.
x represents the values in the sample

$$\hat{\mu} = \bar{x}$$

When is it used?

If you want a good estimate for the true value of the population mean, you can use the mean of the sample.

estimate the population variance

What

Rather than take the variance of all the data in the sample to estimate the population variance, if the size of the sample is n, we can estimate the population variance

when

If you need to estimate the variance of a population using sample data

How

Lucy p: The variance of the data in the sample may **not** be the best way of estimating the population variance.

Because
The sample variance tends to be slightly less than the population variance
==> Dividing a set of numbers by n - 1 gives a higher result than dividing by n

σ^2 is for: estimate for the population variance

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

ste each item in the sample, subtract the sample mean, square the result, then add the lot together.

Divide by the number in the sample minus 1.

The formula for the population variance point estimator is usually written as

$$\hat{\sigma}^2 = s^2$$

Predicting population proportion

when

we use the sample data to predict the proportion of successes in the population

$$\hat{p} = p_s$$

how

p^* - Point estimate for the proportion of successes in the population
p_s- Proportion of successes in the sample

p_s = number of successes / number in sample

you derive the probability of getting a success in exactly the same way as you derive the proportion of successes:

p = probability = proportion

The sampling distribution of proportions

what

The sampling distribution of proportions is what you get if you consider all possible samples of size n taken from the same population and form a distribution out of their proportions. We use Ps to represent the sample proportion random variable.

How

the probability distribution for the sample proportion

1. Look at all possible samples the same size as the one we're investigating
For instance, for a sample of size n, we need to consider all possible samples of size n. There are 100 gumballs in the box, so in this case n is 100.
2. Look at the distribution formed by all the samples, and find the expectation and variance for the proportion. Every sample is different, and the number of gumballs of red gumballs in each box of gumballs will probably vary.
3. Once we know how the proportion is distributed, use it to find probabilities.
Knowing how the proportion of successes in a sample is distributed, we can use it to find probabilities for the value of a proportion in a random samples—in this case, a jumbo box of gumballs.

$E[P_s] = p$

$Var[P_s] = \frac{pq}{n}$

p is the population proportion

the expectation & variance of Ps

$P_s \sim N(p, \frac{pq}{n})$

the standard error of proportion, as it tells you what the error for the proportion is likely to be in the sample.

$Var[P_s]$

It allows you to work out probabilities for the proportion of successes taken from a known population. It gives you an idea of how reliable a sample to be like

when

$$P_s \sim N\left(p, \frac{pq}{n}\right)$$

Ps follows a normal distribution
n > 30
continuity correction required

As Ps follows a normal distribution for n > 30, this means that we can use the normal distribution to solve our problem.

The number of successes in each of the samples is discrete, and as it's used to calculate proportion, you need to apply a continuity correction when you use the normal distribution to find probabilities.

apply a continuity correction of ±1/2n

Continuity correction = 1/2n

$\frac{1}{2n}$

$\pm\frac{1}{2n}$

probabilities for the sample mean

This is a slightly different problem from last time. We're told what the population mean and variance are for the packets of gumballs, but we're not told anything about how they're spread out probabilities for. Instead of working out probabilities for the sample proportion, this time we need to work out probabilities for the sample mean.

what is it? & when

Before we can use the probabilities for the sample mean, we need to figure out the probability distribution.

1. Look at all possible samples the same size as the one we're investigating
If we have a sample of size n, we need to consider all possible samples of size n. There are 30 packets of gumballs, so in this case n is 30.
2.Look at the distribution formed by all the samples, and find the expectation and variance for the sample mean. Every sample is different, and the number of gumballs in each packet varies.
3. Once we know how the sample mean is distributed, use it to find probabilities.
If we know how the means of all possible samples are distributed, we can use it to find probabilities for the mean in a random sample, in this case the, packets of gumballs.

How

Each Xi is an **independent observation** of X, which means that they follow the same distribution

$$\text{Sample of } X \begin{cases} X_1 \sim N(\mu, \sigma^2) \\ X_2 \sim N(\mu, \sigma^2) \\ \dots \\ X_n \sim N(\mu, \sigma^2) \end{cases}$$

each Xi has an expectation of μ and variance σ2

The sampling distribution of the mean.

X bar: the mean of the sample, the mean number of gumballs in the packets.

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

We can form a distribution out of all the sample means from all possible samples. This is called the sampling distribution of means, or the distribution of $\bar{X}$.

what

The sampling distribution of the mean is also a normal distribution for the mean of a sample.

when

the average of all the possible sample means of size n is the mean of the population they're taken from

Ex: If the population mean is 10 gumballs per packet, you can expect the sample mean to be 10 gumballs per packet too.

the expectation for X bar:

$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

the the variance of X bar

$$\text{Standard error of the mean} = \frac{\sigma}{\sqrt{n}}$$

The standard error of the mean gets smaller as n gets larger. This means that the more items there are in your sample, the more reliable your sample mean becomes as an estimate of the population mean.

The standard deviation tells you how far from the mean a value is likely to be. so it's known as the standard error of the mean.

The binomial distribution

Imagine you have a population represented by X ~ B(n, p) where n is greater than 30. As we've seen before, μ = np, and σ2 = npq

The central limit theorem tells us that in this situation, X bar ~ N(np, npq)

The central limit theorem tells us that in this situation, X bar ~ N(np, pq)

Now suppose you have a population that follows a Poisson distribution of X ~ Po(λ). again where n is greater than 30. For the Poisson distribution μ = λ0 = λ

The Poisson distribution

X bar~ N(λ, λ/n)

Now is it fair distributed

Since X follows a normal distribution, this means that you can use standard normal probability tables to look up probabilities. In other words, you can use the normal distribution as an approximation for you would for any other normal distribution.

If X ~ N(μ, σ2), then X bar ~ N(μ, σ2/n)

As n gets larger, X gets closer and closer to a normal distribution.

9 using the normal distribution ii Beyond Normal

find the probability distribution for the combined weights of the bride and groom

What is it used for?

The geometric distribution, on the other hand, never looks like the normal, so the normal can never effectively approximate it.

The variance of X + Y is greater than the variance of X and also greater than the variance of Y, which means that the curve of X + Y is more elongated than either.

$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$

X and Y are independent
$X \sim N(\mu_X, \sigma_X^2)$
$Y \sim N(\mu_Y, \sigma_Y^2)$

$X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$

If you need to find probabilities involving the **difference** between two variables, you'll need to use $X - Y$.

the shape of the distribution gets flatter so that the **total area under the curve is still 1**.

How

Linear transforms affect the underlying values in your probability distribution.
an example, if you have a length of rope of a particular length, then applying a linear transform affects the length of the rope.

$aX + b \sim N(a\mu + b, a^2\sigma^2)$

$Var(aX + b) = a^2 \sigma^2 = a^2 \sigma^2$

$E(aX + b) = a\mu + b$.

a, b are constant values
$a\mu + b$
$a^2 \sigma^2$

Ex: The distribution of Ri is actually a **linear transform** of X.

The #A probability distribution describes adults whose weights have been multiplied by 4. The weight is changed, not the number of adults.

$X1 + X2 + \dots + Xn \sim N(n\mu, n\sigma^2)$

$E(X1 + X2 + \dots Xn) = nE(X)$

$Var(X1 + X2 + \dots + Xn) = nVar(X)$

X1 + X2 + … + Xn
X1, X2, … Xn are independent observations of X
$X \sim N(\mu, \sigma 2)$

Independent observations have to do with the quantity of things you observe. For example, if you have 10 independent observations of a piece of rope, then you're talking about 4 pieces of rope.

when np and nq are both greater than 5.

Under certain circumstances, the shape of the binomial distribution starts to look like the normal distribution.

Normal approximation of X.
$X \sim N(\mu, \sigma)$
$np > 5$ and $nq > 5$
Continuity correction required

continuity correction

m is the number of values, **p** is the probability of success, and **q is 1 - p.**

$X \sim N(np, npq)$

$\mu = np$ and $\sigma^2 = npq$

The binomial is discrete, but the normal is continuous
==> Continuity correction required

As we're approximating a discrete probability distribution with a continuous one, we have to apply a continuity correction

Normal approximation of X
$X \sim Po(\lambda)$
$\lambda > 15$
Continuity correction required

Situations where you're dealing with things breaking down at a **mean rate** follow a Poisson distribution, being a parameter of the mean. If X represents the number of breakages in a week, then $X \sim Po(4.5)$

make sure that you choose your range so that it includes the discrete value a.

On a continuous scale, the discrete value a goes up to (a + 0.5)

using the normal distribution to find P(X = a), you actually need to calculate **P(X < a + 0.5)**

P(X < a - 0.5) P(X < a)

P(X > b - 0.5) P(X ≥ b)

P(X > b - 0.5)

P(X > b)

P(a - 0.5 < X < b + 0.5) P(a ≤ X ≤ b)

P(a < X < b)

As we're approximating a discrete probability distribution with a continuous one, we have to apply a continuity correction

A continuity correction is the small adjustment that needs to be made when you **translate discrete values onto a continuous scale.**

continuity correction

Muc dich

how continuous probability distributions work

introduce the normal distribution.

8 using the normal distribution Being Normal

continuous probability distributions

what is it?

For discrete probability distributions, we look at the probability of getting a particular value. For continuous probability distributions, we look at the probability of getting **a particular range.**

Ex: measure pieces of string that are between 10 inches and 11 inches long

Continuous data is like a smooth, continuous path

Continuous data can take countless values.

a lot of real-world problems involve continuous data

what is it used for?

find probabilities for continuous data - which happen in real life

How

probabilities are given by area

use probability density functions

f(x) is a function that you can use to find the probabilities of a **continuous** variable across a **range** of values.

It tells us what the **shape** of the probability distribution is.

start off by sketching the probability density function, -> need to find f(x)

The total area under the line must be equal to 1

the normal distribution.

Chú ý: No matter how far you go out on the graph, the probability density never equals 0.
The probability density gets closer and closer to 0, but never quite reaches it.

What it is used for?

Another name for the normal distribution is the **Gaussian distribution**

$X \sim N(\mu, \sigma 2)$.

The variable X follows a normal distribution, and has a mean of μ and a variance of $\sigma 2$.

Both the mean and median are at the center and have the highest probability density.

The normal distribution is in the shape of a **bell curve.**

The curve is **symmetrical** at

μ tells you where the **center** of the curve is, and it gives you the **spread**

The **larger** σ2 becomes, the **wider** and **flatter** the curve becomes.

What it is

How to calculate

calculating the area under the curve of the distribution

Step 1: determine the distribution of the data

the mean and standard deviation

$X \sim N(\mu, \sigma 2)$.

$Z \sim N(0, 1)$ where
$Z = \frac{X - \mu}{\sigma}$

Step 2: Standardize to N(0, 1).

μ: The mean of X.

z: Standard core

σ: The standard deviation of X.

X is the variable we're trying to find probabilities for (Value)

how do you use probability tables to find P(Z < z)

B1: calculating z to 2 decimal places.

B2: ma cửu z tiến lăng, va the second decimal place for z... **pages 688-689**

$P(Z > z) = 1 - P(Z < z)$

Finding P(Z > z)

$P(a < Z < b) = P(Z < b) - P(Z < a)$

Finding P(a < Z < b)

Three steps

Step 3: Look up the probability in your handy table

How to choose your sample

systematic sampling

With systematic sampling, you list the population in some sort of order, and then choose a number, k, and sample every kth one.

disadvantage: if k is a multiple of a hidden pattern in the population, your sample will be biased

Simple random sampling

Sampling with replacement

when you've selected each unit and recorded relevant information about it, you put it back into the population

Sampling without replacement

the sampling unit isn't replaced back into the population.

Drawing lots

just like pulling names out of a hat.

How to choose

Random number generators or random number tables

have a large sampling frame -> you give each member of the sampling frame a number, generate a set of n random numbers, and then choose members of the population whose numbers correspond to the random numbers that were generated.

make sure that each number has an equal chance of occuring.

stratified sampling

the population is split into similar groups that share similar characteristics.

each individual group is called a **stratum.**

These characteristics or groups are called **strata.**

1. As an example, we could split up the gumballs into the different colors, yellow, red, and pink, so that each color forms a different stratum.
Once you've done that, you perform simple random sampling on each stratum to ensure that each group is represented in your overall sample.

To do step 2, look at the proportions of each stratum within the overall population and take a proportionate number from each.

Vd: As an example, if 50% of the gumballs that Mighty Gumball produce are red, half of your sample should consist of red gumballs.

cluster sampling

advantage: don't need a sampling frame of the whole population in order to achieve it

Cluster sampling is useful if the population has a number of similar groups or clusters.

you draw a simple random sample of clusters, and then survey everything within each of these clusters

The problem with cluster sampling is that it might not be entirely random.

As an example, it's likely that all of the gumballs in a package have been produced by the same factory. If there are differences between the factories, you may not pick these up.

12 using statistical sampling Taking Samples

When sampling goes wrong

using a misleading sample can actually lead you to draw the wrong conclusions about your population.

wrong conclusions about population parameters, such as the mean or standard deviation

How sampling works

The key to creating a good sample is to choose one that is as close a match to your population as possible.

If your sample is representative, this means it has similar characteristics to the population.

you can use your sample to predict what characteristics the population will have.

the more gumballs you try, the clearer the shape is.

How to design a sample

how you can effectively gather data in the real world

1. Define your target population
2. Define your sampling units
what sort of object you're going to sample
3. Define your sampling frame
you need a list of all the sampling units within your target population, preferably with each sampling unit either named or numbered

Unbiased samples

If we know the shape of the sample distribution, we can use it to predict that of the population.

An unbiased sample is **representative** of the target population

it has similar characteristics to the population -> **use these to make inferences about the population itself.**

If you don't design your sample well, your sample may not be unbiased.
A poorly designed sample can introduce **bias**

Biased samples

A biased sample is **not** representative of the target population

Sources of bias

Poorly designed questions in a questionnaire. Design your questions so that they're neutral and everyone can answer them.

A sampling frame where items have been left off, such that not everything in the target population is included. If it's not in your sampling frame, it won't be in your sample.

An incorrect sampling unit. Instead of individual gumballs, maybe the sampling unit should have been boxes of gumballs instead.

Individual sampling units you chose for your sample weren't willing to take part. For instance, you might have chosen to send out a questionnaire that not everybody responds to.

Samples that aren't random. As an example, if you're conducting a survey in person, you might unconsciously choose anyone that looks too busy to stop, or too aggressive. This means that you exclude aggressive or busy-looking people from your survey.

Definition

Stratified sampling: the population is split into similar groups that share similar characteristics.

These characteristics or groups are called **strata.**

each individual group is called a **stratum.**

Sampling with replacement: means that the sampling unit isn't replaced back into the population.

Sampling with replacement: when you've selected each unit and recorded relevant information about it, you put it back into the population

Bias is a sort of favoritism that you can unwittingly (or maybe knowingly) introduce into your sample

target population: the group that you're researching and want to collect results for.

A statistical sample is a selection of items taken from a population

A statistical population: the entire group of things that you're trying to measure, study, or analyze

A census is a study or survey involving the entire population

A sample survey: A study or survey involving a selection of the population

the sampling frame: a list of all the sampling units

Each packet would form a **cluster.**

Imagine you have a population of 5 sampling units, and you need to pick a sample of n sampling units. **Simple random sampling** is where you choose your sample by using a random process, and all possible samples of size n are equally likely to be selected.

7 geometric, binomial, and poisson distributions Keeping Things Discrete

Tom tắt

show you some **special probability distributions** to calculate probabilities, expectations, and variances in record time.

the geometric distribution

the mode of any geometric distribution is **always 01**.

If you're dealing with a situation that can't be modelled by the geometric distribution, **don't use these shortcuts.**

the expectation and variance

$E(X) = 1 / p$

$Var(X) = q / p^2$

When do I use it?

you're interested in how many trials are needed to get the first successful outcome

running independent trials

There can be either a success or failure for each trial, and the probability of success is the same for each trial

How do I calculate probabilities?

$P(X = r) = p^r (1 - p)$

The probability you'll need r trials or less to get your first success

$P(X <= r) = p (1 - p^r) / (1 - p)$

The probability of the first success being on the rth trial

For the number of trials needed for a success to be greater than r there must have been r failures.

$P(X > r) = q^r$
The probability you'll need more than r trials to get your first success

We don't need p in this formula because we don't need to know exactly which trial was successful, just that there must be more than r trials.

X ~ Geo(p)

X is the number of trials needed in order to get the first successful outcome

X follows a geometric distribution where the probability of success is p

p is the probability of success in a trial

the probability of failure is **q = 1 - p**

Mode

$P(X = 0) = P(X > 0) = 1$
$P(X ≤ 0) = 1 - P(X > 0)$
$P(X = 0) = 1 - p^1$

the binomial distribution

$X \sim B(n, p)$

p is the probability of a successful outcome in each trial, and **n** is the number of trials

X is the number of successful outcomes out of n trials

the expectation and variance

$E(X) = np$

$Var(X) = npq$

When do I use it?

There can be either a success or failure for each trial, and the probability of success is the same for each trial.

There are a finite number of trials.

You're running a series of independent trials.

How do I calculate probabilities?

$P(X = r) = nCr \times p^r \times q^{n-r}$ - 0

$nCr = n! / r! (n - r)!$

Mode

If p is 0.5 and n is even, the mode is np. If p is 0.5 and n is odd it has two modes, the two values either side of np. For other values of p and q, finding the mode is a matter of trial and error. It's generally fairly close to np.

the Poisson distribution

The trouble with this sort of problem is that while we know the mean number of popcorn machine malfunctions per week, the actual number of breakdowns varies each week. On the whole we can expect 3 or 4 malfunctions per week, but it's by no means certain. There could be 5 malfunctions one week. We need to find the probability that the popcorn machine won't break down next week. Sound difficult? Don't worry, there's a probability distribution that can help us in this situation. It's called the Poisson distribution.

What is it?

X follows a Poisson distribution with a mean of occurrences per interval or rate **X ~ Po(λ)**

X is the number of occurrences in a particular interval

λ is the rate of occurrences

What does the Poisson distribution look like?

If λ is **small**, then the distribution is skewed to the **right**, but it **becomes more symmetrical as λ gets larger**.

If λ is an **integer**, then there are two modes, **λ and λ - 1.**
If λ is **not an integer**, then the mode is λ.

What connection does it have to the binomial distribution?

If X ~ B(n, p), where n is large and p is small **(n > 50 & p < 0.1)**, then it can be approximated using **X ~ Po(np)**

the expectation and variance?

$E(X) = Var(X) = \lambda$

When do I use it?

Individual events occur at **random** and **independently** in a **given interval**. This can be in terms of time or space—for example, during a week, or per mile.

You know **the mean number** of occurrences in the interval or the rate of occurrences, and **it's finite**.

The mean number of occurrences, **letter λ (lambda).**

Poisson distribution doesn't involve a series of trials like other probability distributions

$P(X = r) = e^{\lambda} \lambda^r / r!$

probabilities

e = 2.718

How do I combine independent random variables?

If $X \sim Po(\lambda_1)$, $Y \sim Po(\lambda_2)$ and X and Y are independent, then **X + Y ~ Po(λ1 + λ2)**

$P(X + Y) = P(X) \times P(Y)$

$E(X + Y) = E(X) + E(Y)$

X and Y are independent random variables

5 Using discrete probability distributions Manage Your Expectations

Definition

Linear transforms are when a variable X is transformed into aX + b, where a and b are constants. The expectation and variance of aX + b are:
E(aX + b) = aE(X) + b
Var(aX + b) = a^2Var(X)

Probability distributions describe the probability of all possible outcomes of a given variable.

The **variance** and **standard deviation** of a probability distribution measure how the probabilities of particular values are dispersed.

a random variable

is a variable that can takes on a set of values, where each value is associated with a specific probability.

represent it by a capital letter, like X or Y

the probability that the variable X takes a particular value x

P(X = x)

The expectation is the expected average **long-term** outcome. It's represented as either E(X) or μ, and is calculated using $E(X) = \sum P(X=x)$.

Venn diagrams and probability trees are useful if you want to calculate probabilities. With a probability distribution, the probabilities have already been calculated.

Independent observations of X are different instances of X. Each observation has the same probability distribution, but the outcomes can be different.

what is it used for?

to predict long-term outcomes

measure the certainty of these predictions.

How

B1: calculate the probability distribution of a random variable
B2: determine the expected outcome
B3: use variance to measure this spread

Now

X is ~ B ... , ... as an **independent observations** of Xmore

B1: probability distribution
p(x)

B2: find the expectation.
$E(X) = \sum P(X = x)$

B3: use variance to measure this spread
$Var(X) = \sum P(X = x)(x - \mu)^2$ or $= \sum P(X = x)x^2 - \mu^2$

a standard deviation $\sigma = \sqrt{Var(X)}$

$E(X_1 + X_2 + \dots + X_n) = nE(X)$

$Var(X_1 + X_2 + \dots + X_n) = nVar(X)$

In this situation, we have two independent probability distributions **X and Y are independent.**

LOLI 1: If X and Y are not independent, then $Var(X + Y)$ is no longer equal to $Var(X) + Var(Y)$.

The expectation and variance of linear transforms of X and Y are given by

$Var(aX + bY) = a^2Var(X) + b^2Var(Y)$

$Var(aX - bY) = a^2Var(X) + b^2Var(Y)$

$E(aX + bY) = aE(X) + bE(Y)$

$E(aX - bY) = aE(X) - bE(Y)$

$X - Y$ is really useful if you want to find the difference between two variables.

$E(X - Y)$ is a bit like saying "What do you expect the difference between X and Y to be", and $Var(X - Y)$ tells you the variance.

If X and Y are independent **random variables**

E(X + Y) = E(X) + E(Y)
E(X - Y) = E(X) - E(Y)

Var(X + Y) = Var(X) + Var(Y)

Var(X - Y) = Var(X) + Var(Y)

Remember to add the variances.

Subtracting independent random variables still increases the variance.

6 permutations and combinations Making Arrangements

How

combinations

$nCr = n! / [r! (n-r)!]$

permutations

$nPr = n! / (n-r)!$

Going round in circles

Calculate the number of arrangements **or** $n \times n-1 \times n-2 \times \dots \times 3 \times 2 \times 1$.

The number of ways of arranging n items in a circle is $(n - 1)!$

find the number of arrangements

There are n objects in total

r of the objects are alike

There are n objects in total.

1 of one type of object are alike, and so are **r** of another type.

Definition

the end result of combination is **smaller** than permutation.

This type of calculation is called the factorial of a number. The factorial of 3 is written as 3!, and the factorial of n is n!. This pronounce n! **"n factorial."**

Permutation: order matters.

Combination: order doesn't matter.

what is it used for?

to count the possibilities.

1. Visualizing Information: First Impressions

Data refers to raw facts and figures that have been collected.
Information is data that has some sort of added meaning.

Statistics

make objective decisions, make accurate predictions that seem inspired, and convey the message you want in the most effective way possible.

Statistics are numbers that summarize raw facts and figures in order to help us.
The study of statistics covers where statistics come from, how to calculate them, and how you can use them effectively.

what sort of data you're dealing with?

categorical (qualitative)

quantities

Charts

visualize your data

segmented bar chart

If you want to show frequencies and percentages

Horizontal bar charts

useful if the names of your categories are long

bar charts

Vertical bar charts

Pie chart

useful if you want to compare basic proportions

less useful if all the slices have similar sizes

Features

Values are shown on a continuous numeric scale.
grouped data
the area of each bar is proportional to the frequency
there are no gaps between the bars on the chart

Histogram

to make a histogram

2. Measuring Central Tendency: The Middle Way

effectively **summarize** data so concisely and usefully as possible.

Find the most **representative values** in your data and draw important conclusions.

the mean

to add all the numbers together, and then divide by how many numbers there are.

$\mu = \Sigma x / n$

$\mu = 2k / 2l$

When to use it?
When the data is fairly symmetric and shows just the one trend.

The median

Outlier is An extreme high or low value that stands out from the rest of the data

Skewed Data: When outliers "pull" the data to the left or right. Ex: data is skewed to the right

The median is always in the middle. It's the middle value.

How to find the median in three steps:
1. **Line your numbers up in order,** from smallest to largest.
2. **Find the middle number.** Make sure that there's the same # in the middle. If you have n numbers, the middle number is at position $(n + 1) / 2$.
3. **If you have an even number** of values, get the median by adding the two middle ones together and dividing by 2. You can find the midpoint by calculating $(n + 1) / 2$. The two middle numbers are on either side of this point.

Ord CHU:
1. The data is skewed to the right, which pulls the mean higher.
2. The data is skewed to the left, so the mean is pulled lower.
3. The data is skewed to the left, so the mean is pulled to the left. Therefore, the mean is lower than the median.

When to use it?
When the data is skewed because of outliers.

the mode

the value with the highest frequency.

if a set of data has two modes, then we call the data **bimodal.**

"-modal"

The mode doesn't just work with numeric data; it works with categorical data too.

it's the only sort of average that works with categorical data.

The category or group with the highest frequency is called the **modal class.**

Three steps for finding the mode:
1. Find all the distinct categories or values in your set of data.
2. Count the number of times each one appears or occurs.
3. Pick out the one(s) with the highest frequency to get the mode.

When to use it?
When you're working with categorical data.
When the data shows two or more clusters of data.

3. Measuring Spread: Bower Rangers

Measuring the range

Range = upper bound - lower bound

The range is only really good for describing the difference between the highest and lowest number

The range is a way of measuring how spread out a set of values are

the lowest value is **the lower bound.**

the largest value is **the upper bound.**

average distance

measure variability

to measure the variability of the players' scores, look at how far away each value is from the mean

The average distance of values from the mean is always 0.

The variance is a way of measuring spread, and it's the average of the distance of values from the mean squared.

$Variance = \frac{\Sigma(x - \mu)^2}{n}$ or $\frac{\Sigma(x - \bar{x})^2}{n - 1}$

σ = variance
$\sigma 2$ = variance

If you need to measure distances from the mean -> use the variance

The problem with the variance is that it can be quite difficult to think about spread in terms of distances squared -> **standard deviation**

Statisticians use the variance a lot as a means of **measuring the spread of data.**

Standard core

$z = (x - \mu) / \sigma$

Interquartile range

A box and whisker diagram **shows the range, interquartile range, and median of a set of data**

So box and whisker diagrams are really just **a neat way of showing ranges and quantiles.**

1. lining up the values in ascending order
2. then splitting the data into four equally sized chunks -> **quartiles**

DEFINITION:
The **lower quartile**, **the lower quartile**, or first quartile(Q_1) is the number that lies between the lowest number and the median. It's also the median of the data's lower half.
The range of the values in these two quartiles is called **the interquartile range** (IQR).

Finding the position of the lower quartile
1. calculating **$n \div 4$**
2. **an integer**, then the lower quartile is positioned halfway between this position and the next one. Take the average of the numbers in these two positions to get your lower quartile.
3. **not an integer**, then round it up. This gives you the position of the lower quartile.
As an example, if you have 9 numbers, start off by calculating $9 \div 4$, which gives you 2.5. Rounding this number up gives you 2, which means that the lower quartile is at position 2.

Interquartile range = Upper quartile – Lower quartile

Finding the position of the upper quartile
1. calculating **$3n \div 4$**
2. **an integer**, then the upper quartile is positioned halfway between this position and the next. Add the two numbers at these positions together and divide by 2.
3. **not an integer**, then round it up. This new number gives you the position of the upper quartile.

Problems

outliers

We need to get away from outliers

interpercentile range

DEFINITION:
1. **Percentiles** are values that **split your data into percentages** in the same way that quartiles split data into quarters.
2. our data is split into tenths, so the values are called **deciles.**
3. The interpercentile range is sometimes a new range called the **interpercentile range.**
4. The **10th percentile** is the value that's **1/**% of the way through your data. It's denoted by **P_{10}**.

Percentile uses

the interpercentile range isn't that commonly used

Percentile are useful for **benchmarking and determining rank or position**
They enable you to determine how high a particular value is relative to all the others

Finding percentiles
1. line all your values up in ascending order.
2. To find the position of the **kth percentile out of n numbers,** start off by calculating **k(n+1)**

```
your percentile is _k_ **an integer**, then
the percentile is the number that's
below_at_position_k)_ **and_then_the_next
number_along_ **the_average** of the
two_numbers_ Do_the_percentiles_to_give_you
your_percentile
```

3. **not an integer, then round it up**. This then gives you the position of the percentile.

4. Calculating Probabilities: Taking Chances

definition

P(A | B): The probability of A given that we know B has happened
the probability of event A occurring given event B

A ∩ B: the **union** of A and B - meaning **"or."**

A ∩ B: the **intersection** between A and B - meaning **"and."**

Probability is a way of measuring the chance of something happening.

Probability is measured on a scale of 0 to 1. If an event is impossible, it has a probability of 0. If it's an absolute certainty, then the probability is 1

Event is An outcome or occurrence that has a probability assigned to it.

If events A and B are exclusive, then
P(A ∩ B) = 0
If events A and B are exhaustive, then
P(A ∪ B) = 1

what it use for?

The important thing to remember is that **a probability indicates a long-term trend only.**

make predictions about your data and see patterns

predict the future

make informed decisions

How?

add probabilities: These two events are **mutually exclusive**, so they can't occur.

P(A) + P(B)

$P(A) = n(A)/n(S)$

n(S): The number of possible outcomes

P(A): Probability of event A occurring

n(A):Number of ways of getting an event A

S is known as the possibility space, or sample space.

P(A) + P(A') = 1

P(A') = 1 - P(A)

The two events intersect.

P(Black or Even) = P(Black) + P(Even) - P(Black and Even)
P(A ∪ B) = P(A) + P(B) - P(A ∩ B)

P(A ∩ B) = P(A) + P(B)

P(A ∩ B) = P(A ∩ B) / P(B)

P(B ∩ A) = P(B ∩ A) / P(A)

$= P(A) \times P(B \mid A) / [P(A) \times P(B \mid A) + P(A') \times P(B \mid A')]$

p159

Conditional probabilities = probability tree

P(A | B) = P(A)

P(A ∩ B) = P(A) x P(B)

If two events A and B are independent.