## **DATA ANALYSIS AND FUNDAMENTALS**

The table below summarizes our data types. To expand on the information in the table, you can look through the text that follows.

| **Data Types** |  |  |
| --- | --- | --- |
| **Quantitative:** | **Continuous** | **Discrete** |
|  | Height, Age, Income | Pages in a Book, Trees in Yard, Dogs at a Coffee Shop |
|  |  |  |
| **Categorical:** | **Ordinal** | **Nominal** |
|  | Letter Grade, Survey Rating | Gender, Marital Status, Breakfast Items |

Below is a little more detail of the information shared in the above table.

Another Look

To break down our data types, there are two main blocks:

**Quantitative** and **Categorical**

**Quantitative** can be further divided into Continuous or Discrete.

**Categorical** data can be divided into Ordinal or Nominal.

You should have now mastered what types of data in the world around us falls into each of these four buckets: Discrete, Continuous, Nominal, and Ordinal. In the next sections, we will work through the numeric summaries that relate specifically to quantitative variables.

Quantitative vs. Categorical

Some of these can be a bit tricky - notice even though zip codes are a number, they aren’t really a quantitative variable. If we add two zip codes together, we do not obtain any useful information from this new value. Therefore, this is a categorical variable.

**Height**, **Age**, the **Number of Pages in a Book**, and **Annual Income** all take on values that we can add, subtract and perform other operations with to gain useful insight. Hence, these are quantitative.

**Gender**, **Letter Grade**, **Breakfast Type**, **Marital Status**, and **Zip Code** can be thought of as labels for a group of items or individuals. Hence, these are categorical.

Continuous vs. Discrete

To consider if we have continuous or discrete data, we should see if we can split our data into smaller and smaller units. Consider time - we could measure an event in years, months, days, hours, minutes, or seconds, and even at seconds we know there are smaller units we could measure time in. Therefore, we know this data type is continuous. **Height**, **age**, and **income** are all examples of continuous data. Alternatively, the **number of pages in a book**, **dogs I count outside a coffee shop**, or **trees in a yard** are discrete data. We would not want to split our dogs in half.

Ordinal vs. Nominal

In looking at categorical variables, we found **Gender**, **Marital Status**, **Zip Code**, and your **Breakfast items** are nominal variables where there is no order ranking associated with this type of data. Whether you ate cereal, toast, eggs, or only coffee for breakfast; there is no rank-ordering associated with your breakfast.

Alternatively, the **Letter Grade** or **Survey Ratings** have a rank ordering associated with it, as ordinal data. If you receive an A, this is higher than an A-. An A- is ranked higher than a B+, and so on... Ordinal variables frequently occur on rating scales from very poor to very good. In many cases, we turn these ordinal variables into numbers, as we can more easily analyze them, but more on this later!

Final Steps for Calculating the Mean

To finalize our calculation of the mean, we introduce **n** as the total number of values in our dataset. We can use this notation both at the top of our summation, as well as for the value that we divide by when calculating the mean.

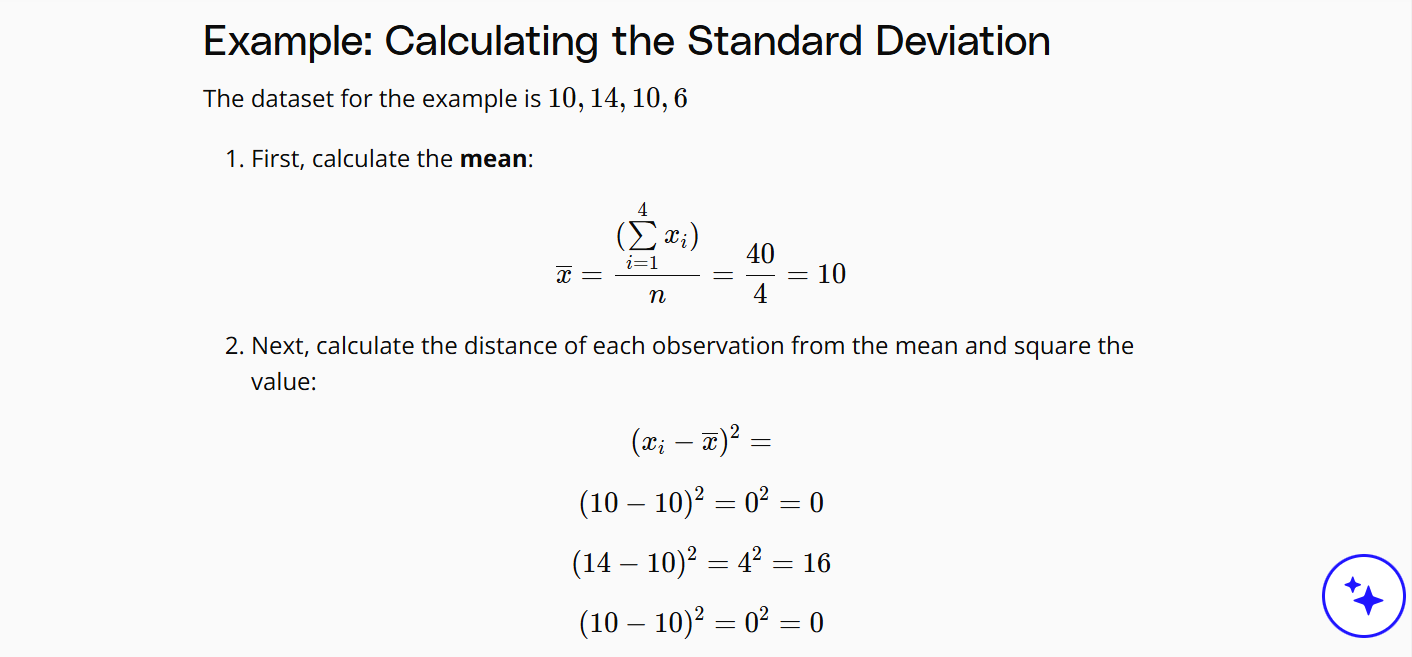
1n∑i=1nxi*n*1​*i*=1∑*n*​*xi*​

Instead of writing out all of the above, we commonly write xˉ*x*ˉ to represent the mean of a dataset. Although similar to the first video, we could use any variable. Therefore, we might also write yˉ*y*ˉ​, or any other letter.

We also could index using any other letter, not just i*i*. We could just as easily use j*j*, k*k*, or m*m* to index each of our data values. The quizzes on the next concept will help reinforce this idea.

Notice

At second 0:12, this should say ∑i=15xi=x1+x2+x3+x4+x5*i*=1∑5​*xi*​=*x*1​+*x*2​+*x*3​+*x*4​+*x*5​. The xi*xi*​ is missing here in front of the summation.



A math problem with numbers and equations

Description automatically generated with medium confidence

Important Final Points

1. The variance is used to compare the spread of two different groups. A set of data with higher variance is more spread out than a dataset with lower variance. Be careful though, there might just be an outlier (or outliers) that is increasing the variance when most of the data are actually very close.
2. When comparing the spread between two datasets, the units of each must be the same.
3. When data are related to money or the economy, higher variance (or standard deviation) is associated with higher risk.
4. The standard deviation is used more often in practice than the variance because it shares the units of the original dataset.

**Use in the World**

The standard deviation is associated with risk in finance, assists in determining the significance of drugs in medical studies, and measures the error of our results for predicting anything from the amount of rainfall we can expect tomorrow to your predicted commute time tomorrow.

These applications are beyond the scope of this lesson as they pertain to specific fields, but know that understanding the spread of a particular set of data is extremely important to many areas. In this lesson, you mastered the calculation of the most common measures of spread.

Variable Types

We have covered a lot up to this point! We started with identifying data types as either categorical or quantitative. We then learned we could identify quantitative variables as either continuous or discrete. We also found we could identify categorical variables as either ordinal or nominal.

Categorical Variables

When analyzing categorical variables, we commonly just look at the count or percent of a group that falls into each **level** of a category. For example, if we had two **levels** of a dog category: lab and not lab. We might say, 32% of the dogs were lab (percent), or we might say 32 of the 100 dogs I saw were labs (count).

However, the 4 aspects associated with describing quantitative variables are not used to describe categorical variables.

Quantitative Variables

Then we learned there are four main aspects used to describe quantitative variables:

1. Measures of **Center**
2. Measures of **Spread**
3. **Shape** of the Distribution
4. **Outliers**

Measures of Center

We looked at calculating measures of Center

1. **Means**
2. **Medians**
3. **Modes**

Measures of Spread

We also looked at calculating measures of Spread

1. **Range**
2. **Interquartile Range**
3. **Standard Deviation**
4. **Variance**

Shape

We learned that the distribution of our data is frequently associated with one of the three **shapes**:

**1. Right-skewed**

**2. Left-skewed**

**3. Symmetric** (frequently normally distributed)

Depending on the shape associated with our dataset, certain measures of center or spread may be better for summarizing our dataset.

When we have data that follows a **normal** distribution, we can completely understand our dataset using the mean and standard deviation.

However, if our dataset is **skewed**, the 5 number summary (and measures of center associated with it) might be better to summarize our dataset.

Outliers

We learned that outliers have a larger influence on measures like the mean than on measures like the median. We learned that we should work with outliers on a situation by situation basis. Common techniques include:

**1.** At least note they exist and the impact on summary statistics.

**2.** If typo - remove or fix

**3.** Understand why they exist, and the impact on questions we are trying to answer about our data.

**4.** Reporting the 5 number summary values is often a better indication than measures like the mean and standard deviation when we have outliers.

**5.** Be careful in reporting. Know how to ask the right questions.

Histograms and Box Plots

We also looked at histograms and box plots to visualize our quantitative data. Identifying outliers and the shape associated with the distribution of our data are easier when using a visual as opposed to using summary statistics.

Inferential Statistics

Inferential Statistics**is about using our collected data to draw conclusions about a larger population**.

We looked at specific examples that allowed us to identify the

1. **Population** - our entire group of interest.
2. **Parameter** - numeric summary about a population
3. **Sample** - a subset of the population
4. **Statistic** numeric summary about a sample

Marketing Funnel Metrics

**Impressions & Reach** – building brand and product awareness using ad platforms and search engine optimization (SEO). SEO allows ads to show up for the right mix of search terms as people search online

* **Impressions** – an instance of an advertisement appearing on a website when it is viewed by a visitor.

**Lead generation** – measures how many visits are made to the website.

* **Click –** every time a website visitor views the ad and clicks it
* **Click Thru Rate** – number of users that clicked an ad or clicked a link sent via email
* **Cost Per Click**
* **Cost Per Lead** – indicates a user has become a potential customer or **lead** because they have expressed interest in the company by downloading a document, creating an account, or providing an email address.

**Conversion** – when a lead converts to a paid customer

* **Customer Acquisition Cost**

Two Additional Levels

Before we move on, I wanted to share 2 more measures that companies use.

Loyalty

To grow their revenue and company profits, companies don’t just want their customers to buy once from them, but to come back to their website. Especially if the product is not a high-priced product. That customer loyalty allows you to track how many revisits a customer is making after their first purchase, or how many of the customers have continued shopping after their first purchase.

**Metrics:** Some commonly used metrics include **Repeat Purchase Rate** and **Net Promoter Score**. We will not be going in-depth with these, but please do check out the resources below to learn more about them.

Advocacy

Another level companies sometimes track is whether their customer is advocating for their company. That is, saying good things about the product and services. Leaning on social media provides a great opportunity to do just that.

**Metrics:** Some commonly used metrics include **Customer Referrals** and **Leads from Social Media**. For example, as the paid customer tweets about the company, likes the product on FB, provides a good rating on Amazon or the company website, analysts can use those metrics, such as ratings and likes to show how many of the customers serve as advocates.

We will not be going in-depth with these last two stage levels, but we have provided some resources below to help you understand these more.

Lifetime Value

When deciding how to spend the marketing budget, you want to focus on some of your best customers – those that will stay for the long term and continue to generate revenue for the company.

These are your **high-value customers** and you want to bring in more of them.

Your goal should be for every dollar spent on marketing efforts. It should provide a higher rate of return and generate revenue multiple times over.

Terms needed to calculate Lifetime Value:

**Purchase Cycle**: The time increment adopted for business calculations

* **Total Sale Revenue Per Cycle**: Revenue earned from a customer per purchase cycle
* **Number of Sales Per Purchase Cycle**: Number of times customer buys during the purchase cycle
* **Cost Per Acquisition**: (Cost of marketing and sales)/ number of new leads
* **Expected Retention Time**: Amount of time (measured in purchasing cycles) you expect to retain the customer.
* **Average Sale Revenue**: (Total customer revenue/ Number of purchases in the cycle) OR Average revenue received from the customer per transaction during the cycle
* **Profit Margin Per Customer**: ((Average Sale - Average Cost of Sale) / Average Sale)

Lifetime Value (LTV) = Average Sale x Number of Repeat Sales x Expected Retention Time x Profit Margin

Sales metrics borrow some terminology from marketing metrics. Sales can not only focus on the end customer, such as the consumer but also a company that will likely generate customers.

* B2C
  + In the case of WeCart from marketing, we have thus far focused on the end consumer and their order as our marketing focus, or *unit of analysis*. This type of model is called a **business-to-consumer** model or *B2C*.
* B2B
  + Let's say that we add an additional focus on generating sales by reaching out to grocery stores. This is where We Card partners with the local grocery stores. We can now gain access to their customers and the local grocery stores can access a delivery service. This type of business model is called a **business-to-business** model or *B2B*. So, part of our sales team can focus on the B2B model, where each sales rep is trying to create a sales lead. A sales lead would be a grocery store interested in partnering with We Card. If a grocery store signs a deal, then it becomes a booking or a closed deal.

New Vocabulary

* **Business to Business (B2B)**: When one business makes a business transaction (goods or services) with another business. Often takes place when one business is providing source materials to the other business to in turn finally sell it to the consumer.
* **Business to Consumer (B2C)**: When a business sells products and services to the final consumer.

**Measuring your company's growth**

Growth for a website or app is counted in the number of users.

* Are we seeing the number of people actually using the site increasing or decreasing?
* If you see your website use as high, are they unique users or the same people coming back?

An important aspect of growth is not just whether you have users but whether they continue to actively use or engage with the website.

In the next few pages, we'll talk about the following metrics:

* Active users
* Stickiness
* Churn rate
* **Stickiness** = Daily Active Users/ Monthly Active Users
* It is a useful KPI for management and investors. It tells the management what is their company's growth rate. Investors want to know if this online app or website has the potential to make money in the future. For example, if the plan is to introduce advertising into the app, the potential valuation will depend on whether the app has a large number of users that keep coming back to it.

Churn Rate = (Customers at start of usage interval - Customers at end of usage interval)​  
 customers at start of usage interval

**Profit and Loss Statement**

(*aka income statement*)

The following list is a breakdown of the individual items within the Profit and Loss Statement.

* **Revenues**: The money a company makes from the sales of its products and services.
* **Cost of Goods Sold (COGS) or Cost of Sales**: These are the direct costs the company incurs to develop the product or service being sold.
* **Gross Profit:** The difference between the revenue earned and the costs summarized in COGS. **Gross Profit = Revenue - COGS**
* **Selling, General, and Administrative expenses (SGAs):** Includes the following expenses:
  + Marketing, sale commissions
  + Salaries for office staff
  + Supplies and computer hardware
  + Note: Some companies list total operating expenses separately from SGAS while others treat them as synonymous with SGAS.
* **Operating expenses**: Expenses incurred outside of direct manufacturing costs:
  + Overhead costs
  + Legal
  + Rent
  + Utilities
  + Taxes
  + Interest
  + R&D expenses.
* **Total Operating Expenses** = Sum of SGAs and Operating expenses **Total Operating Expenses= SGAs + Operating Expenses**
* **Operating Income**: The difference between Gross profit and Total operating expenses **Operating Income = Gross Profit - Total Operating Expenses**
  + Note: **Operating Income** is also referred to as **Earnings Before Interest and Tax (EBIT)**
* **Net Income:** Subtracting the Interest and Tax from Operating Income gives the Net Income **Net Income = Operating Income - (Interest and Taxes)**

**Gross Margin**

Gross Margin is a statement about the overall profitability of the company.

Calculation

Gross Margin = (Total Sales Revenue – Cost of Goods Sold) / Total Sales Revenue

which is the same as for Gross Profit / Total Sales Revenue

This metric identifies the revenue that remains after accounting for direct costs of production.

* **Gross Margin** = (Total Sales Revenue – Cost of Goods Sold) / Total Sales Revenue
* Can also be represented in percentage by multiplying it by 100
  + **Gross Margin (in %)** = [ (Total Sales Revenue – Cost of Goods Sold) / Total Sales Revenue] \*100
* Gross Margin tells business executives what percentage of each revenue dollar is available to cover operating expenses after the COGS have been accounted for.

Costs to a company can be split into two major groups;

* **Fixed costs**
  + Fixed costs are expenses that you will incur on a regular, perhaps monthly basis, such as rent, utilities, and employee salaries.
* **Variable costs**
  + Variable costs are expenses that move up and down in response to production output. This interpretation is particularly helpful for companies to determine the pricing of the product.

In other words, it helps them find the breakeven point where the pricing will cover fixed overhead costs for sure.

Contribution margin can also be helpful as a useful tool to dive into the P&L statement! While the typical P&L statement line items tell us the overall profitability of our business, contribution margin can be used to identify which product or product line is contributing the most to our profit margin.

(Looking at the graph in the video)

First of all, we want to calculate the overall contribution margin, the difference between the *sales revenue* and *variable cost*.

Once you get the difference for the total contribution margin, you divide it by the units sold, and that gives you the contribution margin per unit.

You can think of the contribution margin as a percentage of your revenue that'll cover your fixed costs, which you have little control over and have to incur.

* If the contribution margin per unit > fixed cost, this means you're making a profit on each sale.
* If the contribution margin per unit < fixed cost, this means you're making a loss on each sale.

So in other words, if you want to make a profit, you have to sell your product with at least your contribution margin covered for each unit sale.

Contribution Margin

Contribution Margin tells us the amount of revenue that covers the variable costs and is now available to cover the fixed costs and generate profits. Companies use it to identify which product or product line is contributing the most to the profit margin. It also helps determine the break-even point where the pricing will cover fixed overhead costs and leave enough for profits too.

Fixed costs are also called **sunk costs**. A good caution to keep in mind is that fixed or sunk costs can increase (for e.g., unexpected rent increases, machinery replacement costs), which is why operational managers prefer the term sunk costs. These sunk costs can prove tricky, because a small increment when taken in bulk, can turn out to be catastrophic for companies, especially start-ups.

Citation for the graph depicting [**Contribution Margin(opens in a new tab)**](https://www.dummies.com/business/operations-management/how-to-prepare-a-cost-volume-profit-analysis/).

In the graph depicted in the video **Total Contribution Margin is indicated by the red line.**

Terminology

* **Fixed costs:** Expenses incurred on a regular basis, such as monthly rent, utilities, and employee salaries.
* **Variable costs:** Expenses that move up and down in response to production output.
* **Contribution Margin:** The amount of revenue that covers the variable costs and is able to cover the fixed costs.

Calculation

* **Total Contribution Margin** = Total Sales Revenue - Total Variable Cost
* **Contribution Margin Per Unit:** Total Contribution Margin / Number of Units Sold

The following video describes how to calculate Total Contribution Margin and Contribution Margin Per Unit.

Topics Covered

**Key Performance Indicators:** We discussed key performance indicators and how they differ by industry.

**Business Process Flow**: We went through the business process flow across various divisions. This provided the context for learning about the business metrics.

**Business metric**s: We took on each business area, such as marketing and growth, and introduced you to a metric commonly used to measure success in that business area. We discussed what each meant and how to calculate it. We practiced calculating the metrics and applying the metrics, as well as when and where to use the metric. To do this, we focused on 3 main elements related to metrics:

* Evaluate important business metrics
* Interpret and analyze these metrics
* Create visualizations of these metrics

**Distribution and central tendency**: We circled back to the topic of data distribution that you learned about in the previous lesson and discussed why paying attention to the distribution of the data and the choice of the measure of central tendency was important.

**Grouping data**: We ended with a discussion on how to look at the data across groups, cohorts, and time.

Let's take a moment to summarize the key take-aways from this lesson.

Overarching Themes Summary

* Businesses use **Key Performance Indicators** to track how they are performing on key goals or objectives.
* The **Marketing Funnel** captures the various stages in the customer's journey. At the top of the funnel, it captures the impressions, clicks, leads, and conversions at the bottom of the funnel.
* **Optimizing the funnel** refers to maximizing the conversion rate at each level of the funnel.
* The **Sales Funnel** captures the various stages in the sales cycle. At the top of the funnel, it captures the prospects, then the leads and qualified leads, and ends with bookings or closed deals at the bottom of the funnel.
* It is important to look at the **distribution of the data** to understand if the measures of central tendency represent a normal distribution. Looking at the distribution and measures of central tendency is a critical step of the data analysis process.
* Data should be examined split across cohorts, business cycles, time, product lines, regions, and other **grouping** criteria to fully understand the data. It is critical to slice the data across various factors to make sense of the data and make recommendations.

**Metrics**

*Marketing*

* **Click Through Rate (CTR)** is an indication of whether the ad campaign is generating enough interest in potential customers. When the CTR increases, it is an indicator of effective and interesting content in your ad campaign, and that maybe you should increase the number of impressions for that ad.
* **Cost Per Click (CPC)** is an indicator of the cost-effectiveness of the ad platform and a useful tool to compare and strategize about which marketing platform is yielding higher impression and reach and resulting in potential leads.
* **Cost Per Lead (CPL)** is an indicator of the cost-effectiveness of the ad platform and a useful tool to compare and strategize about which marketing platforms yielded more leads.
* **Customer Acquisition Cost (CAC)** is a useful metric used to get an estimate of how much it cost us to acquire the customer in the period the money was spent to reach out to them.

*Marketing and Financial*

* **Cost Per Acquisition (CPA)** allows a business to gauge whether the marketing campaign is generating enough potential leads.
* **Life Time Value (LTV)** allows you to focus on audiences and potential customers that will generate higher LTVs with minimum customer acquisition cost. There are several ways to calculate the Life Time Value and it is best to calculate the LTV using different ways to arrive at the average LTV for a customer.

*Growth*

* **Stickiness** indicates whether the customers are staying and returning to the website frequently enough. It is a good measure of the potential growth of the business.
* **Churn rate** is a measure of declining growth and businesses aim to have a higher growth rate than churn rate. It is a measure of whether the business is retaining the acquired customers.

*Financial*

* The **Profit and Loss Statement** also called an income statement, is one type of financial statement that shows a company's performance and financial position. needed to create the P&L statement are:
* **Revenue** is the money that your company makes from the sales of your products and services
* **Cost of Goods Sold OR Cost of Sales** are the direct costs the company incurs to develop and produce the product or service being sold
* **Gross Profit** is the difference between the revenue and COGS
* **Selling, General, and Administrative expenses** capture a wide range of expenses, from administrative, sales commissions, supplies, legal fees, rent, utilities, taxes, and interests. It is used synonymously with **Operating expenses**. SG&A typically excludes research and development expenses.
* **Operating Profit** is the difference between gross profit and total operating expenses.
* **Net Income** is operating profit minus interest and tax expenses.
* **Gross Margin** tells business executives what percentage of each revenue dollar is available to cover operating expenses after the COGS have been accounted for.
* **Contribution Margin** provides the break-even point where the pricing of a product will cover fixed overhead costs.

Historical Data

Historical data is about what your performance metrics show for the past. For sales, for example, we look at prior sales data from the previous year or months. For financial modeling, we look at prior financial statements, as well as quarterly and monthly results. Above we are using prior year data (e.g., Revenue, COGS) to get our operating income, which we then use to calculate the historical operating margin.

Formulas for Calculating Historical Financial Metrics

Typically, the historical statistics or metrics used to forecast financial metrics in an Income Statement are:

1. Revenue Growth
2. Gross Margin
3. Operating Margin
4. Historical Tax Rate
5. Historical Interest Expense Rate

The following list provides more information about calculating the historical statistics.

1. **Revenue Growth (in %) = (Current Year's Revenue / Previous year's revenue) - 1**
2. **Gross Margin = 1 - (Current Year's Cost of COGS / Current Year's Total Revenue)**

Keep in mind the two terms COGS and Cost of Revenue can be used interchangeably.

1. **Operating Margin = Current Year's Operating Income / Current Year's Total Revenue**
2. **Historical Tax Rate** is the tax rate from the companies previous year's tax rate.
3. **Historical Interest Rate** is the interest rate coming from the previous year's Debt Schedule.