Now, this result doesn't imply that the estimate will be really within 1% of the true mean (it's a random variable), but it's possible to compute the probability it will be. To do this we would need its probability density function, which we don't have, but if we assume a large sample, we can apply the Central Limit Theorem and conclude it has Normal distribution, thus:

$$Pr(9.9 \le \bar{X} \le 10.1) = Pr\left(\frac{9.9-10}{0.1} \le \frac{\bar{X}-10}{0.1} \le \frac{10.1-10}{0.1}\right) = Pr(-1 \le Z \le 1)$$

where $Z \sim N(0,1)$

$$= \Phi(1) - \Phi(-1) = 2\Phi(1) - 1 = 2 \cdot 0.8413 = 0.6826$$

so, there is a probability of 0.3174 that the estimate is actually more than 1% away of the true value

We are now interested in estimating the variance. It's necessary to control it because, if it's too large, many of our samples can be far away from the mean. It will also be useful to determine sample sizes when the population variance is unknown

Since the population variance is defined as $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

it's natural to define the sample variance as $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

for a sample $X_1, X_2, \ldots, X_n$

In the same way we did for the population variance, is possible to show that $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$, then

$Var(X) = E[X^2] - E[X]^2$ and sample with replacement

$$E[\hat{\sigma}^2] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i^2] - E[\bar{X}^2] \downarrow = \frac{1}{n}\sum_{i=1}^{n}(\sigma^2+\mu^2) - \left(\frac{\sigma^2}{n} + \mu^2\right)$$

$$= \sigma^2 + \mu^2 - \frac{\sigma^2}{n} + \mu^2 = \frac{n-1}{n}\sigma^2$$ thus $\hat{\sigma}^2$ is a biased estimate of σ^2

which can be made easily unbiased defining $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

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

Exercise: Prove that $E[S^2] = \sigma^2$, that is, that S^2 is an unbiased estimator of σ^2

With sample without replacement (or small population) we define

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

Observe that it tends to the previous result when $N \to \infty$

Some distributions derived from the Normal distribution

Recall that, if $Z \sim N(0,1) \Rightarrow U = Z^2 \sim \chi^2$

If $U_1, U_2, \dots, U_n$ are independent chi-square random variables, then

$$V = \sum_{i=1}^{n} U_i \sim \chi^2_{(n)}$$

Proposition. Let $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $X_i \sim N(\mu, \sigma^2)$ for $i = 1, 2, \dots n$ independent. Then $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$ $\left[\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$ ⌟

Proof. As $X_i \sim N(\mu, \sigma^2) \Rightarrow \frac{X_i - \mu}{\sigma} \sim N(0,1) \Rightarrow \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2_{(1)} \Rightarrow$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2_{(n)}$$

On the other hand $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}[(X_i - \bar{X}) + (\bar{X} - \mu)]^2$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[(X_i - \bar{X})^2 + 2(X_i - \bar{X})(\bar{X} - \mu) + (\bar{X} - \mu)^2\right] =$$

$$= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2 + 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \bar{X}) + \sum_{i=1}^{n}(\bar{X} - \mu)^2\right]$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 + \frac{n}{\sigma^2}(\bar{X} - \mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2$$

$$\Rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 - \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2$$

$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2$: $\chi^2_{(n)}$ (first thing we proved)

$\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2$: $\chi^2_{(1)}$ (because $X_i \sim N(\mu, \sigma^2)$ and $E[\bar{X}] = \mu$, $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$)

$$\Rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2_{(n-1)}$$

Proposition. If $Z \sim N(0,1)$ and $U \sim \chi^2_{(n)}$ are independent, then $T = \dfrac{Z}{\sqrt{\frac{U}{n}}} \sim t_{(n)}$

Proof. By standard method of transformations of random variables

The density function of the t distribution with n degrees of freedom is

$$f(x) = \frac{\Gamma[(n+1)/2]}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})}\left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

Proposition. Let $X_i \sim N(\mu, \sigma^2)$ $i = 1, 2, \ldots, n$ independent and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$, then $T = \dfrac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{(n-1)}$

Proof. $T = \dfrac{\bar{x} - \mu}{s/\sqrt{n}} = \dfrac{\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}}{\sqrt{s^2/\sigma^2}}$ but $\dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$\frac{s^2}{\sigma^2} = \frac{\frac{(n-1)s^2}{\sigma^2}}{n-1} \quad \text{and} \quad \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

so, taking $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ and $U = \dfrac{(n-1)s^2}{\sigma^2}$ we can apply the previous proposition and get the desired result.

* Sampling distributions and confidence intervals

As we saw before, to find the probability that estimates of parameters are within certain bounds we need to know their distributions.

Consider a sample $X_1, X_2, \ldots, X_n$. If the X_i are Gaussian and independent then $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ and this is valid regardless of the size of the sample n. If the population is not Gaussian but n is large then the Central Limit Theorem assures us that the result is valid (approximately).

If the population variance is not known and we have to replace it with its estimate s^2 then $T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{(n-1)}$ where $t_{(n-1)}$ is the Student's t, or simply t, distribution with $n-1$ degrees of freedom.

The sample mean is a point estimate of the population mean. An alternative to this is a interval estimate, which is an interval where the true value of the parameter falls with certain probability.

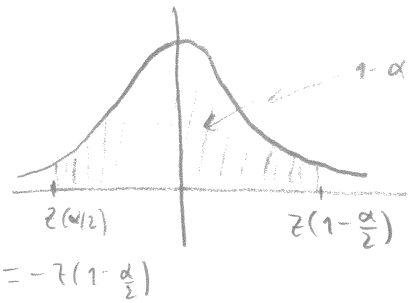
Definition. Confidence interval. Let μ be a population parameter. A $100(1-\alpha)\%$ confidence interval is an interval which contains the true value of μ with probability $1-\alpha$. The limits of the interval are called confidence limits and $1-\alpha$ is the confidence level.

Suppose a sample $X_1, X_2, \ldots X_n$ where X_i are Gaussian and independent then $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$, thus

$$Pr\left(z(\alpha/2) \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z(1-\tfrac{\alpha}{2}) \right) = 1-\alpha$$

$$\Leftrightarrow$$

$$Pr\left(-z(1-\tfrac{\alpha}{2}) \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z(1-\tfrac{\alpha}{2}) \right) = 1-\alpha$$



$$Pr\left(-z(1-\tfrac{\alpha}{2})\frac{\sigma}{\sqrt{n}} \le \bar{X}-\mu \le z(1-\tfrac{\alpha}{2})\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$Pr\left(\bar{X}-z(1-\tfrac{\alpha}{2})\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}+z(1-\tfrac{\alpha}{2})\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

So, the probability that μ is in $\left[\bar{X}-z(1-\frac{\alpha}{2})\frac{\sigma}{\sqrt{n}},\ \bar{X}+z(1-\frac{\alpha}{2})\frac{\sigma}{\sqrt{n}}\right]$ is $1-\alpha$

By a similar procedure we find that $\left[\bar{X}-t_{(n-1)}(1-\frac{\alpha}{2})\frac{S}{\sqrt{n}},\ \bar{X}+t_{(n-1)}(1-\frac{\alpha}{2})\frac{S}{\sqrt{n}}\right]$ is also a $100(1-\alpha)\%$ confidence interval for μ when σ^2 is unknown.

Example. For a certain population with parameters $\mu = 500$ and $\sigma = 10$ a sample of size $n = 25$ yields $\bar{x} = 480$, then a confidence interval at 95% confidence level for μ is

$$\left[\bar{X} - z\left(1 - \frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} , \bar{X} + z\left(1 - \frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} \right] = \left[480 - 1.96 \cdot \frac{10}{\sqrt{25}} , 480 + 1.96 \cdot \frac{10}{\sqrt{25}} \right] = [476.08, 483.92]$$

which actually doesn't contain the true value of μ.

To illustrate the concept of confidence we repeat the procedure 99 more times, that is, we draw another 99 samples of size $n = 25$.
Each one of those samples will yield an estimate $\bar{x}_i$, $i = 1, 2, \ldots, 100$ counting the first one, and in general they will be different. If we then use those estimates to construct intervals, all of them at 95% confidence, then in average 95 of them will contain the true value of μ.