Computing joint confidence intervals

Abstract

Assume you have two data sets from two independent runs. In each run you have an evaluation group to be compared against the baseline. You can compute the confidence interval for the difference in each data set, however, there may be reasons you do not want to combine the two data sets to compute and overall confidence interval. Then how would you combine statistics you get from each run to compute an overall confidence interval? That is the question we want to address.

Index Terms

Normal distribution, confidence intervals, central limit theorem

Contents

I. Introduction

The goal is to combine information from separate experimental runs to quote a single number to assess the difference of two groups in the data. This would be a case where one cannot combine the data sets due to confounding factors. We will discuss two ways of combining statistics:

- Computing joint p-values,
- Computing joint confidence intervals.

II. Sampling statistics

The end goal of statistics is to estimate the true values of the population parameters, such as the mean value μ and the variance σ^2. We try to get a sense of what μ could be by pulling N samples from the population and studying it. Each sample is a random variable, S_i, and we can compute their average

$$\bar{S} = \frac{1}{N}\sum_{i=1}^{N} S_i. \tag{1}$$

S itself is a random variable: it will be different if you pick up another set of N samples. We can also compute its expected value

$$E[\bar{S}] = \frac{1}{N}\sum_{i=1}^{N} E[S_i] = \mu, \tag{2}$$

and its variance

$$\mathrm{Var}[\bar{S}] = \frac{1}{N^2}\sum_{i=1}^{N} \mathrm{Var}[S_i] = \frac{\sigma^2}{N}. \tag{3}$$

email: quarktetra@gmail.com
Find the interactive HTML-document here.

Furthermore, by central limit theorem, we know that for large N $\bar{S}$ will be normally distributed, denoted as

$$\bar{S} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right). \tag{4}$$

III. p-Values

The P value is the probability of obtaining a result equal to or larger than what was actually observed assuming the the null hypothesis is true. It does not give the probability of the null hypothesis is true or not. Figure 1 illustrates the $p-$value.

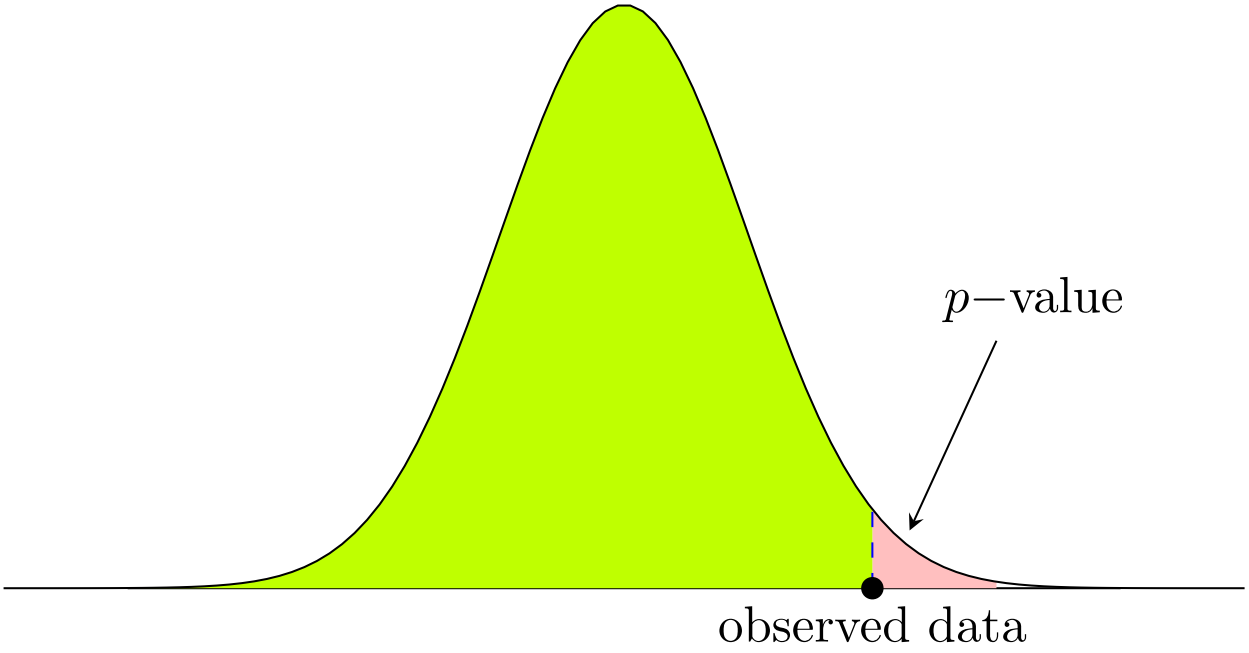

Figure 1: Illustration of $p-$value.

Such a number can be computed for any distribution, and it is supposed to be compared against α, which sets how extreme the data must be for the null hypothesis to be rejected. For example, for $\alpha = 5\%$, $p < 0.05$ will result in the rejection of the null hypothesis.

A. Pooling $p-$values

A way of pooling $p-$values is proposed in [1] as a harmonic sum

$$\frac{1}{\bar{p}} = \frac{\sum_{i=1}^{n} \frac{Ni}{p_i}}{\sum_{i=1}^{n} N_i}, \tag{5}$$

where N_i's are the samples sizes, and p_i's are the individual $p-$values computed for each experiment, and n is the number of experiments to be pooled. For equal sample sizes, this simplifies to

$$\frac{1}{\bar{p}} = \frac{\sum_{i=1}^{n} \frac{1}{p_i}}{n}. \tag{6}$$

This is will result in a single number that will represent the results combining from all experimental runs.

IV. Confidence intervals

When the true values of the population parameters, μ and σ^2 are unknown, we can replace them with the sample statistics $\hat{\mu}$ and $\hat{\sigma}^2$

$$\bar{S} \sim \mathcal{N}\left(\hat{\mu}, \frac{\hat{\sigma}^2}{N}\right). \tag{7}$$

We are trying to estimate μ and σ^2 based on the data we sample. This is illustrated in Fig. (2).

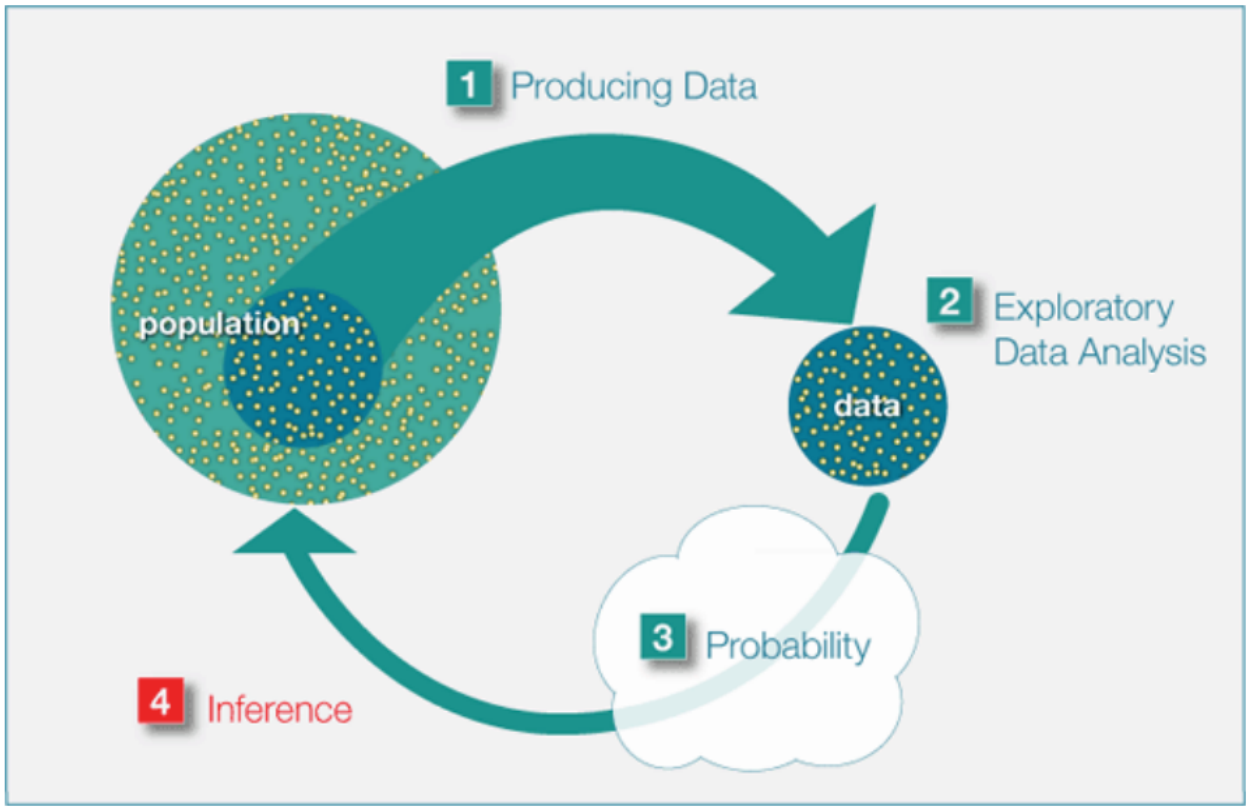

Figure 2: Population vs Sample statistics. Image taken from [UF Biostatistics text book](https://bolt.mph.ufl.edu/6050-6052/unit-1/)

Equation (7) tells us that the sample mean value, $\hat{\mu}$, will vary from as we get another set of N samples. And we do not necessarily know its relative position with respect to the population mean μ. We want to create an interval around the $\hat{\mu}$ such that we can estimate whether μ will happen to be in that interval. It is important to notice that this is more about creating a procedure to create an interval which, if repeated, will contain the population mean value μ. Assume you have a method of creating confidence intervals(CI), say with 95% confidence, which we will discuss later, below is what it means:

- You pull N samples and compute the interval with the data.
  - μ is not guaranteed to be in the interval.
  - You cannot say it will be in the interval with 95% probability. It is either in or out. This is not probabilistic.
- You go back and pull N new samples.
  - Note that you will have a new $\hat{\mu}$ and a new confidence interval. μ may or may not be in it.
- If you repeat the process many times, if your method of computing intervals is correct, 95% of the intervals you computed will include the true value μ.
- The process is illustrated in Fig 3.

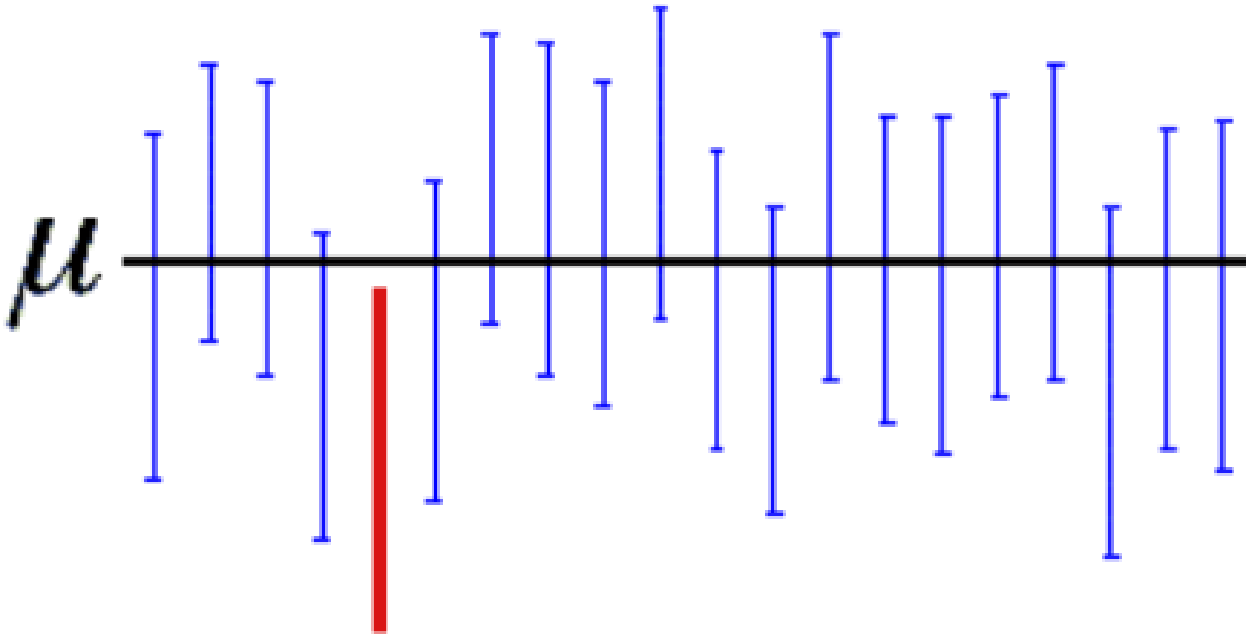

Figure 3: Repeting a 95% confidence interval computation 20 times will result in CIs containing the population parameter 19 out of 20 times. Image taken minitab [blog](https://blog.minitab.com/blog/adventures-in-statistics-2/understanding-hypothesis-tests-confidence-intervals-and-confidence-levels).

To compute the confidence intervals, it is convenient to define the following random variable:

$$Z = \frac{\bar{S} - \hat{\mu}}{\frac{\hat{\sigma}}{\sqrt{N}}}. \tag{8}$$

If the sample size is large ($N >\sim 30$), this particular random variable will have standard normal distribution[1]. A confidence level of c corresponds to a z^* value such that the area under the standard normal distribution between $-z^*$ and z^* is c. Equivalently, the area outside on each side will be $\frac{1-c}{2} = \frac{\alpha}{2}$, where $\alpha \equiv 1 - c$. Take an example of 95% confidence level, which yields $\alpha = 0.05$. The value of z^* in this case will be 1.96 so that 2.5% of the total area lies under the tails, i.e., $|z| > 1.96$. Reverting Eq. (8) shows that 95% CI is from $\hat{\mu} - 1.96\frac{\hat{\sigma}}{\sqrt{N}}$ to $\hat{\mu} + 1.96\frac{\hat{\sigma}}{\sqrt{N}}$. Figure 4 shows this for generic c.

[1]for $N <\sim 30$, it will be t-distribution. Here we will assume sample size is large enough

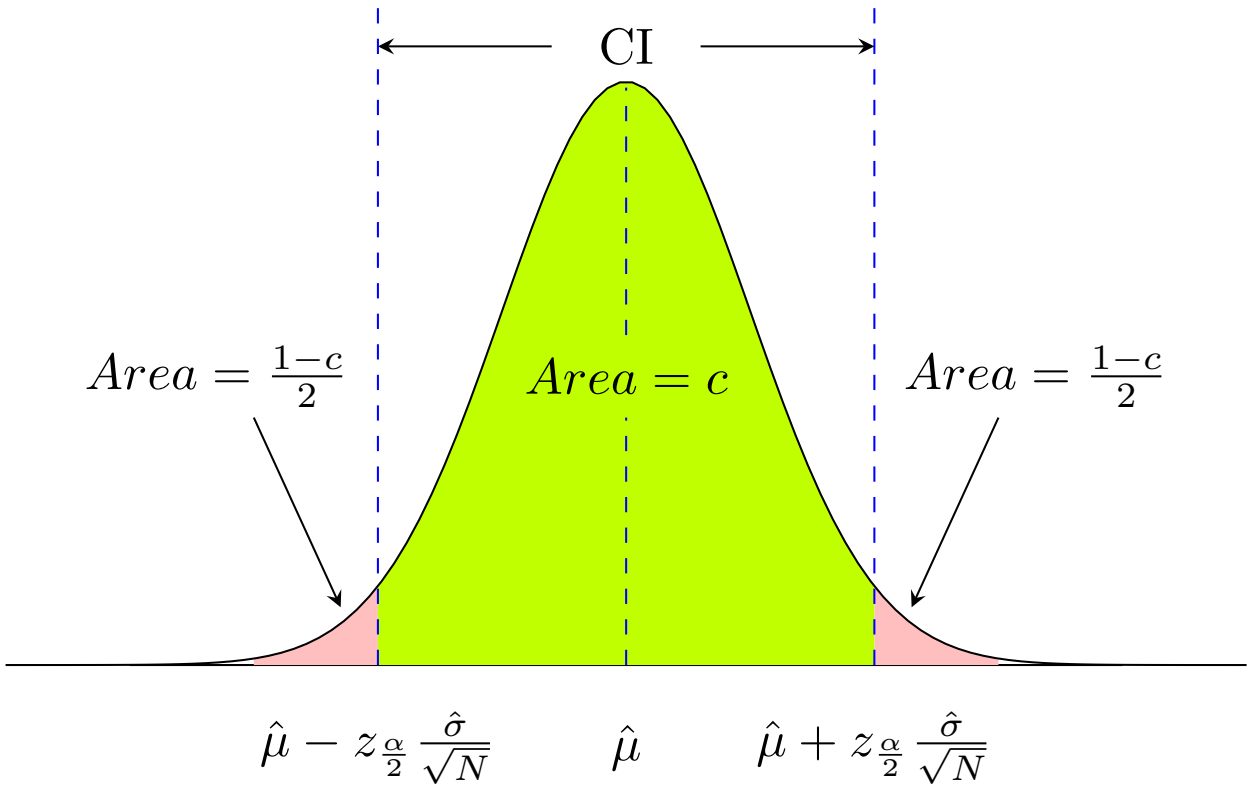

Figure 4: Illustration of confidence intervals with confidence level of c.

A. Pooling multiple confidence levels

Assume you have two sets of data, and you can compute $\hat{\mu}$, $\hat{\sigma}$ and the corresponding CIs for each set of data. How would you combine these two sets to produce joint CIs? We can do this by pooling the estimates as follows:

$$\bar{S} = \frac{N_1 \bar{S}_1 + N_2 \bar{S}_2}{N_1 + N_2}, \tag{9}$$

where subscripts label the data sets. $\bar{S}$ is yet another normal random variable, and we can compute its expected value as

$$E[\bar{S}] = \frac{N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2}{N_1 + N_2}, \tag{10}$$

and its variance:

$$\mathrm{Var}[\bar{S}] = \frac{N_1^2 \mathrm{Var}[\bar{S}_1] + N_2^2 \mathrm{Var}[\bar{S}_2]}{(N_1 + N_2)^2} = \frac{N_1^2 \frac{\hat{\sigma}_1^2}{N_1} + N_2^2 \frac{\hat{\sigma}_2^2}{N_2}}{(N_1 + N_2)^2} = \frac{N_1 \hat{\sigma}_1^2 + N_2 \hat{\sigma}_2^2}{(N_1 + N_2)^2}. \tag{11}$$

Therefore, the pooled estimator becomes

$$\bar{S} \sim \mathcal{N}\left(\frac{N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2}{N_1 + N_2}, \frac{N_1 \hat{\sigma}_1^2 + N_2 \hat{\sigma}_2^2}{(N_1 + N_2)^2} \right), \tag{12}$$

which completely defines the joint distribution. One can easily compute the corresponding CI as

$$\left[\frac{N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2}{N_1 + N_2} - z_{\frac{\alpha}{2}} \frac{\sqrt{N_1 \hat{\sigma}_1^2 + N_2 \hat{\sigma}_2^2}}{N_1 + N_2}, \frac{N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2}{N_1 + N_2} + z_{\frac{\alpha}{2}} \frac{\sqrt{N_1 \hat{\sigma}_1^2 + N_2 \hat{\sigma}_2^2}}{N_1 + N_2} \right]. \tag{13}$$

For equal sample sizes, $N_1 = N_2 = N$, we get

$$\left[\frac{\hat{\mu}_1 + \hat{\mu}_2}{2} - z_{\frac{\alpha}{2}} \frac{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}}{2}, \frac{\hat{\mu}_1 + \hat{\mu}_2}{2} + z_{\frac{\alpha}{2}} \frac{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}}{2}\right], \tag{14}$$

which is the final result that expresses the CI in terms of the statistical parameters of both experiments.

V. Visualization

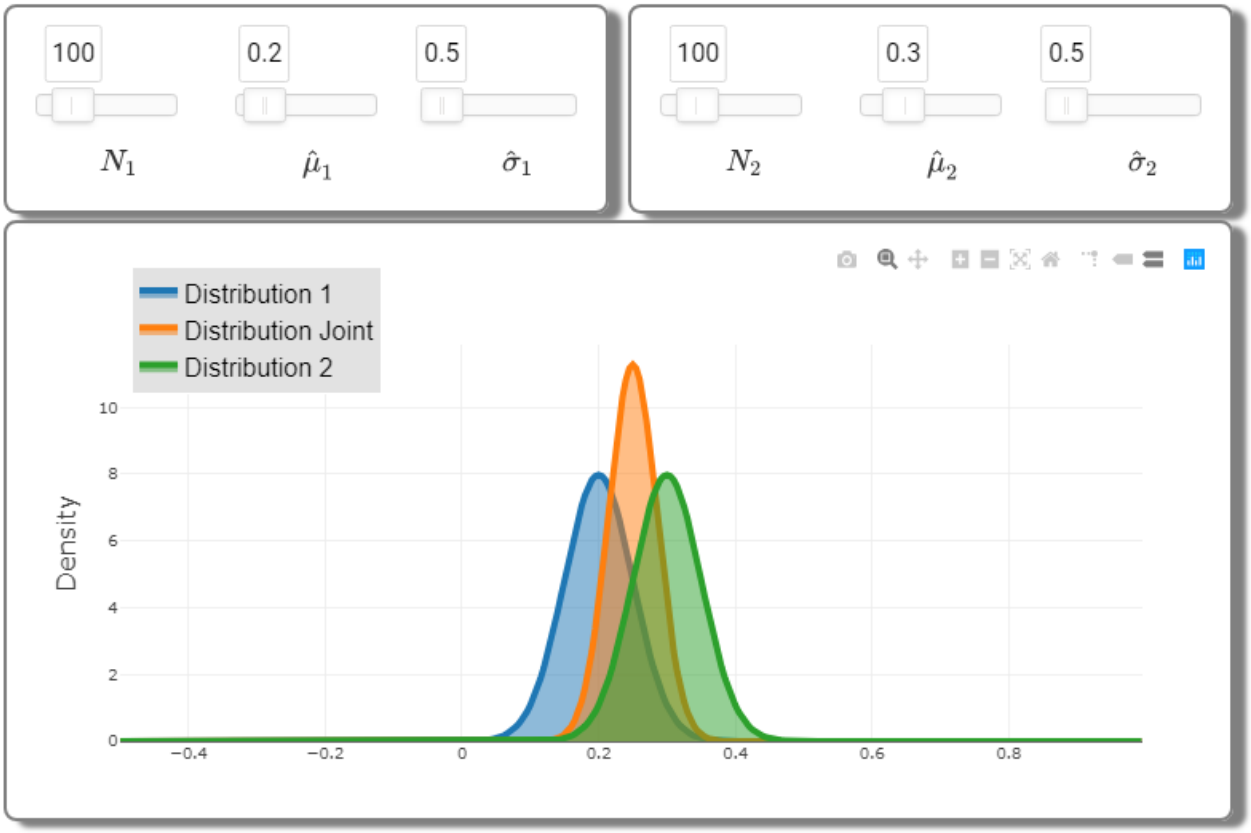

Figure 5: Curves showing the distribution of the mean value based on sample 1 and 2 data and the computed joint distribution.

[1] D. J. Wilson, "The harmonic mean p-value for combining dependent tests," *Proceedings of the National Academy of Sciences*, vol. 116, no. 4, pp. 1195–1200, 2019, doi: 10.1073/pnas.1814092116. [Online]. Available: https://www.pnas.org/content/116/4/1195