Confidence Intervals

The CLT tells us:

as the sample size *n* increases, the sample mean is approximately Normal with mean μ and standard deviation $\sigma/\sqrt{n}$.

Thus, we have a standard normal variable

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

If the underlying population is Normally distributed, we don't need CLT or large sample size for the sample mean to be Normally distributed – normality is guaranteed.

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Confidence interval for sample mean

Because the area under the standard normal curve between –1.96 and 1.96 is .95, we know:

$$P\left(-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = .95$$

This is equivalent to:

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = .95$$

which can be interpreted as the probability that the interval

$$\left(\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}},\quad \overline{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right)$$

includes the true mean μ is 95%.

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Confidence interval for sample mean

The interval

$$\left(\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}},\quad \overline{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right)$$

is thus called the 95% confidence interval for the mean.

This interval varies from sample to sample, as the sample mean varies.

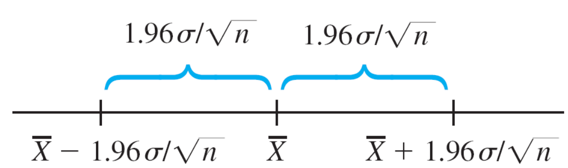
So the interval itself is a random interval: its bounds are random variables.

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Confidence interval for sample mean

The CI interval is centered at the sample mean and extends $1.96\sigma/\sqrt{n}$ to each side of the sample mean.

Thus the interval's width is 2 (1.96)$\sigma/\sqrt{n}$ and is not random; only the interval boundaries are random



Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Basic Properties of Confidence Intervals

For a given sample, the CI can be expressed either as

$$\left(\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) \text{ is a 95\% CI for } \mu$$

or as

$$\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} \text{ with 95\% confidence}$$

A concise expression for the interval is $\bar{x} \pm 1.96\, \sigma/\sqrt{n}$
where – gives the left endpoint (lower limit) and + gives the right endpoint (upper limit).

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Interpreting a Confidence Level

We started with an event (that the random interval captures the true value μ) whose probability was .95

It is tempting to say that μ *lies* within this fixed interval with probability 0.95.

μ is a constant (unfortunately unknown to us). It is therefore *incorrect* to write the statement

P(μ lies in (*a*, *b*)) = 0.95

-- since μ either is in (*a*,*b*) or isn't.

Basically, μ is not random (it's a constant), so it can't have a probability associated with its behavior.

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Interpreting a Confidence Level

Instead, a correct interpretation of "95% confidence" relies on the long-run relative frequency interpretation of probability.

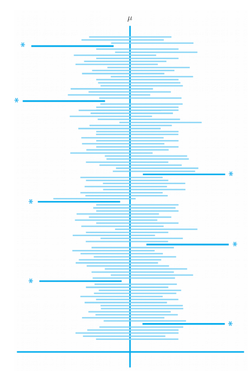
To say that an event *A* has probability .95 is to say that if the same experiment is performed over and over again, in the long run *A* will occur 95% of the time.

So the right interpretation is to say that in repeated sampling, 95% of the confidence intervals obtained from all samples will actually contain μ. The other 5% of the intervals will not.

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Interpreting a Confidence Level

Example: the vertical line cuts the measurement axis at the true (but unknown) value of μ.



One hundred 95% CIs (asterisks identify intervals that do not include μ).

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Interpreting a Confidence Level

Notice that 7 of the 100 intervals shown fail to contain μ.

In the long run, only 5% of the intervals so constructed would fail to contain μ.

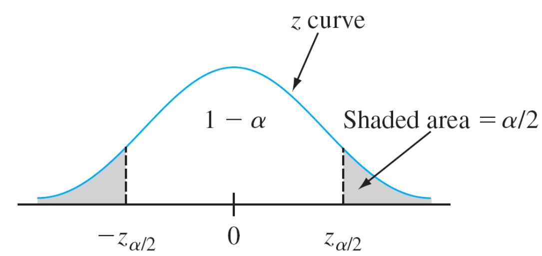
According to this interpretation, the confidence level is not a statement about any particular interval, eg (79.3, 80.7).

Instead it pertains to what would happen if a very large number of like intervals were to be constructed using the same CI formula.

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Other Levels of Confidence

Probability of $1 - \alpha$ is achieved by using $z_{\alpha/2}$ in place of 1.96



$$P(-z_{\alpha/2} \leq Z < z_{\alpha/2}) = 1 - \alpha$$

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Other Levels of Confidence

A **100(1 – α)% confidence interval** for the mean μ when the value of σ is known is given by

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

or, equivalently, by

$$\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}.$$

The formula for the CI can also be expressed in words as

Point estimate ± (z critical value) (standard error).

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Example

A sample of 40 units is selected and diameter measured for each one. The sample mean diameter is 5.426 mm, and the standard deviation of measurements is 0.1mm.

Let's calculate a confidence interval for true average diameter using a confidence level of 90%. This requires that 100(1 – α) = 90, from which α = .10.

Using `qnorm(0.05)`

$z_{\alpha/2} = z_{.05} = 1.645$

(corresponding to a cumulative *z*-curve area of .95).

The desired interval is then

$$5.426 \pm (1.645)\frac{.100}{\sqrt{40}} = 5.426 \pm .026 = (5.400, 5.452)$$

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Interval width

Since the 95% interval extends $1.96\sigma/\sqrt{n}$ to each side of

$\overline{x}$, the width of the interval is $2(1.96)\sigma/\sqrt{n}$ = 3.92

Similarly, the width of the 99% interval is (using `qnorm(0.005)`)

2(2.58) = 5.$\sigma/\sqrt{n}$ $\sigma/\sqrt{n}$

We have more confidence that the 99% interval includes the true value precisely because it is wider.

The higher the desired degree of confidence, the wider the resulting interval will be.

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Sample size computation

For each desired confidence level and interval width, we can determine the necessary sample size.

Example: A response time is Normally distributed with standard deviation 25 millisec. A new system has been installed, and we wish to estimate the true average response time μ for the new environment.

Assuming that response times are still normally distributed with σ = 25, what sample size is necessary to ensure that the resulting 95% CI has a width of (at most) 10?

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Example

cont'd

The sample size *n* must satisfy

$$10 = 2 \cdot (1.96)(25/\sqrt{n})$$

Rearranging this equation gives

$\sqrt{n}$ = 2 (1.96)(25)/10 = 9.80

So

$n = (9.80)^2 = 96.04$

Since *n* must be an integer, a sample size of 97 is required.

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Unknown mean and variance

We know that
- a CI for the mean μ of a normal distribution
- a large-sample CI for μ for any distribution

with a confidence level of 100(1 – α) % is:

$$\overline{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$$

A practical difficulty is the value of σ, which will rarely be known. Instead we work with the standardized variable

$$(\overline{X} - \mu)/(S/\sqrt{n})$$

Where the sample standard deviation *S* has replaced σ.

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Unknown mean and variance

Previously, there was randomness only in the numerator of Z by virtue of $\overline{X}$, the estimator.

In the new standardized variable, both $\overline{X}$ *and* S vary in value from one sample to another.

$$(\overline{X} - \mu)/(S/\sqrt{n})$$

Thus the distribution of this new variable should be wider than the Normal to reflect the extra uncertainty. This is indeed true when *n* is small.

However, for large *n* the substitution of S for σ adds little extra variability, so this variable also has approximately a standard normal distribution.

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

A Large-Sample Interval for μ

If *n* is sufficiently large, the standardized variable

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has approximately a standard normal distribution. This implies that

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

is a **large-sample confidence interval for μ** with confidence level approximately $100(1 - \alpha)$ %.

This formula is valid regardless of the population distribution.

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

A Large-Sample Interval for μ

In words, the CI is

point estimate of $\mu \pm$ (*z* critical value) (estimated standard error of the mean).

Generally speaking, $n > 40$ will be sufficient to justify the use of this interval.

This is somewhat more conservative than the rule of thumb for the CLT because of the additional variability introduced by using *S* in place of σ.

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Small sample intervals for the mean

- The CI for μ presented in earlier section is valid provided that *n* is large
  - Rule of thumb: n>40
  - The resulting interval can be used whatever the nature of the population distribution.
- The CLT cannot be invoked when *n* is small
  - Need to do something else when n<40
- When n<40, we have to
  - make a specific assumption about the form of the population distribution and
  - then derive a CI tailored to that assumption.

- For example, we could develop a CI for μ when the population is described by a Normal, or gamma distribution, or a Weibull distribution, and so on.

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

t Distributions

Small Sample Intervals Based on a Normal Population Distribution

The result on which inference is based introduces a new family of probability distributions called *t distributions.*

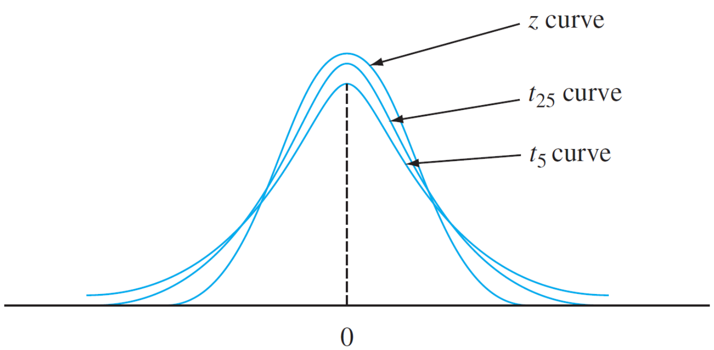
When $\overline{X}$ is the sample mean of a random sample of size n from a **normal distribution** with mean μ, the rv

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has a probability distribution called a t distribution with $n - 1$ degrees of freedom (df).

Properties of t Distributions

Figure below illustrates some members of the t-family



Properties of t Distributions

Properties of t Distributions

Let t_ν denote the t distribution with ν df.

1. Each t_ν curve is bell-shaped and centered at 0.

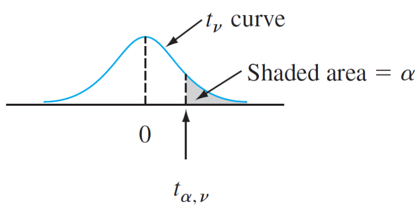
2. Each t_ν curve is more spread out than the standard normal (z) curve.

3. As ν increases, the spread of the corresponding t_ν curve decreases.

4. As $\nu \to \infty$, the sequence of t_ν curves approaches the standard normal curve (so the z curve is the t curve with df = ∞).

Properties of *t* Distributions

Let $t_{\alpha,\nu}$ = the number on the measurement axis for which the area under the *t* curve with ν df to the right of $t_{\alpha,\nu}$ is α; $t_{\alpha,\nu}$ is called a ***t* critical value**.



For example, $t_{.05,6}$ is the *t* critical value that captures an upper-tail area of .05 under the *t* curve with 6 df

Because *t* curves are symmetric about zero, $-t_{\alpha,\nu}$ captures lower-tail area α.

The One-Sample *t* Confidence Interval

Let $\overline{X}$ and *s* be the sample mean and sample standard deviation computed from the results of a random sample from a normal population with mean μ.

Then a **100(1 – α)% confidence interval for μ** is

$$\left(\bar{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}},\ \bar{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}\right)$$

or, more compactly

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot s/\sqrt{n}.$$

Example

cont'd

A dataset on the modulus of material rupture (psi):

6807.99	7637.06	6663.28	6165.03	6991.41	6992.23
6981.46	7569.75	7437.88	6872.39	7663.18	6032.28
6906.04	6617.17	6984.12	7093.71	7659.50	7378.61
7295.54	6702.76	7440.17	8053.26	8284.75	7347.95
7422.69	7886.87	6316.67	7713.65	7503.33	7674.99

There are 30 observations.

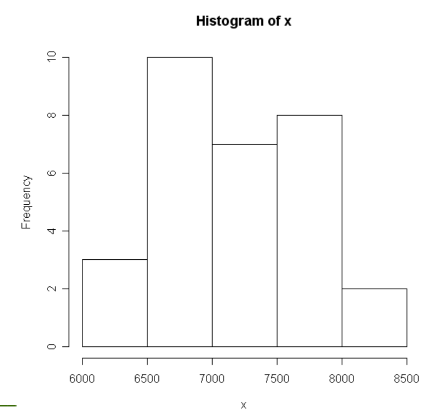
The sample mean is 7203.191

The sample standard deviation is 543.5400.

Example

cont'd

The histogram provides support for assuming that the population distribution is at least approximately normal.



Example

cont'd

Recall the sample mean and sample standard deviation are 7203.191 and 543.5400, respectively. The 95% CI is based on $n - 1 = 29$ degrees of freedom, so the necessary critical value is $t_{.025,29} = 2.045$. The interval estimate is now

$$\bar{x} \pm t_{.025,29} \cdot \frac{s}{\sqrt{n}} = 7203.191 \pm (2.045) \cdot \frac{543.5400}{\sqrt{30}}$$

$$= 7203.191 \pm 202.938$$

$$= 7203.191 \pm 202.938 = (7000.253, 7406.129)$$

Intervals Based on Nonnormal Population Distributions

The one-sample t CI for μ is robust to small or even moderate departures from normality unless n is quite small.

By this we mean that if a critical value for 95% confidence, for example, is used in calculating the interval, the actual confidence level will be reasonably close to the nominal 95% level.

If, however, n is small and the population distribution is nonnormal, then the actual confidence level may be considerably different from the one you think you are using when you obtain a particular critical value from the t table.

General parameter Confidence Interval

A General Large-Sample Confidence Interval

The large-sample intervals

$$\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n} \quad \text{and} \quad \bar{x} \pm z_{\alpha/2} \cdot S/\sqrt{n}$$

are special cases of a general large-sample CI for a parameter θ.

Suppose that $\hat{\theta}$ is an estimator such that:
(1) It has approximately a normal distribution;
(2) it is (at least approximately) unbiased;
(3) an expression for $\sigma_{\hat{\theta}}$, the standard deviation of $\hat{\theta}$, is available.

Then,

$$P\left(-z_{\alpha/2} < \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} < z_{\alpha/2}\right) \approx 1 - \alpha$$

And the CI for θ is: $\hat{\theta} \pm z_{\alpha/2} \cdot s_{\hat{\theta}}$

A Confidence Interval for a Population Proportion

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

A Confidence Interval for a Population Proportion

Let *p* denote the proportion of "successes" in a population, where *success* identifies an individual or object that has a specified property (e.g., individuals who graduated from college, computers that do not need warranty service, etc.).

A random sample of *n* individuals is to be selected, and *X* is the number of successes in the sample.

X can be thought of as a sum of all X_i's, where 1 is added for every *success* that occurs and a 0 for every *failure*, so $X_1 + \ldots + X_n = X$).

Thus, *X* can be regarded as a binomial rv with mean *np* and $\sigma_X = \sqrt{np(1-p)}$.

Furthermore, if both $np \geq 10$ and $n(1-p) \geq 10$, *X* has approximately a normal distribution.

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

A Confidence Interval for a Population Proportion

The natural estimator of *p* is $\hat{p} = X/n$, the sample fraction of successes.

Since $\hat{p}$ is the sample mean, $(X_1 + \ldots + X_n)/n$

It has approximately a normal distribution. As we know that, $E(\hat{p}) = p$ (unbiasedness) and

$$\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$$

The standard deviation $\sigma_{\hat{p}}$ involves the unknown parameter *p*.

Standardizing $\hat{p}$ then implies that

$$P\left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} < z_{\alpha/2}\right) \approx 1 - \alpha$$

And the CI is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$$

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

One-Sided Confidence Intervals

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

One-Sided Confidence Intervals (Confidence Bounds)

The confidence intervals discussed thus far give both a lower confidence bound *and* an upper confidence bound for the parameter being estimated.

In some circumstances, an investigator will want only one of these two types of bounds.

For example, a psychologist may wish to calculate a 95% upper confidence bound for true average reaction time to a particular stimulus, or a reliability engineer may want only a lower confidence bound for true average lifetime of components of a certain type.

One-Sided Confidence Intervals (Confidence Bounds)

Because the cumulative area under the standard normal curve to the left of 1.645 is .95, we have

$$P\left(\frac{\overline{X} - \mu}{S/\sqrt{n}} < 1.645\right) \approx .95$$

One-Sided Confidence Intervals (Confidence Bounds)

Starting with $P(-1.645 < Z) \approx .95$ and manipulating the inequality results in the upper confidence bound. A similar argument gives a one-sided bound associated with any other confidence level.

Proposition

A **large-sample upper confidence bound for** μ is

$$\mu < \overline{x} + z_{\alpha} \cdot \frac{s}{\sqrt{n}}$$

and a **large-sample lower confidence bound for** μ is

$$\mu > \overline{x} - z_{\alpha} \cdot \frac{s}{\sqrt{n}}$$

Confidence Intervals for Variance of a normal population

Confidence Intervals for the Variance of a Normal Population

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with parameters μ and σ^2. Then

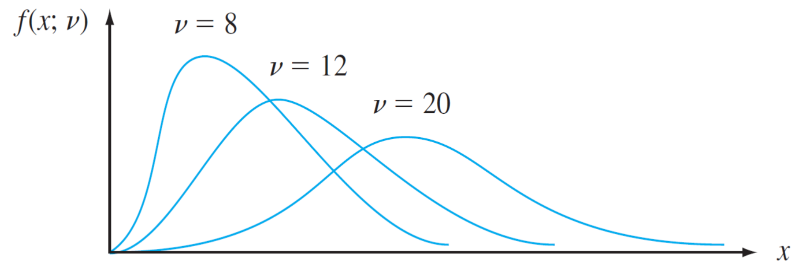
$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum(X_i - \overline{X})^2}{\sigma^2}$$

has a chi-squared (χ^2) probability distribution with $n - 1$ df.

We know that the chi-squared distribution is a continuous probability distribution with a single parameter v, called the number of degrees of freedom, with possible values 1, 2, 3, ...

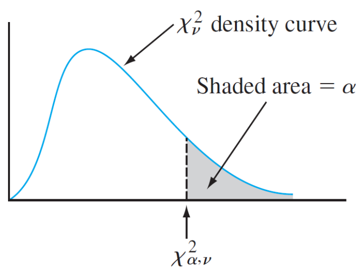
Confidence Intervals for the Variance of a Normal Population

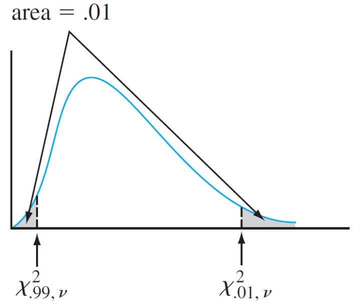
The graphs of several χ^2 probability density functions are



Confidence Intervals for the Variance of a Normal Population

The chi-squared distribution is not symmetric, so need values of $\chi^2_{\alpha,\nu}$ both for α near 0 and 1





Confidence Intervals for the Variance of a Normal Population

As a consequence

$$P\left(\chi^2_{1-\alpha/2,n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2,n-1}\right) = 1 - \alpha$$

Or equivalently

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

Substituting the computed value s^2 into the limits gives a CI for σ^2

Taking square roots gives an interval for σ.

Confidence Intervals for the Variance of a Normal Population

A **100(1 – α)% confidence interval for the variance σ^2 of a normal population** has lower limit

$$(n-1)s^2/\chi^2_{\alpha/2,n-1}$$

and upper limit

$$(n-1)s^2/\chi^2_{1-\alpha/2,n-1}$$

A **confidence interval for** σ has lower and upper limits that are the square roots of the corresponding limits in the interval for σ^2.

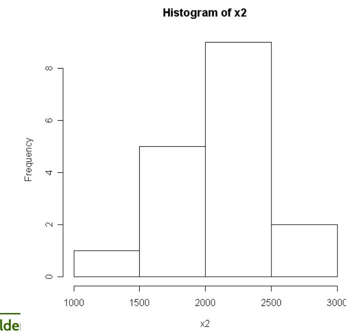
Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Example

The data on breakdown voltage of electrically stressed circuits are:

1470	1510	1690	1740	1900	2000	2030	2100	2190
2200	2290	2380	2390	2480	2500	2580	2700	

breakdown voltage is approximately normally distributed.



Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulde

Example

cont'd

Let σ^2 denote the variance of the breakdown voltage distribution. The computed value of the sample variance is s^2 = 137,324.3, the point estimate of σ^2.

With df = n – 1 = 16, a 95% CI requires $\chi^2_{.025,16}$ = 6.908 and $\chi^2_{.975,16}$ = 28.845.

The interval is

$$\left(\frac{16(137{,}324.3)}{28.845}, \frac{16(137{,}324.3)}{6.908}\right) = (76{,}172.3,\ 318{,}064.4)$$

Taking the square root of each endpoint yields (276.0, 564.0) as the 95% CI for σ.

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder

Probability intervals

Very different than confidence intervals

We need to make a probability statement about the random quantity you are predicting

For example, you have a random sample of size 10, and each Xi is iid Normal.

You can find the sample mean, and the CI for the true population mean

Or you can give a 95% interval for a new data point (X11) – that is a prediction interval, describing where X11 will be with 95% probability.

Copyright Prof. Vanja Dukic, Applied Mathematics, CU-Boulder