CONFIDENCE INTERVALS

Simply reporting the value of an estimator e.g. $(\bar{X})$,...

Simply reporting the value of an estimator e.g. $(\bar{X})$,...

tells us nothing about the magnitude of the discrepancy that may exist between the estimator and the estimated parameter $(E\{X\} = \mu)$. What would be the "**confidence interval"** of the estimated PARAMETER? WE MAY DEFINE the confidence interval in the following manner:

Simply reporting the value of an estimator e.g. $(\bar{X})$,...

tells us nothing about the magnitude of the discrepancy that may exist between the estimator and the estimated parameter $(E\{X\} = \mu)$. What would be the "**confidence interval"** of the estimated PARAMETER?

WE MAY DEFINE the confidence interval in the following manner:

1. we choose a value of the *confidence level*:
   $1 - \alpha; \quad 0 < \alpha < 1$, or α. Usually: $\alpha = 0,01; \ \mathbf{0,05}; \ 0,1$

Simply reporting the value of an estimator e.g. $(\bar{X})$,...

tells us nothing about the magnitude of the discrepancy that may exist between the estimator and the estimated parameter $(E\{X\} = \mu)$. What would be the "**confidence interval**" of the estimated PARAMETER?
WE MAY DEFINE the confidence interval in the following manner:

1. we choose a value of the *confidence level*:
   $1 - \alpha; \quad 0 < \alpha < 1$, or α. Usually: $\alpha = 0,01; \;\; \mathbf{0,05}; \;\; 0,1$
2. In practice we are looking for a distribution parameter λ

our data: a random sample $X_1, X_2, \ldots, X_n$
We form **two** statistics:

$$\begin{aligned} \lambda_1 &= \lambda_1(X_1, X_2, \ldots, X_n; \alpha) \\ \lambda_2 &= \lambda_2(X_1, X_2, \ldots, X_n; \alpha) \end{aligned}$$

The Confidence Interval $\stackrel{\text{def}}{=} \Delta = \lambda_2 - \lambda_1$
λ_1 i λ_2 HAVE BEEN CHOSEN IS SUCH A WAY THAT THE PROBABILITY FOR Δ TO "COVER" THE UNKNOWN λ IS $1 - \alpha$
In other words: we are allowed to think that n repetitions of the same procedure of estimating the Confidence Interval will produce n (different) confidence intervals of which 100(1-α) percent will contain the (looked for) parameter λ.

EXAMPLE

The distribution of RV X in a given population is normal: $N(\mu, \sigma)$
μ — unknown and we want to construct its confidence interval at **the confidence level** $1 - \alpha$; the smd σ is known (e.g. – it may be the error of our single measurement)
The random sample is : $X_1, X_2, \ldots, X_n$
The point estimator of μ is the $\bar{X}$ statistic,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{its pdf is} \quad N(\mu, \frac{\sigma}{\sqrt{n}})$$

The standardised statistic

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad \text{has the pdf} \quad N(0, 1)$$

EXAMPLE

Let z_1 i z_2 be the two quantiles of the STANDARDISED NORMAL DISTRIBUTION for which

$$\mathcal{P}(z_1 < Z < z_2) = F_N(z_2) - F_N(z_1) = 1 - \alpha$$

where F_N is the cumulative distribution of the STANDARDISED NORMAL VARIABLE, whose distrb. function is $f_N(z)$

$$\begin{aligned}
\alpha_1 &= F_N(z_1) = \int_{-\infty}^{z_1} f_N(z)dz; \quad z_1 \equiv z(\alpha_1) \\
1 - \alpha_2 &= F_N(z_2) = \int_{-\infty}^{z_2} f_N(z)dz; \quad z_2 \equiv z(1 - \alpha_2) \\
1 - \alpha &= F_N(z_2) - F_N(z_1) = \int_{z_1}^{z_2} f_N(z)dz
\end{aligned}$$

EXAMPLE

$$
\begin{aligned}
\alpha_1 &= F_N(z_1) = \int_{-\infty}^{z_1} f_N(z)dz; \quad z_1 \equiv z(\alpha_1) \\
1-\alpha_2 &= F_N(z_2) = \int_{-\infty}^{z_2} f_N(z)dz; \quad z_2 \equiv z(1-\alpha_2) \\
1-\alpha &= F_N(z_2) - F_N(z_1) = \int_{z_1}^{z_2} f_N(z)dz
\end{aligned}
$$

EXAMPLE

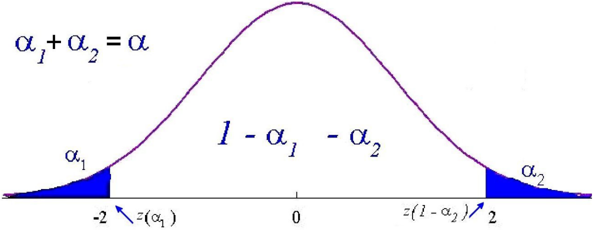
$$\begin{aligned}
\alpha_1 &= F_N(z_1) = \int_{-\infty}^{z_1} f_N(z)dz; \quad z_1 \equiv z(\alpha_1) \\
1-\alpha_2 &= F_N(z_2) = \int_{-\infty}^{z_2} f_N(z)dz; \quad z_2 \equiv z(1-\alpha_2) \\
1-\alpha &= F_N(z_2) - F_N(z_1) = \int_{z_1}^{z_2} f_N(z)dz
\end{aligned}$$

here comes the graph of the standardised normal cummulative distribution and the $\alpha_{1,2}$ regions

EXAMPLE

$$\begin{aligned} \alpha_1 &= F_N(z_1) = \int_{-\infty}^{z_1} f_N(z)dz; \quad z_1 \equiv z(\alpha_1) \\ 1-\alpha_2 &= F_N(z_2) = \int_{-\infty}^{z_2} f_N(z)dz; \quad z_2 \equiv z(1-\alpha_2) \\ 1-\alpha &= F_N(z_2) - F_N(z_1) = \int_{z_1}^{z_2} f_N(z)dz \end{aligned}$$

here comes the graph of the standardised normal cummulative distribution and the $\alpha_{1,2}$ regions



formulae . . .

$$P\left[z(\alpha_1) < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z(1-\alpha_2)\right] = 1-\alpha$$

$$\frac{z(\alpha_1)\sigma}{\sqrt{n}} < \bar{X}-\mu < \frac{z(1-\alpha_2)\sigma}{\sqrt{n}}$$

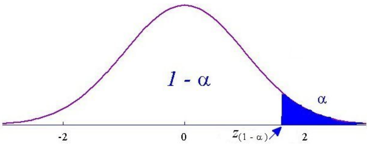
$$\bar{X} - \frac{z(\alpha_1)\sigma}{\sqrt{n}} > \mu > \bar{X} - \frac{z(1-\alpha_2)\sigma}{\sqrt{n}}$$

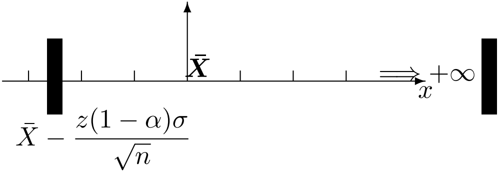
We may have 3 cases:

We may have 3 cases. . .

1. LOWER one-sided confidence interval: $\alpha_1 = 0 \quad z(\alpha_1) = -\infty$ $z(\alpha_2) = z(1-\alpha)$; the interval is:

$$\left(\bar{X} - z(1-\alpha)\frac{\sigma}{\sqrt{n}},\ +\infty \right)$$



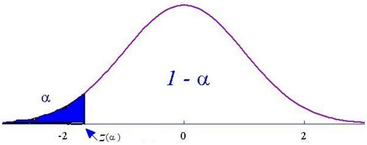


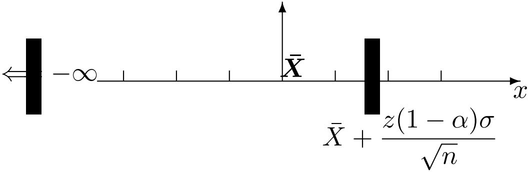
we may be $1-\alpha$ certain that μ is **no less** than $\bar{X} - \dfrac{z(1-\alpha)\sigma}{\sqrt{n}}$

We may have 3 cases...

2. UPPER one-sided confidence interval $\alpha_2 = 0 \quad z(1-\alpha_2) = \infty$ the interval is:

$$\left(-\infty,\ \bar{X} - z(\alpha)\frac{\sigma}{\sqrt{n}}\right) \equiv \left(-\infty,\ \bar{X} + z(1-\alpha)\frac{\sigma}{\sqrt{n}}\right)$$



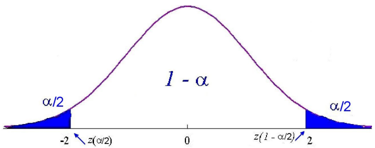


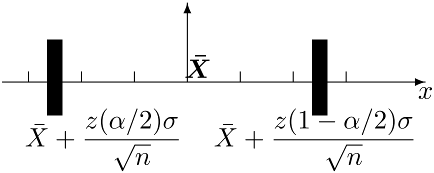
we may be $1-\alpha$ certain that μ is **not greater** than $\bar{X} + \dfrac{z(1-\alpha)\sigma}{\sqrt{n}}$

We may have 3 cases. . .

3. two-sided (symmetric) confidence interval (most frequent)
   $\alpha_1 = \alpha_2 = \dfrac{\alpha}{2}$
   the interval is:

$$\left(\bar{X} + z(\frac{\alpha}{2})\frac{\sigma}{\sqrt{n}},\ \bar{X} + z(1-\frac{\alpha}{2})\frac{\sigma}{\sqrt{n}}\right) \equiv \left(\bar{X} \mp z(1-\frac{\alpha}{2})\frac{\sigma}{\sqrt{n}}\right)$$





the former formulae assumed σ to be known (given). What if we don't know (have) it?

1. **big sample**; $n \geq 30 - 100$
   we may estimate σ with a fair accuracy by its unbiased estimator:
   $$\sigma \approx S^* = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
   so the two-sided (symmetric) interval will be
   $$\left(\bar{x} - z(1-\frac{\alpha}{2})\frac{S^*}{\sqrt{n}},\ \bar{x} + z(1-\frac{\alpha}{2})\frac{S^*}{\sqrt{n}}\right)$$
2. **the sample is not too numerous.**
   We introduce the new RV t:
   $$t = \frac{\bar{X}-\mu}{S}\sqrt{n-1} = \frac{\bar{X}-\mu}{S^*}\sqrt{n}$$
   let's recall: $$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$
   The new RV has the so–called STUDENT's t DISTRIBUTION

The new RV has the so–called STUDENT's DISTRIBUTION

(or t distribution) with $\nu = n - 1$ *degrees of freedom*. The only parameter of this distribution is n (ν).

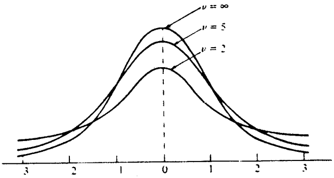
Note: this is the case most frequently met in practice. That's why the t-distribution is so very important. The STUDENT's distribution or, simply, the t distribution is given by:

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

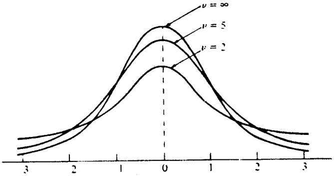
RV's Exp.Val. is: $E\{t\} = 0$; and its variance $VAR\{t\} = \frac{\nu}{\nu - 2}$; $(\nu > 2)$

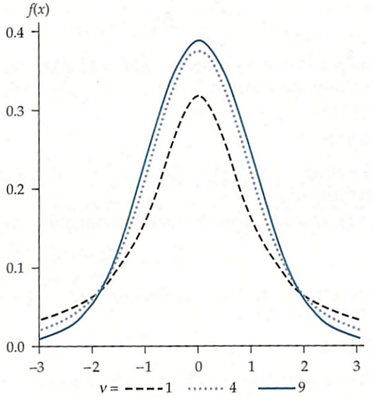
Note: by convention (tradition) a variable having the Student's distribution is denoted by (small !) t.

The STUDENT's distribution



The STUDENT's distribution





Returning to the problem of interval estimation:

the quantiles $z(\alpha_1)$ and $z(\alpha_2)$ of the standardised normal distribution have to be replaced by analogous quantiles: $t(\alpha_1)$ and $t(\alpha_2)$ of the Student's distribution, so the two-sided (symmetric) interval will be

$$1-\alpha = P\left[|t| < t(1-\frac{1}{2}\alpha, n-1)\right] = P\left[\left|\frac{\bar{X}-\mu}{S}\sqrt{n-1}\right| < t(1-\frac{1}{2}\alpha, n-1)\right]$$

$$\bar{X} - t(1-\frac{1}{2}\alpha, n-1)\frac{S}{\sqrt{n-1}} < \mu < \bar{X} + t(1-\frac{1}{2}\alpha, n-1)\frac{S}{\sqrt{n-1}}$$

Dstb	Alpha value — $\alpha =$				
	0.90	0.95	0.975	0.99	0.995
t(10)	1.37	1.81	2.23	2.76	3.17
t(30)	1.31	1.70	2.04	2.46	2.75
t(100)	1.29	1.66	1.99	2.37	2.67
N	1.28	1.64	1.96	2.33	2.56

The RV X of our population follows a normal distribution – $N(\mu, \sigma)$ – we ignore both distribution parameters. The sample size is ≤ 30:
We introduce the "chi-square" STATISTIC:

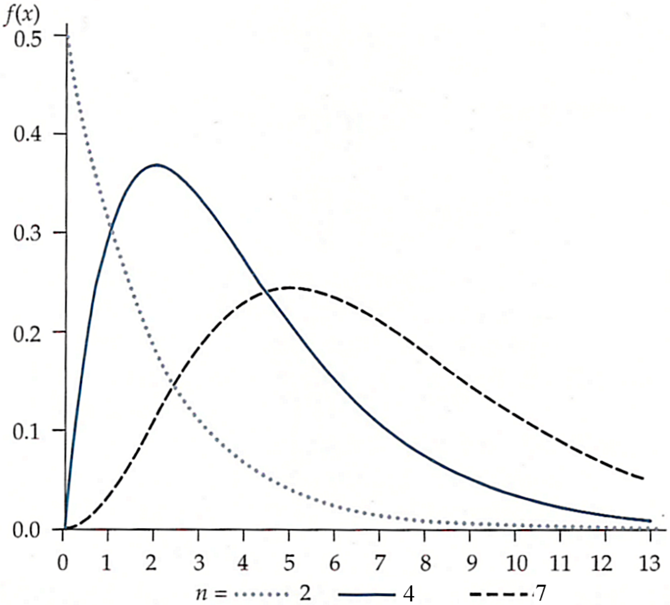
$$\chi^2 = \frac{nS^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$$

This statistic (RV) has a certain distribution – the so–called "chi-square" distribution— again its only parameter is the number of degrees of freedom: $\nu = n - 1$
THE CHI–SQUARE DISTRIBUTION FUNCTION is given by the formula:

$$f(\chi^2) = \frac{1}{\Gamma(\nu)2^\nu} \left(\chi^2\right)^{\nu-1} e^{-\frac{1}{2}\chi^2}$$

$$E\{\chi^2\} = \nu; \quad VAR\{\chi^2\} = 2\nu$$

the "chi-square" distribution:



Unlikely to the most RV distribution functions the distribution χ^2 is not symmetric so even if constructing a two–sided (symmetric) confidence interval we need TWO quantiles: $\chi^2(\alpha/2)$ i $\chi^2(1-\alpha/2)$.
the two–sided (symmetric) confidence interval will be given by

$$1-\alpha = P[\chi^2(\frac{1}{2}\alpha, n-1) < \chi^2 < \chi^2(1-\frac{1}{2}\alpha, n-1)] \quad \text{or}$$

$$1-\alpha = P[\chi^2(\frac{1}{2}\alpha, n-1) < \frac{nS^2}{\sigma^2} < \chi^2(1-\frac{1}{2}\alpha, n-1)]$$

so we have

$$\frac{nS^2}{\chi^2(1-\frac{1}{2}\alpha, n-1)} < \sigma^2 < \frac{nS^2}{\chi^2(\frac{1}{2}\alpha, n-1)}$$

For big sample sizes

we may make use of the fact that the χ^2 distributions tends (for big n) to a normal distribution:

$$\sqrt{2\chi^2} = \sqrt{2n}\frac{S}{\sigma} \to N(\sqrt{2n-3}, 1)$$

Consequently, the two–sided (symmetric) confidence interval for the msd σ (the square-root of variance) will be given by:

$$\frac{S\sqrt{2n}}{\sqrt{2n-3} + z(1-\alpha/2)} < \sigma < \frac{S\sqrt{2n}}{\sqrt{2n-3} - z(1-\alpha/2)}$$

The χ^2 distribution

should be always associated with a RV which describes the dispersion of the **square of the deviations** of an RV around a fixed point. A natural question would be: what if this central point is the "true" expected value of X, μ_X (and not its estimator $\bar{X}$. The answer is: The variable

$$\chi^2 = \sum_{i=1}^{n} \frac{(X_i - E\{X\})^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \mu_X)^2}{\sigma^2}$$

has indeed a χ^2 distribution with $\nu = n$ (!) degrees of freedom.