1. (biology)
   1. We know that for a given element i in the vector : so in order to get the average value of , we need to sum these values for i = 1 to i = n and divide by n: , we know from the problem statement that , so we can finish our previous equation as: . Given is an IID random variable the variance of is as its standard deviation is so we can now calculate the variance of : . Using these values we can now calculate the expected value of as: , since is the average of all values in its expected value should be zero as this number will approach zero as n increases so we can finish defining the above expected value as: . Now we need to calculate the variance of as: . The variance of should be smaller than that of any given as it is the average of all values in and would therefore be much less likely to vary than any given element. Given that tends to zero as n increases, and , will only converge to the truth when b = 0 as this amount of error doesn’t arrive from random events but instead bias in your method. The typical amount that the expected value of differs from the truth is defined as the root mean squared error: , however can further simply this given our above information as:

as should be zero given a large enough n. This function also goes to zero when b = 0 and a large enough sample size n is used.

* 1. As stated in paragraph four the kit is known to give “unbiased measurements that fluctuate around the true value with a SD of 0.15 and an approximately normal distribution for its measurement errors”. Therefore I can write my measurement and we know that is a normally distributed variable with a SD of 0.15. Given this we need to calculate . We will define a new variable , is now a normally distributed random variable with mean zero and SD 1, I can now use a standard normal distribution table to calculate the above probability. From looking at the table we can conclude the probability of to be 0.2546.
  2. We need to calculate the sample size such that and accomplish this by furthering simplifying this probability as . We know each has a mean of zero and a SD of 0.15 so as we calculated earlier the variance of is . We now know that we need to calculate n such that . . We will now define a new variable which a normal random variable with mean 0 and SD . We can now rewrite our objective probability as . And therefore using a standard normal distribution table must be greater than 2.57. Given all this we now must solve for n and since our table only covers positive z values and the distribution is symetric about zero therefore we can solve for n using, . So you must take at least 15 samples in order for your probability of incorrectly measuring the Ph to be less than 5 to be less than 0.5%.

1. (medicine)
   1. Given the data from the mid-1970s the average effect of Captopril on the systolic blood pressure is to reduce the value by about 18.6 mmHg for people deemed hypertensive. At that point in time hypertensive was defined as a systolic blood pressure of 140 mmHg of mercury or higher. The effect when compared to the average systolic blood pressure of the subjects is about a 10% reduction (). This is medically significant as a 10% difference is quite rare in medicine. In order to calculate the standard error we will use the inverse t equation in our case which gives us a standard error of 6.4 mmHg. Using this we can construct a 95% confidence interval with a lower bound of 12.2 mmHg, center of 18.6 mmHg, and upper bound of 25 mmHg. We are assuming that this group of 12 subjects is representative of the entire hyptertensive population on the United States and that the equipment used to measure their blood pressure is unbiased. This estimated effect is statistically significant because zero change or no effect due to Captopril is not within the 95% confidence interval. Given all this Captopril seems to be a good treatment option for hypertension.
      1. We will define the variance of the completely randomized design as: . We will now do the same for the repeated measures design:

. Now we need to find the covariance of using the provided probability . We can now finish our variance calculation: . We will now define the efficiency of RM compared to CR as: . This means the RM model is five times as efficient as the CR model.

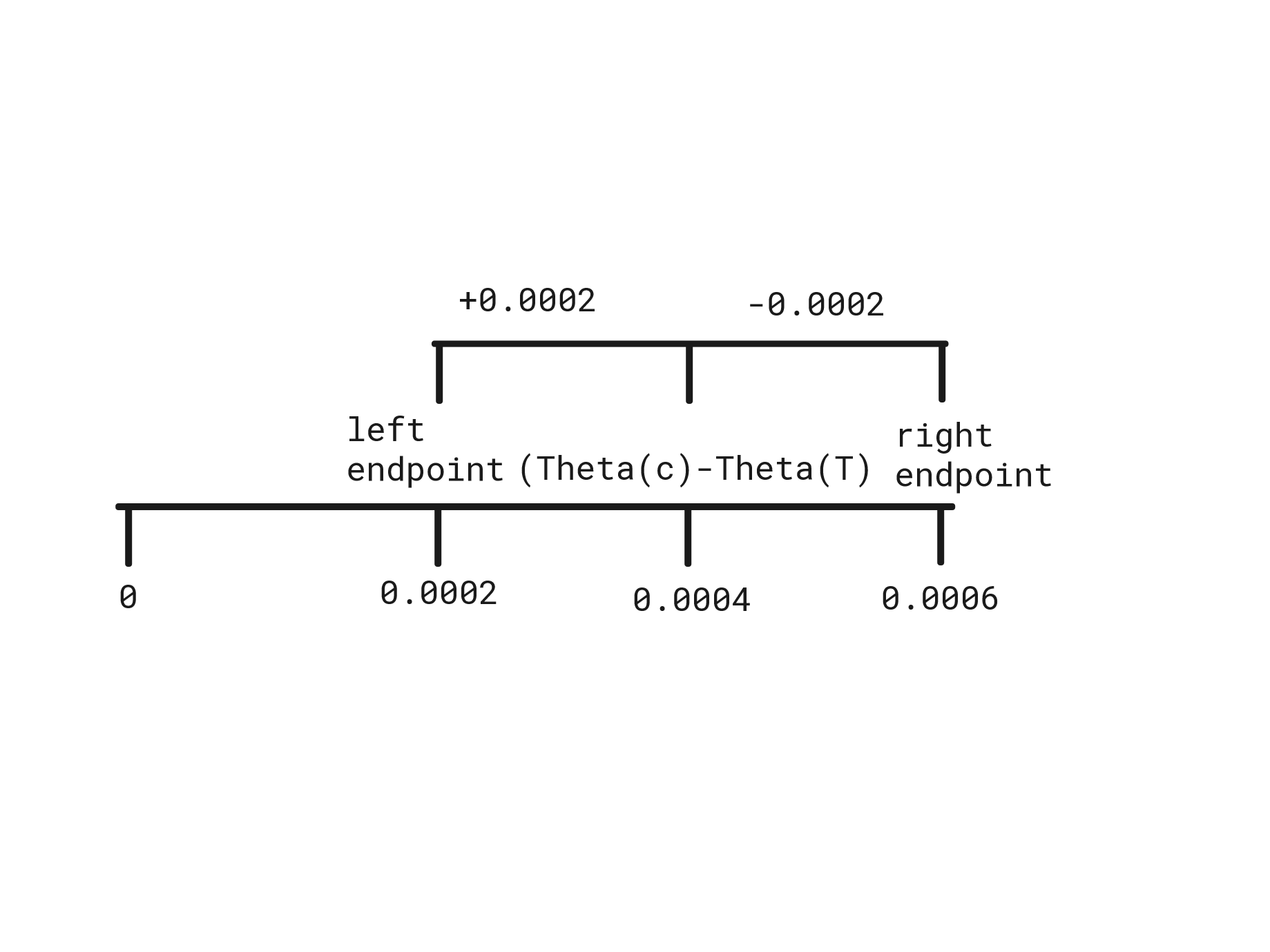
* + 1. It seems to have a stronger effect on people who have a higher initial blood pressure. This is supported by figure 1 as there is a 0.501 correlation between the before blood pressure of patients and their resulting difference after taking the drug.

1. (binomial and negative binomial sampling)
   1. We need to use a Binomial model for S because it matches the sampling method for S. The proportion of 1s in the data is given by, which is a good guess for . The expected value of is as the expected value of is because S is a binomial distribution. Given this must be unbiased for . The standard error for is . S is defined as therefore . In order to find the variance of we need to find the expected value of it and it squared. , . Now we can calculate the variance of as: . Using this we can finish our calculation of the variance of S: . Using all this we can now calculate the standard error of as: . should be approximately normal when the skewness goes away so when n is sufficiently large or .
      1. Given X is the total of number of failures before the sth success and s is the number of successful tests, the the total number of tests N is: .
      2. In order to find the PMF of N we need to work forward from the definition of PMF provided for X.

. Using the theorem for we will set , giving us: . This allows us to finish the PMF of N as: .

* + 1. We know that . Given X is a negative binomial distribution with inputs s and its variance is: which is also the variance of N given our previous findings. Using this and the fact that the expected value of X is: we can now calculate the expected value of N as: . We know that , this makes calculating the expected value and variance of it very hard so we will use the delta method. We will define a function g and its derivative as that we can now take the expected value of with N as its parameter. so . We will define a new variable W and find its expected value and variance and then sub in into the resulting simpler equations. . We now need to find the variance of g when its parameter is N: . Given this . Using this we can now define the standard error of as: .
    2. Jensen’s inequality tells us for a linear convex function g(x), , and for a non-linear convex function its the same except it is a hard inequality. so its expected value is: . is a nonlinear convex function so by Jensen’s inequality: . We are assuming without proof that . We now need to find the difference between andso we can later calculate the expected value of this quantity. . Now that we have simplified the difference we can use it to better understand the relationship between and . . Sincewill always be positive, must be true. Additionally since as goes to zero goes to infinity must be biased on the high side. Therefore as N gets sufficiently large becomes unbiased. The order of bias magnitude for .
  1. Yes because the expected value and standard error are very similar.

1. (public health)
   1. This helps combat possible non-placebo effects in the population that didn’t receive the vaccine, as if they know they didn’t receive the vaccine they would take more precautions.
   2. You could run this experiment in a double blind fashion by writing down which syringe numbers contain what and then locking this away until the trial period is over. This would help in the case of this question as the person administering the syringe could somehow indicate to the patient that they are receiving a placebo.
   3. This experiment consists of two sample sets, those who were actually treated with the vaccine and those whose received a saline solution in the control group. We will define the random variables for each as follows: vaccine = and control = . Based on the information provided in the problem we know that and . Using this we can calculate the difference between the two quantities as: . Now we need to calculate the standard error of this rough estimation: , . We now need to calculate the variance for each random variable, however the resulting equation will hold for both if we replace the corresponding n and values so I will only show the math for .

. Based on this we can finish our above variance calculation as: . We can also now finish our standard error calculation of , . This is a simple pythagorean calculation so we can complete it based on the following values: , so the standard error for is: . We will now sketch the 99.9% confidence interval for the our approximation , . 

This difference is very statistically significant as zero is not in the 99.9% confidence interval and there is a whole interval step between the left endpoint and zero meaning the treatment is highly effective.

* 1. In order to find a proper sample size n we must first calculate the distance between our center and 0 in order to find the correct n: . We should construct a scenario in which the treatment and control sample spaces are each half of the total sample space i.e. .We will now refine our above calculation using this fact:

. Therefore the optimal sample size is 95,402 kids.