Statistics

Juan Arellano

KSM Consulting

jarellano@ksmconsulting.com

November 10, 2020

Overview

Descriptive Statistics: Fundamentals

Types of Data

- Categorical: Car brands, answers to yes/no questions
- Numerical:
  - Discrete: Number of children you want to have, SAT score
  - Continuous: weight, height

Descriptive Statistics: Fundamentals

Levels of Measurement

- Qualitative
  - Nominal: four seasons
  - Ordinal: rating your meal (disgusting, unappetizing, neutral, tasty, and delicious)
- Quantitative
  - Interval: degrees Celsius and Fahrenheit
  - Ration: degrees Kelvin, length

Descriptive Statistics: Fundamentals

Graphs and tables that can represent categorical variables:

- Frequency distribution tables: show the category and its corresponding absolute frequency
- Bar charts: each bar represents a category, and on the y-axis we have the absolute frequency
- Pie charts: used when we want to see the share of an item as part of the total
- Pareto diagrams: a special type of bar chart where the categories are shown in descending order of frequency, and a separate curve shows the cumulative frequency

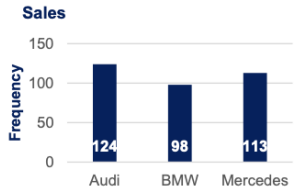
Descriptive Statistics: Fundamentals

Graphs and tables that can represent categorical variables:

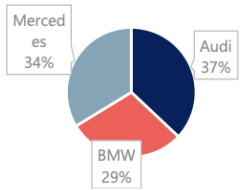
- Frequency distribution tables:

	Frequency
Audi	124
BMW	98
Mercedes	113
Total	335

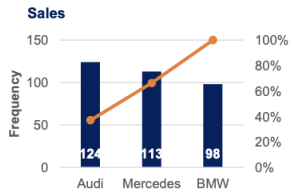
- Bar charts:



- Pie charts:



- Pareto diagrams:



Descriptive Statistics: Fundamentals

Graphs and tables that can represent numerical variables:

- Frequency distribution tables: Frequency distribution tables for numerical variables are different than the ones for categorical. Usually, they are divided into intervals of equal (or unequal) length. The tables show the interval, the absolute frequency and sometimes it is useful to also include the relative (and cumulative) frequencies
- Histograms: Histograms are the one of the most common ways to represent numerical data. Each bar has width equal to the width of the interval. The bars are touching as there is continuation between intervals: where one ends, the other begins.

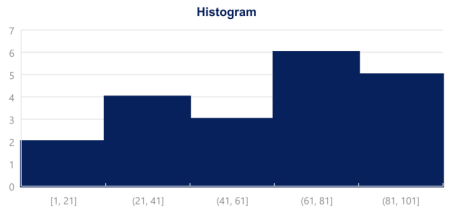
Descriptive Statistics: Fundamentals

Graphs and tables that can represent numerical variables:

- Frequency distribution tables:

Interval start	Interval end	Frequency	Relative frequency
1	21	2	0.10
21	41	4	0.20
41	61	3	0.15
61	81	6	0.30
81	101	5	0.25

- Histograms:



Descriptive Statistics: Fundamentals

Graphs and tables that can represent relationships between variables:

- Cross tables: Cross tables (or contingency tables) are used to represent categorical variables. One set of categories is labeling the rows and another is labeling the columns. We then fill in the table with the applicable data. It is a good idea to calculate the totals. Sometimes, these tables are constructed with the relative frequencies as shown in the table below.
- Scatter plots: When we want to represent two numerical variables on the same graph, we usually use a scatter plot. Scatter plots are useful when we doing regression analysis, as they help us detect patterns (linearity, homoscedasticity). Scatter plots usually represent lots and lots of data. Typically, we are not interested in single observations, but rather in the structure of the dataset.

Descriptive Statistics: Fundamentals

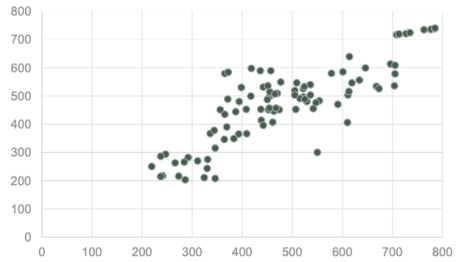
Graphs and tables that can represent relationships between variables:

- Cross tables:

Type of investment \ Investor	Investor A	Investor B	Investor C	Total
Stocks	96	185	39	320
Bonds	181	3	29	213
Real Estate	88	152	142	382
Total	365	340	210	915

Type of investment \ Investor	Investor A	Investor B	Investor C	Total
Stocks	0.10	0.20	0.04	0.35
Bonds	0.20	0.00	0.03	0.23
Real Estate	0.10	0.17	0.16	0.42
Total	0.40	0.37	0.23	1.00

- Scatter plots:



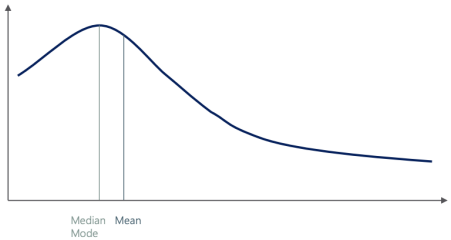
Descriptive Statistics: Fundamentals

The three measure of central tendency are mean, median, and mode.

- Mean: the most widely spread measure of central tendency. It is the simple average of the dataset $\frac{\sum_{i=1}^{N} X_i}{N}$
- Median: the midpoint of the ordered dataset. It is not as popular as the mean, but is often used in academia and data science since it is not affected by outliers. In an ordered dataset, the median is the number at position $\frac{n+1}{2}$
- Mode: the value that occurs most often. A dataset can have 0 modes, 1 mode or multiple modes. The mode is calculated simply by finding the value with the highest frequency.

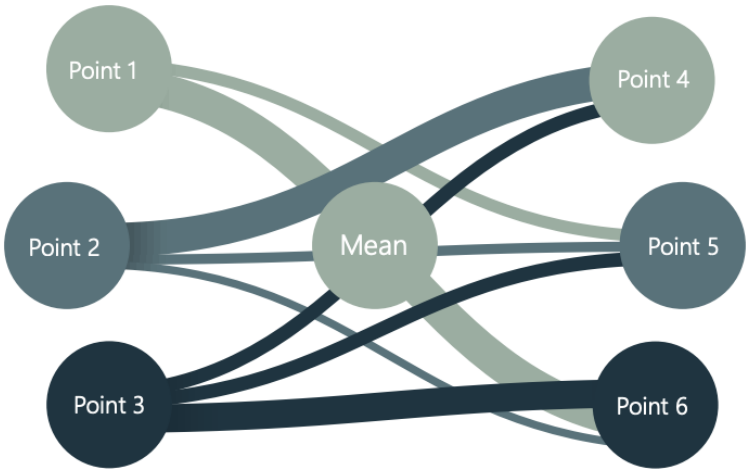
Descriptive Statistics: Fundamentals

Skewness is a measure of asymmetry that indicates whether the observations in a dataset are concentrated on one side. Right (positive) skewness looks like the one in the graph. It means that the outliers are to the right (long tail to the right). Left (negative) skewness means that the outliers are to the left. Usually, you will use software to calculate skewness.



Descriptive Statistics: Fundamentals

Variance and standard deviation measure the dispersion of a set of data points around its mean value. There are different formulas for population and sample variance and standard deviation. This is due to the fact that the sample formulas are the unbiased estimators of the population formulas



Descriptive Statistics: Fundamentals

- Covariance: a measure of the joint variability of two variables. A positive covariance means that the two variables move together. A covariance of 0 means that the two variables are independent. A negative covariance means that the two variables move in opposite directions.
- Correlation: measure of the joint variability of two variables. Unlike covariance, correlation could be thought of as a standardized measure. It takes on values between -1 and 1, thus it is easy for us to interpret the result. A correlation of 1, known as perfect positive correlation, means that one variable is perfectly explained by the other. A correlation of 0 means that the variables are independent. A correlation of -1, known as perfect negative correlation, means that one variable is explaining the other one perfectly, but they move in opposite directions.

Descriptive Statistics: Practical Example

Imagine we have access to the database of a real estate company operating in California. The company is launching a marketing campaign. We are the data analysts and we must identify which group of people will be more likely to buy our product. We have two tables in our database, Product information and Customer information. Customer information is only available for some products. One row contains a customer name, customer age, apartment id, building name, price they paid for apartment, square foot, year of purchase, customer location, mortgage and where they found the product(such as website). ID variables are like names, so they are categorical. Id is qualitative and nominal. Age is quantitative ratio. Age is a continuous variable. Age interval is ordinal because it represents different categories people fall in. Price is always numerical, can be discrete or continuous. Gender is a categorical, nominal variable. Location is a categorical variable.

Descriptive Statistics: Practical Example

365 DataScience RE California Database

Practical example

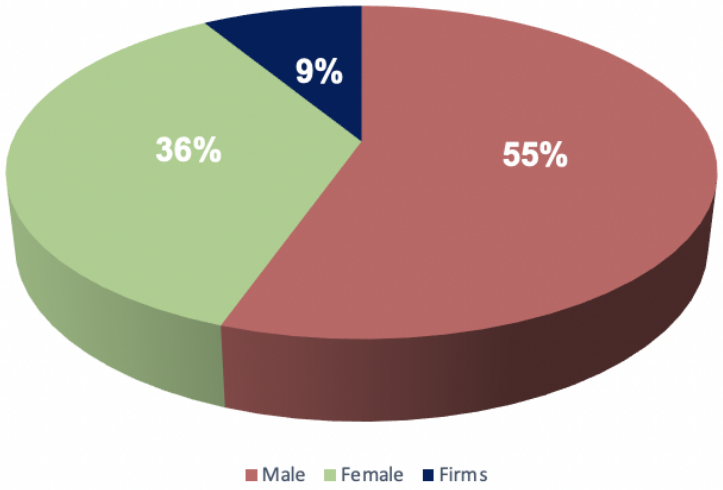
Product								
ID	Building	Year of sale	Month of sale	Type of property	Property #	Area (ft.)	Price	Status
1030	1	2005	11	Apartment	30	743.09	$ 246,172.68	Sold
1029	1	2005	10	Apartment	29	756.21	$ 246,331.90	Sold
2002	2	2007	7	Apartment	2	587.28	$ 209,280.91	Sold
2031	2	2007	12	Apartment	31	1604.75	$ 452,667.01	Sold
1049	1	2004	11	Apartment	49	1375.45	$ 467,083.31	Sold
3011	3	2007	9	Apartment	11	675.19	$ 203,491.85	Sold
3026	3	2007	9	Apartment	26	670.89	$ 212,520.83	Sold
3023	3	2008	1	Apartment	23	720.81	$ 198,591.85	Sold
1031	1	2006	6	Apartment	31	782.25	$ 265,467.68	Sold
4023	4	2006	3	Apartment	23	794.52	$ 235,633.26	Sold
1036	1	2004	10	Apartment	36	1160.36	$ 317,473.86	Sold
1046	1	2006	8	Apartment	46	1942.50	$ 503,790.23	Sold
4035	4	2007	10	Apartment	35	794.52	$ 217,786.38	Sold
2036	2	2006	11	Apartment	36	1109.25	$ 460,001.26	Sold
2056	2	2007	4	Apartment	56	1400.95	$ 460,001.26	Sold
1047	1	2007	12	Apartment	47	1479.72	$ 448,134.27	Sold
5051	5	2006	3	Apartment	51	790.54	$ 249.591.99	Sold

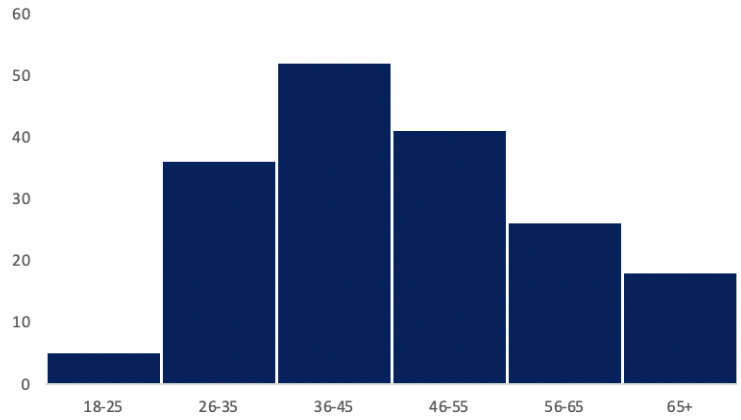
Customer													
Customer ID	Entity	Name	Surname	Age at time of purchase	Interval	Y	Gender	Country	State	Purpose	Deal satisfaction	Mortgage	Source
C0028	Individual	Madalyn	Mercer	19	18-25	1986	F	USA	California	Home	5	No	Website
C0027	Individual	Lara	Carrillo	22	18-25	1983	F	USA	California	Home	5	No	Website
C0112	Individual	Donavan	Flowers	22	18-25	1985	M	USA	California	Home	1	Yes	Client
C0160	Individual	Darien	Dorsey	22	18-25	1985	M	USA	California	Investment	3	Yes	Website
C0014	Individual	Alessandra	Perry	25	18-25	1979	F	USA	California	Home	4	No	Agency
C0125	Individual	Kaitlin	Owen	26	26-35	1981	F	USA	Virginia	Investment	5	No	Client
C0125	Individual	Kaitlin	Owen	26	26-35	1981	F	USA	Virginia	Investment	5	No	Agency
C0166	Individual	Terry	Forbes	26	26-35	1982	M	USA	California	Home	5	No	Client
C0034	Individual	Kole	Shannon	27	26-35	1979	M	USA	Arizona	Home	2	Yes	Website
C0170	Individual	Emmy	Singh	27	26-35	1979	F	USA	Virginia	Investment	3	Yes	Agency
C0009	Individual	Arabella	Ferrell	28	26-35	1976	F	USA	Oregon	Home	1	No	Agency
C0041	Individual	Christian	Costa	26	26-35	1980	M	USA	California	Home	5	No	Website
C0057	Individual	Michelle	Cameron	29	26-35	1978	F	USA	Nevada	Home	3	Yes	Website
C0061	Individual	Enrique	Cardenas	29	26-35	1977	M	USA	California	Home	2	No	Website
C0089	Individual	Amanda	Simon	29	26-35	1978	F	USA	California	Home	5	No	Agency
C0159	Individual	Kamden	Stewart	29	26-35	1978	M	USA	California	Home	5	No	Website
C0171	Individual	Skylar	Buchanan	29	26-35	1977	M	USA	Nevada	Home	4	Yes	Website

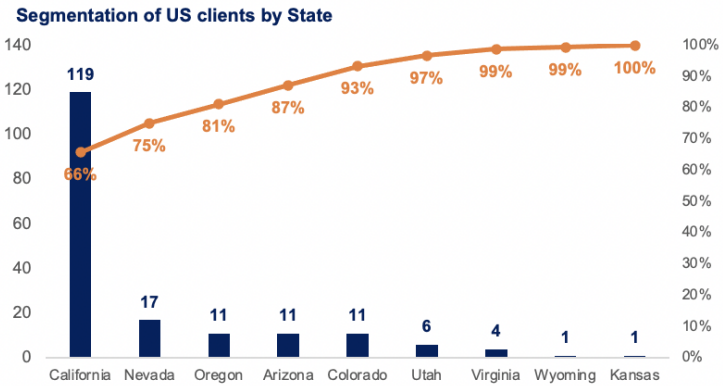
Descriptive Statistics: Practical Example

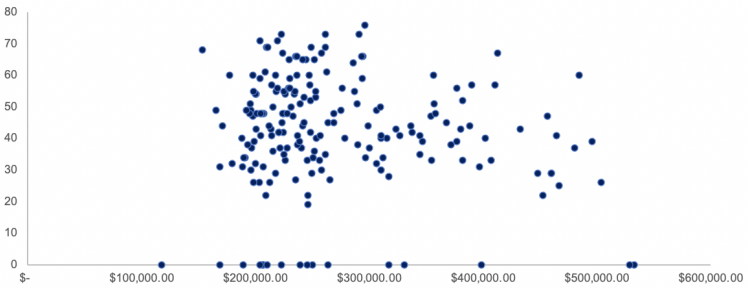
We must identify the groups of people that are most likely to buy our product. Gender is categorical. Most of our customers are male. When we analyze age, we see most of our data falls between 25 and 65 years. Since our data set is only a sample of customers who will ever buy our products, we use sample formulas. Let's see if age and price have a relationship. When we calculate the covariance we get a huge number, our correlation coefficient is -0.17. A common practice is to disregard correlations which are less than abs(0.2). We tell the marketing team they should target males. 68 % of sales come from California. 71% of sales were made with customers aged between 26 and 55 years old, and the age variable was right skewed meaning younger people buy more property than older people. There is no relationship between the age of a given customer and price.

Descriptive Statistics: Practical Example









Inferential Statistics: Fundamentals

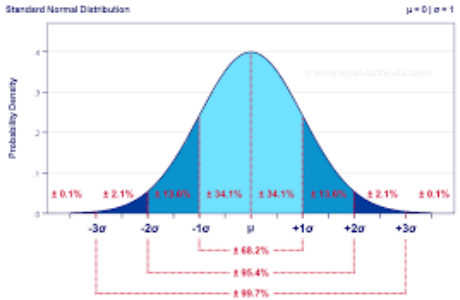
Distributions:

- A distribution is a function that shows the possible values for a variable and how often they occur
- It is a common mistake to believe that the distribution is the graph. In fact the distribution is the 'rule' that determines how values are positioned in relation to each other.
- Very often, we use a graph to visualize the data. Since different distributions have a particular graphical representation, statisticians like to plot them

Inferential Statistics: Fundamentals

The Normal distribution is also known as Gaussian distribution or the Bell curve. It is one of the most common distributions due to the following reasons:

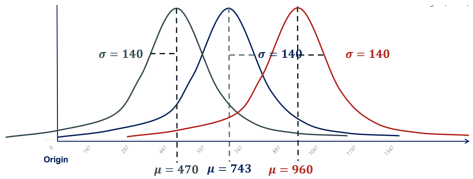
- Noted as $N(\mu, \sigma^2)$ where N stands for normal, stands for distribution, μ is the mean, and σ^2 is the variance
- It approximates a wide variety of random variables
- Distributions of sample means with large enough samples sizes could be approximated to normal
- All computable statistics are elegant
- Heavily used in regression analysis
- Good track record



Inferential Statistics: Fundamentals

Keeping the standard deviation constant, the graph of a normal distribution with:

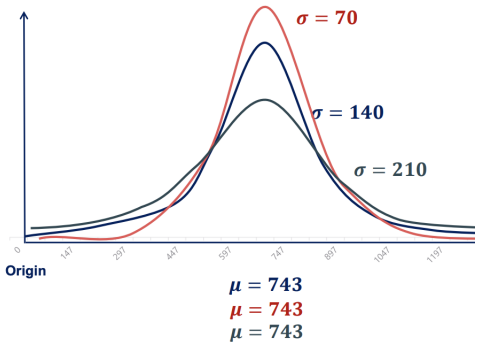
- a smaller mean would look in the same way, but be situated to the left (in gray)
- a larger mean would look in the same way, but be situated to the right (in red)



Inferential Statistics: Fundamentals

Keeping the mean constant, a normal distribution with:

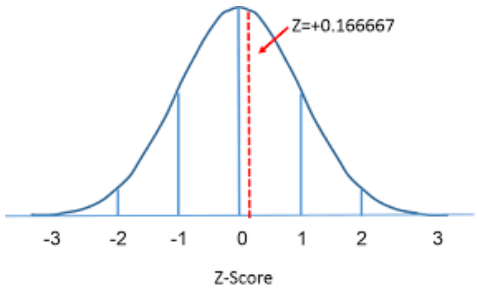
- a smaller standard deviation would be situated in the same spot, but have a higher peak and thinner tails (in red)
- a larger standard deviation would be situated in the same spot, but have a lower peak and fatter tails (in gray)



Inferential Statistics: Fundamentals

The Standard Normal distribution is a particular case of the Normal distribution. It has a mean of 0 and a standard deviation of 1.

- Noted as $N(0,1)$ where N stands for normal, stands for distribution, 0 is the mean, and 1 is the variance
- Standardization allows us to:
  - compare different normally distributed datasets
  - detect normality
  - detect outliers
  - create confidence intervals
  - test hypotheses
  - perform regression analysis



Inferential Statistics: Fundamentals

The Central Limit Theorem (CLT) is one of the greatest statistical insights. It states that no matter the underlying distribution of the dataset, the sampling distribution of the means would approximate a normal distribution. Moreover, the mean of the sampling distribution would be equal to the mean of the original distribution and the variance would be n times smaller, where n is the size of the samples. The CLT applies whenever we have a sum or an average of many variables (e.g. sum of rolled numbers when rolling dice).

- The theorem: No matter the distribution, the distribution of $\bar{x_1}, \bar{x_2}, \bar{x_3} ... \bar{x_4}$ would tend to $N\ (\mu, \frac{\sigma^2}{n}$. The more the samples the closer to normal, the bigger the samples the closer to normal.

Inferential Statistics: Fundamentals

The Central Limit Theorem

- Why is it useful? The CLT allows us to assume normality for many different variables. That is very useful for confidence intervals, hypothesis testing, and regression analysis. In fact, the Normal distribution is so predominantly observed around us due to the fact that following the CLT, many variables converge to Normal
- Where can we see it? Since many concepts and events are a sum or an average of different effects, CLT applies and we observe normality all the time. For example, in regression analysis, the dependent variable is explained through the sum of error terms.

Inferential Statistics: Fundamentals

Estimators

- Broadly, an estimator is a mathematical function that approximates a population parameter depending only on sample information Examples of estimators and the corresponding parameters:
  - Mean Estimator: $\bar{x}$
  - Variance Estimator: s^2
  - Correlation Estimator: r
- Estimators have two important properties
  - Bias: The expected value of an unbiased estimator is the population parameter. The bias in this case is 0. If the expected value of an estimator is (parameter + b), then the bias is b.
  - Efficiency: The most efficient estimator is the one with the variance.

Inferential Statistics: Fundamentals

Estimates

- An estimate is the output that you get from the estimator (when you apply the formula). There are two types of estimates: point estimates and confidence interval estimates.
  - Point estimates: A single value (1, 5, 122.67, 0.32)
  - Confidence intervals: An interval ([1,5], [12,33], [221.78, 745.66])
- Confidence intervals are much more precise than point estimates. That is why they are preferred when making inferences.

Inferential Statistics: Confidence Intervals

A confidence interval is an interval within which we are confident (with a certain percentage of confidence) the population parameter will fall. We build the confidence interval around the point estimate. $(1-\alpha)$ is the level of confidence. We are $(1-\alpha)*100\%$ confident that the population parameter will fall in the specificed interval. Common alphas are $0.01, 0.05, 0.1$.

General formula:

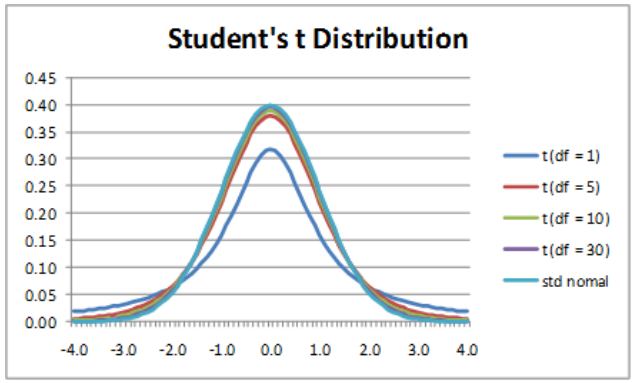
- $[\bar{x} - ME, \bar{x} + ME]$, where ME is the margin of error
- $ME =$ reliability factor $* \frac{\text{standard deviation}}{\sqrt{\text{sample size}}}$
  - $z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$
  - $t_{v,\alpha/2} * \frac{s}{\sqrt{n}}$

Inferential Statistics: Confidence Intervals

A student's T distribution is used predominantly for creating confidence intervals and testing hypotheses with normally distributed populations when the sample sizes are too small. It is particularly useful when we don't have enough information or it is too costly to obtain it.

All else equal, the student's T distribution has fatter tails than the Normal distribution and a lower peak. This is to reflect the higher level of uncertainty, caused by the small sample size.

A random variable following the t-distribution is denoted $t_{v,\alpha}$ where v are the degrees of freedom. We can obtain the student's T distribution for a variable with a Normally distributed population



using the formula: $t_{v,\alpha} = \frac{|barx - \mu}{s/\sqrt{n}}$

Inferential Statistics: Confidence Intervals

# populations	Population variance	Samples	Statistic	Variance	Formula
One	known	-	z	σ^2	$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
One	unknown	-	t	s^2	$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$
Two	-	dependent	t	$s^2_{difference}$	$\bar{d} \pm t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}}$
Two	Known	independent	z	σ_x^2 , σ_y^2	$(\bar{x} - \bar{y}) \pm z_{\alpha/2} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$
Two	unknown, assumed equal	independent	t	$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$	$(\bar{x} - \bar{y}) \pm t_{n_x+n_y-2,\alpha/2} \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$
Two	unknown, assumed different	independent	t	s_x^2 , s_y^2	$(\bar{x} - \bar{y}) \pm t_{\upsilon,\alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$

Inferential Statistics: Practical Example

You are a data analyst for Al Bundy's shoe shop. The firm sells mid to high end shoes ranging from 120 to 200 dollars. You have a lot of inventory that never sells.One way to solve this problem is to use confidence intervals. We have data on invoice no, date, country, product id, gender, shoe size, unit price, discount and sale price. Men and women shoes are different so should not be bundled together. Segment the data by shoe size, country, and gender. We can assume normality due to the Central Limit Theorem. If we wanted to estimate the number of shoes that are likely to be sold we can find the 95 percent confidence interval. We can segment the data to men's shoe sales in 2016. Once we find the mean and standard error for each shoe size over the course of the year, we can calculate out Confidence Interval. Within our confidence interval, we then know how many pairs to get.

Inferential Statistics: Practical Example

InvoiceNo	Date	Country	ProductID	Shop	Gender	Size (US)	Size (Europe)	Size (UK)	UnitPrice	Discount	SalePrice
52389	1/1/14	United Kingdom	2152	UK2	Male	11	44	10.5	$ 159.00	0%	$ 159.00
52390	1/1/14	United States	2230	US15	Male	11.5	44-45	11	$ 199.00	20%	$ 159.20
52391	1/1/14	Canada	2160	CAN7	Male	9.5	42-43	9	$ 149.00	20%	$ 119.20
52392	1/1/14	United States	2234	US6	Female	9.5	40	7.5	$ 159.00	0%	$ 159.00
52393	1/1/14	United Kingdom	2222	UK4	Female	9	39-40	7	$ 159.00	0%	$ 159.00
52394	1/1/14	United States	2173	US15	Male	10.5	43-44	10	$ 159.00	0%	$ 159.00
52395	1/2/14	Germany	2200	GER2	Female	9	39-40	7	$ 179.00	0%	$ 179.00
52396	1/2/14	Canada	2238	CAN5	Male	10	43	9.5	$ 169.00	0%	$ 169.00
52397	1/2/14	United States	2191	US13	Male	10.5	43-44	10	$ 139.00	0%	$ 139.00
52398	1/2/14	United Kingdom	2237	UK1	Female	9	39-40	7	$ 149.00	0%	$ 149.00
52399	1/2/14	United States	2197	US1	Male	10	43	9.5	$ 129.00	0%	$ 129.00
52399	1/2/14	United States	2213	US11	Female	9.5	40	7.5	$ 169.00	10%	$ 152.10
52399	1/2/14	United States	2206	US2	Female	9.5	40	7.5	$ 139.00	0%	$ 139.00

	United States, 2016												Mean	Standard error	ME	95% CI		Number of pairs
US	1	2	3	4	5	6	7	8	9	10	11	12	2016	2016	2016	2016		
6	4	1	3	1	3	3	3	4	3	7	3	0	2.92	0.51	1.13	1.78	4.05	4
6.5	3	2	0	1	0	0	1	7	2	1	2	1	1.67	0.56	1.22	0.45	2.89	3
7	0	0	1	0	6	4	4	2	3	0	0	0	1.67	0.61	1.34	0.33	3.00	3
7.5	3	2	3	1	7	0	7	3	4	6	1	1	3.17	0.69	1.53	1.64	4.70	5
8	7	9	7	3	12	2	9	4	7	5	2	6	6.08	0.88	1.94	4.14	8.03	8
8.5	12	12	8	8	15	9	17	17	6	9	10	6	10.75	1.12	2.47	8.28	13.22	13
9	17	13	13	11	21	22	25	30	26	25	13	10	18.83	1.97	4.33	14.50	23.17	23
9.5	19	25	27	24	26	33	25	47	31	44	37	26	30.33	2.45	5.39	24.95	35.72	36
10	17	26	26	19	16	31	25	24	23	31	15	20	22.75	1.57	3.45	19.30	26.20	26
10.5	13	16	22	14	28	19	18	15	19	21	16	10	17.58	1.37	3.01	14.57	20.59	21
11	5	16	13	10	10	11	15	8	9	7	6	7	9.75	1.01	2.22	7.53	11.97	12
11.5	4	3	6	3	3	5	6	4	5	12	13	5	5.75	0.96	2.12	3.63	7.87	8
12	3	0	0	4	4	4	3	12	4	9	2	1	3.83	1.01	2.23	1.60	6.06	6
13	1	1	2	0	3	2	1	0	0	4	3	2	1.58	0.38	0.83	0.75	2.42	2
14	2	6	3	3	5	3	2	1	0	1	2	1	2.42	0.50	1.10	1.32	3.52	4
15	0	0	0	1	1	0	4	0	0	0	0	2	0.67	0.36	0.78	-0.12	1.45	1
16	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0
Total	110	132	134	103	160	148	165	178	142	182	125	98						