# Data Science 1
# Statistical Inference
# Estimation for One Population

**Ann Maharaj**

# Statistical Inference - Estimation for One Population

# Introduction

- Statistical inference is the process by which conclusions are drawn about a population from analysing a representative sample drawn from this population. This involves estimation and hypothesis testing.

- In this part of the topic, we focus on estimation of parameters for a single population.

- A point estimator is a single value that estimates an unknown population parameter such as the population mean, variance, proportion, and it is the corresponding sample statistic.

| | **Population Parameter** | **Sample Statistic** |
|---|---|---|
| Mean | $\mu$ | $\overline{X}$ |
| Variance | $\sigma^2$ | $s^2$ |
| Standard Deviation | $\sigma$ | $s$ |
| Proportion | $\pi$ | $p$ |

# Confidence interval estimation of the population mean

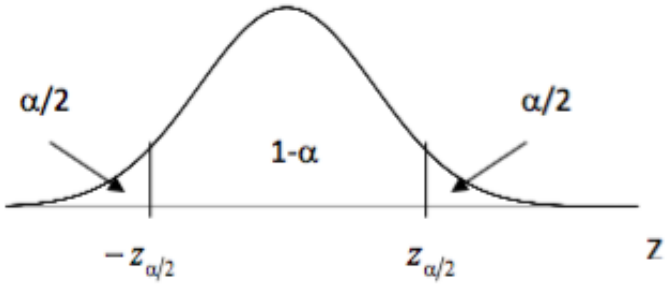
If the population standard deviation $\sigma$ is known, and $X$ is normally distributed or $n$ is large then the standardised sample mean follows a standard normal distribution.

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

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

$$\Rightarrow P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha/2}\right) = 1 - \alpha$$

$$\Rightarrow P\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$



# Confidence interval estimation of the population mean

- A $100(1-\alpha)\%$ confidence interval estimator of $\mu$ when $\sigma$ **is known** is

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- If the population size $N$ is known and $n/N > 0.05$ then including the finite population correct factor ($fpc$), the $100(1-\alpha)\%$ confidence interval estimator of $\mu$ becomes

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

- $z_{\alpha/2}$ is referred to as a multiplier or a critical value. It is a percentile of the standard normal distribution.

# Example 1

- To determine the mean waiting time for his customers to be served, a manager of a large fast food take away outlet at a particular location, took a random sample of 60 customers over a period of a week and found that the mean waiting time was 6.5 minutes.

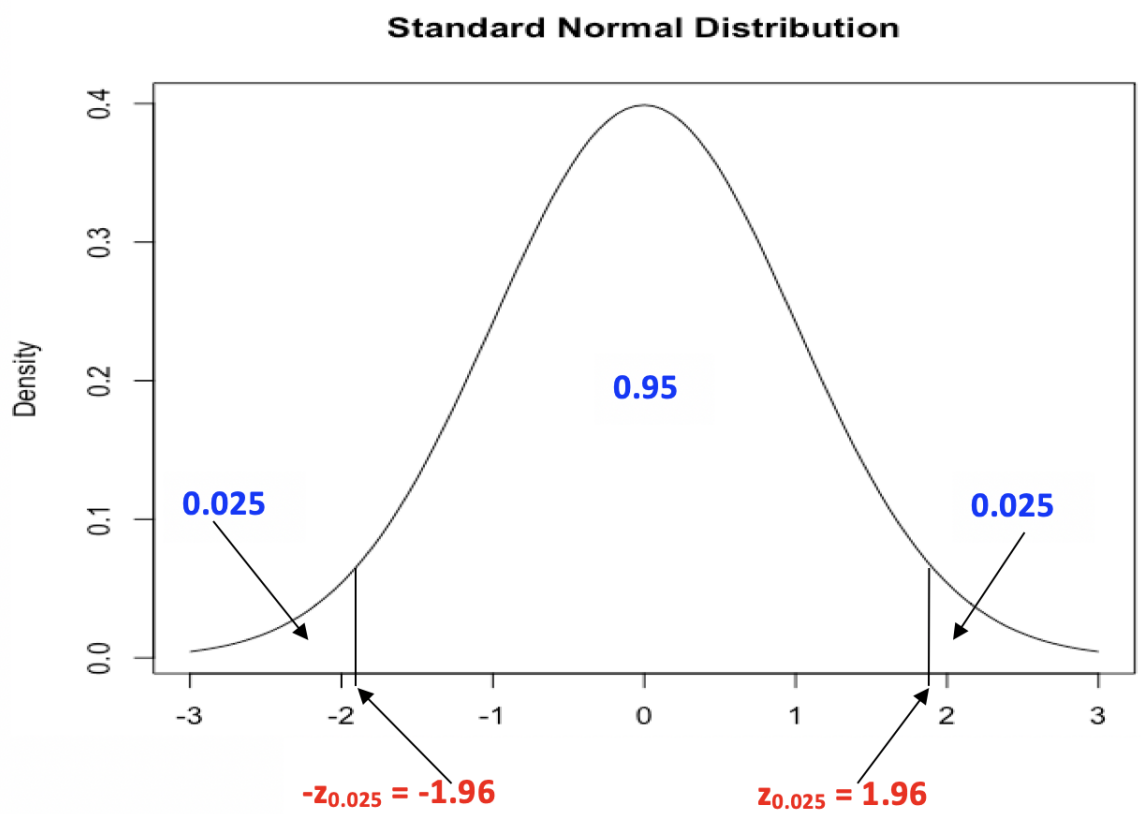
- Assuming that the population standard deviation is known from a previous study to be to 4 minutes, obtain and report the 95% confidence interval estimate of the mean waiting time for all customers of this fast food outlet at this location.

$n = 60, \overline{X} = 6.5, \sigma = 4$, (1 - $\alpha$) $= 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$.

# Example 1

$z_{\alpha/2} = z_{0.025}$ is the 97.5th percentile of the standard normal distribution.

$-z_{\alpha/2} = -z_{0.025}$ is the 2.5th percentile of the standard normal distribution.



# Example 1 - R code

```
> #sample size
> n <- 60
> #sample mean
> xbar <- 6.5
> #population standard deviation
> sigma <- 4
> #97.5th percentile of standard normal distribution
> z025 <- qnorm(.975)
> z025
[1] 1.959964
> #standard error
> se <- sigma/sqrt(n)
> se
[1] 0.5163978
> #Lower confidence limit
> LCL <- xbar - (z025 * se)
> LCL
[1] 5.487879
> #Upper confidence limit
> UCL <- xbar + (z025 * se)
> UCL
[1] 7.512121
> #95% confidence interval estimate of mean
> cbind(LCL, UCL)
          LCL      UCL
[1,] 5.487879 7.512121
```

# Example 1

- The 95% confidence interval is [5.49, 7.51]
- We estimate the mean waiting time for all customers at this particular fast food outlet to be between 5.49 and 7.51 minutes with 95% confidence.
- If we were to take another sample the confidence limits will not be the same.
- However, if we repeatedly sample, and for each sample, we obtain a 95% confidence interval estimate of the mean, then we expect that 95% of these confidence intervals will contain the true mean.

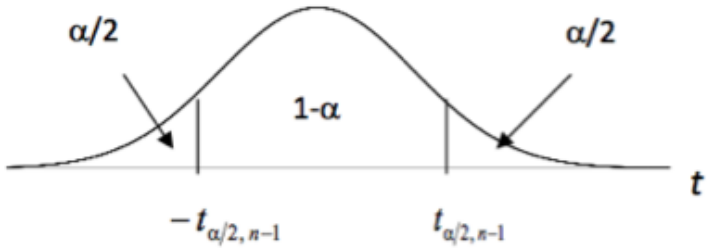
# Confidence interval estimation of the population mean

If the population standard deviation is unknown, and $X$ is normally distributed, the standardised sample mean follows a $t$-distribution with $n$ -1 degrees of freedom.

$$t = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

$$P\left(-t_{\alpha/2,n-1} < t < t_{\alpha/2,n-1}\right) = 1 - \alpha$$

$$\Rightarrow P\left(-t_{\alpha/2,n-1} < \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} < t_{\alpha/2,n-1}\right) = 1 - \alpha$$




$$\Rightarrow P\left(\overline{X} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right) = 1-\alpha$$

# Confidence interval estimation of the population mean

- A $100(1-\alpha)\%$ confidence interval estimator of $\mu$ when $\sigma$ **is unknown** is

$$\overline{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

- If the population size $N$ is known and $n/N > 0.05$ then including the *fpc*, the $100(1-\alpha)\%$ confidence interval estimator of $\mu$ becomes

$$\overline{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

- It is assumed that the population from which sample has been drawn, has an approximate normal distribution.

- $t_{\alpha/2,n-1}$ is a percentile of the $t$-distribution.

# Example 2

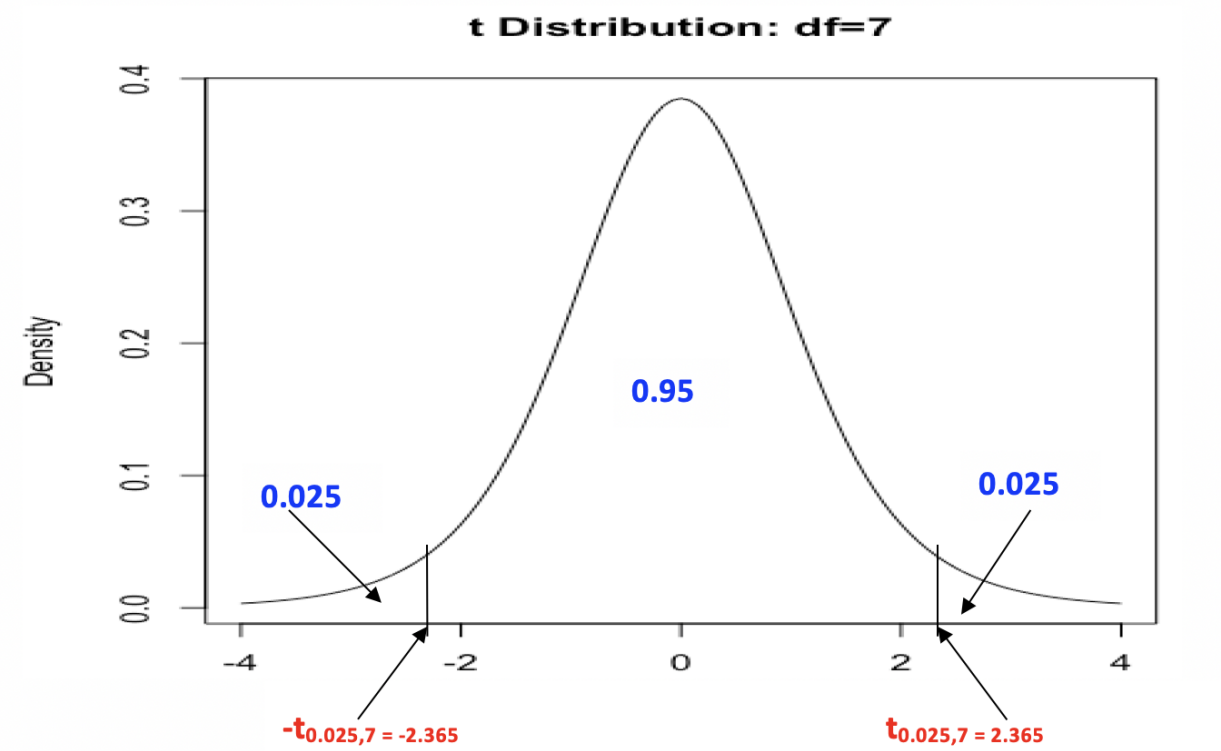
- The owner of a large fleet of taxis is trying to estimate his costs for next years operations.
- One major cost is for fuel purposes. Because of the high cost of petrol, and the need to reduce emissions, the owner has changed his entire fleet to hybrid vehicles.
- In order to estimate the fuel consumption of his hybrid fleet of taxis, he took a random sample of 8 taxis and measured the kilometres per litre achieved by each.
- The results are as follows: 3.5, 4.2, 5.3, 4.7, 3.4, 4.6, 4.9, 3.7. Obtain and report a 95% confidence interval estimate of the mean fuel consumption of all taxis in the fleet.
- Assume that the distribution of fuel consumption is approximately normal.

$n = 8$, $\overline{X}$ and $s$ must be computed, (1 - $\alpha$) $= 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$.

# Example 2

$t_{\alpha/2,n-1} = t_{0.025,7}$ is the 97.5th percentile of t-distribution with 7 degrees of freedom.

$-t_{\alpha/2,n-1} = -t_{0.025,7}$ is the 2.5th percentile of t-distribution with 7 degrees of freedom.



# Example 2 - R code

```
> #sample size
> n <- 8
> #create vector of numbers
> x <- c(3.5, 4.2, 5.3, 4.7, 3.4, 4.6, 4.9, 3.7)
> #sample mean
> xbar <- mean(x)
> #sample standard deviation
> s <- sd(x)
> #97.5th percentile of the t-distribution with 7 degrees of freedom
> t025_7 <- qt(.975,7)
> t025_7
[1] 2.364624
> #standard error
> se <-  (s/sqrt(n))
> se
[1] 0.2474423
> #Lower confidence limit
> LCL <- xbar - (t025_7 * se)
> LCL
[1] 3.702392
> #Upper confidence limit
> UCL <- xbar + (t025_7 * se)
> UCL
[1] 4.872608
> #95% confidence interval estimate of mean
> cbind(LCL, UCL)
          LCL      UCL
[1,] 3.702392 4.872608
```

# Example 2

- The 95% confidence interval is [3.70, 4.87]
- We estimate the mean fuel consumption for all taxis in the fleet to be between 3.70 and 4.87 kilometres per litre with 95% confidence.

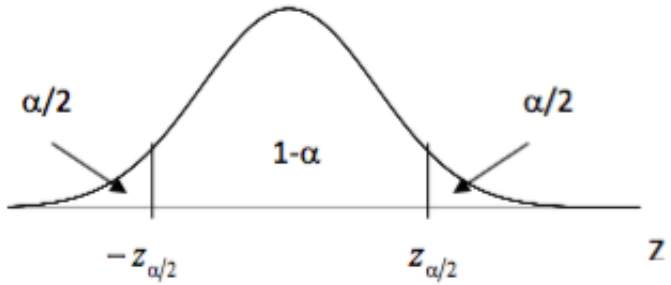
# Confidence interval estimation of the population proportion

Since the standardised sample proportion follows a standard normal distribution, i.e.,

$$Z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0,1)$$

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

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




$$\Rightarrow P\left(p - z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} < \pi < p + z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}\right) = 1-\alpha$$

# Confidence interval estimation of the population proportion

- A $100(1-\alpha)\%$ confidence interval estimator of $\pi$ is

$$p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

- If the population size $N$ is known and $n/N > 0.05$ then including the *fpc*, the $100(1-\alpha)\%$ confidence interval estimator of $\pi$ becomes

$$p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$

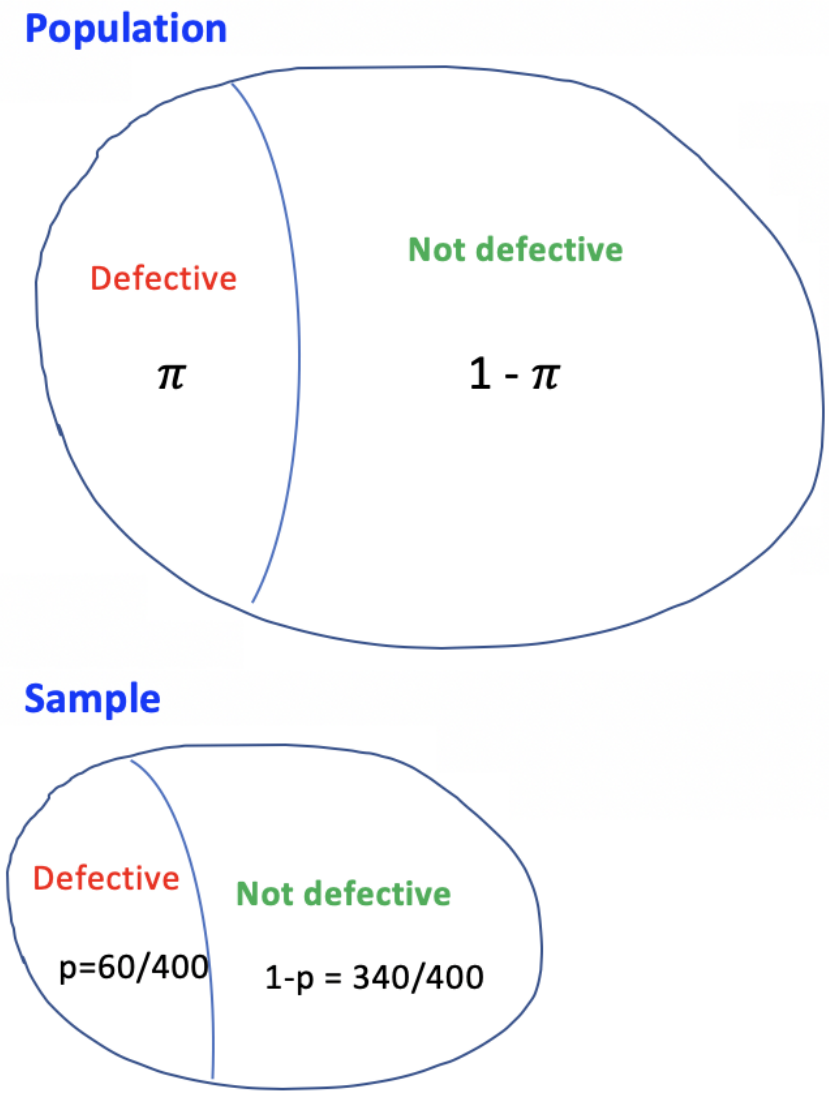
# Example 3

- A factory produces a component used in manufacturing computers.
- Components are tested prior to shipment to determine whether or not they are defective.
- In a random sample of 400 units, 60 were found to be defective.
- Obtain and report a 99% confidence interval estimate of the true proportion of defective components produced by the factory.

n $= 400$, x $= 60$, p $= 60/400$,

(1 - $\alpha$) $= 0.99 \Rightarrow \alpha = 0.01 \Rightarrow \alpha/2 = 0.005$.

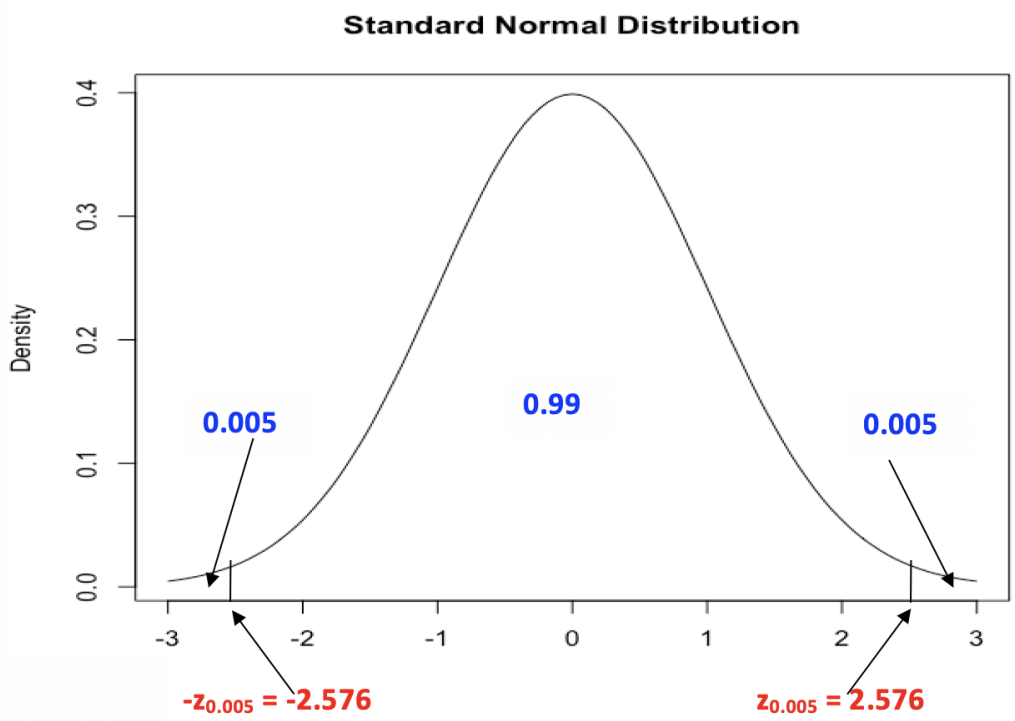
# Example 3


Population
Defective
$\pi$
Not defective
1 - $\pi$
Sample
Defective
p=60/400
Not defective
1-p = 340/400

# Example 3

$z_{\alpha/2} = z_{0.005}$ is the 99.5th percentile of the standard normal distribution.

$-z_{\alpha/2} = -z_{0.005}$ is the 0.5th percentile of the standard normal distribution.



# Example 3 - R code

```
> #sample size
> n <- 400
> # number of defective component
> x <- 60
> #sample proportion
> p <- x/n
> p
[1] 0.15
> #99.5th percentile of the standard normal distribution
> z005 <- qnorm(.995)
> z005
[1] 2.575829
> #standard error
> se <- sqrt(p*(1-p)/n)
> se
[1] 0.01785357
> #Lower confidence limit
> LCL <- p - (z005 *se)
> LCL
[1] 0.1040122
> #Upper confidence limit
> UCL <- p + (z005 *se)
> UCL
[1] 0.1959878
> #99% confidence interval estimate of proportion
> cbind(LCL, UCL)
           LCL       UCL
[1,] 0.1040122 0.1959878
```

# Example 3

- The 99% confidence interval is [0.1040, 0.1959]
- We estimate with 99% confidence that the percentage of defective components is between 10% and 20%.

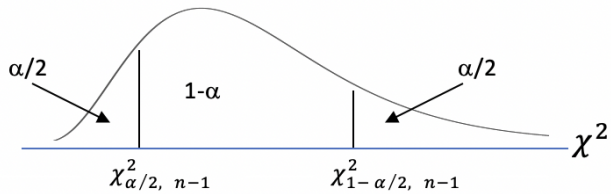
# Confidence interval estimation of the population variance

Given that a sample of size $n$ is drawn from a normal distribution with mean $\mu$ and variance $\sigma^2$, the expression $\chi^2$ follows a chi- square distribution with n-1 degrees of freedom, i.e.,

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

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

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



# Confidence interval estimation of the population variance

- A $100(1-\alpha)\%$ confidence interval estimator of $\sigma^2$ is

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\ ,\ \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}\right]$$

- A $100(1-\alpha)\%$ confidence interval estimator of $\sigma$ is

$$\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}}\ ,\ \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}}\right]$$

- $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$ are percentiles of the chi-square distribution.

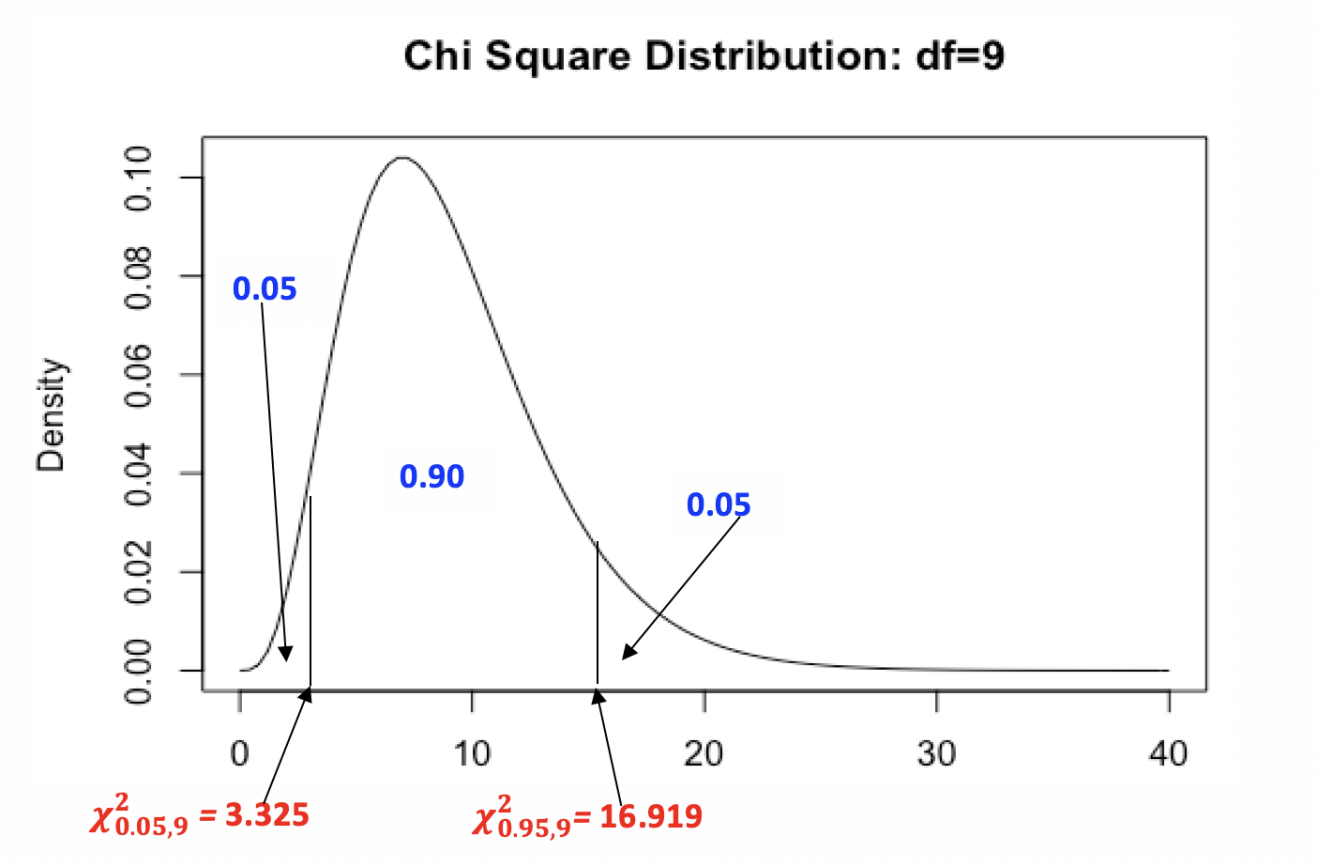
# Example 4

- A company manufactures steel shafts for use in engines.
- One method of judging inconsistencies in the production process is to determine the variance of the lengths of the shafts.
- On a particular day, a random sample of 10 shafts produced the following measurements of the lengths in centimeters: 20.5, 19.8, 21.1, 20.2, 18.9, 19.6, 20.7, 20.1, 19.8, 19.0.
- Assuming that the lengths of the shafts are normally distributed, obtain and report a 90% confidence interval estimate of the variance of the lengths of all shafts produced on that day.

$n$ =10, $s$ must be computed, (1 - $\alpha$) = 0.90 $\Rightarrow$ $\alpha = 0.10 \Rightarrow \alpha/2 = 0.05$.

# Example 4

$\chi^2_{\alpha/2,n-1} = \chi^2_{0.05,9}$ is the 5th percentile of the chi-square distribution.
$\chi^2_{1-\alpha/2,n-1} = \chi^2_{0.95,9}$ is the 95th percentile of the chi-square distribution.



# Example 4 - R code

```
> #sample size
> n <- 10
> #create vector of numbers
> x <- c(20.5, 19.8, 21.1, 20.2, 18.9, 19.6, 20.7, 20.1, 19.8, 19.0)
> #sample standard deviation
> s <- sd(x)
> s
[1] 0.702456
> #95th percentile of the chi-square distribution with 9 degrees of freedom
> chi95_9 <- qchisq(.95,9)
> chi95_9
[1] 16.91898
> #5th percentile of the chi-square distribution with 9 degrees of freedom
> chi5_9 <- qchisq(.05,9)
> chi5_9
[1] 3.325113
> #Lower confidence limit for variance
> LCL <- ((n-1)*s^2)/chi95_9
> LCL
[1] 0.2624863
> #Upper confidence limit for variance
> UCL <- ((n-1)*s^2)/chi5_9
> UCL
[1] 1.335594
> #90% confidence limit estimate of variance
> cbind(LCL,UCL)
           LCL      UCL
[1,] 0.2624863 1.335594
> #90% confidence limit estimate of standard deviation
> cbind(sqrt(LCL), sqrt(UCL))
          [,1]     [,2]
[1,] 0.5123342 1.155679
```

# Example 4

- The 90% confidence interval of the variance is [0.26, 1.34].
- We estimate the variance of the lengths of all the shafts produced on the day is between 0.26 and 1.34 $cm^2$ with 90% confidence.
- The 90% confidence interval of the standard deviation is [0.51, 1.16].
- We estimate the of the standard deviation of the lengths of all the shafts produced on the day is between 0.51 and 1.16 cm with 90% confidence.

# Controlling the confidence interval length

A more precise confidence interval estimate of a parameter (e.g., mean, proportion, variance, etc.) can be obtained by:

- **Decreasing the level of confidence $1 - \alpha$.**
  - This results in the multiplier (e.g. $z_{\alpha/2}$ or $t_{\alpha/2,n-1}$ for estimating the population mean) decreasing and hence resulting in a narrower interval $\Rightarrow$ increased accuracy.
- **Increasing the sample size $n$**
  - For example, in estimating the population mean, assuming that $\overline{X}$ remains relatively unchanged and if $\sigma$ has to be estimated, $s$ remains relatively unchanged) , resulting in a narrower interval $\Rightarrow$ increased accuracy
- We will demonstrate these concepts with an example in the following slides.

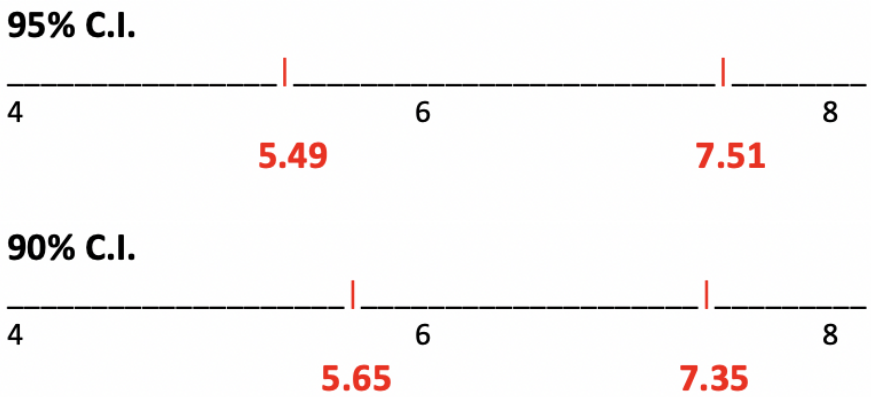
# Controlling the confidence interval length

- Consider Example 1 where 95% confidence interval estimate of the mean waiting time of all the fast food outlet customers was required. Suppose we also require a 90% C.I. but keep all the other values fixed

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

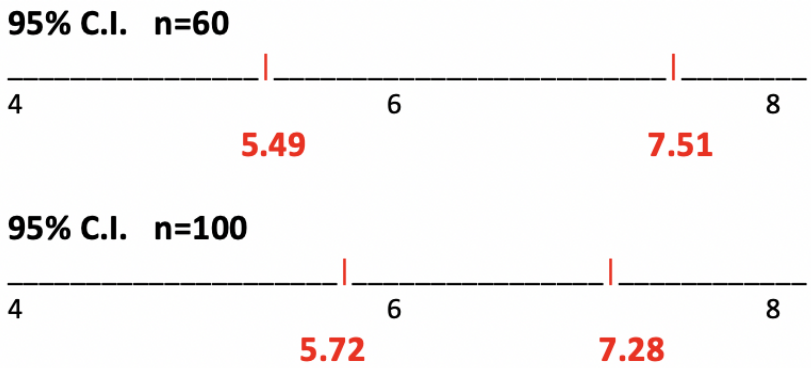
- $n = 60, \overline{X} = 6.5, \sigma = 4$
  - For a 95% C.I. we use $z_{\alpha/2} = 1.96$ the C.I. is [5.49, 7.51]
  - For a 90% C.I. we use $z_{\alpha/2} = 1.645$ the C.I. is [5.65, 7.35]

- The 90% C.I. is narrower than the 95% C.I.



# Controlling the confidence interval length

- Consider Example 1 again but this time change the sample size to 100, but keep all the other values fixed.
  - For a 95% C.I. with $n = 60$ the C.I. is [5.49, 7.51]
  - For a 95% C.I. with $n = 100$ the C.I. is [5.72, 7.28]
  - The 95% C.I. with $n = 100$ is narrower than the 95% C.I. with $n = 60$



# Summary

- A $100(1-\alpha)\%$ confidence interval estimator of the **population mean** $\mu$ when $\sigma$ **is known**

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

- A $100(1-\alpha)\%$ confidence interval estimator of the **population mean** $\mu$ when $\sigma$ **is unknown**

$$\overline{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

$$\overline{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

# Summary

- A $100(1-\alpha)\%$ confidence interval estimator of the **population proportion** $\pi$

$$p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

$$p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$

# Summary

- A $100(1-\alpha)\%$ confidence interval estimator of the population variance $\sigma^2$ is

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \ , \ \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}\right]$$

- A $100(1-\alpha)\%$ confidence interval estimator of population standard deviation $\sigma$ is

$$\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}} \ , \ \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}}\right]$$