# Minilab 8a Worksheet

## Statistical Testing: Confidence Intervals and Hypothesis Tests

Remember the idea that from a randomly drawn **sample** (the data) we wish to draw conclusions about a **population** (some larger group).  The sample mean $\bar{x}$ gives an **estimate** of the population mean $\mu$ (this is known as a *point estimate*)/

In this minilab, we will see how to use the sample mean $\bar{x}$  and the sample standard deviation $s$ to construct an *interval estimate* for $\mu$ (called a *confidence interval*).  A confidence interval gives a "*likely range*" for $\mu$ and so it also tells us the *unlikely ranges* for $\mu$. We can then ask about whether a given particular value is a likely value for $\mu$.

### 1.  What is a Confidence Interval?



Image from: https://pballew.blogspot.com/2011/03/100-confidence-interval.html

Suppose we have sample data available and wish to estimate the population mean $\mu$.

- We can give a *point estimate*, i.e., a single number.  In this case, the best point estimate of the population mean $\mu$ is the sample mean $\bar{x}$.

- We can give an *interval estimate*, i.e., a range of possible values between a lower value and an upper value.  In this case, a *confidence interval* for the population mean $\mu$ is:

$$\bar{x} - z \times \frac{s}{\sqrt{n}} < \mu < \bar{x} + z \times \frac{s}{\sqrt{n}}$$

In this formula, $\bar{x}$ is the sample mean, $s$ is the sample standard deviation, $n$ is the sample size, and $z$ is a special (magic) number that we will investigate further.

(1) Suppose (just as an example) we use a random sample of size $n$ from a continuous uniform distribution $U(a, b)$. Once we have a random sample (the variable x in the R code below), we can calculate the sample mean $\bar{x}$ (we say "x bar").

```
# Endpoints of Uniform distribution
a = 0      # feel free to change
b = 50     # feel free to change but need a<b
# Sample size
n = 40     # feel free to change but need n>=30
# Random sample
x = runif(n,a,b)
x
# Sample mean
xbar = mean(x)
xbar
```

So $\bar{x}$ is a point estimate of $\mu$ and in the case of the uniform distribution, $\mu = \frac{1}{2}(a + b)$. How close is your point estimate?

(2) Including the sample standard deviation $s$, and the sample size $n$, we can form the confidence interval using the formulas for the lower and upper ends given above. All when need to do is select a value for $z$.

```
a = 0
b = 50
n = 40
x = runif(n,a,b)
xbar = mean(x)
# Sample standard deviation
s = sd(x)
```

2

```
# Magic value
z = 2
# Confidence interval
lower = xbar - z*s/sqrt(n)
upper = xbar + z*s/sqrt(n)
print(c(lower,upper))
```

Each time this R code is run, we get a different random sample each time, and hence the sample mean and the sample standard deviation change from sample to sample, and we construct a different confidence interval. Each of these confidence intervals is an interval estimate of the true population mean, which in the case of a uniform distribution is $\mu = \frac{1}{2}(a + b)$.

(3) But, also notice that the width of the confidence interval clearly also depends on this magic value $z$. For a particular choice of $z$, we need to investigate how often the true population mean $\mu$ is within the interval.

- If $z$ is small then the confidence interval is quite narrow (precise) and the true population mean $\mu$ will hardly ever be in this interval.

- If $z$ is large then the confidence interval is quite wide (imprecise) and the true population mean $\mu$ will very often be in this interval.

Let us try taking lots of random samples, say 100000 of them, each one of size $n$. For each random sample, we will construct a 95% confidence interval, and check whether the true population mean is within the confidence interval.

```
z = 2     # magic value
a = 0
b = 50
n = 40
replications = 100000
outside = rep(FALSE,replications)
for (i in 1:replications)
{
  x = runif(n,a,b)
  xbar = mean(x)
```

```
  s = sd(x)

  lower = xbar - z*s/sqrt(n)

  upper = xbar + z*s/sqrt(n)

  print(c(lower,upper))

  truemean = (a+b)/2

  outside[i] = (truemean<lower) | (truemean>upper)

}

summary(outside)

failure = sum(outside)/length(outside)

failure
```

When the code is run, you should see 100000 different confidence intervals go past in the output, followed by a summary table with counts of TRUE (the true population mean is OUTSIDE the interval) and FALSE. Finally, the proportion of times the true population mean is outside the interval is reported in the variable `failure`.

*Exercise.* For different values of $z$ record the proportion of failures (`failure`). For example, when $z = 2$ the result is approximately 0.05337 (which will vary each time you run due to random variation) which means approximately 5.337% of confidence intervals did **not** contain true population mean $\mu$.

| z | failure | prob |
|---|---|---|
| 2.5 | | |
| 2.0 | 0.05337 | 0.0455 (4dp) |
| 1.5 | | |
| 1.0 | | |
| 0.5 | | |

The third column records the variable `prob` from the R code below, corresponding to each value of $z$. Calculate the missing values of `prob`.

```
z = 2.0

prob = (1-pnorm(z,0,1))+pnorm(-z,0,1)

prob
```

What do you notice when you compare `failure` with `prob`, for each different value of $z$?

Try changing the values $a$, $b$ and $n$ (but make sure you have $a < b$ and $n \geq 30$). Does your conclusion regarding `failure` and `prob` change?

(4) What have found that, for a given $z$, the probability (`failure`) of a confidence interval not containing true population mean comes from the *standard normal distribution*. In particular it is $P(X < -z) + P(X > z)$. This value is commonly known as $\alpha$ ("alpha").

- If $\alpha = 0.05$ (or 5%) the corresponding confidence interval is called a 95% confidence interval for the population mean.

- If $\alpha = 0.01$ (or 1%) the corresponding confidence interval is called a 99% confidence interval for the population mean.

We can therefore use the R function `qnorm()` to find the $z$ value corresponding to a given probability $\alpha$.

```
alpha = 0.05

qnorm(1-alpha/2,0,1)
```

This gives $z = 1.96$, i.e., a 95% confidence interval for the population mean is

$$\bar{x} - 1.96 \times \frac{s}{\sqrt{n}} < \mu < \bar{x} + 1.96 \times \frac{s}{\sqrt{n}}$$

We expect that out of all such 95% confidence intervals, approximately 5% of them will **not** contain the population mean $\mu$. **This only works for sample size $n \geq 30$.**

Exercise. What value of $z$ would be used for a 99% confidence interval? What value of $z$ would be used for a 90% confidence interval?

## 2. Confidence intervals for "small" samples

For small samples, i.e., where $n \leq 29$, we can form a confidence interval only when the

**data comes from an approximately normally distributed population**.

Then $\alpha$ and $z$ are related by "Student's t-distribution" with $n - 1$ "degrees of freedom". This distribution is similar in shape to the normal distribution, but the height of the peak depends on the degrees of freedom parameter. For further information see
https://en.wikipedia.org/wiki/Student%27s_t-distribution.

---

You might want to read more about William Gosset (1876–1937) (also known as "Student") who was head brewer for the Guinness brewery.

---

(1) The R code below plots the standard normal distribution (the highest peak) and some example t-distributions. Of course R know all about the t-distribution, so it has functions dt, pt, qt and rt available.

```
x = seq(-3.5,3.5,0.01)
ggplot(NULL) +
  geom_line(aes(x,y=dnorm(x,0,1))) +
  geom_line(aes(x,y=dt(x,30)), colour="red") +
  geom_line(aes(x,y=dt(x,8)), colour="green") +
  geom_line(aes(x,y=dt(x,3)), colour="blue")
```
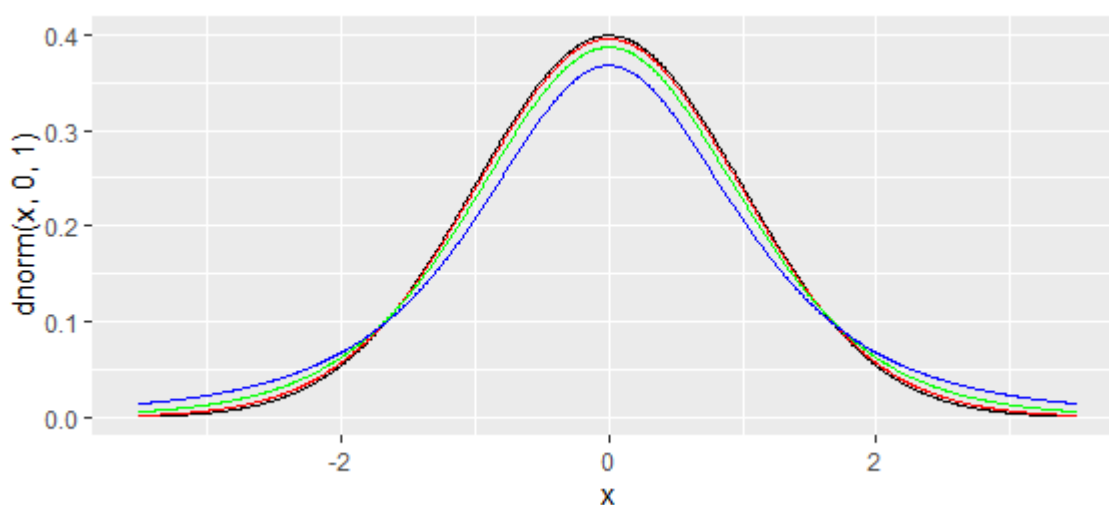
(2) Therefore, we change our confidence interval demo R code so that the random samples come from a normal distribution, and the "1.96" is replaced by a quantile from the correct t-distribution, i.e., the one with value $n - 1$ for the degrees of freedom parameter.

```
alpha = 0.05
mu = 25
sigma = 8
n = 5
replications = 100000
outside = rep(FALSE,replications)
for (i in 1:replications)
{
  x = rnorm(n,mu,sigma)
  xbar = mean(x)
  s = sd(x)
  lower = xbar - qt(1-alpha/2,n-1)*s/sqrt(n)
  upper = xbar + qt(1-alpha/2,n-1)*s/sqrt(n)
  print(c(lower,upper))
  truemean = mu
  outside[i] = (truemean<lower) | (truemean>upper)
}
summary(outside)
failure = sum(outside)/length(outside)
failure
```

Similarly to before, when this code is run, you should see 100000 different confidence intervals go past in the output, followed by a summary table where TRUE has frequency approximately 5000 (i.e. 5% of 100000) and FALSE has frequency approximately 95000 (i.e. 95% of 100000).

*Exercise.* Try different values for $\alpha$, mu, sigma and $n$, and check the corresponding value of `failure`. (Remember that $\alpha$ is a probability.)

## 3. Hypothesis Testing using `t.test()`

Given data in a sample, we often ask questions such as: "Is this given particular value a likely value for $\mu$?"  It is easy to answer this question using a confidence interval, i.e., see if the given particular value is in the confidence interval.

Alternatively, we can measure the difference between the sample mean and the given particular value (in appropriate units) and ask whether or not this observed difference is too large for $\mu$ to actually be equal to the given particular value.  This is called a **hypothesis test** for the population mean, where we are testing whether the population mean $\mu$ is equal to some specific value.

(1) We will look at data from across the different states of the USA.  The data set "state.x77" is built into R.  The variable Murder is the murder rate in each state per 100,000 people (in the year 1976).

```
library(tidyverse)
state = as_tibble(state.x77)
summary(state)
```

(2) Suppose we wish to test whether the mean murder rate in the population is 8.5.

We write this formally as the null hypothesis $H_0 : \mu = 8.5$.

Of course, R knows all about hypothesis testing, and has one main command that does all the work.

```
t.test(state$Murder, mu=8.5)
```

The output is as follows.

```
        One Sample t-test

data:  state$Murder
t = -2.1492, df = 49, p-value = 0.03659
alternative hypothesis: true mean is not equal to 8.5
95 percent confidence interval:
 6.328876 8.427124
sample estimates:
mean of x
    7.378
```

(3) Let us see where all of these numbers come from.

```
x = state$Murder
t = (mean(x)-8.5)/(sd(x)/sqrt(length(x)))
t
df = length(x)-1
df
p_value = 2*pt(t,df)
p_value
lower = mean(x)-qt(1-0.05/2,df)*sd(x)/sqrt(length(x))
upper = mean(x)+qt(1-0.05/2,df)*sd(x)/sqrt(length(x))
c(lower,upper)
```

So the R function t.test() calculates the test statistic $t$ (based on the sample mean $\bar{x}$, sample standard deviation $s$, sample size $n$, and the value of $\mu$ proposed in the null hypothesis), the degrees of freedom $n-1$, a confidence interval for $\mu$ and a "$p$-value". Clearly t.test() saves us from having to type out a lot of formulas to calculate all of these values.

The $p$-value is a conversion of the test statistic $t$ into a probability of observing a worse value, i.e., $P(T < -|t| \text{ or } T > |t|)$ if the null hypothesis is true.

We interpret the $p$-value by comparing to 0.05.

- If $p$-value $\leq 0.05$ then we **reject** the null hypothesis. It is very unlikely that the proposed value of the population mean is correct.

- If $p$-value $> 0.05$ then we **cannot reject** the null hypothesis. The proposed value of the population mean is actually a likely (plausible) value for $\mu$.

In this example, the $p$-value is 0.03659 so we we **reject** the null hypothesis that the true population murder rate is 8.5. Notice that 8.5 is not in the 95% confidence interval from 6.328876 to 8.427124.

To make a conclusion from a *hypothesis test* (e.g., a $t$-test) we really only need to check that the $p$-value is $\leq$ our chosen *level of significance* $\alpha$. We usually have $\alpha = 0.05$.

9

Data Science Minilabs (2021/22)

Exercise. Try the R code below. What confidence interval is reported? What is the null hypothesis of the hypothesis test carried out?

```
t.test(state$Income, conf.level=0.90)
```

## Summary

In this minilab, we have looked at confidence interval for the population mean $\mu$, for both large ($n \geq 30$) and small ($n \leq 29$) samples, and how to conduct a hypothesis test for the population mean $\mu$. It is possible to calculate everything we need ourselves, but actually the R function t.test() does all the calculations required and we can then make conclusions based either on the $p$-value or on the confidence interval.
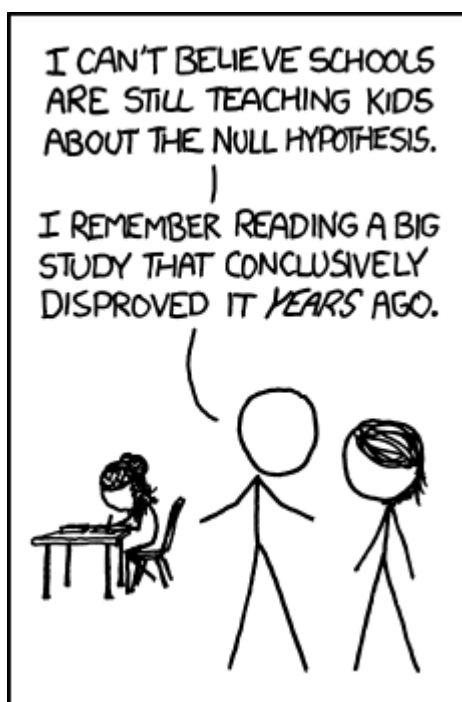


Image from: https://xkcd.com/892/

10