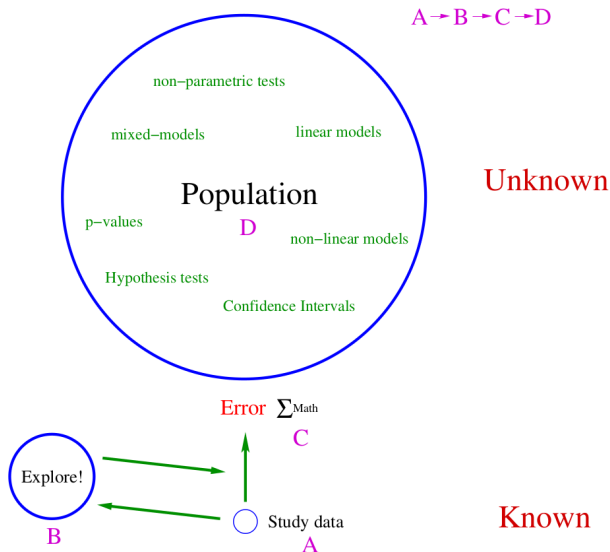# **Data Analysis with R:**
Day 4

Sonja Hartnack, Terence Odoch & Muriel Buri

October 2018

Basic Statistical Tests

Study data is collected for a purpose - to answer one or more specific scientific questions. The classical way to perform a formal statistical analyses of these data is to formulate these research questions into statistical **hypothesis tests**.

In this section we will go through a simple example in detail to highlight some of the important concepts - the general approach for more complex analyses is exactly same. *Note: the precise technical details are much less important than the concepts!*

Simple Example - One Population

After six weeks will the mean weight of a chicken be more than 250 grams?

There are 71 observations in `chickwts` from which to answer this question. This can be formulated into a statistical hypothesis test. A hypothesis test has two parts, the null hypothesis and the alternative hypothesis. This is typically written as follows:

$$H_0 : \quad \mu \quad \leq \quad 250$$
$$H_A : \quad \mu \quad > \quad 250$$

where μ is the mean weight in the **population** of chickens from which the sample of 71 chickens was drawn. Remember - we know the mean weight in the sample of chickens is greater than 250 it is the **population** of chickens which we are interested in.

Simple Example - One Population

After six weeks will the mean weight of a chicken be at least 250 grams?

$$H_0 : \quad \mu \leq 250$$
$$H_A : \quad \mu > 250$$

The null hypothesis (H_0) is the default situation, sometimes called the "state of nature". In a treatment-control trial, H_0 is typically that the effect of the treatment is not different from the control. In this example our default position is that the mean weight of chickens is ≤ 250. This is called a single-sided hypothesis test.

We now analyse the 71 observations to see whether there is evidence to **REJECT** the null hypothesis H_0, and if the null hypothesis is rejected then we can conclude that the available evidence supports the alternative hypothesis.

```
t.test(chickwts$weight, mu = 250, alternative = "greater")
t.test(chickwts$weight, mu = 250, alternative = "less")
```

Note that hypothesis testing is concerned with finding evidence in support of the null hypothesis H_0 - the default situation - rather than evidence in favour of the alternative hypothesis.

One Sample t-test

For the chicken weights data an appropriate formal analyses is to use a **one-sample t-test**, why this test is appropriate will be discussed later. This analysis involves calculating a simple summary statistic - called a *t*-statistic - which we do entirely from the observed data.

$$T_{obs} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where $\bar{x}$ is the sample mean, s the sample standard deviation and μ is the population mean in the null hypothesis which we wish to test for. We then look up the value of T_{obs} in a set of statistical tables/computer to see what the answer is to our research question.
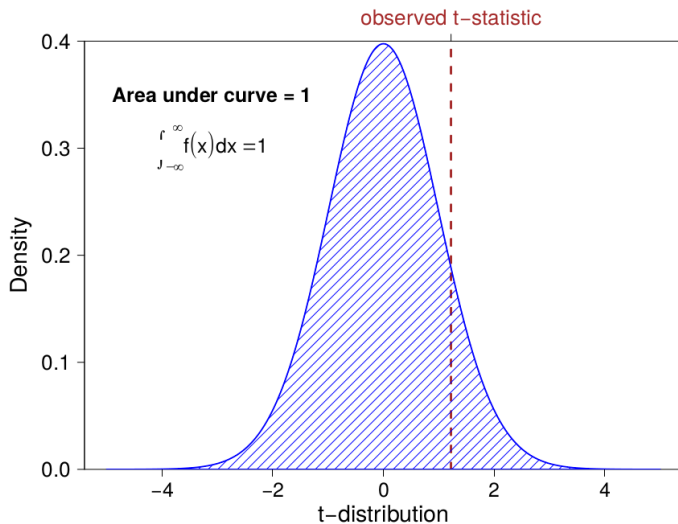
Important concept - sampling

Why is $T_{obs} = \frac{\bar{x} - \mu}{s/\sqrt{(n)}}$ called a t-statistic?

If another sample of 71 chickens from the same population were weighed then the values for $\bar{x}$ and s would be different, and hence the value for T_{obs}. If this was repeated many times and a histogram/Q-Q/P-P plot produced of the values for T_{obs} then this would follow the shape of a known distribution - **student-t probability distribution**. It is this piece of mathematics - knowing what the sampling distribution of T_{obs} is - which allows us to infer information about the population of chickens from which our original 71 chickens were sampled - without actually having to collect lots and lots of other samples of chickens! Mathematical theory is used to fill this data gap.

chickwts: t-test

$$T_{obs} = \frac{261.31 - 250}{78.07 / \sqrt{71}} = 1.22$$

Put the values for the sample mean and standard deviation into the t-statistic formula along with the $\mu = 250$. We now look up the value of this in a t-distribution reference table. All this calculation will be done for you in R but it is important to understand the general process as this is the same for hypothesis testing in other more complex analyses.

One Sample, one-sided, t-test

Important concept - p-values

- The result of a hypothesis test is usually communicated in the form of a **p-value**

- The interpretation of a p-value is of crucial importance - it is the *probability that the test statistic takes values at least as extreme as that observed* **assuming that H_0 is true**

- Exactly what **at least as extreme as** refers to depends on the alternative hypothesis H_A.

- This may sound rather abstract but it is usually obvious in practice

Simple Example - One Population

After six weeks will the mean weight of a chicken be at least 250 grams?

$$H_0 : \quad \mu \leq 250$$
$$H_A : \quad \mu > 250$$

The alternative hypothesis is $\mu > 250$ so in this test **at least as extreme as** in the definition of a p-value is the probability of observing a t-statistic which is > 1.22 **assuming that H_0 is true** - this is why 250 was used for μ when calculating T_{obs}.

One Sample, one-sided, t-test
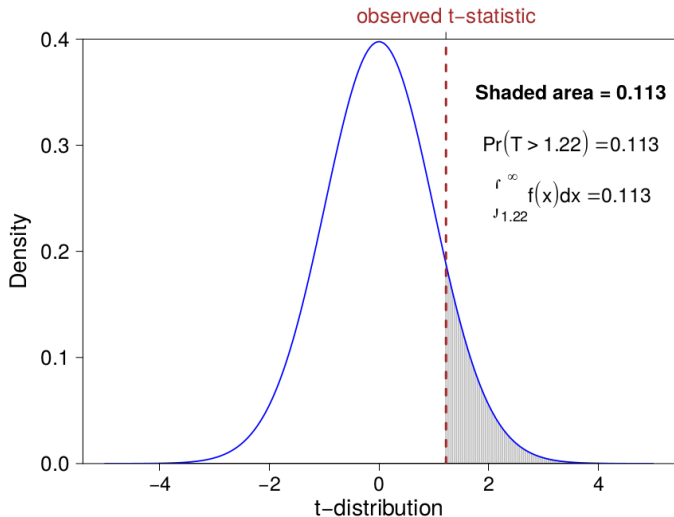
The purpose of this hypothesis test analysis is to answer a very specific scientific question:

After six weeks will the mean weight of a chicken be more than 250 grams?

So what is our answer?

The *p*-value for this hypothesis test is 0.113. Based on this value we can either **reject** H_0 and conclude that the mean weight of chickens in the population is likely to be greater than 250 grams or else we can **accept** H_0 where the mean chicken weight is less than 250 grams.

Research Question - be pragmatic with p-values

By convention a p-value of less than 0.05 is considered to provide **reasonable evidence** for rejecting H_0. A p-value of between 0.05 and 0.1 might be considered as **weak evidence** against H_0. Values of less that 0.01 are generally considered as **very strong evidence** for rejecting H_0. It is **always** best to provide a p-value in any analyses to let the reviewer/client see the strength of evidence rather than simply claiming statistically significant findings!

Communicating Results of Hypothesis Tests

Transparency is essential - the devil can be in the detail - which at the very least should comprise:

  i. what hypothesis was being tested - be clear and precise
  ii. what statistical test was used
  iii. what the *p*-value is
  iv. what the treatment effect is (more later).

This is particularly crucial if the analyses are to be given to someone *else* to then make a judgment on the scientific significance.

After six weeks will the mean weight of a chicken be equal to 250 grams?

This is now a two sided hypothesis test:

$$H_0 : \quad \mu \ = \ 250$$
$$H_A : \quad \mu \ \neq \ 250$$

This time the hypothesis test is asking how much evidence is there in our sample data to conclude that in the population of all chickens the mean weight is not equal to 250 grams.
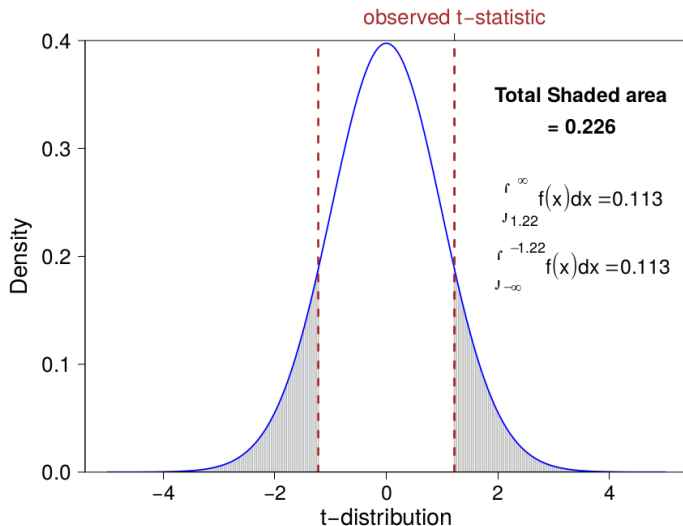
Two-sided Tests: One Population

```r
# t.test(chickwts$weight, mu = 250, alternative = "two.sided")
t.test(chickwts$weight, mu = 250)

##
##  One Sample t-test
##
## data:  chickwts$weight
## t = 1.2206, df = 70, p-value = 0.2263
## alternative hypothesis: true mean is not equal to 250
## 95 percent confidence interval:
##  242.8301 279.7896
## sample estimates:
## mean of x
##  261.3099
```

Two-sided Tests

A two-sided test is similar to a one-sided test - the key difference is in what is now defined as **at least as extreme** in the definition of the *p*-value. This time the alternative hypothesis refers to observing a value of **either $\bar{x} > 250$ or $\bar{x} < -250$ assuming that H_0 is true**, which using the t-test approach is equivalent to the probability of observing $T_{obs} > 1.22$ or $T_{obs} < -1.22$ which we can again look up in reference tables.

One Sample, two-sided, t-test

observed t−statistic

Total Shaded area = 0.226

$\int_{1.22}^{\infty} f(x)dx = 0.113$

$\int_{-\infty}^{-1.22} f(x)dx = 0.113$

Two-sided Tests

- The two-sided t-test has a *p*-value which is exactly double the single sided test.
- Think! - Intuitively the *p*-value should be less for a single sided test as the research question you are asking is much narrower e.g. greater than 250 grams, as opposed to whether the mean chicken weight might be **either** less than 250 grams **or greater** than 250 grams.

  $\rightarrow$ You are using the same amount of information (71 observations) to answer a narrower research question and so all else being equal you should expect a "more powerful" analyses (e.g. a lower *p*-value all else being equal)

Exercise 11

Chi-square Test

There are two very commonly used statistical tests for testing dependence between two categorical variables: Chi-squared test & Fisher's exact test.

To test independence of rows and columns

Risk Factor	Disease		Total
	+	-	
+	a	b	a+b
-	c	d	c+d
Total	a+c	b+d	n = a+b+c+d

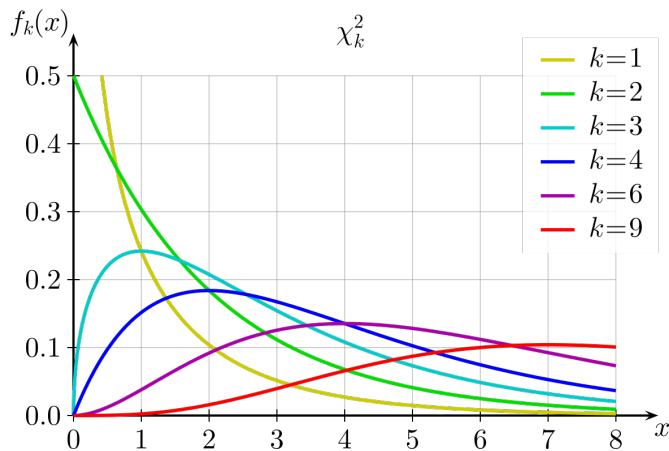- Assumptions: a, b, c, d must have at least 5 observations!

$$\chi^2 = \frac{n * (a * d - b * c)^2}{(a + b) * (c + d) * (a + c) * (b + d)}$$

Test statistic chi-square

$$\chi^2 = \Sigma \frac{(O - E)^2}{E^2}$$

The Chi-square Distribution

Exact Fisher Test: Permutation Test

	Success	Failure	Total
Therapy	7	2	9
New Therapy	2	8	10
Total	9	10	19

7	2
2	8

P = 9!*10!*9!*10! / 19!*7!*2!*2!*8! = 0.01754

8	1
1	9

P = 9!*10!*9!*10! / 19!*8!*1!*1!*9! = 0.00097

9	0
0	10

P = 9!*10!*9!*10! / 19!*9!*0!*0!*10! = 0.00001

For a one-sided test: p = 0.01754+0.00097+0.00001 = 0.01852

Exercise 12

Simulating in R

R

- Continuous variable

```r
n <- 100
rnorm(n, mean = 0, sd = 1)       # Normal distribution
```

- Binary variable

```r
rbinom(n, size = 1, prob = 0.4)  # Binomial distribution
```

- Count variable

```r
rpois(n, lambda = 7)             # Poisson distribution
```

- Other options

```r
seq(from = 0, to = 100, by = 1)          # ID
sample(3:30, size = n, replace = TRUE)   # herd size
c(rep(c(1, 2, 3), times = 33), 1)
c(rep(c(1, 2, 3), each = 33), 1)
```

- ...