Parameter Estimation and Hypothesis Testing

Ryan G.

July 31, 2017

A Note on the use of tables of the Standard Normal Distribution

Its easy to forget that a standard normal distribution has a function of:

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \times e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

And hence the probability would be the area under the curve:

$$P(a < x < b) = \int_a^b [\frac{1}{\sqrt{2\pi\sigma^2}} \times e^{-\frac{(x-\mu)^2}{2\sigma^2}}]dx \tag{2}$$

The whole point of using **R** or tables of a Standard Normal distribution is to totally bypass that function and just use, in the case of a table, pre-allocated solutions for a standard distribution of $\mu = 0$, $\sigma = 1$.

Normal Distribution Commands in **R**:

`dnorm`............ Outputs the value (height) of the curve (frequency), given input of an x-value

`pnorm`............ Outputs the area under the curve (probability), given input of an x-value

`qnorm`............ Outputs the x-value (relevant measurement), given input of probability (area)

`rnorm`............ Generates a random normal distribution.

The same principles apply to the `dbinom`, `pbinom`, `qbinom` and `rbinom` commands as well as the `dt`, `pt`, `qt`, `rt`, with reference to Binomial and t distributions respectively.

Exercise 2.1

Exercise

2.1 Describe some statistical experiment. Define the population, and how you will take physical samples. What is the random variable? What kind of shape and characteristics do you think the probability distribution of this random variable will have?

Solution

Description; Measuring the average size of a type of fish in a river.

Population; Every fish in the river would represent the population of fish.

Physical Samples; Samples will be taken by measuring a collection of fish at various locations.

Random Variable; The random variable is the length of that type of fish.

Probability Distribution This experiment would be normally distributed,

- Because, the observations are independent of one another, the values are continuous rather than discrete and the samples are random.

Exercise 2.2

Exercise

2.2 A hazardous waste facility monitors the groundwater adjacent to the facility to ensure no chemicals are seeping into the groundwater. Several wells have been drilled, and once a month groundwater is pumped from each well and tested for certain chemicals. The facility has a permit specifying an Alternate Concentration Limit (ACL) for aldicarb of 30 ppb. This limit cannot be exceeded more than 5% of the time. If the natural distribution of aldicarb at a particular well can be modelled as a normal distribution with a mean of 20 and a standard deviation of 5, how often will this well exceed the ACL (i.e., what is the probability that the aldicarb level at that well will be greater than 30 ppb on any given sampling occasion)?

Solution

Emperically a 95% confidence interval will span approximately two standard deviations from the mean (technically it is $1.96{\times}\sigma$), hence 5% of the data under the distribution will be less than 10 ppb or greater than 30 ppb. Thus approximately 2.5% of the time the well will naturally have more than 30 ppb of the chemical.

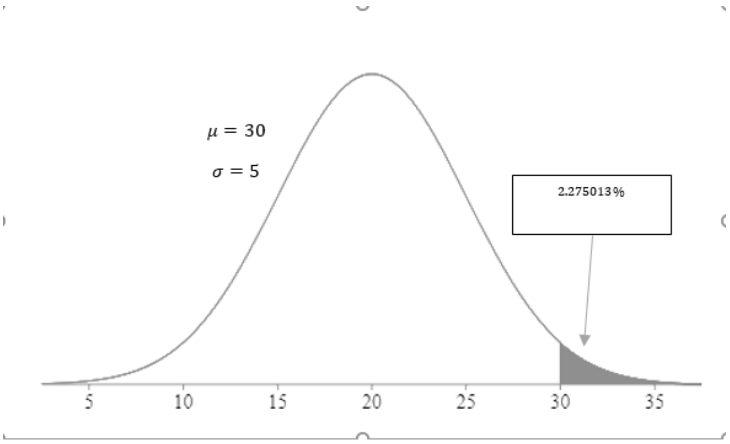

Figure 1: Normal Distribution of aldicarb in the well

Using R the question becomes:

> Determine the probability that a sample will have a mean value $\bar{x} > 30$ if the population has a mean value of $\mu = 20$ and a standard deviation of $\sigma = 5$.

In **R** the `pnorm` command can provide the area under the curve (i.e. the probablity) fig:[1]:

```
> z <- (30-20)/5
>    pnorm(z, mean=0, sd=1, lower.tail=FALSE, log.p=FALSE)
[1] 0.02275013
```

Thus, given there is a 2.275% probability of the sample mean being greater than 30 ppb of aldicarb, 2.275% of the time the chemical levels will be above 30 ppb in the well.

Exercise 2.3

Exercise

2.3 Suppose that the concentration of 1,2,3,4-tetrachlorobenze (TcCB) in soil at a background site follows a normal distribution with mean 0.5 and standard deviation 0.8. Suppose further that a regulation states that if any concentrations of TcCB at a remediated site are larger than the 95th percentile at the background site, then the remediated site will be declared to be still contaminated. Suppose the probability distribution of TcCB in the remediated site is exactly the same as the distribution of TcCB in the background site.
a. Find the 95th percentile of TcCB concentration at the background site.
b. If you take n = 1 soil sample from the remediated site, what is the probability that the TcCB concentration will exceed the 95th percentile of the background site?
c. If you take n = 10 soil samples, what is the probability that at least two of the 10 TcCB concentrations will exceed the 95th percentile of the background site?

Solution

Part A; To find the value on the normal distribution given a percentile we could use a *Standard Normal Distribution* table, one of *StatTrek's* calculators[1] or equivalently use the `qnorm` command in **R**:

```
> qnorm(0.05,mean=0.5,sd=0.8, lower.tail=FALSE, log.p=FALSE)
[1] 1.815883
```

Thus at the 95^{th} percentile the concentration value of TcCB is 1.815883.

Part B; The question, can best be rephrased as:
What is the probability that the sample mean of $\bar{x}$ from the sample of 1 is greater than 1.815883?

Using the Detailed Formulas;

The distribution of sample means drawn from a population tend to be normally distributed, this is provided by the *Central Limit Theorem* [2]

If the relevant population distribution is *non-normal*, a sampling distribution of the sample mean $\bar{x}$ would be approximately normally distributed for large samples by the *Central Limit Theorem*[3] $(n \geq 30)$

However if the relevant population has a *normal* distribution, the sampling distribution of $\bar{x}$ will be exactly normal no matter what the sample size is.

As the concentration value of TcCB follows a normal distribution, the sample means are also normally distributed. Thus we can form a normal distribution of samples.

observe that σ_{SE} is used and not σ:

σ is used when finding an area under the probability distribution of x, that is the distribution of actual values centred about the population mean value μ.

σ_{SE} is used when finding the area under the curve for a sampling distribution of $\bar{x}$, that is the distribution of all sample mean values centred about the mean value of sample means $\mu_{\bar{x}}$:

The average mean value $\mu_{\bar{x}}$ of a normal sampling distribution, is actually the population mean $/mu$, so $\mu_{\bar{x}} = \mu$.

[1] http://stattrek.com/online-calculator/normal.aspx

[2] Mendenhall, W. and Beaver, R. (2013). *Introduction to probability and statistics.* Boston, MA: CL-Wadsworth, pp.254-258.

[3] Mendenhall, W. and Beaver, R. (2013). *Introduction to probability and statistics.* Boston, MA: CL-Wadsworth, pp.251.

Steps to find the Solution

Let:

x .... be the TcCB concentration at the background site
$\sigma = 0.8$ .... be the standard deviation of the concentration of TcCB at the background site
μ .... be the mean value of concentration of TcCB at the backtround site.

$\bar{x}$ .... be the sample mean of the TcCB concentration
$n = 1$ .... be the number of samples taken
σ_{SE} .... be the standard deviation of all sample means from a sample of size 1.

1. Find the corresponding Z-Value for the sample mean:

$$z = \frac{\bar{x} - \mu}{\sigma_{SE}} \tag{3}$$

Find the Value of σ_{SE}

$$\begin{aligned} \sigma_{SE} &= \frac{\sigma}{\sqrt{n}} && (4) \\ &= \frac{0.8}{\sqrt{1}} && (5) \\ &= 0.8 && (6) \end{aligned}$$

Substitute the value in (We are concerned with a value sample value greater than 1.815883, hence that is the relevant $\bar{x} value$):

$$\begin{aligned} z &= \frac{\bar{x} - \mu}{\sigma_{SE}} && (7) \\ &= \frac{1.815883 - 0.5}{0.8} && (8) \\ &= 1.64485375 && (9) \end{aligned}$$

2. Find the probability of an x value exceeding this value under a *Standard Normal Distribution*:

$$\begin{aligned} P(\bar{x} > 1.815883) &= P(z > 1.64485375) && (10) \\ &\approx P(z > 1.64) && (11) \\ &\text{Refer to a Standard Table} \\ &= 0.0505 && (12) \end{aligned}$$

3. Formulate a Conclusion:

Thus, there is a 5.05% probability that a sample size of 1 will yield a sample mean with a value greater than the 95^{th} percentile of background site ($\bar{x} > 1.815883$).

■

Solving the Problem in R

In **R** the procedure is more or less the same, the difference being the use of the `pnorm` command:

```
###Begin
###Assign Variables####
        sd  <- 0.8                   #Std. Deviation of Background Site TcCB
        mu  <- 0.5                   #Mean of Background Site TcCB

        n  <- 1                      #Sample Size for part B question
        SE <- s.d._2.3/sqrt(n_2.3b)  #Standard Error of the Sample Mean

        prcnt <- 95                  #Relevant Percentile
        alpha  <- 1-prcnt_23/100     #Relevant alpha value


###Find the value of the 95th Percentile

        TcCB_95th_backup <-  qnorm(alpha, mean=mu, sd, lower.tail=FALSE, log.p=FALSE)


###Find the probability

        answer <- pnorm(1.815883, mean=0.5, sd = 0.8, lower.tail = FALSE, log.p = FALSE)


#Probability of sample exceeding 95th percentile of backup site:
percent(answer)
[1] "5%"
###End
```

or, in short:

```
> pnorm(1.815883, mean=0.5, sd = 0.8, lower.tail = FALSE, log.p = FALSE)
[1] 0.04999999
```

Part C; This question, relates to a Binomial Distribution.

Binomial Distribution; A statistical experiment will follow a binomial distribution where:[4]

- the experiment consists of n repeated trials (In this case 10 Samples).
- Each trial can result in either success or failure (in this case above or below the 95% threshold)
- The probability of success, denoted by P, is the same on every trial (In this case 5.05% from Part B).
- Each trial or sample is independent.

Detailed Forumula

$$\begin{aligned} P(x=k) &= C^n_k p^k q^{n-k} \qquad (13) \\ &= \frac{n!}{k!(n-k)!} \times p^k q^{n-k} \qquad (14) \end{aligned}$$

Where:
n .... Is the number of trials
k .... Is the number of successes
p .... Is the probability of Success
q .... Is the probability of failure, which is equivilant to $1-p$

[4] http://stattrek.com/probability-distributions/binomial.aspx

1. State the equation:

$$P(x = k) = \frac{n!}{k!(n-k)!} \times p^k q^{n-k} \tag{15}$$

2. Substitute the values

$$P(x \geq 2) = 1 - P(x < 2) \tag{16}$$

$$= 1 - P(x = 2) - (x = 1) - (x = 0) \tag{17}$$

$$\tag{18}$$

$$= 1 - \frac{10!}{2!(10-2)!} \times 0.05^2 \times 0.95^{10-2} \tag{19}$$

$$- \frac{10!}{1!(10-1)!} \times 0.05^1 \times 0.95^{10-1} \tag{20}$$

$$- \frac{10!}{0!(10-0)!} \times 0.05^0 \times 0.95^{10-0} \tag{21}$$

$$\tag{22}$$

$$= 1 - 0.0746 - 0.315 - 0.5987 \tag{23}$$

$$= 0.012 \tag{24}$$

3. Conclusion
   Thus their is a 1.2% probability of finding 2 samples above the 95^{th} percentile conentration of TcCB from a batch of 10 samples taken from a remediated site, given that the background site has TcCB concentrations distributed normally with $\sigma = 0.8$ and $\mu = 0.5$.

Solving in R To solve in **R**, the dbinom command is required:

```
###Begin

#We need to use a binomial distribution to determine this probability

####Variables
        p <- answer_23b   #The probability success, i.e. that a Sample will be above
            the 95th percentile
        q <- 1-p_23c      #The probability of failure, as above.

        k <- 2            #The relevant number of successes, the relevant number of
            samples above the 95th percentile
        n <- 10           #The number of samples taken

###Find the probability
        answer <- pbinom(k_23c, n_23c, p_23c, lower.tail=FALSE, log.p=FALSE)

##The probability, of 2 samples of the 10, being above the 95th percentile is:
        percent(answer)
[1] "1.15%"

#End
```

Thus there is a 1.15% probability of finding 2 samples above the 95^{th} percentile from a batch of 10. Observe that this probability suffers from less rounding error and is hence more accurate.

Exercise 2.4

Normal Distribution

The Normal Distribution is a bell curve distribution of continuous data, a population may be normally distributed with a population mean of μ and population standard deviation of σ.
The *Central Limit Theorem* provides that the sums, means and sampling statistics taken from random samples of measurments will tend to have a normal distribution [5], thus many continuous values measured from large populations, such as height, weight, etc. will tend to be normally distributed.

The Student's t distribution

The Central Limit Theorem provides that the sampling distriubtion of a statistic (e.g. a mean value), will follow a normal distribution, hence,
as long as the population mean μ and the population standard deviation σ is known a *Standard Normal Distribution* can be used with a z-score to calculate the probability of finding a sample with that statistic (e.g. mean). [6]
However, if:

1. The population σ and μ is unknown, OR
2. A population is not normally distributed, AND

   The sample size is not sufficiently large (Usually $n \geq 30$)

Then the Normal distribution could not be used, hence
the t distribution is relied upon:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{25}$$

Where s is the sample standard deviation and $\bar{x}$ is the sample mean.
Relative to the sample size taken, the t distribution comes in many different forms, determined by the *degress of freedom.*

The t distribution can be used in any of the following situations :

1. The population distribution is normal
2. The sample size is sufficiently large (Usually $n \geq 30$)

[5] Mendenhall, W. and Beaver, R. (2013). *Introduction to probability and statistics.* Boston, MA: CL-Wadsworth, pp.251.

[6] http://stattrek.com/probability-distributions/t-distribution.aspx

Confidence Intervals

Normal Distribution

a $(1-\alpha)\%$ confidence interval of a normal population is provided by:

$$\mu \pm \sigma \times Z_{\alpha/2} \tag{26}$$

Where:

μ .... Is the Population Mean

σ .... Is the Population standard deviation

Z_α .... Is the x-value on a standard normal distribution such that the area under the curve preceeding it is equivalent to the α value.

α .... Is the probability of incorrectly rejecting the null hypothesis, i.e. the probability of incorrectly assuming that somthing happened when it didn't.

Student's t-Distribution

A $(1-\alpha)\%$ confidence interval of a normal population is provided by:

$$\bar{x} \pm (\frac{s}{\sqrt{n}}) \times t_{\alpha/2,d.f.} \tag{27}$$

Where: $\bar{x}$ .... Is the sample mean

s .... Is the sample Standard Deviation

n .... Is the sample size

$d.f.$ .... Is the degrees of freedom of the t-distribution: $d.f. = n-1$

$t_{\alpha,d.f.}$ .... Is the x-value on a t-distribution (which is normal) of given degrees of freedom, which is such that the area under the curve is equivalent to α

Assuming,

- The samples are random
- The observations are independent of each other
- The sample was sufficiently large ($n \geq 30$)

Exercise 2.5

Exercise

2.5 Use R to find the confidence interval for the average maximum temperature in January in Melbourne, and for the average minimum temperature.

Solution

Detailed Equation; To create a 95% confidence Interval for that sample data we need to use the equation:

$$\bar{x} \pm (\frac{s}{\sqrt{n}}) \times t_{\alpha/2,d.f.} \tag{28}$$

Using **R** we can solve the descriptive statistics:

```
x_mean <- mean(Temperature$‘Max Temperature‘)
s <- sd(Temperature$‘Max Temperature‘) #observe, sample std.dev
n <- length(Temperature$‘Max Temperature‘)
df <- n-1

statistics_1 <- c("Sample Mean"=x_mean, "Sample Standard Deviation"=s)
statistics_2 <- c("Sample Size"=n, "Degrees of Freedom"=df, "t-value, 95%, d.f.=364"=t)

> statistics_1
Sample Mean          Sample Standard Deviation
 20.512329                  6.581775

>   statistics_2
Sample Size     Degrees of Freedom       t-value, 95%, d.f.=364
365.000000             364.000000                -1.966503
```

By substituting the values:

$$20.512329 \pm (\frac{6.581775}{\sqrt{365}}) \times -1.966503 = T : \tag{29}$$

$$(19.83486 < T < 21.1898) \tag{30}$$

Entirely via R; this can be solved by allowing **R** to calculate the confidence interval for us:

```
##Max Temp Confidence interval
#-------------------------------#

###State the equation
        #upper <-  x_sample + (s/sqrt(n)) * t_25
###Variables
        x_mean <- mean(Temperature$‘Max Temperature‘)
        s <- sd(Temperature$‘Max Temperature‘) #observe, sample std.dev
        n <- length(Temperature$‘Max Temperature‘)
        df <- n-1
###Calculate the t-value
        t <- qt(0.05/2, df_25,  lower.tail=TRUE, log.p=FALSE)
#Calculate the Upper Limit
        upper <- x_mean + (s/sqrt(n)*t)
#Calculate the Lower Limit
        lower <- x_mean - (s/sqrt(n)*t)
#Print the limits
 upper
[1] 21.1898
> lower
[1] 19.83486
```

Thus the 95% confidence interval for the maximum temperature in Melbourne is $T\ ^{o}C$:
. $(19.83486 < T < 21.1898)$

Exercise 2.6

Exercise

2.6 Use R to find the confidence interval for the mean difference between the maximum temperature and the minimum temperature in January in Melbourne, using the matched-data approach.

Solution

Paired Means; The process of paired means involves first finding the difference between two data points and then applying statistical tests (e.g. a t-distribution) to that distribution of differences, the differences are compared rather than the sample confidence intervals for a reason best illustrated by way of example:

> Take a sample of people and measure their heights with and without shoes, assume we wish to statistically prove that with shoes they are taller. If the 95% confidence intervals of both samples are compared with one another, a statistical difference probably won't be found, because the increase in height is such a small proportion. If however the heights are compared and a confidence interval drawn up from the differences it will be readily observed that there is indeed an increase in height caused by the shoes.

Detailed Equation; this can be solved the same way as the last exercise, but for the use of a new data set, namely a data set containing the differences between maximum and minimum temperatures, this new set of data would be manipulated identically to the last set using the following descriptive statistics:

```
###Mean difference between Maximum and Minimum Temperatures, Jan, Melb.
#------------------------------#
defecit <- Temperature$`Max Temperature`-Temperature$`Min Temperature`
defecit

statistics_1 <- c("Sample Mean"=x_mean, "Sample Standard Deviation"=s)
statistics_2 <- c("Sample Size"=n, "Degrees of Freedom"=df, "t-value, 95%, d.f.=364"=t)

 statistics_1
Sample Mean        Sample Standard Deviation
8.995068                  4.395687

> statistics_2
Sample Size     Degrees of Freedom       t-value, 95%, d.f.=364
365.000000             364.000000              -1.966503
```

$$8.995068 \pm (\frac{4.395687}{\sqrt{365}}) \times -1.966503 = \Delta : \tag{31}$$

$$(8.542613 < T < 9.447523) \tag{32}$$

Entirely via R; This can be solved in R using identical code as last time:

```
###Mean difference between Maximum and Minimum Temperatures, Jan, Melb.
#---------------------------------------------------------------------------------------------------


###Create the defecit vector from the data frame
defecit <- Temperature$‘Max Temperature‘-Temperature$‘Min Temperature‘
defecit


###State the equation
#upper <-  x_sample + (s/sqrt(n)) * t_25
#lower <-  x_sample - (s/sqrt(n)) * t_25

###Variables
x_mean <- mean(defecit)
s <- sd(defecit) #This is the sample standard deviation
n <- length(defecit)
df <- n-1

###Calculate the t-value
t <- qt(0.05/2, df_25,  lower.tail=TRUE, log.p=FALSE) #This value will be negative

#Calculate the Upper Limit
upper <- x_mean - (s/sqrt(n)*t)

#Calculate the Lower Limit
lower <- x_mean + (s/sqrt(n)*t)

#Print the limits
> upper
[1] 9.447523

> lower
[1] 8.542614
```

Thus the 95% confidence interval, for this sample of data, is such that the mean difference between Maximum and Minimum temperatures within Melbourne for the month of January is Δ_t ^oC:

. $(8.5 < \Delta_t < 9.4)$

Exercise 2.7

Exercise

2.7 Use R to find the confidence interval for the proportion of contaminated sites, defined as a certain concentration being more than 5. (Data in ChemConc.)

Solution

Detailed Equation; A confidence interval for proportions can be used whenever: [7]

- The sampling method is simple random sampling

- The sample is sufficientl large (10 successes and 10 failures)

In this context, a success is a concentration value > 5 and a failure is a concentration value < 5; sorting the data frame by maximum values shows > 10 values greater than 5.

Standard Error; Where the population size is unknown and s Where the true population proportion P is unknown the standard deviation of the sampling distribution cannot be calculated, thus we will need to use the standard error.
If the population size is unknown and the population size is ufficiently large (20 times larger than the sample size), the standard error can be approximated by:

$$SE_p = \sqrt{p \times \frac{(1-p)}{n}} \tag{33}$$

Where:
P .... Is the Population Proportion
N .... Is the Population Size
p .... Is the sample proportion
n .... Is the Sample Size

Confidence Interval; can be calculated by using SE_P with the margin of Error $t_{\alpha/2,d.f.}$:

$$p \pm t_{\alpha/2,d.f.} \times SE_p \tag{34}$$

[7] http://stattrek.com/estimation/confidence-interval-proportion.aspx?Tutorial=AP

In **R** we can solve the descriptive statistics:

```
contam <- with(ChemConc, Concentration>5)
no._contam <- length(contam[contam==TRUE]) # I just found the test==true on Stack Exchange,
    apparently better than table(contam)["TRUE"]

n <- length(Concentration)
p <- no._contam/n
df <- n-1
SE <- sqrt(p*(1-p)/n)
alpha <- 0.05
t <- qt(alpha/2, df, lower.tail=FALSE, log.p=FALSE )

statistics_1 <- c("Sample Mean"=x_mean, "Sample Standard Deviation"=s)
statistics_2 <- c("Sample Size"=n, "Degrees of Freedom"=df, "t-value, 95%, d.f.=364"=t)

> statistics_1
Sample Proportion    Standard Error
  0.35000000           0.06157651

> statistics_2
Sample Size         Degrees of Freedom     t-value, 95%, d.f.=364
 60.000000              59.000000               2.000995
```

By Substituting the values:

$$\begin{aligned} SE_p &= \sqrt{0.35 \times \frac{(1-0.35)}{60}} && (35) \\ &= 0.061577 && (36) \end{aligned}$$

Entirely via R; this can be solved by allowing **R** to calculate the confidence interval for us:

```
###Mean difference between Maximum and Minimum Temperatures, Jan, Melb.
#-----------------------------------------------------------------------------------------------------


###Create the Concentration vector from the data frame
Concentration_vector <- c(ChemConc$Concentration) #So I can use the length command to count
    the observations


###State the equation
#upper <-  p + t * SE
#lower <-  p - t * SE

###Variables
###The How many contaminated site?
contam <- with(ChemConc, Concentration>5)
no._contam <- length(contam[contam==TRUE]) # I just found the test==true on Stack Exchange,
    apparently better than table(contam)["TRUE"]

n <- length(Concentration)
p <- no._contam/n
df <- n-1
SE <- sqrt(p*(1-p)/n)
alpha <- 0.05


###Calculate the t-value
t <- qt(alpha/2, df, lower.tail=FALSE, log.p=FALSE )

#Calculate the Upper Limit
upper <-  p + t * SE

#Calculate the Lower Limit
lower <-  p - t * SE

#Print the limits
> upper
[1] 0.4732143

> lower
[1] 0.2267857
```

Therefore the 95% confidence interval for the proprtion of contaminated sites, (defined as > 5) from the *ChemConc* data set ranges from 22.7% to 47.3%.

Exercise 2.8

Exercise

2.8 Re Example 2.5, use R to test whether the average aldicarb concentrations in wells 2 and 3 exceeds the MCL, respectively.

Descriptive Equation;

State the Hypotheses:

1. H_0: The mean concentration of aldicarb is less than or equal to 7 ppb
2. H_a: The mean concentration of aldicarb is above 7ppb

Significance level; use $\alpha = 0.05$, that is a 5% probability that the null hypothesis will be incorrectly rejected, i.e. it will be incorrectly concluded that the levels are below the MCL.

Test Method; The test method used will be a t-distribution, because the population descriptive statistics are unknown.
Although there are only 4 observations, because the population of Aldicarb levels are normally distributed, a t-distribution can still be used.

Solving in R; The p-value could be obtained by finding a standard t-value and using a table or using the `pt` command, then that p-value could be compared, however, **R** allows this to be done in one command:

```
### View the Data Frame
Aldicarb

##Assign the Variables
x_well1 <- Aldicarb$well1
x_well2 <- Aldicarb$well2
X_well3 <- Aldicarb$well3

###Well 1
>   t.test(x_well1, alternative ="greater", mu=7,  conf.level = 0.95) #99.6% confidence level
    of exceeding MCL

One Sample t-test

data:  x_well1
t = 6.5249, df = 3, p-value = 0.003657
alternative hypothesis: true mean is greater than 7
95 percent confidence interval:
17.29318      Inf
sample estimates:
mean of x
23.1

>
> ###Well 2
>   t.test(x_well2, alternative ="greater", mu = 7, conf.level = 0.95) #99.97% confidence
    level of exceeding MCL

One Sample t-test

data:  x_well2
t = 15.465, df = 3, p-value = 0.0002937
alternative hypothesis: true mean is greater than 7
95 percent confidence interval:
21.96417      Inf
sample estimates:
mean of x
24.65

>
> ###Well 3
>   t.test(X_well3, alternative ="greater", mu = 7, conf.level = 0.95) #99.97% confidence
    level of exceeding MCL

One Sample t-test

data:  X_well3
t = -2.3556, df = 3, p-value = 0.9501
alternative hypothesis: true mean is greater than 7
95 percent confidence interval:
2.052335      Inf
sample estimates:
mean of x
4.525



####Perform the test
###Given that the wells are measured identically and are operated in the same fashion, pooled
    variance would probably be more accurate,
###But the exemplar did not use it so we won't
```

Thus, by interpreting the p-values, it can be concluded, at a 99% confidence level, that the first two wells have an average aldicarb concentration over the Maximum Contaminant Levl (MCL).

Observe that at a 95% confidence interval, the third well could have an average concentration value as low as 2 (the upper level is infinite because this is a one sided test).
The third well has a very high *p*-value, such that there would be a 95% probability of incorrectly assuming that the well was above the MCL, if the test is reversed, it could be concluded at a 95% confidence level that the MCL had **not** been exceeded at the third well.

Exercise 2.9

Exercise

2.9 Re Example 2.6, test
a. whether there is a significant difference between the two means in wells 1 and 3.
b. whether the mean concentration of well 1 exceeds that of well 3 by 20 ppb.

Solution

As with Exercise 2.6, the data frame values could be subtracted and the mean differences examined, however, we can do all of this in **R** in one fell swoop:

Part A;

```
##Assign the Variables
x_w1 <- Aldicarb$well1
x_w2 <- Aldicarb$well2
x_w3 <- Aldicarb$well3



####Perform the test
###Well 1 vs Well 3
t.test(x_w1, x_w3, alternative ="two.sided", mu=7,  conf.level = 0.95)

data:  x_w1 and x_w3
t = 4.3161, df = 4.0533, p-value = 0.01213
alternative hypothesis: true difference in means is not equal to 7
95 percent confidence interval:
11.16747 25.98253
sample estimates:
mean of x mean of y
23.100     4.525
```

Thus, by observing the p-value, it can be concluded at a 98.787% significance level that the mean values of Wells 1 and 3 differ.

Part B; By running a t distribution test, on the mean value of differences, against a hypothesis mean of 20:

```
> t.test(x_w1, x_w3, alternative ="greater", mu=20,  conf.level = 0.95)

Welch Two Sample t-test

data:  x_w1 and x_w3
t = -0.53135, df = 4.0533, p-value = 0.6885
alternative hypothesis: true difference in means is greater than 20
95 percent confidence interval:
12.87925      Inf
sample estimates:
mean of x mean of y
23.100     4.525
```

Observe that a one-sided 95% confidence interval could go as low as a mean difference of 12.87925 and that the p-value is large, suggesting a low confidence-level would be required to determine that the difference is greater than 20 ppb.
Thus it cannot be concluded at a 95% confidence interval that the mean values differ by more than 20 ppb, the confidence level would have to be 31.15%, that is a 68.85% probability of incorrectly assuming the difference in mean values is not less than 20.

Testing for a difference of 15 ppb:

```
> t.test(x_w1, x_w3, alternative ="greater", mu=15,  conf.level = 0.95)

Welch Two Sample t-test

data:  x_w1 and x_w3
t = 1.333, df = 4.0533, p-value = 0.1263
alternative hypothesis: true difference in means is greater than 15
95 percent confidence interval:
12.87925      Inf
sample estimates:
mean of x mean of y
23.100     4.525
```

Would only provide a mean difference of 15ppb at a confidence level of 87.37%

By trial and error, the largest average difference between the well's, that can be established at a 95% confidence interval, is 12.8793:

```
> t.test(x_w1, x_w3, alternative ="greater", mu=12.8793,  conf.level = 0.95)

Welch Two Sample t-test

data:  x_w1 and x_w3
t = 2.1238, df = 4.0533, p-value = 0.05
alternative hypothesis: true difference in means is greater than 12.8793
95 percent confidence interval:
12.87925      Inf
sample estimates:
mean of x mean of y
23.100     4.525
```