home / math / sample size calculator This calculator computes the minimum number of necessary samples to meet the desired statistical constraints. Find Out the Margin of Error This calculator gives out the margin of error or confidence interval of observation or survey. RelatedStandard Deviation Calculator | Probability Calculator In statistics, information is often inferred about a population by studying a finite number of individuals from that population, i.e. the population is sampled, and it is assumed that characteristics of the sample are representative of the overall population. For the following, it is assumed that there is a population of individuals where some proportion, p, of the population is distinguishable from the other 1-p in some way; e.g., p may be the proportion of individuals who have brown hair, while the remaining 1-p have black, blond, red, etc. Thus, to estimate p in the population, a sample of n individuals could be taken from the population, and the sample proportion, p̂, calculated for sampled individuals who have brown hair. Unfortunately, unless the full population is sampled, the estimate p̂ most likely won't equal the true value p, since p̂ suffers from sampling noise, i.e. it depends on the particular individuals that were sampled. However, sampling statistics can be used to calculate what are called confidence intervals, which are an indication of how close the estimate p̂ is to the true value p. Statistics of a Random Sample The uncertainty in a given random sample (namely that is expected that the proportion estimate, p̂, is a good, but not perfect, approximation for the true proportion p) can be summarized by saying that the estimate p̂ is normally distributed with mean p and variance p(1-p)/n. For an explanation of why the sample estimate is normally distributed, study the Central Limit Theorem. As defined below, confidence level, confidence intervals, and sample sizes are all calculated with respect to this sampling distribution. In short, the confidence interval gives an interval around p in which an estimate p̂ is "likely" to be. The confidence level gives just how "likely" this is – e.g., a 95% confidence level indicates that it is expected that an estimate p̂ lies in the confidence interval for 95% of the random samples that could be taken. The confidence interval depends on the sample size, n (the variance of the sample distribution is inversely proportional to n, meaning that the estimate gets closer to the true proportion as n increases); thus, an acceptable error rate in the estimate can also be set, called the margin of error, ε, and solved for the sample size required for the chosen confidence interval to be smaller than e; a calculation known as "sample size calculation." Confidence Level The confidence level is a measure of certainty regarding how accurately a sample reflects the population being studied within a chosen confidence interval. The most commonly used confidence levels are 90%, 95%, and 99%, which each have their own corresponding z-scores (which can be found using an equation or widely available tables like the one provided below) based on the chosen confidence level. Note that using z-scores assumes that the sampling distribution is normally distributed, as described above in "Statistics of a Random Sample." Given that an experiment or survey is repeated many times, the confidence level essentially indicates the percentage of the time that the resulting interval found from repeated tests will contain the true result. Confidence Interval In statistics, a confidence interval is an estimated range of likely values for a population parameter, for example, 40 ± 2 or 40 ± 5%. Taking the commonly used 95% confidence level as an example, if the same population were sampled multiple times, and interval estimates made on each occasion, in approximately 95% of the cases, the true population parameter would be contained within the interval. Note that the 95% probability refers to the reliability of the estimation procedure and not to a specific interval. Once an interval is calculated, it either contains or does not contain the population parameter of interest. Some factors that affect the width of a confidence interval include: size of the sample, confidence level, and variability within the sample. There are different equations that can be used to calculate confidence intervals depending on factors such as whether the standard deviation is known or smaller samples (n where z is z score p̂ is the population proportion n and n' are sample size N is the population size Within statistics, a population is a set of events or elements that have some relevance regarding a given question or experiment. It can refer to an existing group of objects, systems, or even a hypothetical group of objects. Most commonly, however, population is used to refer to a group of people, whether they are the number of employees in a company, number of people within a certain age group of some geographic area, or number of students in a university's library at any given time. It is important to note that the equation needs to be adjusted when considering a finite population, as shown above. The (N-n)/(N-1) term in the finite population equation is referred to as the finite population correction factor, and is necessary because it cannot be assumed that all individuals in a sample are independent. For example, if the study population involves 10 people in a room with ages ranging from 1 to 100, and one of those chosen has an age of 100, the next person chosen is more likely to have a lower age. The finite population correction factor accounts for factors such as these. Refer below for an example of calculating a confidence interval with an unlimited population. EX: Given that 120 people work at Company Q, 85 of which drink coffee daily, find the 99% confidence interval of the true proportion of people who drink coffee at Company Q on a daily basis. Sample Size Calculation Sample size is a statistical concept that involves determining the number of observations or replicates (the repetition of an experimental condition used to estimate the variability of a phenomenon) that should be included in a statistical sample. It is an important aspect of any empirical study requiring that inferences be made about a population based on a sample. Essentially, sample sizes are used to represent parts of a population chosen for any given survey or experiment. To carry out this calculation, set the margin of error, ε, or the maximum distance desired for the sample estimate to deviate from the true value. To do this, use the confidence interval equation above, but set the term to the right of the ± sign equal to the margin of error, and solve for the resulting equation for sample size, n. The equation for calculating sample size is shown below. where z is the z score ε is the margin of error N is the population size p̂ is the population proportion EX: Determine the sample size necessary to estimate the proportion of people shopping at a supermarket in the U.S. that identify as vegan with 95% confidence, and a margin of error of 5%. Assume a population proportion of 0.5, and unlimited population size. Remember that z for a 95% confidence level is 1.96. Refer to the table provided in the confidence level section for z scores of a range of confidence levels. Thus, for the case above, a sample size of at least 385 people would be necessary. In the above example, some studies estimate that approximately 6% of the U.S. population identify as vegan, so rather than assuming 0.5 for p̂, 0.06 would be used. If it was known that 40 out of 500 people that entered a particular supermarket on a given day were vegan, p̂ would then be 0.08. From the description you provided, your first question is about the distribution of people's age. Normal (i.e. Gaussian) distribution applies to such kind of applications. It will be helpful if you know how the confidence interval (CI) was calculated. For instance, if the distribution is of normal distribution, and the CI was calculated using t-test, then the SD can be estimated with following equation: SD = sqrt(n)*(ci_upper - ci_lower)/(2 * tinv((1-CL)/2; n-1)), where CL is the confidence level, 'ci_upper' and 'ci_lower' are the upper and lower limits of CI respectively, and 'tinv()' is the inverse of Student's T cdf. Otherwise, if it is of normal distribution, but a known SD was used in calculating CI, then the SD can be calculated with following equation: SD = sqrt(n)*(ci_upper - ci_lower)/(sqrt(8) * erfinv(CL)), where 'erfinv ()' is the inverse error function. Your second question is about the distribution of people's sex (i.e. male or female). From the data you provided, it sounds that there are k=274 males among n=427 of whole samples. Bernoulli distribution applies to this application. In this case, the variance (of male's population) = p*(1-p) = 0.2299, and SD = sqrt(0.2299) = 0.4795, where p is the mean value. Note that "valiance = mean*(1-mean)" is applicable to Bernoulli distribution only. From the description you provided, your first question is about the distribution of people's age. Normal (i.e. Gaussian) distribution applies to such kind of applications. It will be helpful if you know how the confidence interval (CI) was calculated, because there are many different possible ways that the CI was calculated. For instance, if the distribution is of normal distribution, and the CI was calculated using t-test, then the SD can be estimated with following equation: SD = sqrt(n)*(ci_upper - ci_lower)/(2 * tinv((1-CL)/2; n-1)), where CL is the confidence level, 'ci_upper' and 'ci_lower' are the upper and lower limits of CI respectively, and 'tinv()' is the inverse of Student's T cdf. Otherwise, if it is of normal distribution, but a known SD was used in calculating CI, then the SD can be calculated with following equation: SD = sqrt(n)*(ci_upper - ci_lower)/(sqrt(8) * erfinv(CL)), where 'erfinv ()' is the inverse error function. Your second question is about the distribution of people's sex (i.e. male or female). From the data you provided, it sounds that there are k=274 males among n=427 of whole samples. Bernoulli distribution applies to this application. In this case, the variance (of male's population) = p*(1-p) = 0.2299, and SD = sqrt(0.2299) = 0.4795, where p is the mean value. Note that "valiance = mean*(1-mean)" is applicable to Bernoulli distribution only. Planning research, estimating parameters, and quantifying uncertainty are essential components of statistical analysis. While the theoretical principles are well-established, applying these concepts requires careful consideration and precise calculations. Statology's suite of calculators transforms complex statistical processes into accessible tools for researchers and analysts. These specialized calculators help users plan their research by determining appropriate sample sizes and precision levels. They guide the estimation of population parameters through various methods, from basic point estimates to sampling distributions. They also provide tools for quantifying uncertainty through confidence intervals, allowing researchers to measure precision and compare populations with clarity. Each calculator balances statistical rigor with straightforward implementation. From initial study design through analysis and comparison, these calculators deliver immediate, practical solutions for statistical challenges. Let's explore how each calculator supports these goals, starting with essential planning tools and progressing through increasingly sophisticated analytical methods. Sample Size Calculator for a Mean This calculator determines the exact sample size needed to estimate a population mean with desired confidence and precision. Based on your specified confidence level, margin of error, and population variability, it provides the minimum sample size required for reliable estimation. The tool simplifies sample size planning by requiring just three straightforward inputs: population standard deviation, confidence level, and margin of error. Calculates minimum required sample size using z-critical values Works with common confidence levels (e.g., 95%) Adapts to different precision requirements and population variability Sample Size Calculator for a Proportion This calculator determines the sample size needed to estimate a population proportion with specified confidence and precision. It provides flexibility by allowing you to either specify an expected proportion or use a conservative estimate of 0.5. The tool requires just three simple inputs: confidence level, margin of error, and expected proportion. Calculates required sample size for proportion studies Offers default 0.5 proportion for unknown populations Adapts to different confidence levels and precision needs Point Estimate Calculator This calculator determines the best point estimate for population parameters based on sample data. It computes multiple estimation methods (MLE, Wilson, Jeffrey, and Laplace) and automatically selects the most appropriate one based on your data characteristics. The tool provides a clear comparison of different estimation approaches while recommending the optimal choice. Calculates multiple types of point estimates Automatically selects the best estimation method Shows comparisons of different approaches Central Limit Theorem Calculator This calculator demonstrates how sample means behave according to the Central Limit Theorem. It calculates the mean and standard deviation of the sampling distribution using population parameters and sample size, helping illustrate key properties of sampling distributions. The tool shows how the standard error decreases as sample size increases. Illustrates sampling distribution properties Shows effect of sample size on precision Calculates both mean and standard error Confidence Interval for Mean Calculator This calculator constructs confidence intervals for a population mean using your sample statistics. It provides intervals at multiple confidence levels (90%, 95%, and 99%) simultaneously using the formula based on sample mean, standard deviation, and size. The tool helps you understand how different confidence levels affect the width of your intervals. Calculates intervals at multiple confidence levels Requires only basic sample statistics (mean, SD, size) Shows how confidence level impacts interval width Confidence Interval for Proportion Calculator This calculator creates confidence intervals for population proportions using sample data. It uses the standard formula with sample proportion and size to generate intervals at multiple confidence levels. The calculator requires just two inputs and provides a clear interpretation of the results. Calculates intervals for population proportions Provides plain-language interpretation of results Works with any confidence level specified Confidence Interval for Standard Deviation Calculator This calculator constructs confidence intervals for population standard deviations using chi-square distribution properties. It creates intervals based on your sample standard deviation and size, providing both the numerical interval and its interpretation. The tool offers a straightforward way to estimate population variability with a specified level of confidence. Calculates intervals for population standard deviation Provides clear interpretation of results Requires only sample size, standard deviation, and confidence level Confidence Interval for Correlation Coefficient Calculator This calculator creates confidence intervals for population correlation coefficients using sample data. It generates intervals that estimate the true strength of relationships between variables, providing both numerical results and plain-language interpretation. The calculator requires just three inputs: sample correlation coefficient, sample size, and desired confidence level. Estimates population correlation from sample data Provides interpreted results with confidence bounds Works with any sample size and correlation value Confidence Interval for the Difference Between Means Calculator This calculator constructs confidence intervals for comparing means between two groups. It computes intervals using sample statistics from both groups, incorporating a pooled variance approach for more precise estimation. The tool provides both numerical results and plain-language interpretation of the interval's meaning for comparing population means. Calculates intervals for mean differences between groups Uses pooled variance for improved precision Provides clear interpretation of results Tolerance Interval Calculator This calculator creates intervals that contain a specified proportion of a population with a given confidence level. Unlike confidence intervals, it helps predict where future observations will fall based on your sample data. The tool accepts raw data values, desired population proportion, and confidence level to generate appropriate bounds. Creates intervals for predicting future observations Allows custom population proportion settings Accepts comma-separated sample values directly Standard Error of the Proportion Calculator This calculator determines the standard error of a sample proportion, which measures the variability in proportion estimates. It uses a simple formula based on the sample proportion and size to quantify sampling uncertainty. The tool provides both the numerical result and shows the calculation steps for transparency. Calculates sampling variability for proportions Shows detailed calculation steps Requires only proportion and sample size Putting It All Together Statistical analysis progresses through distinct stages, with each calculator supporting specific analytical needs. This integrated approach ensures thorough and reliable results at every step. Planning Your Research Use sample size calculators to determine study requirements Apply point estimation tools to understand population parameters Set appropriate precision levels and confidence bounds Analyzing Your Data Generate confidence intervals for key parameters Compare groups using difference-based intervals Assess reliability through multiple estimation approaches Interpreting Results Combine insights from multiple interval types Consider both precision and practical significance Document uncertainty in your estimates Through this collection of calculators, Statology transforms complex statistical procedures into straightforward tools. This allows researchers to focus on understanding their results and making informed decisions rather than managing computational details. You can use this free sample size calculator to determine the sample size of a given survey per the sample proportion, margin of error, and required confidence level.You can calculate the sample size in five simple steps:Choose the required confidence level from the dropdown menuInput the margin of errorInput the proportion of the total population (%)If required, specify the population sizeClick on the "Calculate" button to generate the results.The sample size of a survey is the total number of complete responses that were received during the survey process. It is referred to as a sample because it does not include the full target population; it represents a selection of that population. For example, many studies involve random sampling by which a selection of a target population is randomly asked to complete a survey.Some basic terms are of interest when calculating sample size. These are as follows:Confidence level: The level of confidence of a sample is expressed as a percentage and describes the extent to which you can be sure it is representative of the target population; that is, how frequently the true percentage of the population who would select a response lies within the confidence interval. For example, if you have a confidence level of 90%, if you were to conduct the survey 100 times, the survey would yield the exact same results 90 times out of those 100 times.Margin of Error: Margin of error is also measured in percentage terms. It indicates the extent to which the outputs of the sample population are reflective of the overall population. The lower the margin of error, the nearer the researcher is to having an accurate response at a given confidence level. To determine the margin of error, take a look at our margin of error calculator.Percentage of population selecting a given choice: The accuracy of the research outputs also varies according to the percentage of the sample that chooses a given response. If 98% of the population select "Yes" and 2% select "No," there is a low chance of error. However, if 35% of the population select "Yes" and 65% select "No", there is a higher chance an error will be made, regardless of the sample size. When selecting the sample size required for a given level of accuracy, researchers should use the worst-case percentage; i.e., 50%.Population Size: The population size is the total number of people in the target population. For example, if you were performing research that was based on the people living in the UK, the full population would be approximately 66 million. Likewise, if you were conducting research on an organization, the total size of the population would be the number of employees who work for that organization.Sample Size FormulaThe Sample Size Calculator uses the following formulas:1. n = z2 * p * (1 - p) / e22. n (with finite population correction) = [z2 * p * (1 - p) / e2] / [1 + (z2 * p * (1 - p) / (e2 * N))]Where:n is the sample size,z is the z-score associated with a level of confidence,p is the sample proportion, expressed as a decimal,e is the margin of error, expressed as a decimal,N is the population size.Example of a Sample Size Calculation: Let's say we want to calculate the proportion of patients who have been discharged from a given hospital who are happy with the level of care they received while hospitalized at a 90% confidence level of the proportion within 4%. What sample size would we require?The sample size (n) can be calculated using the following formula:n = z2 * p * (1 - p) / e2where z = 1.645 for a confidence level (α) of 90%, p = proportion (expressed as a decimal), e = margin of error.z = 1.645, p = 0.5, e = 0.04n = 1.6452 * 0.5 * (1 - 0.5) / 0.042n = 0.6765 / 0.0016 = 422.816n ≈ 423 patients.Desired Confidence LevelZ-Score70%1.0475%1.1580%1.2885%1.4490%1.64591%1.7092%1.7593%1.8194%1.8895%1.9696%2.0597%2.1798%2.3399%2.57699.5%2.80799.9%3.2999.99%3.89Reference: Daniel WW (1999). Biostatistics: A Foundation for Analysis in the Health Sciences. 7th edition. New York: John Wiley & Sons.You may also be interested in our Effect Size (Cohen's d) Calculator or Relative Risk Calculator home / math / probability calculator To find out the union, intersection, and other related probabilities of two independent events. Please input values between 0 and 1. Please provide any 2 values below to calculate the rest probabilities of two independent events. Please input values between 0 and 1. Use the calculator below to find the area P shown in the normal distribution, as well as the confidence intervals for a range of confidence levels. RelatedStandard Deviation Calculator | Sample Size Calculator | Statistics Calculator Probability of Two Events Probability is the measure of the likelihood of an event occurring. It is quantified as a number between 0 and 1, with 1 signifying certainty, and 0 signifying that the event cannot occur. It follows that the higher the probability of an event, the more certain it is that the event will occur. In its most general case, probability can be defined numerically as the number of desired outcomes divided by the total number of outcomes. This is further affected by whether the events being studied are independent, mutually exclusive, or conditional, among other things. The calculator provided computes the probability that an event A or B does not occur, the probability A and/or B occur when they are not mutually exclusive, the probability that both event A and B occur, and the probability that either event A or event B occurs, but not both. Complement of A and B Given a probability A, denoted by P(A), it is simple to calculate the complement, or the probability that the event described by P(A) does not occur, P(A'). If, for example, P(A) = 0.65 represents the probability that Bob does not do his homework, his teacher Sally can predict the probability that Bob does his homework as follows: P(A') = 1 - P(A) = 1 - 0.65 = 0.35 Given this scenario, there is, therefore, a 35% chance that Bob does his homework. Any P(B') would be calculated in the same manner, and it is worth noting that in the calculator above, can be independent; i.e. if P(A) = 0.65, P(B) does not necessarily have to equal 0.35, and can equal 0.30 or some other number. Intersection of A and B The intersection of events A and B, written as P(A ∩ B) or P(A AND B) is the joint probability of at least two events, shown below in a Venn diagram. In the case where A and B are mutually exclusive events, P(A ∩ B) = 0. Consider the probability of rolling a 4 and 6 on a single roll of a die; it is not possible. These events would therefore be considered mutually exclusive. Computing P(A ∩ B) is simple if the events are independent. In this case, the probabilities of events A and B are multiplied. To find the probability that two separate rolls of a die result in 6 each time: The calculator provided considers the case where the probabilities are independent. Calculating the probability is slightly more involved when the events are dependent, and involves an understanding of conditional probability, or the probability of event A given that event B has occurred, P(A|B). Take the example of a bag of 10 marbles, 7 of which are black, and 3 of which are blue. Calculate the probability of drawing a black marble if a blue marble has been withdrawn without replacement (the blue marble is removed from the bag, reducing the total number of marbles in the bag): Probability of drawing a blue marble: P(A) = 3/10 Probability of drawing a black marble: P(B) = 7/10 Probability of drawing a black marble given that a blue marble was drawn: P(B|A) = 7/9 As can be seen, the probability that a black marble is drawn is affected by any previous event where a black or blue marble was drawn without replacement. Thus, if a person wanted to determine the probability of withdrawing a blue and then black marble from the bag: Probability of drawing a blue and then black marble using the probabilities calculated above: P(A ∩ B) = P(A) × P(B|A) = (3/10) × (7/9) = 0.2333 Union of A and B In probability, the union of events, P(A U B), essentially involves the condition where any or all of the events being considered occur, shown in the Venn diagram below. Note that P(A U B) can also be written as P(A OR B). In this case, the "inclusive OR" is being used. This means that while at least one of the conditions within the union must hold true, all conditions can be simultaneously true. There are two cases for the union of events; the events are either mutually exclusive, or the events are not mutually exclusive. In the case where the events are mutually exclusive, the calculation of the probability is simpler: A basic example of mutually exclusive events would be the rolling of a dice, where event A is the probability that an even number is rolled, and event B is the probability that an odd number is rolled. It is clear in this case that the events are mutually exclusive since a number cannot be both even and odd, so P(A U B) would be 3/6 + 3/6 = 1, since a standard dice only has odd and even numbers. The calculator above computes the other case, where the events A and B are not mutually exclusive. In this case: P(A U B) = P(A) + P(B) - P(A ∩ B) Using the example of rolling dice again, find the probability that an even number or a number that is a multiple of 3 is rolled. Here the set is represented by the 6 values of the dice, written as: S = {1,2,3,4,5,6} Probability of an even number:P(A) = {2,4,6} = 3/6 Probability of a multiple of 3:P(B) = {3,6} = 2/6 Intersection of A and B: P(A ∩ B) = {6} = 1/6 P(A U B) = 3/6 + 2/6 -1/6 = 2/3 Exclusive OR of A and B Another possible scenario that the calculator above computes is P(A XOR B), shown in the Venn diagram below. The "Exclusive OR" operation is defined as the event that A or B occurs, but not simultaneously. The equation is as follows: As an example, imagine it is Halloween, and two buckets of candy are set outside the house, one containing Snickers, and the other containing Reese's. Multiple flashing neon signs are placed around the buckets of candy insisting that each trick-or-treater only takes one Snickers OR Reese's but not both! It is unlikely, however, that every child adheres to the flashing neon signs. Given a probability of Reese's being chosen as P(A) = 0.65, or Snickers being chosen with P(B) = 0.349, and a P(unlikely) = 0.001 that a child exercises restraint while considering the detriments of a potential future cavity, calculate the probability that Snickers or Reese's is chosen, but not both: 0.65 + 0.349 - 2 × 0.65 × 0.349 = 0.999 - 0.4537 = 0.5453 Therefore, there is a 54.53% chance that Snickers or Reese's is chosen, but not both. Normal Distribution The normal distribution or Gaussian distribution is a continuous probability distribution that follows the function of: where μ is the mean and σ2 is the variance. Note that standard deviation is typically denoted as σ. Also, in the special case where μ = 0 and σ = 1, the distribution is referred to as a standard normal distribution. Above, along with the calculator, is a diagram of a typical normal distribution curve. The normal distribution is often used to describe and approximate any variable that tends to cluster around the mean, for example, the heights of male students in a college, the leaf sizes on a tree, the scores of a test, etc. Use the "Normal Distribution" calculator above to determine the probability of an event with a normal distribution lying between two given values (i.e. P in the diagram above); for example, the probability of the height of a male student is between 5 and 6 feet in a college. Finding P as shown in the above diagram involves standardizing the two desired values to a z-score by subtracting the given mean and dividing by the standard deviation, as well as using a Z-table to find probabilities for Z. If, for example, it is desired to find the probability that a student at a university has a height between 60 inches and 72 inches tall given a mean of 68 inches tall with a standard deviation of 4 inches, 60 and 72 inches would be standardized as such: Given μ = 68; σ = 4 (60 - 68)/4 = -8/4 = -2(72 - 68)/4 = 4/4 = 1 The graph above illustrates the area of interest in the normal distribution. In order to determine the probability represented by the shaded area of the graph, use the standard normal Z-table provided at the bottom of the page. Note that there are different types of standard normal Z-tables. The table below provides the probability that a statistic is between 0 and Z, where 0 is the mean in the standard normal distribution. There are also Z-tables that provide the probabilities left or right of Z, both of which can be used to calculate the desired probability by subtracting the relevant values. For this example, to determine the probability of a value between 0 and 2, find 2 in the first column of the table, since this table by definition provides probabilities between the mean (which is 0 in the standard normal distribution) and the number of choices, in this case, 2. Note that since the value in question is 2.0, the table is read by lining up the 2 row with the 0 column, and reading the value therein. If, instead, the value in question were 2.11, the 2.1 row would be matched with the 0.01 column and the value would be 0.48257. Also, note that even though the actual value of interest is -2 on the graph, the table only provides positive values. Since the normal distribution is symmetrical, only the displacement is important, and a displacement of 0 to -2 or 0 to 2 is the same, and will have the same area under the curve. Thus, the probability of a value falling between 0 and 2 is 0.47725 , while a value between 0 and 1 has a probability of 0.34134. Since the desired area is between -2 and 1, the probabilities are added to yield 0.81859, or approximately 81.859%. Returning to the example, this means that there is an 81.859% chance in this case that a male student at the given university has a height between 60 and 72 inches. The calculator also provides a table of confidence intervals for various confidence levels. Refer to the Sample Size Calculator for Proportions for a more detailed explanation of confidence intervals and levels. Briefly, a confidence interval is a way of estimating a population parameter that provides an interval of the parameter rather than a single value. A confidence interval is always qualified by a confidence level, usually expressed as a percentage such as 95%. It is an indicator of the reliability of the estimate. Z Table from Mean (0 to Z) Calculate the minimum sample size required to estimate a population parameter. Usually we have no control over the sample size of a data set. However, if we are able to set the sample size, as in cases where we are taking a survey, it is very helpful to know just how large it should be to provide the most information. Sampling can be very costly, in both time and product. Simple telephone surveys will cost approximately $30.00 each, for example, and some sampling requires the destruction of the product. Selecting a sample that is too large is expensive and time consuming. But selecting a sample that is too small can lead to inaccurate conclusions. We want to find the minimum sample size required to achieve the desired level of accuracy in the confidence interval. Calculating the Sample Size for a Population Mean The margin of error [latex]E[/latex] for a confidence interval for a population mean is [latex]\displaystyle{E=\frac{z \times \sigma}{\sqrt{n}}}[/latex] where [latex]z[/latex] is the [latex]z[/latex]-score so that the area under the standard normal distribution in between [latex]-z[/latex] and [latex]z[/latex] is the confidence level [latex]C[/latex]. Rearranging this formula for [latex]n[/latex] we get a formula for the sample size [latex]n[/latex]: [latex]\displaystyle{n=\left(\frac{z \times \sigma}{E}\right)^2}[/latex] In order to use this formula, we need values for [latex]z[/latex], [latex]E[/latex] and [latex]\sigma[/latex]: The value for [latex]z[/latex] is determined by the confidence level of the interval, calculated the same way we calculate the [latex]z[/latex]-score for a confidence interval. The value for the margin of error [latex]E[/latex] is set as the predetermined acceptable error, or tolerance, for the difference between the sample mean [latex]\overline{x}[/latex] and the population mean [latex]\mu[/latex]. In other words, [latex]E[/latex] is set to the maximum allowable width of the confidence interval. An estimate for the population standard deviation [latex]\sigma[/latex] can be found by one of the following methods: Conduct a small pilot study and use the sample standard deviation from the pilot study. Use the sample standard deviation from previously collected data. Although crude, this method of estimating the standard deviation may help reduce costs significantly. Use [latex]\displaystyle{\frac{\mbox{Range}}{4}}[/latex] where [latex]\mbox{Range}[/latex] is the difference between the maximum and minimum values of the population under study. Although we do not know the population standard deviation when calculating the sample size, we do not use the [latex]t[/latex]-distribution in the sample size formula. In order to use the [latex]t[/latex]-distribution in this situation, we need the degrees of freedom [latex]n-1[/latex]. But [latex]n[/latex] is the sample size we are trying to estimate. So, we must use the normal distribution to determine the sample size. The value of [latex]n[/latex] determined from the formula is the minimum sample size required to achieve the desired level of confidence. The sample size [latex]n[/latex] is a count, and so is an integer. It would be unusual for the value of [latex]n[/latex] generated by the formula to be an integer. Because [latex]n[/latex] is the minimum sample size required, we must round the output from the formula up to the next integer. If we round the value of [latex]n[/latex] down, the sample size will be below the minimum required sample size. After we have found the sample size [latex]n[/latex] and collected the data for the sample, we use the appropriate confidence interval formula and the sample standard deviation from the actual sample (assuming [latex]\sigma[/latex] is unknown), and not the estimate of the standard deviation used in the calculation of the sample size. To find the [latex]z[/latex]-score to calculate the sample size for a confidence interval with confidence level [latex]C[/latex], use the norm.s.inv(area to the left of z) function. For area to the left of z, enter the entire area to the left of [latex]z[/latex]-score you are trying to find. For a confidence interval, the area to the left of [latex]z[/latex] is [latex]\displaystyle{C+\frac{1-C}{2}}[/latex]. The output from the norm.s.inv function is the value of [latex]z[/latex]-score needed to find the sample size. We want to estimate the mean age of Foothill College students. From previous information, an estimate of the standard deviation of the ages of the students is 15 years. We want to be 95% confident that the sample mean age is within two years of the population mean age. How many randomly selected Foothill College students must be surveyed to achieved the desired level of accuracy? Solution: To find the sample size, we need to find the [latex]z[/latex]-score for the 95% confidence interval. This means that we need to find the [latex]z[/latex]-score so that the entire area to the left of [latex]z[/latex] is [latex]\displaystyle{0.95+\frac{1-0.95}{2}=0.975}[/latex]. Function norm.s.inv Answer Field 1 0.975 1.9599... So [latex]z=1.9599....[/latex]. From the question [latex]\sigma \simeq 15[/latex] and [latex]E=2[/latex]. [latex]\begin{eqnarray*}\&=&\left(\frac{z \times\sigma}{E}\right)^2\\&=&\left(\frac{1.9599...\times 15}{2}\right)^2\\&=&216.08...\\&\Rightarrow&217\mbox{ students}\\\\end{eqnarray*}[/latex] 217 students must be surveyed to achieve the desired accuracy. Remember to round the value for the sample size UP to the next integer. This ensures that the sample size is an integer and is large enough. Do not forget to include appropriate units with the sample size. You want to estimate the height of all high school basketball players. You want to be 98% confident with a margin of error of 1.5. From a small pilot study, you estimate the standard deviation to be 3 inches. How large a sample do you need to take to achieve the desired level of accuracy? Click to see Solution Function norm.s.inv Answer Field 1 0.99 2.3263... [latex]\begin{eqnarray*} n & = & \left(\frac{z \times \sigma}{E}\right)^2 \\ & = & \left(\frac{2.3263... \times 3}{1.5}\right)^2 \\ & = & 21.6487... \\ & \Rightarrow & 22 \mbox{ high school basketball players}\end{eqnarray*}[/latex] Calculating the Sample Size for a Population Proportion The margin of error [latex]E[/latex] for a confidence interval for a population proportion is [latex]\displaystyle{E=z \times \sqrt{\frac{p \times (1-p)}{n}}}[/latex] where [latex]z[/latex] is the [latex]z[/latex]-score so that the area under the standard normal distribution in between [latex]-z[/latex] and [latex]z[/latex] is the confidence level [latex]C[/latex]. Rearranging this formula for [latex]n[/latex] we get a formula for the sample size [latex]n[/latex]: [latex]\displaystyle{n=p \times (1-p) \times \left(\frac{z}{E}\right)^2}[/latex] In order to use this formula, we need values for [latex]z[/latex], [latex]E[/latex] and [latex]p[/latex]: The value for [latex]z[/latex] is determined by the confidence level of the interval, calculated the same way we calculate the [latex]z[/latex]-score for a confidence interval. The value for the margin of error [latex]E[/latex] is set as the predetermined acceptable error, or tolerance, for the difference between the sample proportion [latex]\hat{p}[/latex] and the population proportion [latex]p[/latex]. In other words, [latex]E[/latex] is set to the maximum allowable width of the confidence interval. An estimate for the population proportion [latex]p[/latex]. If no estimate for the population proportion is provided, we use [latex]p=0.5[/latex]. The value of [latex]n[/latex] determined from the formula is the minimum sample size required to achieve the desired level of confidence. The sample size [latex]n[/latex] is a count, and so is an integer. It would be unusual for the value of [latex]n[/latex] generated by the formula to be an integer. Because [latex]n[/latex] is the minimum sample size required, we must round the output from the formula up to the next integer. If we round the value of [latex]n[/latex] down, the sample size will be below the minimum required sample size. After we have found the sample size [latex]n[/latex] and collected the data for the sample, we use the appropriate confidence interval formula and the sample proportion from the actual sample. By using [latex]0.5[/latex] as an estimate for [latex]p[/latex] in the sample size formula we will get the largest required sample size for the confidence level and margin of error we selected. This is true because of all combinations of two fractions (the values of [latex]p[/latex] and [latex]1-p[/latex]) that add to one, the largest multiple is when each is 0.5. Without any other information concerning the population parameter [latex]p[/latex], this is the common practice. This may result in oversampling, but certainly not under sampling. There is an interesting trade-off between the level of confidence and the sample size that shows up here when considering the cost of sampling. The table below shows the appropriate sample size at different levels of confidence and different margins of error, assuming [latex]p=0.5[/latex]. Looking at each row, we can see that for the same margin of error, a higher level of confidence requires a larger sample size. Similarly, looking at each column, we can see that for the same level of confidence, a smaller margin of error requires a larger sample size. Required Sample Size (90%) Required Sample Size (95%) Margin of Error 1691 2401 2% 752 1067 3% 271 384 5% 68 96 10% Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ who use text messaging on their cell phones. How many customers aged 50+ should the company survey in order to be 90% confident that the estimated (sample) proportion is within 3 percentage points of the true population proportion of customers aged 50+ who use text messaging on their cell phones? Solution: To find the sample size, we need to find the [latex]z[/latex]-score for the 90% confidence interval. This means that we need to find the [latex]z[/latex]-score so that the entire area to the left of [latex]z[/latex] is [latex]\displaystyle{0.90+\frac{1-0.90}{2}=0.95}[/latex]. Function norm.s.inv Answer Field 1 0.95 1.6448... So [latex]z=1.6448....[/latex]. From the question [latex]E=0.03[/latex]. Because no estimate of the population proportion is given, [latex]p=0.5[/latex]. [latex]\begin{eqnarray*}\&=&p\times(1-p)\times\left(\frac{z }{E}\right)^2\\&=&0.5\times(1-0.5)\times\left(\frac{1.6448...}{0.03}\right)^2\\&=&751.539...\\&\Rightarrow&752\mbox{ customers aged 50+}\\\\end{eqnarray*}[/latex] 752 customers aged 50+ must be surveyed to achieve the desired accuracy. Remember to round the value for the sample size UP to the next integer. This ensures that the sample size is large enough. Do not forget to include appropriate units with the sample size. Suppose an internet marketing company wants to determine the current percentage of customers who click on ads on their smartphones. How many customers should the company survey in order to be 94% confident that the estimated proportion is within 5 percentage points of the true population proportion of customers who click on ads on their smartphones? Click to see Solution Function norm.s.inv Answer Field 1 0.97 1.8807... [latex]\begin{eqnarray*} n & = & p \times (1-p) \times \left(\frac{z}{E}\right)^2 \\ & = & 0.5 \times (1-0.5) \times \left(\frac{1.8807...}{0.05}\right)^2 \\ & = & 353.738... \\ & \Rightarrow & 354 \mbox{ customers}\end{eqnarray*}[/latex] Watch this video: Sample Size for Confidence Intervals by ExcelIsFun [7:54] Concept Review In order to construct a confidence interval, a sample is taken from the population under study. But collecting sample information is time consuming and expensive. The minimum sample size required to achieve the desired level of accuracy is determined before collecting the sample data. Sample size for population means: [latex]\displaystyle{n=\left(\frac{z \times \sigma}{E}\right)^2}[/latex] Sample size for population proportions: [latex]\displaystyle{n=p \times (1-p) \times \left(\frac{z}{E}\right)^2}[/latex] After calculating the value of [latex]n[/latex] from the formula, round the value of [latex]n[/latex] up to the next integer. Attribution "7.2 The Central Limit Theorem for Sums" in Introductory Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License. "8.4 Calculating the Sample Size n: Continuous and Binary Random Variables" in Introductory Business Statistics by OpenStax is licensed under a Creative Commons Attribution 4.0 International License. home / math / sample size calculator This calculator computes the minimum number of necessary samples to meet the desired statistical constraints. Find Out the Margin of Error This calculator gives out the margin of error or confidence interval of observation or survey. RelatedStandard Deviation Calculator | Probability Calculator In statistics, information is often inferred about a population by studying a finite number of individuals from that population, i.e. the population is sampled, and it is assumed that characteristics of the sample are representative of the overall population. For the following, it is assumed that there is a population of individuals where some proportion, p, of the population is distinguishable from the other 1-p in some way; e.g., p may be the proportion of individuals who have brown hair, while the remaining 1-p have black, blond, red, etc. Thus, to estimate p in the population, a sample of n individuals could be taken from the population, and the sample proportion, p̂, calculated for sampled individuals who have brown hair. Unfortunately, unless the full population is sampled, the estimate p̂ most likely won't equal the true value p, since p̂ suffers from sampling noise, i.e. it depends on the particular individuals that were sampled. However, sampling statistics can be used to calculate what are called confidence intervals, which are an indication of how close the estimate p̂ is to the true value p. Statistics of a Random Sample The uncertainty in a given random sample (namely that is expected that the proportion estimate, p̂, is a good, but not perfect, approximation for the true proportion p) can be summarized by saying that the estimate p̂ is normally distributed with mean p and variance p(1-p)/n. For an explanation of why the sample estimate is normally distributed, study the Central Limit Theorem. As defined below, confidence level, confidence intervals, and sample sizes are all calculated with respect to this sampling distribution. In short, the confidence interval gives an interval around p in which an estimate p̂ is "likely" to be. The confidence level gives just how "likely" this is – e.g., a 95% confidence level indicates that it is expected that an estimate p̂ lies in the confidence interval for 95% of the random samples that could be taken. The confidence interval depends on the sample size, n (the variance of the sample distribution is inversely proportional to n, meaning that the estimate gets closer to the true proportion as n increases); thus, an acceptable error rate in the estimate can also be set, called the margin of error, ε, and solved for the sample size required for the chosen confidence interval to be smaller than e; a calculation known as "sample size calculation." Confidence Level The confidence level is a measure of certainty regarding how accurately a sample reflects the population being studied within a chosen confidence interval. The most commonly used confidence levels are 90%, 95%, and 99%, which each have their own corresponding z-scores (which can be found using an equation or widely available tables like the one provided below) based on the chosen confidence level. Note that using z-scores assumes that the sampling distribution is normally distributed, as described above in "Statistics of a Random Sample." Given that an experiment or survey is repeated many times, the confidence level essentially indicates the percentage of the time that the resulting interval found from repeated tests will contain the true result. Confidence Interval In statistics, a confidence interval is an estimated range of likely values for a population parameter, for example, 40 ± 2 or 40 ± 5%. Taking the commonly used 95% confidence level as an example, if the same population were sampled multiple times, and interval estimates made on each occasion, in approximately 95% of the cases, the true population parameter would be contained within the interval. Note that the 95% probability refers to the reliability of the estimation procedure and not to a specific interval. Once an interval is calculated, it either contains or does not contain the population parameter of interest. Some factors that affect the width of a confidence interval include: size of the sample, confidence level, and variability within the sample. There are different equations that can be used to calculate confidence intervals depending on factors such as whether the standard deviation is known or smaller samples (n where z is z score p̂ is the population proportion n and n' are sample size N is the population size Within statistics, a population is a set of events or elements that have some relevance regarding a given question or experiment. It can refer to an existing group of objects, systems, or even a hypothetical group of objects. Most commonly, however, population is used to refer to a group of people, whether they are the number of employees in a company, number of people within a certain age group of some geographic area, or number of students in a university's library at any given time. It is important to note that the equation needs to be adjusted when considering a finite population, as shown above. The (N-n)/(N-1) term in the finite population equation is referred to as the finite population correction factor, and is necessary because it cannot be assumed that all individuals in a sample are independent. For example, if the study population involves 10 people in a room with ages ranging from 1 to 100, and one of those chosen has an age of 100, the next person chosen is more likely to have a lower age. The finite population correction factor accounts for factors such as these. Refer below for an example of calculating a confidence interval with an unlimited population. EX: Given that 120 people work at Company Q, 85 of which drink coffee daily, find the 99% confidence interval of the true proportion of people who drink coffee at Company Q on a daily basis. Sample Size Calculation Sample size is a statistical concept that involves determining the number of observations or replicates (the repetition of an experimental condition used to estimate the variability of a phenomenon) that should be included in a statistical sample. It is an important aspect of any empirical study requiring that inferences be made about a population based on a sample. Essentially, sample sizes are used to represent parts of a population chosen for any given survey or experiment. To carry out this calculation, set the margin of error, ε, or the maximum distance desired for the sample estimate to deviate from the true value. To do this, use the confidence interval equation above, but set the term to the right of the ± sign equal to the margin of error, and solve for the resulting equation for sample size, n. The equation for calculating sample size is shown below. where z is the z score ε is the margin of error N is the population size p̂ is the population proportion EX: Determine the sample size necessary to estimate the proportion of people shopping at a supermarket in the U.S. that identify as vegan with 95% confidence, and a margin of error of 5%. Assume a population proportion of 0.5, and unlimited population size. Remember that z for a 95% confidence level is 1.96. Refer to the table provided in the confidence level section for z scores of a range of confidence levels. Thus, for the case above, a sample size of at least 385 people would be necessary. In the above example, some studies estimate that approximately 6% of the U.S. population identify as vegan, so rather than assuming 0.5 for p̂, 0.06 would be used. If it was known that 40 out of 500 people that entered a particular supermarket on a given day were vegan, p̂ would then be 0.08. When researchers are conducting public opinion polls, they calculate the required sample size based on how precise they want their estimates to be. The sample size is determined by the confidence level, expected proportion and confidence interval needed for the survey. The confidence interval represents the margin of error in the results. For example, if a poll with a confidence interval of plus or minus 3 percentage points showed 56 percent of the people supported a candidate, the true proportion would probably be between 53 and 59 percent. Square the Z-score required for your desired confidence level. For example, if you used a 95 percent confidence level, meaning that you can say with 95 percent certainty that the proportion will fall in your confidence interval, your Z-score would be 1.96, so you would multiply 1.96 times 1.96 to get 3.8416. Estimate the proportion of the largest group. If you are unsure, use 0.5 as the expected proportion because the closer the two proportions, larger the sample size you will need. For example, if you expected 60 percent of people to vote for the incumbent, you would use 0.6. Subtract the expected proportion from 1. Continuing the example, you would subtract 0.6 from 1 to get 0.4. Multiply the result from Step 3 by the proportion from Step 2. In this example, you would multiply 0.4 times 0.6 to get 0.24. Multiply the result from Step 4 by the result from Step 1. Continuing the example, you would multiply 3.8416 by 0.24 to get 0.921984. Square the confidence interval, expressed as a decimal, for your survey. For example, if your confidence interval equals plus or minus 2 percentage points, you would square 0.02 to get 0.0004. Divide the result from step 5 by the confidence interval to calculate the required sample size. In this example, you would divide 0.921984 by 0.0004 to get 2,304.96, meaning that you would need a sample size of 2,305 people for your survey. Kennan, Mark. "How To Calculate Sample Size From A Confidence Interval" sciencing.com, . 24 April 2017. APA Kennan, Mark. (2017, April 24). How To Calculate Sample Size From A Confidence Interval. sciencing.com. Retrieved from Chicago Kennan, Mark. How To Calculate Sample Size From A Confidence Interval last modified March 24, 2022. home / math / sample size calculator This calculator computes the minimum number of necessary samples to meet the desired statistical constraints. Find Out the Margin of Error This calculator gives out the margin of error or confidence interval of observation or survey. RelatedStandard Deviation Calculator | Probability Calculator In statistics, information is often inferred about a population by studying a finite number of individuals from that population, i.e. the population is sampled, and it is assumed that characteristics of the sample are representative of the overall population. For the following, it is assumed that there is a population of individuals where some proportion, p, of the population is distinguishable from the other 1-p in some way; e.g., p may be the proportion of individuals who have brown hair, while the remaining 1-p have black, blond, red, etc. Thus, to estimate p in the population, a sample of n individuals could be taken from the population, and the sample proportion, p̂, calculated for sampled individuals who have brown hair. Unfortunately, unless the full population is sampled, the estimate p̂ most likely won't equal the true value p, since p̂ suffers from sampling noise, i.e. it depends on the particular individuals that were sampled. However, sampling statistics can be used to calculate what are called confidence intervals, which are an indication of how close the estimate p̂ is to the true value p. Statistics of a Random Sample The uncertainty in a given random sample (namely that is expected that the proportion estimate, p̂, is a good, but not perfect, approximation for the true proportion p) can be summarized by saying that the estimate p̂ is normally distributed with mean p and variance p(1-p)/n. For an explanation of why the sample estimate is normally distributed, study the Central Limit Theorem. As defined below, confidence level, confidence intervals, and sample sizes are all calculated with respect to this sampling distribution. In short, the confidence interval gives an interval around p in which an estimate p̂ is "likely" to be. The confidence level gives just how "likely" this is – e.g., a 95% confidence level indicates that it is expected that an estimate p̂ lies in the confidence interval for 95% of the random samples that could be taken. The confidence interval depends on the sample size, n (the variance of the sample distribution is inversely proportional to n, meaning that the estimate gets closer to the true proportion as n increases); thus, an acceptable error rate in the estimate can also be set, called the margin of error, ε, and solved for the sample size required for the chosen confidence interval to be smaller than e; a calculation known as "sample size calculation." Confidence Level The confidence level is a measure of certainty regarding how accurately a sample reflects the population being studied within a chosen confidence interval. The most commonly used confidence levels are 90%, 95%, and 99%, which each have their own corresponding z-scores (which can be found using an equation or widely available tables like the one provided below) based on the chosen confidence level. Note that using z-scores assumes that the sampling distribution is normally distributed, as described above in "Statistics of a Random Sample." Given that an experiment or survey is repeated many times, the confidence level essentially indicates the percentage of the time that the resulting interval found from repeated tests will contain the true result. Confidence Interval In statistics, a confidence interval is an estimated range of likely values for a population parameter, for example, 40 ± 2 or 40 ± 5%. Taking the commonly used

95% confidence level as an example, if the same population were sampled multiple times, and interval estimates made on each occasion, in approximately 95% of the cases, the true population parameter would be contained within the interval. Note that the 95% probability refers to the reliability of the estimation procedure and not to a specific interval. Once an interval is calculated, it either contains or does not contain the population parameter of interest. Some factors that affect the width of a confidence interval include: size of the sample, confidence level, and variability within the sample. There are different equations that can be used to calculate confidence intervals depending on factors such as whether the standard deviation is known or smaller samples (n where z is z score $\hat{p}$ is the population proportion n and n' are sample size N is the population size Within statistics, a population is a set of events or elements that have some relevance regarding a given question or experiment. It can refer to an existing group of objects, systems, or even a hypothetical group of objects. Most commonly, however, population is used to refer to a group of people, whether they are the number of employees in a company, number of people within a certain age group of some geographic area, or number of students in a university's library at any given time. It is important to note that the equation needs to be adjusted when considering a finite population, as shown above. The (N-n)/(N-1) term in the finite population equation is referred to as the finite population correction factor, and is necessary because it cannot be assumed that all individuals in a sample are independent. For example, if the study population involves 10 people in a room with ages ranging from 1 to 100, and one of those chosen has an age of 100, the next person chosen is more likely to have a lower age. The finite population correction factor accounts for factors such as these. Refer below for an example of calculating a confidence interval with an unlimited population. EX: Given that 120 people work at Company Q, 85 of which drink coffee daily, find the 99% confidence interval of the true proportion of people who drink coffee at Company Q on a daily basis. Sample Size Calculation Sample size is a statistical concept that involves determining the number of observations or replicates (the repetition of an experimental condition used to estimate the variability of a phenomenon) that should be included in a statistical sample. It is an important aspect of any empirical study requiring that inferences be made about a population based on a sample. Essentially, sample sizes are used to represent parts of a population chosen for any given survey or experiment. To carry out this calculation, set the margin of error, ε, or the maximum distance desired for the sample estimate to deviate from the true value. To do this, use the confidence interval equation above, but set the term to the right of the ± sign equal to the margin of error, and solve for the resulting equation for sample size, n. The equation for calculating sample size is shown below. where z is the z score ε is the margin of error N is the population size $\hat{p}$ is the population proportion EX: Determine the sample size necessary to estimate the proportion of people shopping at a supermarket in the U.S. that identify as vegan with 95% confidence, and a margin of error of 5%. Assume a population proportion of 0.5, and unlimited population size. Remember that z for a 95% confidence level is 1.96. Refer to the table provided in the confidence level section for z scores of a range of confidence levels. Thus, for the case above, a sample size of at least 385 people would be necessary. In the above example, some studies estimate that approximately 6% of the U.S. population identify as vegan, so rather than assuming 0.5 for $\hat{p}$, 0.06 would be used. If it was known that 40 out of 500 people that entered a particular supermarket on a given day were vegan, $\hat{p}$ would then be 0.08. Copyright © 2007–2025 by Stan Brown, BrownMath.com Contents: If you know the standard deviation σ of the population, and you want to estimate the mean μ to within a given margin of error E in a 1−α confidence interval, here's how to find the required sample size n: $E = z_{\alpha/2} \cdot \sigma/\sqrt{n}$ transforms to $n = (z_{\alpha/2} \cdot \sigma/E)^2$ Example 1: You want to estimate the average hourly output of a machine to within ±1.5, with 90% confidence. Based on historical data, you have reason to believe that the standard deviation of the machine's hourly output is 6.2. How large a sample do you need? Solution: Note first that this is not a realistic situation. It's pretty unlikely that you would know the standard deviation of a population but not know the mean of that population. However, statistics texts always begin with this case because it's the simplest way to demonstrate the principles. You leave Perfectland and enter Realityville in the other cases. With that said— CommentsComputation It's good practice to start any problem by writing down what you know and what you need, with symbols. Given: E = 1.5, σ = 6.2, 1−α = 0.90. Wanted: sample size n The formula wants $z_{\alpha/2}$. How do you compute it? Begin by finding α/2. 1−α = 0.90 ⇒ α = 0.10 ⇒ α/2 = 0.05 Since α/2 = 0.05, $z_{\alpha/2} = z_{0.05}$ z_{rtail} is the critical z, or the z score that divides the normal curve leaving a right-hand tail with an area of rtail. You compute it on your TI-83/84/89 as invNorm(1−rtail). $z_{0.05}$ = invNorm(1−0.05) ≈ 1.6449 Now you have all the pieces. Don't use the rounded value of $z_{\alpha/2}$, but use [2nd (-) makes ANS] to keep full precision. $n = [\, z_{\alpha/2} \times \sigma \div E \,]^2 = (ANS\times 6.2\div 1.5)^2 = 46.2227\ldots \to 47$ Answer: Given a population standard deviation of 6.2 units per hour, if you have a sample size ≥47 the margin of error in a 90% confidence interval will be ≤1.5 units per hour. Why do we round up? After computing 46.2227, why not report a sample size of 46? Well, the computation shows that a sample size of exactly 46.2227… would give a margin of error of exactly 1.5. If you go slightly lower, to 46, the margin of error will be slightly higher than 1.5. Since the sample size must be a whole number, 46 or 47, and your margin of error must not exceed 1.5, you have to choose the slightly higher number 47, which will give a margin of error slightly less than 1.5. Case 1: One Population Mean, Unknown σ Note: Many basic statistics courses skip the material in this section and estimate sample sizes using a z distribution, so the material in this section might be an advanced extra for you. Check your course requirements. This is the realistic case for estimating a population mean. Usually you don't know the standard deviation of the population, so you have to use Student's t distribution instead of the normal (z) distribution. You estimate the standard deviation of the population from the standard deviation of a sample obtained in a prior study or a small pilot study. Here is the formula for sample size: $E = t_{df,\alpha/2} \cdot s/\sqrt{n}$ transforms to $n = t_{df,\alpha/2} \cdot s/E)^2$ There's a certain element of Catch-22 in this formula for n. You don't know n, so you don't know the degrees of freedom df either and you can't compute the critical t for the formula. How do you get around this? Use what NIST/SEMATECH calls an iterative method. First compute the formula using $z_{\alpha/2}$ instead of $t_{df,\alpha/2}$. Then, when you have a preliminary sample size determined by (ab)using z in this way, recompute the formula using that sample size minus 1 for df. The two numbers should not be very different, since t is generally not very different from z; but if they are, you can use the second number to compute t once again. Caution: It may happen that the formula gives a sample size less than 30. But remember that your sample size must always be at least 30 unless you have good reason to believe that the underlying population is normally distributed. See also: Sample Sizes Required in NIST/SEMATECH e-Handbook of Statistical Methods: scroll down to "More often we must compute the sample size with the population standard deviation being unknown" Let's illustrate the method using a modified form of the previous example. Example 2: You want to estimate the average hourly output of a machine to within ±1.5, with 90% confidence. A small pilot study finds a sample standard deviation of the machine's hourly output is 6.2. How large a sample do you need? Solution: Use z instead of t to make a preliminary estimate, then recompute with t. CommentsComputation Marshal your data. Given: E = 1.5, s = 6.2, 1−α = 0.90. Wanted: sample size n The formula wants $t_{df,\alpha/2}$, but we approximate with $z_{\alpha/2}$. Begin by finding α/2. 1−α = 0.90 ⇒ α = 0.10 ⇒ α/2 = 0.05 Since α/2 = 0.05, $z_{\alpha/2} = z_{0.05}$ $z_{0.05}$ is the critical z-score that divides the normal distribution such that the area of the right-hand tail is 0.05, and therefore the area of the left-hand tail is 1−0.05. $z_{0.05}$ = invNorm(1−0.05) ≈ 1.6449 Now you have all the pieces you need for the preliminary sample size. Don't use the rounded value of $z_{\alpha/2}$, but use [2nd (-) makes ANS] to keep full precision. $n = [\, z_{\alpha/2} \times s \div E \,]^2 = (ANS\times 6.2\div 1.5)^2 = 46.2227\ldots \to 47$ Your preliminary sample size is 47, and next you use that to compute t. df = n−1 = 46, so you need $t_{46,0.05}$. On the TI-84 you can use the invT function. $t_{46,0.05}$ = 1.67866 Now recompute the formula using the t value. Remember, always round sample size up. $n = [\, t_{df,\alpha/2} \times s \div E \,]^2 = (ANS\times 6.2\div 1.5)^2 = 48.142\ldots \to 49$ Answer: Given a sample standard deviation of 6.2 units per hour, if you have a sample size ≥49 the margin of error in a 90% confidence interval will be ≤1.5 units per hour. Remark: The sample size of 49 is a bit larger than the Case 0 sample size of 47. This makes sense. When you don't know the standard deviation of the population, you have to use the t distribution. Student's t is more spread out than z, so the confidence intervals are a bit wider, so you have to use a larger sample to keep the confidence interval to the same width. Case 2: One Population Proportion For binomial data with true proportion p, the population standard deviation is $\sigma = \sqrt{p(1-p)}$. Even though you don't know p, the value $\hat{p}(1-\hat{p})$ from your sample — or your prior estimate, if you don't have a sample — will be close to the true value p(1−p) in the population, because the product p(1−p) doesn't vary much as p varies. Replacing p with $\hat{p}$ in a formula may seem like cheating, but n this case it's not, because p(1−p) varies a lot less than p on its own. For instance, suppose the true population proportion is 45% but your estimate is 35%. The true p(1−p) is 0.45×0.55 = 0.2475, and your estimate is 0.35×0.65 = 0.2275. The difference between 0.2475 and 0.2275 is a lot less than the difference between 0.45 and 0.35. Therefore you can use a z function, and the formulas are the same as Case 0 with $\sqrt{\hat{p}(1-\hat{p})}$ substituted for σ: $E = z_{\alpha/2} \cdot \sqrt{\hat{p}(1-\hat{p})/n}$ transforms to $n = (z_{\alpha/2}/E)^2 \cdot \hat{p}(1-\hat{p})$ Because this article helps you,please click to donate!Because this article helps you,please donate atBrownMath.com/donate. If you don't have a sample or any credible estimate, use $\hat{p} = (1-\hat{p}) = 0.5$. This is the conservative procedure because the product $\hat{p}(1-\hat{p})$ takes its highest value when $\hat{p}$ = 0.5. The conservative procedure may give you a sample size larger than necessary, but you can be sure your sample won't be too small, forcing you to throw out your survey and start over. Caution: The sample must not exceed 10% of the population. Another way to look at that is that 10 times sample size must be less than or equal to population size. Example 3: What percent of the voters would vote for your candidate if the election were held today? You want 95% confidence in your answer, with a margin of error no more than 3.5%. Last month's poll showed your candidate had 42% support. How many voters do you need to survey? CommentsComputation Marshal your data. Caution! 3.5% is 0.035, not 0.35. Given: 1−α = 0.95, E = 0.035, $\hat{p}$ = 0.42 Wanted: sample size n To find $z_{\alpha/2}$, first find α/2. 1−α = 0.95 ⇒ α = 0.05 ⇒ α/2 = 0.025 $z_{\alpha/2} = z_{0.025}$, the critical z for a right-hand tail area of 0.025. That's invNorm(1−.025). Divide by E and square. You're going to chain calculations so that you don't have to re-enter any of your intermediate numbers. Press [/], and notice how the calculator responds Ans to let you know it's using the previous answer. Enter .035 for E and press [ENTER]; that gives you the result of the fraction. Press [x²] [ENTER] to square it. The last link in the chain is multiplying by $\hat{p}$ and then by $(1-\hat{p})$. Your result is 764. Remember, always round sample size up, regardless of the decimal part. Answer: To find a 95% CI with a margin of error no more than ±3.5 percentage points, where the true population proportion is around 42%, you must survey at least 764 people. 10×764 = 7640; presumably the electorate is larger than that. Example 4: Suppose you're planning your first poll, and you have no idea of your candidate's level of support. How big a sample would you need to be sure of a margin of error no more than 3.5% in a 95% CI? Solution: Compute $z_{\alpha/2}$ = 1.9600 as in the previous example. But this time use $\hat{p}$ = 0.5 since you have no estimate for p. $n = (\, z_{\alpha/2} \div E \,)^2\, \hat{p}\,(1-\hat{p})$ = (invNorm(1-.025)/.035)² × 0.42 × (1−0.42) = 783.971… → 784 Answer: To find an 95% CI with a margin of error no more than ±3.5 percentage points, where you have no idea of the true population proportion, you must survey at least 784 people. Case 5: Difference of Two Population Proportions When you're comparing two population proportions, it's perfectly legitimate to have different-sized samples. The formula for margin of error, below left, is just an extension of the formula for one population proportion. But when you're planning sample size, you can't solve one equation for two variables n_1 and n_2. (If you had a reason to choose some particular value for one of them, you could solve for the other one.) You can solve for sample size if you decide to use the same size for both samples. transforms to $n = n_1 = n_2 = [\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)] \cdot (z_{\alpha/2}/E)^2$ For the reasons given above, if you have any prior estimates for the population proportions p_1 and p_2 you should use them; otherwise use 0.5. Example 5: You'd like to know how your candidate's support differs between men and women. You know that overall support is 42%. How many of each sex must you survey to answer the question with 90% confidence and a margin of error no more than 3%? CommentsComputation Marshal your data. (Caution! 3% is 0.03 not 0.3.) Given: 1−α = 0.90, E = 0.03 Wanted: sample size $n=n_1=n_2$ Do you have an estimate of p_1 and p_2? Yes, since the overall support is 42% you expect that men's and women's support is not too different from that. (You do expect p_1 and p_2 are somewhat different, or you wouldn't be doing the survey. But remember from one population proportion that p(1−p) doesn't vary much when p varies.) Prior: $\hat{p}_1$ = 0.42 and $1-\hat{p}_1$ = 0.58 $\hat{p}_2$ = 0.42 and $1-\hat{p}_2$ = 0.58 Compute $z_{\alpha/2}$ in the usual way. 1−α = 0.90 ⇒ α = 0.10 ⇒ α/2 = 0.05 $z_{0.05}$ = invNorm(1−0.05) ≈ 1.6449 Finish, using the unrounded value of z. Always remember to round sample sizes up. $[\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)]\ [z_{\alpha/2}\div E]^2$ = (0.42×0.58+0.42×0.58)×(ANS÷0.03)² = 1464.60… → 1465 Answer: To find a 90% CI for the difference in your candidate's support between men and women, with margin of error no more than 3%, you must survey at least 1465 men and at least 1465 women. Remark: You might wonder why the samples must be so large. After all, to estimate one population proportion to ±3% in a 90% CI, with prior estimate $\hat{p}$ = 42%, a sample of 752 is enough. (Check it!) Why do you need over 2900 people in two groups for the same margin of error? The answer is that it's not the same margin of error. If you surveyed 752 men and 752 women you'd have confidence intervals of ±3% for each, but that's an overall margin of error of ±6% — think that the true proportion might be near the bottom of one group's interval and near the top of the other group's, or vice versa. (It's not quite that simple, but that's the basic idea.) To bring down that margin of error, you have to increase the sample size. Example 6: Let's modify the previous example. Suppose you have reason to believe your candidate appeals more strongly to women, with a gap of about 10%? That means you estimate men's support at 37% and women's support at 47%. $\hat{p}_1$ = 0.37, $1-\hat{p}_1$ = 0.63, $\hat{p}_2$ = 0.47, $1-\hat{p}_2$ = 0.52. Your required sample size becomes $z_{0.05}$ = invNorm(1−0.05) ≈ 1.6449 $[\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)]\ [z_{\alpha/2}\div E]^2$ = (0.37×0.63+0.47×0.53)×(ANS÷0.03)² = 1449.57… → 1450 Answer: For a 90% CI with margin of error ≤3%, when you think one population's proportion is 37% and the other's is 47%, you need a sample of at least 1450 from each group. Case 6: Goodness of Fit For χ^2 tests, the requirements are that all of the expected counts must be ≥5. The expected count for each category is sample size n times the model proportion in that category, so to find the the necessary sample size you divide that minimum expected value of 5 by the smallest proportion or percentage in the model: n = 5 / (the smallest proportion in your model) If your model is expressed in ratios, such as 9:3:3:1, the smallest proportion is the smallest number in the model divided by the total in the model, in this case 1/(9+3+3+1) or 1/16. So the required sample size is 5 / (1/16) = 80. Example 7: You expect that customers will choose coffee, tea, bottled water, and Snapple in the proportions of 65%, 15%, 15%, 5%. How large a random sample must you take to test this model? Solution: Take the least likely category and divide 5 by that percentage: n = 5 / 5% = 5 / 0.05 = 100. Answer: You need a random sample of at least 100 to test this model. (As always, the minimum sample will give a significant result only if the null hypothesis is extremely wrong. If the model is only moderately wrong, a larger sample will probably be needed to reveal that.) Example 8: For a kids' play area, you ordered five cartons of blue plastic balls, two of green, and three each of red and yellow. But your assistant dumped them all together after taking delivery. How many balls must you select randomly to see if the proportions are right? Solution: You must divide 5 by the proportion in the smallest category in the model, which is green: smallest proportion = 2 / (5+2+3+3) = 2/13 5 / (2/13) = 32.5 → 33 Answer: To meet the requirements, you need a random sample of at least 33 balls. Example 9: You believe that plain M&Ms are distributed in the proportions 24% blue, 13% brown, 16% green, 18% orange, 15% red, 14% yellow. How large a sample do you need to test this model? Solution: The smallest proportion in the model is 13%, so compute 5/13% = 5/0.13 = 38.46… → 39. (Remember, sample sizes always round up.) Answer: The sample must contain at least 39 M&Ms. What's New? 18 Nov 2021: Updated links here and here. 24 Oct 2020: Convert HTML 4.01 to HTML5. Italicize variable names. Improve formatting of radicals. Replace pictures of most equations with the same equations in text form. (intervening changes suppressed) Fall 2007: New article. From the description you provided, your first question is about the distribution of people's age. Normal (i.e. Gaussian) distribution applies to such kind of applications. It will be helpful if you know how the confidence interval (CI) was calculated, because there are many different possible ways that the CI was calculated. For instance, if the distribution is of normal distribution, and the CI was calculated using t-test, then the SD can be estimated with following equation: SD = sqrt(n)*(ci_upper - ci_lower)/(2 * tinv((1-CL)/2; n-1)), where CL is the confidence level, 'ci_upper' and 'ci_lower' are the upper and lower limits of CI respectively, and 'tinv()' is the inverse of Student's T cdf. Otherwise, if it is of normal distribution, but a known SD was used in calculating CI, then the SD can be calculated with following equation: SD = sqrt(n)*(ci_upper - ci_lower)/(sqrt(8) * erfinv(CL)), where 'erfinv ()' is the inverse error function. Your second question is about the distribution of people's sex (i.e. male or female). From the data you provided, it sounds that there are k=274 males among n=427 of whole samples. Bernoulli distribution applies to this application. In this case, the variance (of male's population) = p*(1-p) = 0.2299, and SD = sqrt(0.2299) = 0.4795, where p is the mean value. Note that "valiance = mean*(1-mean)" is applicable to Bernoulli distribution only.