R Intermediate Short Course
Session 2 - Estimation and Hypothesis Testing

Facilitator: Rajesh Lakhan

The University of the West Indies, St Augustine

Tuesday 19th July 2022
(5:00pm - 7:00pm)
(online)

Estimation of Probability Distributions

A **probability distribution** is a function or rule that assigns probabilities to each value of a random variable.

Probability distributions may be discrete or continuous. The binomial and poisson distributions are discrete.

Six sided die example with sides numbered: 1, 2, 3, 3, 4, and 6

Binomial distribution

A discrete random variable X, with probability mass function f, has a **Binomial** distribution with parameters n and p if:

$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ for x = 0, 1, 2, ... , n

where n = number of trials, p = probability of success

and $\binom{n}{x} = {}^nC_x = \frac{n!}{x!(n-x)!}$

For X $\sim$ Binomial(n,p):

population mean, $\mu = E(X) = np$
population variance, $\sigma^2 = Var(X) = np(1-p)$

In R, the mean and variance of probability distributions are estimated as the sample mean, $\bar{x}$ and the sample variance, s^2.

To generate a sample of random values which all follow a Binomial(n,p):

```
x = rbinom(sample size, n, p)
```

Each time this code is run, a different set of random values may be generated. We can guarantee the same random values are generated using the *set.seed* command eg. `> set.seed(123)`

Example: Let X be a discrete random variable which follows a binomial distribution with parameters $n = 125$ and $p = 0.04$. Estimate the mean and variance for a sample of 10000 values of X.

```
> pop.mean = 125 * 0.04                    [1] 5
> pop.var = 125 * 0.04 * (1-0.04)          [1] 4.8

> set.seed(25)
> x = rbinom(10000,125,0.04)
> mean(x)                                  [1] 5.0117
> var(x)                                   [1] 4.806244
```

Application of the probability mass function, f(x)

Compute the following: a) $P(X = 5)$ and b) $P(X \leq 1)$.

1. Manually:

a) $P(X = 5) = \binom{125}{5} 0.04^5 (1 - 0.04)^{125-5} = 0.1791$

b) $P(X \leq 1) = P(X = 0) + P(X = 1)$

$= \binom{125}{0} 0.04^0 (1 - 0.04)^{125-0} + \binom{125}{1} 0.04^1 (1 - 0.04)^{125-1}$

$= 0.0061 + 0.0317 = 0.0378$

2. Simulated:

```
dbinom(x, n, p, log=FALSE)                          #P(X=x)
pbinom(x, n, p, lower.tail=TRUE, log.p=FALSE)
                                                    #P(X<=x)
```

a)
```
> dbinom(5,125,0.04,log=F)
[1] 0.1790807
```

b)
```
> pbinom(1,125,0.04,lower.tail=TRUE,log.p=F)
[1] 0.03774671
```

Poisson distribution

A discrete random variable X is said to have a **Poisson** distribution with parameter λ ($\lambda > 0$), if the probability mass function of X is:

$f(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$ for x = 0, 1, 2, ...

where λ = mean/average and x can take an infinite number of values.

For X $\sim$ Poisson(λ):

population mean, $\mu = E(X) = \lambda$
population variance, $\sigma^2 = Var(X) = \lambda$

To generate a sample of random values which all follow a Poisson(λ):

```
x = rpois(sample size, lambda)
```

To compute probabilities associated with this distribution:

```
dpois(x, lambda, log=FALSE)                          #P(X=x)
ppois(x, lambda, lower.tail=TRUE, log.p=FALSE)
                                                     #P(X<=x)
```

Example: Let Y be discrete random variable which follows a poisson distribution with parameter $\lambda = 5$.

a) Estimate the mean for a sample of 10000 values of Y.
b) Calculate $P(Y = 2)$ and $P(Y \leq 2)$

a)
```
> set.seed(12)
> y = rpois(10000,5)
> mean(y)
[1] 5.0316
```

b) Manually:

$$P(X = 2) = \frac{e^{-5}5^2}{2!} = 0.0842$$

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= \frac{e^{-5}5^0}{0!} + \frac{e^{-5}5^1}{1!} + \frac{e^{-5}5^2}{2!} = e^{-5}\left[\frac{5^0}{0!} + \frac{5^1}{1!} + \frac{5^2}{2!}\right] \\ &= 18.5e^{-5} = 0.1246 \end{aligned}$$

Simulated:

```
> dpois(2, 5, log=F)
[1] 0.08422434
> ppois(2, 5, lower.tail=T, log.p=F)
[1] 0.124652
```

The binomial distribution $X \sim Binomial(n, p)$ can be approximated to a poisson distribution $X \sim Poisson(\lambda = np)$ if n is large ($n \to \infty$) and p is small ($p \to 0$).

(See R script)

Covariance and Correlation

Covariance is a measure used to determine how the change in one variable affects the other.

$Cov(X, Y) = E(XY) - E(X)E(Y)$

If $Cov(X, Y) = 0$ then X and Y are independent.
[P(A $\cap$ B) = P(A)*P(B)]

If $Cov(X, Y) < 0$ there is a negative or inverse relationship between X and Y (as X increases Y decreases and vice versa).

If $Cov(X, Y) > 0$ there is a positive or direct relationship between X and Y (as X increase Y increases and vice versa).

Correlation is a standardised measure of covariance and is used to determine the strength of the relationship between two variables.

$$Cor(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X) \times Var(Y)}}$$

Correlation values r, lie between -1 and 1: $-1 \leq r \leq 1$

A negative correlation value close to 0 indicates a weak and inverse relationship between X and Y while a correlation value close to -1 indicates a strong and inverse relationship.

A positive correlation value close to 0 indicates a weak and direct relationship between X and Y while a correlation value close to 1 indicates a strong and direct relationship.

In addition to the correlation value, a scatter plot can be used to deduce the strength of the relationship between two variables.

Plots in R:

```
plot(x, y, type="p/l/o")
```

`type="p"` gives a scatter plot (points)
`type="l"` gives a line plot (line)
`type="o"` gives both points and line (over-fitted)

If a linear pattern can be seen in a scatter plot of variables X and Y, then this indicates there is a strong correlation.

The direction of this pattern specifies positive or negative correlation.

No apparent pattern indicates a weak and insignificant correlation.

Example 1: Simulate the covariance and correlation of the samples previously generated for variables X and Y. Also comment on the relationship between X and Y.

```
> cov(x,y)
[1] -0.01347107
> cor(x,y)
[1] -0.002709111
```

The covariance and correlation values indicate a weak and negative relationship between X and Y.

Example 2: Simulate the correlation between vectors x and y defined below. What does this tell you about the relationship between x and y? Construct a scatter plot to support your interpretation.

```
> x = c(106,125,42,51,64,76,72,84,40,171,180,210,
  101,41,70)
> y = c(10,44,0,2,8,14,21,18,24,17,26,52,16,11,37)
> cor(x,y)
[1] 0.6041182
```

There is a fairly strong and positive relationship between x and y.

```
> plot(x, y, type="p", main = "Scatterplot")
```

(See R script for output)

The scatter plot showed a linear relationship in an upward or positive direction. This gives the same interpretation as before.

Hypothesis Testing (one sample T tests)

The **T distribution** commonly referred to as student's t-distribution, can be represented by a symmetric or bell-shaped curve. It is a continuous probability distribution and as such, the area under the curve = 1.

It utilizes degrees of freedom (df) which is the number of values that have the freedom to vary.

The T distribution is used when the sample size $n \leq 30$.

A **hypothesis test** is a statistical test used to determine whether the hypotheses formulated hold true for the entire population.

The hypotheses involved in a hypothesis test are:

1. Null hypothesis, H_0 - a statement regarding the parameter of interest which always contains equality.
2. Alternate hypothesis, H_1 - a statement contradicting the null hypothesis which always contains inequality.

There are three main types of hypothesis test:

1. **Two tailed test** $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$
2. **Upper tailed test** (one tail) $H_0 : \mu \leq \mu_0$ vs $H_1 : \mu > \mu_0$
3. **Lower tailed test** (one tail) $H_0 : \mu \geq \mu_0$ vs $H_1 : \mu < \mu_0$

The type of test is determined by the sign of the alternate hypothesis, H_1.

Note that the null hypothesis may be written as $H_0 : \mu = \mu_0$ regardless of the type of test.

Formulating hypotheses

First identify the null or alternate hypothesis given in the question. Then state along with the opposing hypothesis. Examples:

1. Test the hypothesis that the mean is significantly different from 1.
   i.e Test $H_0 : \mu = 1$ versus $H_1 : \mu \neq 1$
2. Does the average blood pressure exceed 100 mmHg?
   i.e. Required to test $H_0 : \mu \leq 100$ versus $H_1 : \mu > 100$

Testing hypotheses

The decision of a hypothesis test revolves around the null hypothesis; either H_0 is rejected or failed to be rejected. If rejected then H_1 holds true and if failed to be rejected then H_0 is true.

We reject the null hypothesis H_0, if any one of the 3 following conditions holds:

1. The confidence interval does not contain the value at H_0
2. The $|$test statistic$| \geq$ critical value
   i.e. $+$test statistic $\geq$ critical value or -test statistic $\leq$ -critical value
3. P-value $<$ level of significance, α

1. A confidence interval (CI) is a range of values which contains the population value with a given degree of confidence or certainty.

On the other hand, the level of significance α, gives an estimate of the percentage of uncertainty.
It is computed by, $\alpha = 1$ - confidence level $= 1$ - CL

2. The test statistic derived from the data given is denoted by t.

The critical value for a two-tailed test is $t_{\alpha/2,df}$ and for a one-tailed test is $t_{\alpha,df}$ where df = n - 1.

3. A p-value is a probability calculated on the assumption that H_0 is true.

Performing One sample T tests in R:

```
t.test(data, mu=mu0)                          #two tailed
t.test(data, mu=mu0, alternative="greater") #upper
t.test(data, mu=mu0, alternative="less")    #lower
```

The default confidence level is 95%. This can be changed by specifying.

```
t.test(data,mu=mu0,alternative=" ",conf.level=0.99)
                                  #for 99% confidence
```

Critical values:

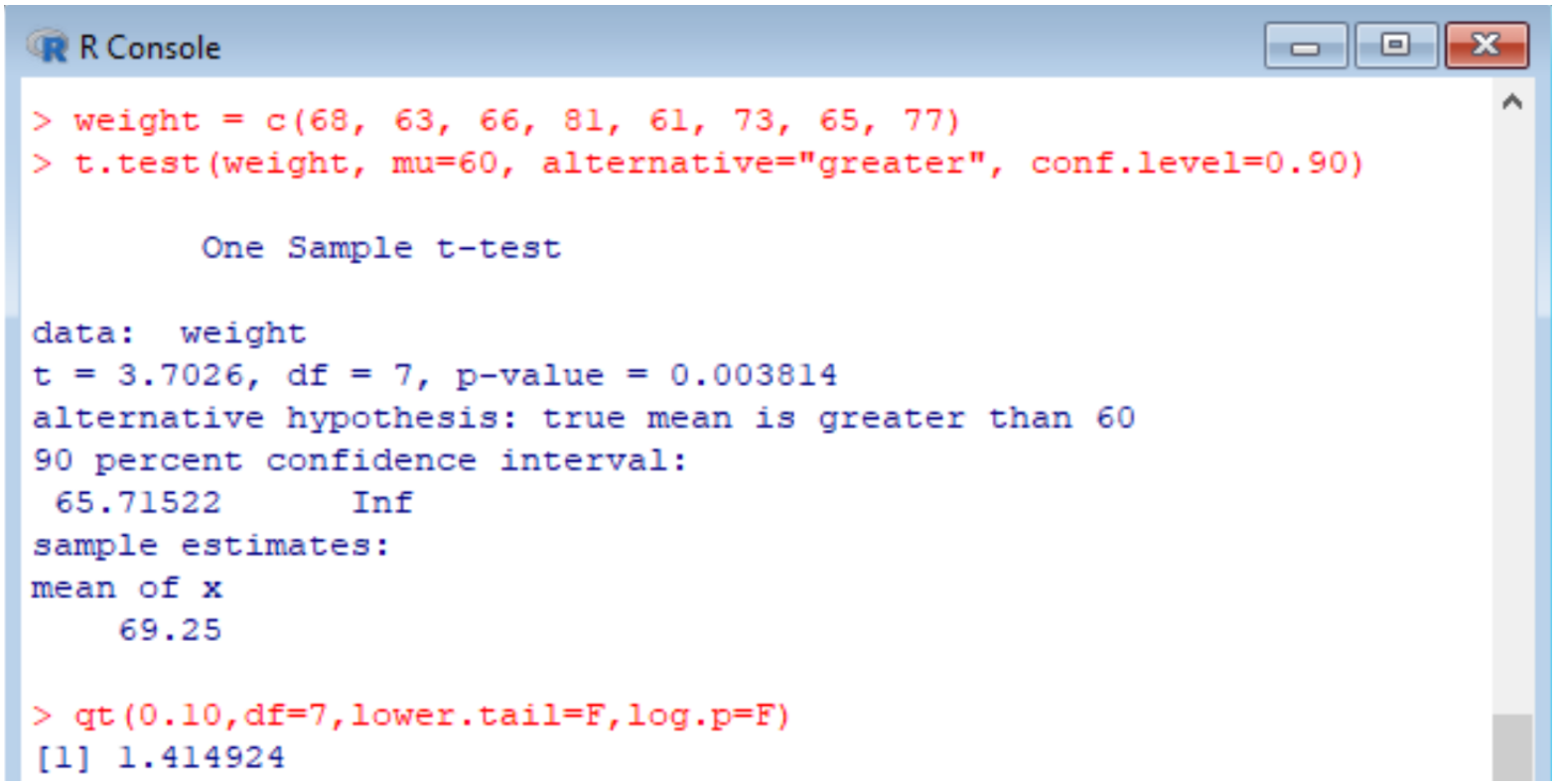
```
qt(alpha/2,df,lower.tail=F,log.p=F)      #t_{α/2,df}
qt(alpha,df,lower.tail=F,log.p=F)        #t_{α,df}
```

Example 1: The weight in kilograms of 8 25-year old men are as follows:

68 63 66 81 61 73 65 77

Test the hypothesis that the average weight of 25-year old men is significantly greater than 60 kg. Use $\alpha = 0.10$.

Required to test H_0: $\mu \leq 60$ vs H_1: $\mu > 60$ at CL = 1 - 0.10 = 0.90



Method 1: The 90% confidence interval is (65.715, ∞). Since this interval does not contain 60 we can reject H_0.

Method 2: The test statistic, t = 3.703 is greater than the critical value, $t_{0.10,7} = 1.415$ so we can reject H_0.

Method 3: The p-value = 0.004 $<$ 0.10 so we can reject H_0.

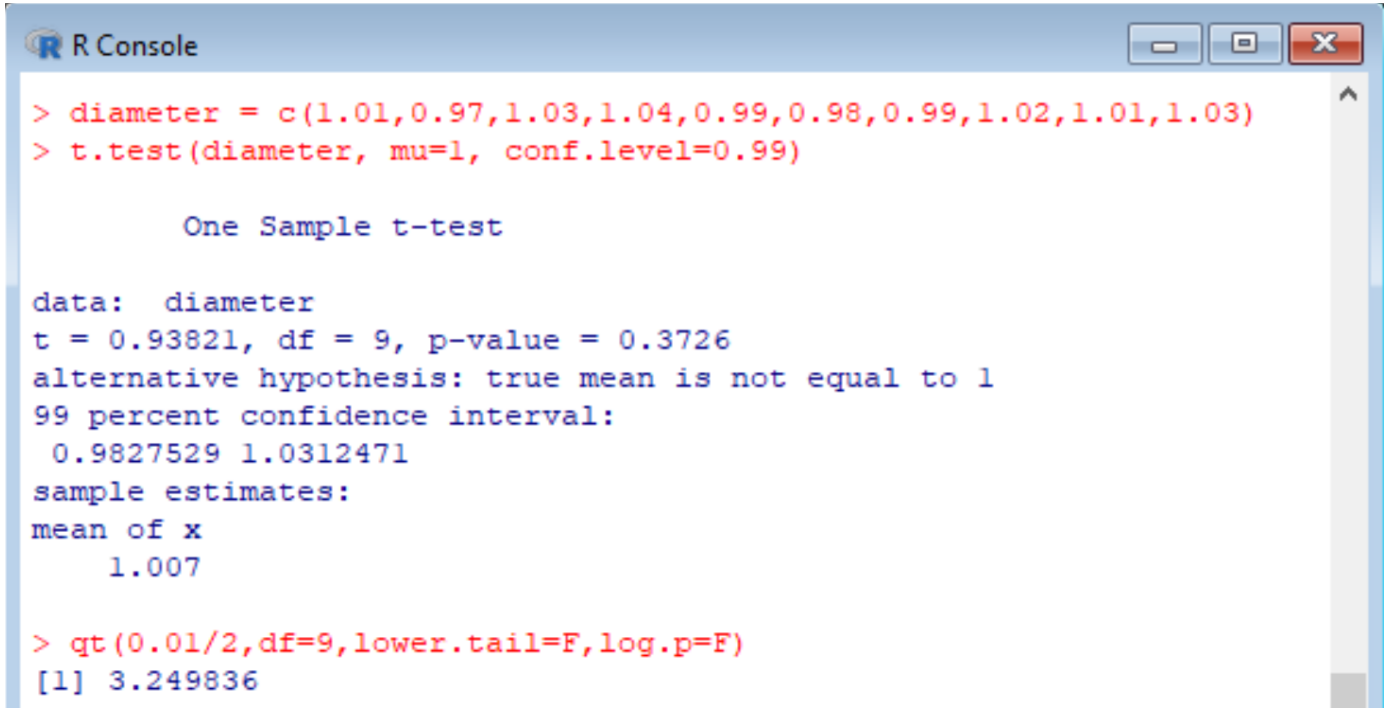
Conclusion: At $\alpha = 0.10$, the average weight of 25-year old men is significantly greater than 60 kg.

Example 2: A machine produces metal pieces that are cylindrical in shape. A sample of the pieces were taken and their diameters in cm are found to be:

1.01 0.97 1.03 1.04 0.99 0.98 0.99 1.02 1.01 1.03

By conducting a hypothesis test at $\alpha = 0.01$, determine whether the mean diameter is equal to 1.00 cm.

Required to test $H_0 : \mu = 1.00$ vs $H_1 : \mu \neq 1.00$ at CL = 1 - 0.01 = 0.99



Method 1: The confidence interval (0.983, 1.031) contains 1

Method 2: The test statistic, t (= 0.938) $< t_{0.01/2,9}$ (= 3.250)

Method 3: The p-value (= 0.373) > 0.01

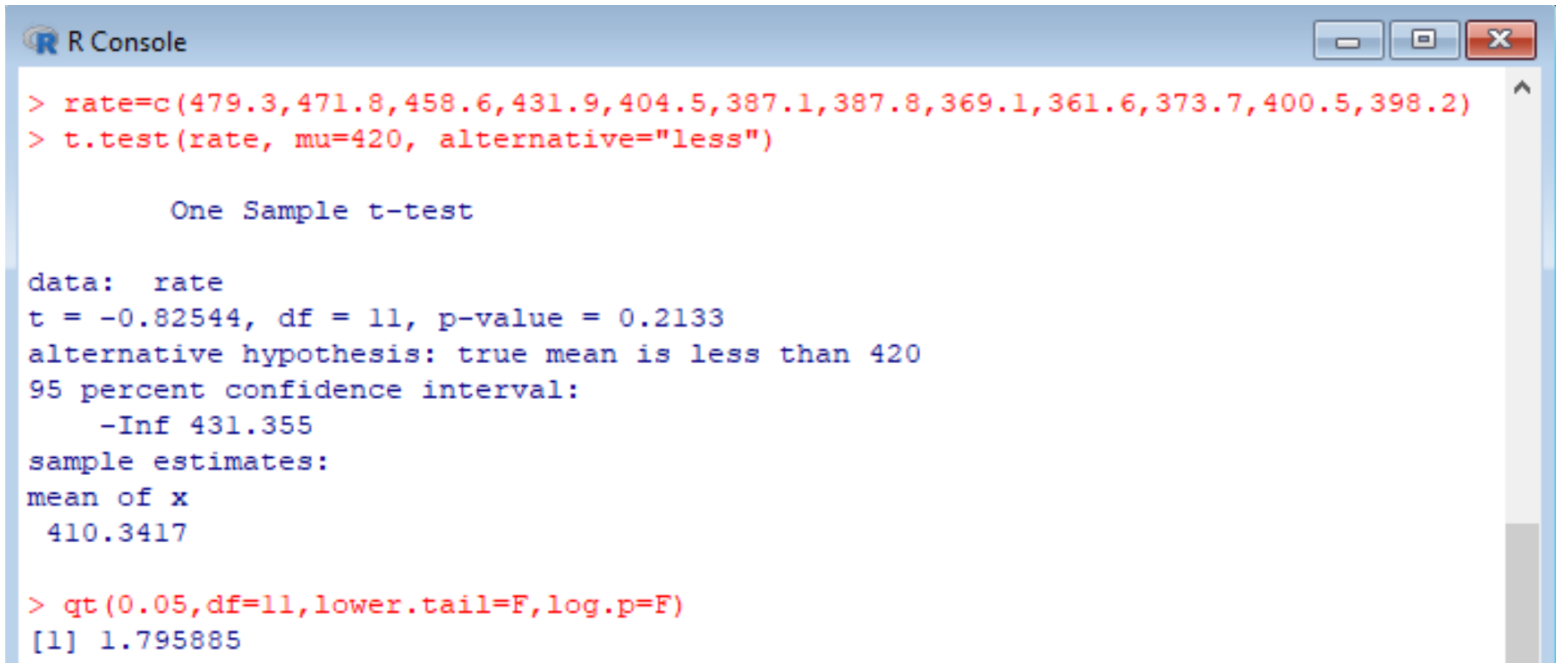
Conclusion: We fail to reject H_0 at $\alpha = 0.01$, therefore the mean diameter is significantly equal to 1.00 cm.

Example 3: The rate of violent crimes committed in the United States over 12 years are as follows:

479.3	471.8	458.6	431.9	404.5	387.1
387.8	369.1	361.6	373.7	400.5	398.2

Test at $\alpha = 0.05$ whether the mean rate of violent crimes is less than 420.

Required to test H_0: $\mu \geq 420$ vs H_1: $\mu < 420$



Method 1: The confidence interval $(-\infty, 431.355)$ contains 420
Method 2: The test statistic, $|t|$ $(= 0.825) < t_{0.05,11}$ $(= 1.796)$
Method 3: The p-value $(= 0.213) > 0.05$

Conclusion: We do not reject H_0 at $\alpha = 0.05$. There is insufficent evidence to conclude that the rate of violent crimes is less than 420.