

MATH 203: PRINCIPLES OF STATISTICS 1
FALL 2024

WHERE DO THE FORMULAS COME FROM ?

All of the one sample and two sample formulas for confidence intervals and hypothesis testing stem from the same basic results from theory related to the Normal distribution.

1. Recall that we saw the sampling distribution result that if $X_1, \dots, X_n$ were a random sample of size n from a distribution with population mean μ and population standard deviation σ , then

$$\bar{X} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right). \quad (1)$$

This result is

- EXACT if the $X_1, \dots, X_n$ themselves have a Normal distribution
- APPROXIMATE if the $X_1, \dots, X_n$ do not have a Normal distribution, but n is large (i.e. $n \geq 30$).

We will proceed treating the two cases identically.

2. We have the same result holding for two independent samples

$$X_1, \dots, X_m \quad Y_1, \dots, Y_n$$

from populations with parameters (μ_1, σ_1) and (μ_2, σ_2) respectively, that is

$$\bar{X} \sim \text{Normal}\left(\mu_1, \frac{\sigma_1^2}{m}\right) \quad \bar{Y} \sim \text{Normal}\left(\mu_2, \frac{\sigma_2^2}{n}\right). \quad (2)$$

We can also prove using distribution theory that, because of (2), we have

$$\bar{X} - \bar{Y} \sim \text{Normal}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right). \quad (3)$$

3. We now consider z-transformations in (1), (2) and (3). Specifically, we can deduce that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1)$$

and

$$\frac{\bar{X} - \mu_1}{\sigma_1/\sqrt{m}} \sim \text{Normal}(0, 1) \quad \frac{\bar{Y} - \mu_2}{\sigma_2/\sqrt{n}} \sim \text{Normal}(0, 1)$$

and

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim \text{Normal}(0, 1)$$

In the last case, we can see that if we know that $\sigma_1 = \sigma_2 = \sigma$, say, then

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim \text{Normal}(0, 1) \quad (4)$$

4. Note that all of the z-transformed quantities can be seen to take the form

$$\frac{\text{Point Estimator} - \text{Thing being estimated}}{\text{Standard deviation of the Point Estimator}}$$

The standard deviation of the point estimator in the denominator is sometimes called the '**standard error**' of the estimator, and we write this expression symbolically as

$$\frac{\hat{\theta} - \theta}{\text{s.e.}(\hat{\theta})}$$

where θ is μ or p . It is this quantity that we denote by T when defining a test statistic.

5. The results in 3. tell us how the z-transformed quantities should behave if our assumptions (random sample, and Normality of the original variables or large sample size) are correct. The z-transformed quantities all depend on the population parameters. We use them to construct confidence intervals and Rejection Regions.

(i) **Confidence Intervals:** Here we use the idea that because of Normality, if z_α is the upper α point of the standard Normal, then, from (1), we have

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha \quad (5)$$

which we rearrange to state

$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

(ii) **Rejection Regions:** Here we use the probability statement in (5) directly, but applied to the proposed value of μ under the null hypothesis, which we denote μ_0 : we write

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

for the test statistic, and then conclude that if P_0 is the probability computed assuming H_0 is true (the **null distribution**)

$$P_0(-z_{\alpha/2} \leq T \leq z_{\alpha/2}) = 1 - \alpha \quad (6)$$

This statement implicitly defines the two-sided version of the Rejection Region, suitable for the alternative hypothesis $H_a : \mu \neq \mu_0$; for the one-sided versions we use the same probability argument to write

$$\begin{aligned} H_a : \mu > \mu_0 & \quad P_0(T \geq z_\alpha) = \alpha \\ H_a : \mu < \mu_0 & \quad P_0(T \leq -z_\alpha) = \alpha \end{aligned}$$

6. In all of the previous formulas, σ , or σ_1 and σ_2 are presumed **known**. In practice, we typically do not know these values, and have to estimate them.

There are two cases, the exact case and the approximate case.

- (i) **Exact case:** This we can apply when n (for (1)) or one of m and n (for (2) and (3)) is small.

- For (1), σ^2 is estimated by

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- For (2) and (3), we **must** assume that $\sigma_1 = \sigma_2 = \sigma$, and estimate σ using

$$S_P^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2} = \frac{1}{m+n-2} \left(\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{i=1}^n (Y_i - \bar{Y})^2 \right)$$

What does this change ? In (1), (5) and (6), the Normal distribution changes to a t-distribution. For the one sample case, the degrees of freedom is $n-1$, for the two sample case, the degrees of freedom is $m+n-2$.

- (ii) **Approximate case:** When n is large, or both m and n are large; we can simply substitute in the estimate of σ , or σ_1 and σ_2 , and still use the Normal distribution instead of the t-distribution. That is

- use S in place of σ for (1)
- use S_1 and S_2 in place of σ_1 and σ_2 for (3)
- use S_P in place of σ for (4).

7. Bernoulli/Binomial special case: For the Bernoulli (binary, Yes/No) outcome case, we use the large sample properties and the Central Limit Theorem described in the Approximate case above.

- In the **one sample case**, we have that $\theta = p$,

$$\hat{p} = \bar{X} \quad \text{s.e.}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

and the Confidence Interval is based on

$$T = \frac{\hat{\theta} - \theta}{\text{s.e.}(\hat{\theta})} = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

and we perform the calculation for the interval based on equation (5). For the hypothesis test Rejection Region, we are proposing a value, p_0 , for p under the null hypothesis, so we instead use

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

and apply the calculation based on (6).

- In the **two sample case**, we have that $\theta = p_1 - p_2$, and

$$\hat{p}_1 = \bar{X} \quad \hat{p}_2 = \bar{Y}$$

with standard errors given by the same formula as for the one sample case. We compute the confidence interval for $p_1 - p_2$ using (3), but replacing σ_1 and σ_2 by the estimates

$$\hat{\sigma}_1 = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{m}} \quad \hat{\sigma}_2 = \sqrt{\frac{\hat{p}_2(1 - \hat{p}_2)}{n}}$$

and base the interval on the quantity

$$T = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{m} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n}}}.$$

For the hypothesis test of $H_0 : p_1 = p_2$, we are effectively saying that $p_1 = p_2 = p$ with p unknown. So we estimate p by

$$\hat{p} = \frac{X + Y}{m + n}$$

which is the point estimate from the pooled samples assuming that the null hypothesis is true. We then proceed as for the one sample case and use

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1 - \hat{p})}{m} + \frac{\hat{p}(1 - \hat{p})}{n}}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{m} + \frac{1}{n} \right)}}$$

to carry out the test, based on the Normal approximation and equation (6). Here the null hypothesis effectively insists that $\sigma_1 = \sigma_2$.