

The basic notation

\bar{X} the sample mean	μ the population mean
s^2 the sample variance	σ^2 the population variance
s the sample standard deviation	σ the population standard deviation
n the sample size	N the population size
p the sample proportion	π the population proportion

Calculating the Sample Mean and Variance. The sample mean is just the sum of the observations divided by the number of observations: $\bar{X} = \sum_{i=1}^n X_i/n$. The sample variance is $s^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$. For population data, divide by n instead of $n-1$. The standard deviation is the positive square root of the variance.

The Distribution of the sample mean. If the sample size is larger than 30 and the variable satisfies certain criteria which you should have memorized, then the sample mean will be normally distributed with mean μ and standard deviation σ/\sqrt{n} , regardless of the distribution of the underlying variable. If σ is not known, we can substitute s for it. The probability that a random variable X with mean μ and standard deviation σ/\sqrt{n} is less than some number t is the same as the probability that a standard normal variable is less than $(t - \mu)/\sigma/\sqrt{n}$.

Confidence intervals for the population mean and proportion. If the sample size is over 30, then our belief about where the population mean lies will follow a normal distribution around the sample mean with mean \bar{X} and standard error $s_{\bar{X}} = s/\sqrt{n}$. This means that our 90% confidence interval goes from $\bar{X} - 1.65s_{\bar{X}}$ to $\bar{X} + 1.65s_{\bar{X}}$; our 95% confidence interval goes from $\bar{X} - 1.96s_{\bar{X}}$ to $\bar{X} + 1.96s_{\bar{X}}$; and our 99% confidence interval goes from $\bar{X} - 2.58s_{\bar{X}}$ to $\bar{X} + 2.58s_{\bar{X}}$.

If the sample size is less than 30 and the underlying variable is normally distributed, the belief about where the population mean lies will follow a t distribution around the sample mean with mean \bar{X} and standard error $s_{\bar{X}} = s/\sqrt{n}$. However, the size of the interval in standard errors, instead of 1.65, 1.96, and 2.58, will have to be gotten from a t distribution table with $n-1$ degrees of freedom.

If a sample proportion has $np > 5$ and $n(1-p) > 5$, then our belief about where the true population proportion lies is normally distributed, with mean p and standard error $s_p = \sqrt{p(1-p)/n}$. The 90%, 95%, and 99% confidence intervals follow the same rule as that for the sample mean with p substituted for \bar{X} and s_p substituted for $s_{\bar{X}}$.

One-sample hypothesis tests. To test the null hypothesis H_0 that a population mean $\mu = \mu(H_0)$, if a sample size is larger than 30, we use the statistic $z = (\bar{X} - \mu(H_0))/(s/\sqrt{n})$. We reject at the 10% level if $|z| > 1.65$; at the 5% level if $|z| > 1.96$; and at the 1% level if $|z| > 2.58$. If the sample size is smaller than 30, we use the statistic $t = (\bar{X} - \mu(H_0))/(s/\sqrt{n})$. It follows a t distribution with $n-1$ degrees of freedom, and the critical values for a given level of significance can be read from a t distribution table.

To test the null hypothesis H_0 that a population proportion $\pi = \pi(H_0)$, we use the statistic $z = (p - \pi)/\sqrt{\pi(1-\pi)/n}$. It has the usual critical values listed above. To use this statistic, it must be the case that $np > 5$ and $n(1-p) > 5$.

Two-sample hypothesis tests. Given two independent samples, subscripted by 1 and 2, to test the hypothesis that $\mu_1 = \mu_2$, we test the hypothesis that $\mu_1 - \mu_2 = 0$ by looking at the statistic $z = (\bar{X}_1 - \bar{X}_2)/\sqrt{s_1^2/n_1 + s_2^2/n_2}$. It has the usual critical values listed above.

Given two paired, dependent samples observations X_1 and X_2 , if we want to test that the mean is the same in both samples, we generate the variable $X = X_1 - X_2$ and then simply use the z statistic for the one-sample hypothesis that the population mean of X is 0. If the sample size of X is less than 30, then we use the t statistic for the one-sample hypothesis test that the population mean of X is 0.

Given two samples of proportion data, subscripted by 1 and 2, to test the hypothesis that $\pi_1 = \pi_2$, we use the statistic $z = (p_1 - p_2)/\sqrt{p_c(1-p_c)/n_1 + p_c(1-p_c)/n_2}$ where $p_c = (n_1p_1 + n_2p_2)/(n_1 + n_2)$. It has the usual critical values listed above.

Regression. Given two variables X and Y , the covariance between X and Y is $\sigma_{XY} = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})/n$. Given population standard deviations σ_X and σ_Y , respectively, and covariance σ_{XY} , the sample correlation coefficient of X and Y is $r_{XY} = \sigma_{XY}/\sigma_X\sigma_Y$. To test the hypothesis that the population correlation coefficient $\rho_{XY} = 0$, use the test statistic $t = r\sqrt{n-2}/\sqrt{1-r^2}$, which follows a t distribution with $n-2$ degrees of freedom.