

# STAT 509: Statistics for Engineers

## Chapter 8: Statistical Intervals for a Single Sample

Dr. Dewei Wang  
Associate Professor  
Department of Statistics  
University of South Carolina  
[deweiwang@stat.sc.edu](mailto:deweiwang@stat.sc.edu)

## Chapter 8: Statistical Intervals for a Single Sample

### Learning Objectives:

1. Construct confidence intervals on the mean of a normal distribution, using either the normal distribution or the  $t$  distribution method
2. Construct confidence intervals on the variance and standard deviation of a normal distribution
3. Construct confidence intervals on a population proportion
4. Construct a prediction interval for a future observation

## An interval estimator

Estimating an unknown parameter  $\theta$  by a point estimator  $\hat{\Theta}_n$  is useful. However, it is like shooting a bird with a pistol. Often  $\hat{\Theta}_n$  has a continuous distribution, and if so, from Chapter 4,  $P(\hat{\Theta}_n = \theta) = 0$ ; i.e., we never capture the true parameter by using a point estimator even though it is an unbiased estimator.

Why not shoot a bird using a shotgun or capture it using a net? Translating to statistical language, why not use an interval to capture the true parameter? This motivates the consideration of interval estimators.

We estimate  $\theta$  by an interval  $[L_n, U_n]$ , where  $L_n$  and  $U_n$  are two statistics computed from a random sample of size  $n$  such that

$$P[L_n \leq \theta \leq U_n] = 1 - \alpha,$$

for some pre-specified  $\alpha \in (0, 1)$ . We call  $[L_n, U_n]$  a  $100(1 - \alpha)\%$  **confidence interval estimator** of  $\theta$  and  $100(1 - \alpha)\%$  the **confidence level** of this interval estimator.

# Confidence Interval on the Mean of a Normal Distribution, Variance Known

A confidence interval estimator is often built from a point estimator and the sampling distribution of the point estimator.

We consider building a confidence interval estimator of  $\mu$  in the normal distribution  $N(\mu, \sigma^2)$  where  $\sigma^2$  is known (based on historical data).

From Chapter 7, a good point estimator of  $\mu$  is  $\bar{X}_n$ , and the sampling distribution of  $\bar{X}_n$  is  $N(\mu, \sigma^2/n)$ . Thus

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

# Confidence Interval on the Mean of a Normal Distribution, Variance Known

For  $a \in (0, 1)$ , define  $z_a$  to be a quantile value of  $Z \sim N(0, 1)$  such that

$$P(Z > z_a) = a$$

Using StatEngine:

$$z_a = \text{normal.quantile}(0, 1, 1 - a).$$

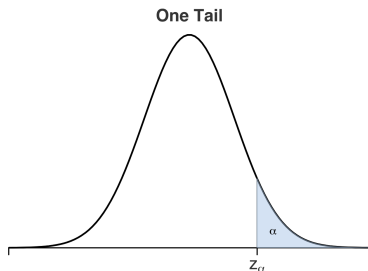
Often used  $z_a$  values:

$$z_{0.005} = 2.5758$$

$$z_{0.025} = 1.96$$

$$z_{0.05} = 1.6449$$

$$z_{0.1} = 1.2816.$$



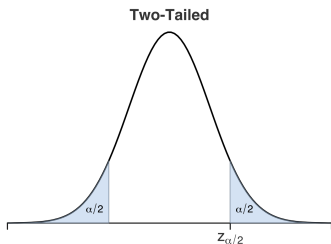
# Confidence Interval on the Mean of a Normal Distribution, Variance Known

For any  $\alpha < 0.5$ , we know that

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

Thus

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



# Confidence Interval on the Mean of a Normal distribution, variance known

## Two-sided confidence interval

If  $\bar{X}_n$  is the sample mean of size  $n$  from a **normal** population with **known variance**  $\sigma^2$ , a  $100(1 - \alpha)\%$  (two-sided) confidence interval estimator on  $\mu$  is given by

$$\left[ L_n = \bar{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, U_n = \bar{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right].$$

where  $z_{\alpha/2} = \text{normal.quantile}(0, 1, 1 - \alpha/2)$ .

Based on an observed sample  $x_1, \dots, x_n$ , a  $100(1 - \alpha)\%$  (two-sided) confidence interval estimate on  $\mu$  is

$$\left[ \bar{x}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

Its length is  $2 \times z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ , which becomes smaller if either  $n$  is larger (more data) or  $\alpha$  is larger (less confidence level).

# One-sided confidence bounds for the Mean of a Normal distribution, variance known

Similarly, we have  $1 - \alpha = P(Z \leq z_\alpha)$  and  $1 - \alpha = P(-z_\alpha \leq Z)$ .

Plugging  $Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ , we obtain

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

One-sided confidence bound

A  $100(1 - \alpha)\%$  **upper-confidence bound** for  $\mu$  is

$$\bar{X}_n + z_\alpha \frac{\sigma}{\sqrt{n}}$$

and a  $100(1 - \alpha)\%$  **lower-confidence bound** for  $\mu$  is

$$\bar{X}_n - z_\alpha \frac{\sigma}{\sqrt{n}}.$$



## Example

ASTM Standard E23 defines standard test methods for notched bar impact testing of metallic materials. The Charpy V-notch (CVN) technique measures impact energy and is often used to determine whether or not a material experiences a ductile-to-brittle transition with decreasing temperature. Ten measurements of impact energy (J) on specimens of A238 steel cut at  $60^{\circ}\text{C}$  are as follows: 64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, 64.3. Assume that impact energy is normally distributed with  $\sigma = 1\text{J}$ . We want to find 95% and 99% CIs for  $\mu$ , the mean impact energy.

```
x=c(64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, 64.3)
Zinterval(level=0.95,sigma=1,sample=x)
Zinterval(level=0.99,sigma=1,sample=x)
```

A 95% CI for  $\mu$  is [63.8402, 65.0798]

A 99% CI for  $\mu$  is [63.6455, 65.2746]

When confidence level increases (or  $\alpha$  decreases), confidence interval becomes wider! (A larger net captures a bird more easily.)

## Interpreting a confidence interval

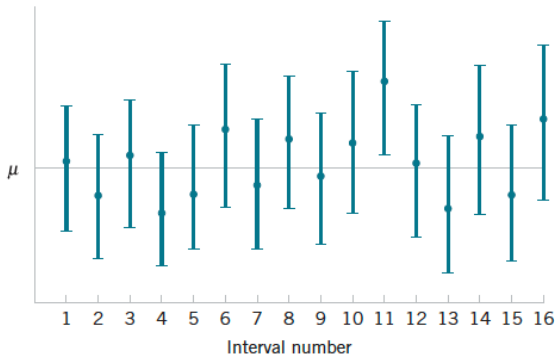
How does one interpret a confidence interval? In the previous example, a 95% CI is  $63.8402 \leq \mu \leq 65.0798$ , so it is tempting to conclude that  $\mu$  is within this interval with probability 0.95.

However, with a little reflection, it is easy to see that this **cannot be correct**; the true value of  $\mu$  is unknown, and the statement  $63.8402 \leq \mu \leq 65.0798$  is either correct (true with probability 1) or incorrect (false with probability 1).

The correct interpretation lies in the realization that a CI is a random interval because in the probability statement defining the end-points of the interval, both  $L_n$  and  $U_n$  are random variables. Consequently, the correct interpretation of a  $100(1 - \alpha)\%$  CI depends on **the relative frequency view of probability**. Specifically, if an infinite number of random samples are collected and a  $100(1 - \alpha)\%$  confidence interval for  $\mu$  is computed from each sample,  $100(1 - \alpha)\%$  of these intervals will contain the true value of  $\mu$ .

## Interpreting a confidence interval

The situation is illustrated in the following figure, which shows several  $100(1 - \alpha)\%$  confidence intervals for the mean  $\mu$  of a normal distribution. The dots at the center of the intervals indicate the point estimate of  $\mu$  (that is,  $\bar{x}_n$ ). Notice that one of the intervals fails to contain the true value of  $\mu$ . If this were a 95% confidence interval, in the long run only 5% of the intervals would fail to contain  $\mu$ .



## Choice of sample size

A  $100(1 - \alpha)\%$  CI takes the form of

$$\left[ \bar{x}_n - \underbrace{z_{\alpha/2} \frac{\sigma}{\sqrt{n}}}_{\text{margin of error}}, \bar{x}_n + \underbrace{z_{\alpha/2} \frac{\sigma}{\sqrt{n}}}_{\text{margin of error}} \right].$$

When  $\alpha$  or the confidence level is fixed, the margin of error becomes smaller if  $n$  increases. It means if we have more data, the  $100(1 - \alpha)\%$  CI has more precision of estimation. Suppose we specify the margin of error to be  $E$ , what is the smallest amount sample to collect?

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \implies n = \left\lceil \left( \frac{z_{\alpha/2} \sigma}{E} \right)^2 \right\rceil,$$

where  $\lceil a \rceil$  means the ceiling of  $a$ .

## Choice of sample size

Consider the previous example and suppose that we want to determine how many specimens must be tested to ensure that the 95% CI on  $\mu$  for A238 steel cut at  $60^\circ\text{C}$  has a length of at most 1.0 J.

**Solution:** The length is at most 1J, meaning the margin of error  $E$  is at most 0.5J. Thus ( $\alpha = 0.05$ ,  $\sigma = 1$ )

$$n = \left\lceil \left( \frac{z_{\alpha/2}\sigma}{E} \right)^2 \right\rceil = \left\lceil \left( \frac{z_{0.025} \times 1}{0.5} \right)^2 \right\rceil = 16.$$

```
sample.size.Zinterval(level=0.95,sigma=1,E=0.5)
```

CI for  $\mu$  when  $\sigma$  is **known** and the distribution is **normal**:

```
Zinterval(level=?,sigma=?,sample=?)
```

```
Zinterval(level=?,sigma=?,n=?,barx=?)
```

```
sample.size.Zinterval(level=?,sigma=?,E=?)
```

What if we do not know the variance  $\sigma^2$ ?

What if it is not normal?

# Confidence Interval on the Mean of a Normal Distribution, Variance Unknown

Let  $X_1, \dots, X_n$  be a random sample from a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . We know that

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

When  $\sigma$  is unknown, we replace  $\sigma$  by its estimator  $S_n$  and obtain

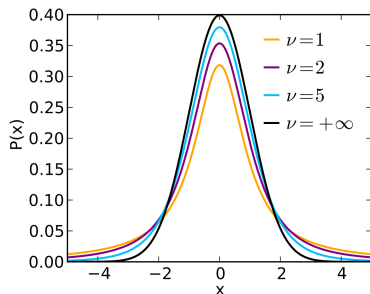
$$T_{n-1} = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1)$$

where  $t(n-1)$  stands for the **student  $t$  distribution** with degree of freedom  $n-1$ .

## Student $t(\nu)$ distribution

It is a continuous distribution with one parameter  $\nu$ , the degree of freedom. Its pdf is

$$f(x) = \frac{\Gamma(\{\nu + 1\}/2)}{\sqrt{\pi\nu}\Gamma(\nu/2)} \frac{1}{[(x^2/\nu) + 1]^{(\nu+1)/2}}, -\infty < x < \infty.$$



- ▶ Similar to normal distributions: bell shape, symmetric with respect to 0
- ▶ Heavier tails than  $N(0, 1)$  when  $\nu$  is small
- ▶ When  $\nu \rightarrow \infty$ ,  $t(\nu)$  converges to  $N(0, 1)$



# Confidence Interval on the Mean of a Normal Distribution, Variance unknown

For  $a \in (0, 1)$ , define  $t_{n-1,a}$  to be a quantile value of  $t(n-1)$  such that

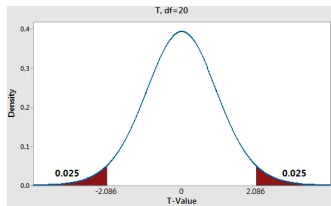
$$P(T_{n-1} > t_{n-1,a}) = a$$

Using StatEngine:

$$t_{n-1,a} = t.quantile(df = n - 1, 1 - a).$$

For any  $\alpha < 0.5$ , we know that

$$P(-t_{n-1,\alpha/2} \leq T_{n-1} \leq t_{n-1,\alpha/2}) = 1 - \alpha$$



$$\begin{aligned} &= P\left(-t_{n-1,\alpha/2} \leq \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \leq t_{n-1,\alpha/2}\right) \\ &= P\left(\bar{X}_n - t_{n-1,\alpha/2} \frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + t_{n-1,\alpha/2} \frac{S_n}{\sqrt{n}}\right) \end{aligned}$$

# Confidence Interval on the Mean of a Normal distribution, variance unknown

## Two-sided confidence interval

If  $\bar{X}_n$  is the sample mean of size  $n$  from a **normal** population with **unknown variance**  $\sigma^2$ , a  $100(1 - \alpha)\%$  (two-sided) confidence interval estimator on  $\mu$  is given by

$$\left[ L_n = \bar{X}_n - t_{n-1, \alpha/2} \frac{S_n}{\sqrt{n}}, U_n = \bar{X}_n + t_{n-1, \alpha/2} \frac{S_n}{\sqrt{n}} \right]$$

where  $t_{n-1, \alpha/2} = t.\text{quantile}(df = n - 1, 1 - \alpha/2)$ .

Based on an observed sample  $x_1, \dots, x_n$ , a  $100(1 - \alpha)\%$  (two-sided) confidence interval estimate on  $\mu$  is

$$\left[ \bar{x}_n - t_{n-1, \alpha/2} \frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{n-1, \alpha/2} \frac{s_n}{\sqrt{n}} \right].$$

Its length is  $2 \times t_{n-1, \alpha/2} \frac{s_n}{\sqrt{n}}$ , which becomes smaller if either  $n$  is larger (more data) or  $\alpha$  is larger (less confidence level).

# One-sided confidence bounds for the Mean of a Normal distribution, variance unknown

## One-sided confidence bound

A  $100(1 - \alpha)\%$  **upper-confidence bound** for  $\mu$  is

$$\bar{x}_n + t_{n-1,\alpha} \frac{s_n}{\sqrt{n}}$$

and a  $100(1 - \alpha)\%$  **lower-confidence bound** for  $\mu$  is

$$\bar{x}_n - t_{n-1,\alpha} \frac{s_n}{\sqrt{n}}.$$

StatEngine:

```
Tinterval(level=?,sample=?)
```

```
Tinterval(level=?,n=?,barx=?,s=?)
```

Remark: T-intervals are quite robust to the normality assumption when  $n$  is small. Thus, **in practice, even if we do not have normality, one can still use T-intervals.**

# Large-Sample Confidence Interval on the Mean of a population

## Two-sided confidence interval

If  $\bar{X}_n$  is the sample mean of size  $n$  from a population with mean  $\mu$  and a finite variance  $\sigma^2$ . When  $n$  is large ( $n \geq 25$ ), the CLT (plus the Slutsky theorem) tells

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim AN(0, 1).$$

Thus, a large-sample confidence interval estimator for  $\mu$  with confidence level of **approximately**  $100(1 - \alpha)\%$  is given by

$$\left[ L_n = \bar{X}_n - z_{\alpha/2} \frac{S_n}{\sqrt{n}}, U_n = \bar{X}_n + z_{\alpha/2} \frac{S_n}{\sqrt{n}} \right].$$

## One-sided confidence bounds for the Mean of a Normal distribution, variance known

Based on an observed sample  $x_1, \dots, x_n$ , a large-sample confidence interval estimator for  $\mu$  with confidence level of approximately  $100(1 - \alpha)\%$  is given by

$$\left[ \bar{x}_n - z_{\alpha/2} \frac{s_n}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2} \frac{s_n}{\sqrt{n}} \right].$$

### One-sided large-sample confidence bound

A  $100(1 - \alpha)\%$  **large-sample upper-confidence bound** for  $\mu$  is

$$\bar{x}_n + z_{\alpha} \frac{\sigma}{\sqrt{n}}$$

and a  $100(1 - \alpha)\%$  **large-sample lower-confidence bound** for  $\mu$  is

$$\bar{x}_n - z_{\alpha} \frac{\sigma}{\sqrt{n}}.$$

## StatEngine summary of CIs on the population mean

CI for  $\mu$  when  $\sigma$  is **known** and the distribution is **normal**:

Zinterval(level=?,sigma=?,sample=?)

Zinterval(level=?,sigma=?,n=?,barx=?)

sample.size.Zinterval(level=?,sigma=?,E=?)

CI for  $\mu$  when  $\sigma$  is **unknown** and the distribution is **normal** (or for any distribution, but in this case, it provides approximated CIs):

Tinterval(level=?,sample=?)

Tinterval(level=?,n=?,barx=?,s=?)

Large-sample CI ( $n \geq 25$ ) for  $\mu$  under any distribution:

AZinterval(level=?,sample=?)

AZinterval(level=?,n=?,barx=?,s=?)

Both T-interval and AZ-interval are approximated CIs when normality does not hold. The T-intervals are more conservative (wider).

## Practice 1

An article in the Journal of Materials Engineering [“Instrumented Tensile Adhesion Tests on Plasma Sprayed Thermal Barrier Coatings” (1989, Vol. 11(4), pp. 275–282)] describes the results of tensile adhesion tests on 22 U-700 alloy specimens. The load at specimen failure is as follows (in megapascals):

$x = c(19.8, 10.1, 14.9, 7.5, 15.4, 15.4, 15.4, 18.5, 7.9, 12.7, 11.9, 11.4, 11.4, 14.1, 17.6, 16.7, 15.8, 19.5, 8.8, 13.6, 11.9, 11.4)$

Find a 95% CI on  $\mu$ , the population mean load at specimen failure.

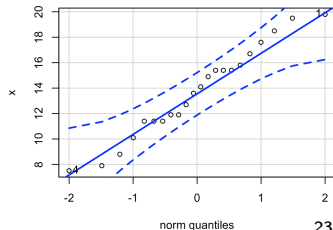
**Solution:** The sample size is  $n = 22 < 25$ , large-sample inference might not work. We do not know the population variance  $\sigma$  nor the type of the population distribution (did not say normal). But we can check normality first using the QQ plot.

```
data.summary(x)
```

It appears that the sample follows a normal distribution. Thus use the T-interval.

```
Tinterval(level=0.95,sample=x)
```

Conclusion: based on the data, we are 95% confident that the population mean load at specimen failure falls between 12.1381 and 15.2892.



## Practice 2

An article in the 1993 volume of the Transactions of the American Fisheries Society reports the results of a study to investigate the mercury contamination in large mouth bass. A sample of fish was selected from 53 Florida lakes, and mercury concentration in the muscle tissue was measured (ppm). The mercury concentration values were

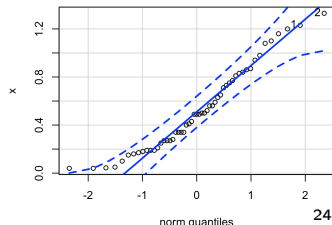
```
x=scan("https://raw.githubusercontent.com/Harrindyl/StatEngine/master/Data/Mercury.csv")
```

Find a 95% confidence interval estimate for  $\mu$ , the population mean mercury concentration.

**Solution:** The dashed lines do not cover all dots; i.e., the sample might not be from a normal distribution. But  $n = 53 > 25$ , we could use a large-sample CI.

```
data.summary(x)
AZinterval(level=0.95,sample=x)
Tinterval(level=0.95,sample=x) #Try this!
```

Conclusion: based on the data, we are 95% confident that the population mean mercury concentration falls between 0.4311 and 0.6188.





## Practice 3

Past experience has indicated that the breaking strength of yarn used in manufacturing drapery material is normally distributed and that  $\sigma = 2$  psi. A random sample of nine specimens is tested, and the average breaking strength is found to be 98 psi. Find a 95% two-sided confidence interval on the true mean breaking strength.

**Solution:** Normality and known  $\sigma = 2$ .

```
Zinterval(level=0.95,sigma=2,n=9,barx=98)
```

Conclusion: based on the data, we are 95% confident that the true mean breaking strength falls between 96.6934 and 99.3066.

## Practice 4

A confidence interval estimate is desired for the gain in a circuit on a semiconductor device. Assume that gain is normally distributed. Consider the following cases where we suppose the sample standard deviation  $s_n$  is always 20.

- (a) Find a 95% CI for  $\mu$  when  $n = 10$  and  $\bar{x}_n = 1000$ .
- (b) Find a 95% CI for  $\mu$  when  $n = 25$  and  $\bar{x}_n = 1000$ .
- (c) Find a 99% CI for  $\mu$  when  $n = 10$  and  $\bar{x}_n = 1000$ .
- (d) Find a 99% CI for  $\mu$  when  $n = 25$  and  $\bar{x}_n = 1000$ .

**Solution:** Normality and  $\sigma$  unknown.

```
Tinterval(level=0.95,n=10,barx=1000,s=20)
```

```
[ 985.6929 , 1014.307 ]
```

```
Tinterval(level=0.95,n=25,barx=1000,s=20)
```

```
[ 991.7444 , 1008.256 ]
```

```
Tinterval(level=0.99,n=10,barx=1000,s=20)
```

```
[ 979.4462 , 1020.554 ]
```

```
Tinterval(level=0.99,n=25,barx=1000,s=20)
```

```
[ 988.8122 , 1011.188 ]
```

# Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

StatEngine:

```
Chi2interval(level=?,sample=?)
```

```
Chi2interval(level=?,n=?,s=?)
```

Reasoning: Let  $X_1, \dots, X_n$  be a random sample from a normal distribution with mean  $\mu$  and variance  $\sigma^2$ , and let  $S_n^2$  be the sample variance. Then the random variable

$$\chi_{n-1}^2 = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$$

where  $\chi^2(n-1)$  stands for the chi-square distribution with  $n-1$  degrees of freedom.

# Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

Let  $\chi^2_{n-1,\alpha/2}$  and  $\chi^2_{n-1,1-\alpha/2}$  be the value such that

$$P(X^2_{n-1} > \chi^2_{n-1,\alpha/2}) = \alpha/2, \text{ and } P(X^2_{n-1} > \chi^2_{n-1,1-\alpha/2}) = 1-\alpha/2,$$

respectively, where  $\chi^2_{n-1,\alpha/2} = \text{Chi2.quantile}(df = n - 1, 1 - \alpha/2)$  and  $\chi^2_{n-1,1-\alpha/2} = \text{Chi2.quantile}(df = n - 1, \alpha/2)$ .

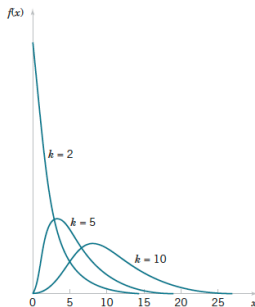
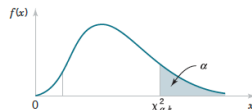
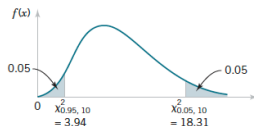


FIGURE 8.8

Probability density functions of several  $\chi^2$  distributions.



(a)



(b)

FIGURE 8.9

Percentage point of the  $\chi^2$  distribution. (a) The percentage point  $\chi^2_{\alpha,k}$ . (b) The upper percentage point  $\chi^2_{0.05,10} = 18.31$  and the lower percentage point  $\chi^2_{0.95,10} = 3.94$ .

# Confidence Interval on the Variance and Standard Deviation of a Normal Distribution

$$\begin{aligned}1 - \alpha &= P(\chi_{n-1, 1-\alpha/2}^2 \leq \mathbf{X}_{n-1}^2 \leq \chi_{n-1, \alpha/2}^2) \\&= P\left(\chi_{n-1, 1-\alpha/2}^2 \leq \frac{(n-1)S_n^2}{\sigma^2} \leq \chi_{n-1, \alpha/2}^2\right) \\&= P\left(\frac{(n-1)S_n^2}{\chi_{n-1, \alpha/2}^2} \leq \sigma^2 \leq \frac{(n-1)S_n^2}{\chi_{n-1, 1-\alpha/2}^2}\right)\end{aligned}$$

If  $s_n^2$  is the sample variance from a random sample of  $n$  observations from a **normal** distribution with unknown variance  $\sigma^2$ , then a  $100(1-\alpha)\%$  confidence interval on  $\sigma^2$  is

$$\frac{(n-1)s_n^2}{\chi_{n-1, \alpha/2}^2} \leq \sigma^2 \leq \frac{(n-1)s_n^2}{\chi_{n-1, 1-\alpha/2}^2}$$

# One-sided confidence bounds on the Variance and Standard Deviation of a Normal Distribution

The  $100(1 - \alpha)\%$  lower and upper confidence bounds on  $\sigma^2$  are

$$\sigma^2 \geq \frac{(n-1)s_n^2}{\chi_{n-1,\alpha}^2}, \text{ and } \sigma^2 \leq \frac{(n-1)s_n^2}{\chi_{n-1,1-\alpha}^2}.$$

respectively. If the parameter of interest is the population standard deviation  $\sigma$  instead of the population variance  $\sigma^2$ , one can take square root of the above results:

$$\sqrt{\frac{(n-1)s_n^2}{\chi_{n-1,\alpha/2}^2}} \leq \sigma \leq \sqrt{\frac{(n-1)s_n^2}{\chi_{n-1,1-\alpha/2}^2}},$$

$$\sigma \geq \sqrt{\frac{(n-1)s_n^2}{\chi_{n-1,\alpha}^2}}, \text{ and } \sigma \leq \sqrt{\frac{(n-1)s_n^2}{\chi_{n-1,1-\alpha}^2}}.$$

Remark: Chi2-intervals are **not** robust to the normality assumption.

## Example

An automatic filling machine is used to fill bottles with liquid detergent. A random sample of 20 bottles results in a sample variance of fill volume of  $s_n^2 = 0.01532^2$  (fluid ounce). If the variance of fill volume is too large, an unacceptable proportion of bottles will be under- or overfilled. We will assume that the fill volume is approximately **normally distributed**. Find a 95% upper confidence bound for  $\sigma$ , the population standard deviation of fill volume.

**Solution:** Normality checked (with data, use QQ plot).

```
Chi2interval(level=0.95,n=20,s=0.01532)
```

The sample standard variance is 0.0002347024 and sample size is 20

A 95% two-sided confidence interval for the population variance is [ 0.0001357391 , 0.0005006835 ]

A 95% upper-confidence bound for the population variance is 0.0004407769

A 95% lower-confidence bound for the population variance is 0.0001479371

The sample standard deviation is 0.01532 and sample size is 20

A 95% two-sided confidence interval for the population standard deviation is [ 0.01165071 , 0.02237596 ]

A 95% upper-confidence bound for the population standard deviation is 0.02099469

A 95% lower-confidence bound for the population standard deviation is 0.01216294

**Conclusion:** based on the data, we are 95% confident that the population standard deviation of fill volume  $\sigma$  is bounded above by 0.021.

# Large-Sample Confidence Interval for a Population Proportion

It is often necessary to construct confidence intervals on a population proportion. For example, suppose that a random sample of size  $n$  has been taken from a large (possibly infinite) population and that  $X(\leq n)$  observations in this sample belong to a class of interest. Then  $\hat{p} = \frac{X}{n}$  is a point estimator of the proportion of the population  $p$  that belongs to this class. Note that  $X \sim \text{Binomial}(n, p)$ .

When  $n$  is large (rule of thumb :  $n\hat{p} \geq 5, n(1 - \hat{p}) \geq 5$ ), we have

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim AN(0, 1).$$

Thus

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



# Large-Sample Confidence Interval for a Population Proportion

After some algebra and approximation, we have

## Approximate Confidence Interval on a Population Proportion

If  $\hat{p}$  is the proportion of observation in a random sample of size  $n$  that belongs to a class of interest, an approximate  $100(1 - \alpha)\%$  confidence interval on the proportion  $p$  of the population that belongs to this class is

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}.$$

Approximate  $100(1 - \alpha)\%$  One-Sided lower and upper Confidence Bounds are

$$p \geq \hat{p} - z_{\alpha} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \text{ and } p \leq \hat{p} + z_{\alpha} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}.$$

StatEngine: *Propinterval*(level =?, n =?, X =?).

## Choice of Sample Size

Suppose we want to choose  $n$  such that  $100(1 - \alpha)\%$  confident that the error is less than some specified value  $E$ , we have

$$n = \left( \frac{z_{\alpha/2}}{E} \right)^2 p(1 - p).$$

Now we have a question, if we know  $p$ , we can calculate  $n$ . However, if we know  $p$ , why do we need to estimate  $p$ ? Too solutions:

1. Suppose we have an initial estimate of  $p$ , denoted by  $\tilde{p}$ :

$$n = \left\lceil \left( \frac{z_{\alpha/2}}{E} \right)^2 \tilde{p}(1 - \tilde{p}) \right\rceil.$$

2. If no information about  $p$  is available, then we use a conservative approach (because  $p(1 - p) \leq 0.25$ )

$$n = \left\lceil \left( \frac{z_{\alpha/2}}{E} \right)^2 0.25 \right\rceil.$$

StatEngine:

*sample.size.Propinterval(level = ?, ini.p = ?, E = ?)*

## Example

In a random sample of 85 automobile engine crankshaft bearings, 10 have a surface finish that is rougher than the specifications allow. Find a 95% two-sided confidence interval for  $p$ , the proportion of bearings in the population that exceeds the roughness specification.

**Solution:**  $x = 10, n = 85, \hat{p} = 10/85, n\hat{p} = 10 \geq 5, n(1 - \hat{p}) = 75 \geq 5$ . Condition checked!

`Propinterval(level=0.95,n=85,X=10)`

Conclusion: based on the data, we are 95% confidence that  $p$  falls between 0.0492 and 0.1861.

Now using  $\tilde{p} = 0.12$  as an initial estimate of  $p$ , how large a sample is required if we want to be 95% confident that the error in using  $\hat{p}$  to estimate  $p$  is less than 0.05? Then redo this problem using the conservative approach (answer: 163 and 385).

`sample.size.Propinterval(level=0.95,ini.p=0.12,E=0.05)`

`sample.size.Propinterval(level=0.95,ini.p=0.5,E=0.05)`

## Prediction Interval

Suppose that  $X_1, \dots, X_n$  is a random sample from a normal population. The sample mean and sample variance are  $\bar{X}_n$  and  $S_n^2$ , respectively. We wish to predict the value  $X_{n+1}$ , a single future observation. A point prediction of  $X_{n+1}$  is  $\bar{X}_n$ , the prediction error is  $X_{n+1} - \bar{X}_n$  and the variance of the prediction error is

$$V(X_{n+1} - \bar{X}_n) = \sigma^2 + \frac{\sigma^2}{n}, \text{ and } X_{n+1} - \bar{X}_n \sim N\left(0, \sigma^2 \left(1 + \frac{1}{n}\right)\right)$$

Estimate  $\sigma$  by  $S_n$ , we have

$$T_{n-1} = \frac{X_{n+1} - \bar{X}_n}{S_n \sqrt{1 + \frac{1}{n}}} \sim t(n-1).$$

## Prediction Interval

Based on a random sample  $x_1, \dots, x_n$ , A  $100(1-\alpha)\%$  **prediction interval (PI)** on a single future observation from a normal distribution is given by

$$\bar{x}_n - t_{n-1, \alpha/2} s_n \sqrt{1 + \frac{1}{n}} \leq X_{n+1} \leq \bar{x}_n + t_{n-1, \alpha/2} s_n \sqrt{1 + \frac{1}{n}} \leq X_{n+1}.$$

One could also compute the upper- and lower- prediction bounded.  
Use StatEngin:

```
Predinterval(level=?,sample=?)
```

```
Predinterval(level=?,n=?,barx=?,s=?)
```

The prediction interval for  $X_{n+1}$  will always be longer than the confidence interval for  $\mu$  because more variability is associated with the prediction error than with the error of estimation.

## Example

Consider the tensile adhesion tests on specimens of U-700 alloy described in Practice 1:

$x=c(19.8, 10.1, 14.9, 7.5, 15.4, 15.4, 15.4, 18.5, 7.9, 12.7, 11.9, 11.4, 11.4, 14.1, 17.6, 16.7, 15.8, 19.5, 8.8, 13.6, 11.9, 11.4)$

A 95% confidence interval on  $\mu$  is  $Tinterval(level = 0.95, sample = x)$  which gives

$$12.1381 \leq \mu \leq 15.2892.$$

We plan to test a 23rd specimen. A 95% prediction interval on the load at failure for this specimen is

$Predinterval(level = 0.95, sample = x)$  which gives

$$6.1575 \leq X_{23} \leq 21.2698$$

Conclusion, we are 95% confident that the next observation will be between 6.1575 and 21.2698.

# StatEngine Summary of One-Sample CIs

CI for  $\mu$  when  $\sigma$  is **known** and the distribution is **normal**:

Zinterval(level=?,sigma=?,sample=?)

Zinterval(level=?,sigma=?,n=?,barx=?)

sample.size.Zinterval(level=?,sigma=?,E=?)

CI for  $\mu$  when  $\sigma$  is **unknown** and the distribution is **normal** (or for any distribution, but in this case, it provides approximated CIs):

Tinterval(level=?,sample=?)

Tinterval(level=?,n=?,barx=?,s=?)

Large-sample CI ( $n \geq 25$ ) for  $\mu$  under any distribution:

AZinterval(level=?,sample=?)

AZinterval(level=?,n=?,barx=?,s=?)

Both T-interval and AZ-interval are approximated CIs when normality does not hold. The T-intervals are more conservative (wider).

# StatEngine Summary of One-Sample CIs

CI for  $\sigma^2$  (or  $\sigma$ ) when the distribution is **normal** (does not work well if lack of normality)

`Chi2interval(level=?,sample=?)`

`Chi2interval(level=?,n=?,s=?)`

Large-sample CI ( $n\hat{p}, n(1 - \hat{p}) \geq 5$ ) for a population proportion  $p$ :

`Propinterval(level=?,n=?,X=?)`

`sample.size.Propinterval(level=?,ini.p=?,E=?)`

Prediction interval of a future observation from **normal** distribution.

`Predinterval(level=?,sample=?)`

`Predinterval(level=?,n=?,barx=?,s=?)`

Common patterns:

- ▶ Based on the same sample, an interval estimate becomes wider if its confidence level increases.
- ▶ When confidence level is fixed, a larger sample size leads to a narrower (more precise) interval estimate.