Ghana Epidemiology using R - Day 2 part 1

Day 2: Introduction to Epidemiology

Epidemiology is the study of disease in populations rather than just in the individuals. It is a critical tool in the process of understanding how disease processes work. In order to understand disease we need to understand what drives disease. In the early part of the last century this was largely driven by the discovery of an ever growing list of pathogens. However, it became clear that disease is driven by varying combinations of hosts, pathogens (not always for environmental diseases) and environment. eg. many cattle are exposed to and become infected with parainfluenza virus 3 but only a few will develop disease.

Epidemiology has developed and continues to develop an array of tools to help both describe the who, what, when and where of a disease but also inferential tools and modelling to help identify and quantify the impact of risk (and protective) factors and to aid decision making for control.

Resources

Websites

- R cran: http://www.r-project.org/
- The epidemiologist R Handbook (excellent resource and they run courses) https://www.epirhandbook.com/en/
- epi info: CDC tool for epidemiological studies (http://wwwn.cdc.gov/epiinfo/)
- epitools: calculators for complex study designs (http://epitools.ausvet.com.au/content.php?page=home)

Books

- Veterinary Epidemiology Research - Dohoo et al. 2009 https://projects.upei.ca/ver/ (free to down load)
- Veterinary Epidemiology - An introduction - Pfeiffer 2009
- Veterinary Epidemiolgy - Thrusfield 2018 4th edition (5th editiion in preparation)

Day 2.1

USEFUL EPI PACKAGES

In this course we will be using mainly a package called (**epiDisplay**) and another called (**epiR**).

And of course we will want **tidyverse** as well and the package **here** that helps with construct the paths to files within your project or outside if needed (but generally good idea to have all necessary files in the project if you can).

Load required packages

```
library(tidyverse)
library(epiDisplay)
library(epiR)
library(epitools)
library(here)
library(survey)
library(skimr)
library(gt)
library(gtsummary)
library(janitor)
library(ggrepel)
```

READ IN THE DATA

```
dat <- read_csv(here("data", "fmd_herd_training.csv"))
```

This is a dataset originally collected from Cameroon in 2000 for an FMD project. A total of 147 herds were examined and herdsmen completed a questionnaire.

The sample was stratified by Division, with sample size proportional to the number of vet centres in each division. - Vet centres were randomly sampled and represented the clustering variable. They were sampled with replacement (i.e. the same centre could be selected more than once - ones with more herds were therefore more likely to get selected more) - 3 herds were randomly sampled per vet centre selected (i.e if a centre was selected twice 6 herds would be selected).

Commonly you want to check basic things about your data and that it has been read in correctly and that the variables are the right type etc.

How many rows and columns does the data frame have?

How many herds are there in the dataset?

How many variables have missing data? (hint in this data set they will be NAs or -888s or -999s)

Data Dictionary

Variable	Description
Ccode	Vet centre code (clusters)
hcode	unique herd ID
Div	Administrative Division
cluster	Which cluster does the herd belong to
vc.herds	Number of herds in each vet centre
weight.h	herd weighting
entries	Number of herds samples in each vet centre
disltyr	FMD reported in herd by herdsman last 12 months
trans1yr	Did you go on transhumance in the last 12 months
buffevr	Does the herd have contact with buffalo at the grazing sites
buycow	Do you buy cattle in
cotton	Do you feed cotton seed cake to this herd

Variable	Description
drkmix	Does the herd mix with otehr herds at the water point
DDlongJitter	longitude (with added noise)
DDlatJitter	latitude (with added noise)

Exploratory Data Analysis

It is an essential part of any data analysis exercise to first explore the data by looking at each variable in turn checking the distribution of continuous plots:

- are they what you expect?
- are there odd negatives like -888 or -999 which are often used for missing values?
- are there NAs which R interprets as missing?
- are there categorical variables with low counts that might be misspelt or that you might merge with other categories for ease of analysis?
- sometimes odd results only appear when you do biplots of two continuous variables against each other as with the GPS coordinates

A good starting point is a simple summary. What can you see from this summary?

```
summary(dat)
```

```
##       ...1           hcode         DDlatJitter     DDlongJitter
##  Min.   :  1.0   Min.   :  1.00   Min.   :6.139   Min.   :11.24
##  1st Qu.: 37.5   1st Qu.: 38.50   1st Qu.:6.538   1st Qu.:12.53
##  Median : 74.0   Median : 78.00   Median :6.836   Median :13.19
##  Mean   : 74.0   Mean   : 77.45   Mean   :6.867   Mean   :13.12
##  3rd Qu.:110.5   3rd Qu.:115.50   3rd Qu.:7.202   3rd Qu.:13.90
##  Max.   :147.0   Max.   :153.00   Max.   :7.558   Max.   :14.97
##                                   NA’s   :1       NA’s   :1
##     disltyr          buycow           cotton          trans1yr
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :1.0000   Median :0.0000   Median :0.0000
##  Mean   :0.5918   Mean   :0.5442   Mean   :0.2993   Mean   :0.4558
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##      drkmix          buffevr           Div               Ccode
##  Min.   :0.0000   Min.   :0.0000   Length:147         Length:147
##  1st Qu.:1.0000   1st Qu.:0.0000   Class :character   Class :character
##  Median :1.0000   Median :0.0000   Mode  :character   Mode  :character
##  Mean   :0.8219   Mean   :0.3265
##  3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000
##  NA’s   :1
##     cluster         vc.herds        weight.h         entries
##  Min.   : 1.00   Min.   : 30.0   Min.   : 10.00   Min.   :1.000
##  1st Qu.:13.25   1st Qu.:103.2   1st Qu.: 24.49   1st Qu.:3.000
##  Median :28.00   Median :166.0   Median : 43.83   Median :3.000
##  Mean   :27.51   Mean   :183.3   Mean   : 50.15   Mean   :4.164
##  3rd Qu.:41.00   3rd Qu.:240.0   3rd Qu.: 63.75   3rd Qu.:6.000
##  Max.   :54.00   Max.   :589.0   Max.   :196.33   Max.   :7.000
##  NA's   :1       NA's   :1       NA's   :1        NA's   :1
```

So there are a number of things to pick up on. There is a strange first column. lots of variables seem to go between 0 and 1. And several variables seem to be missing a row. This is the NAs.

I always like to check the dataframe size. There there the right number of observations? I know there should be 147 herds (I was there!).

```r
dim(dat)
```

```
## [1] 147  16
```

I also like to check the top and the bottom of the table to make sure the data is what i was expecteding it to be. Particularly when you have manipulated the data and made new variables with mutate it is a good idea to check that upi have not accidentally turned everything to NAs etc.

```r
head(dat)
```

```
## # A tibble: 6 x 16
##    ...1 hcode DDlatJitter DDlongJitter disltyr buycow cotton trans1yr drkmix
##   <dbl> <dbl>       <dbl>        <dbl>   <dbl>  <dbl>  <dbl>    <dbl>  <dbl>
## 1     1     7        6.86         13.2       1      0      1        0      1
## 2     2     8        6.81         13.2       0      1      1        0      1
## 3     3     9        7.09         13.9       1      0      0        1      1
## 4     4    10        7.32         14.1       1      1      0        0      1
## 5     5    11        7.18         14.1       1      0      1        0      1
## 6     6    12        7.20         14.1       1      0      1        0      1
## # i 7 more variables: buffevr <dbl>, Div <chr>, Ccode <chr>, cluster <dbl>,
## #   vc.herds <dbl>, weight.h <dbl>, entries <dbl>
```

```r
tail(dat)
```

```
## # A tibble: 6 x 16
##    ...1 hcode DDlatJitter DDlongJitter disltyr buycow cotton trans1yr drkmix
##   <dbl> <dbl>       <dbl>        <dbl>   <dbl>  <dbl>  <dbl>    <dbl>  <dbl>
## 1   142   150        7.46         13.7       1      0      0        1      1
## 2   143   147        7.52         13.4       0      0      0        0      1
## 3   144   148        7.01         13.5       1      1      0        0      1
## 4   145   151        7.49         13.7       1      0      0        0      1
## 5   146   152        7.29         13.3       0      1      1        0      1
## 6   147   153        7.34         13.4       1      1      1        0      1
## # i 7 more variables: buffevr <dbl>, Div <chr>, Ccode <chr>, cluster <dbl>,
## #   vc.herds <dbl>, weight.h <dbl>, entries <dbl>
```

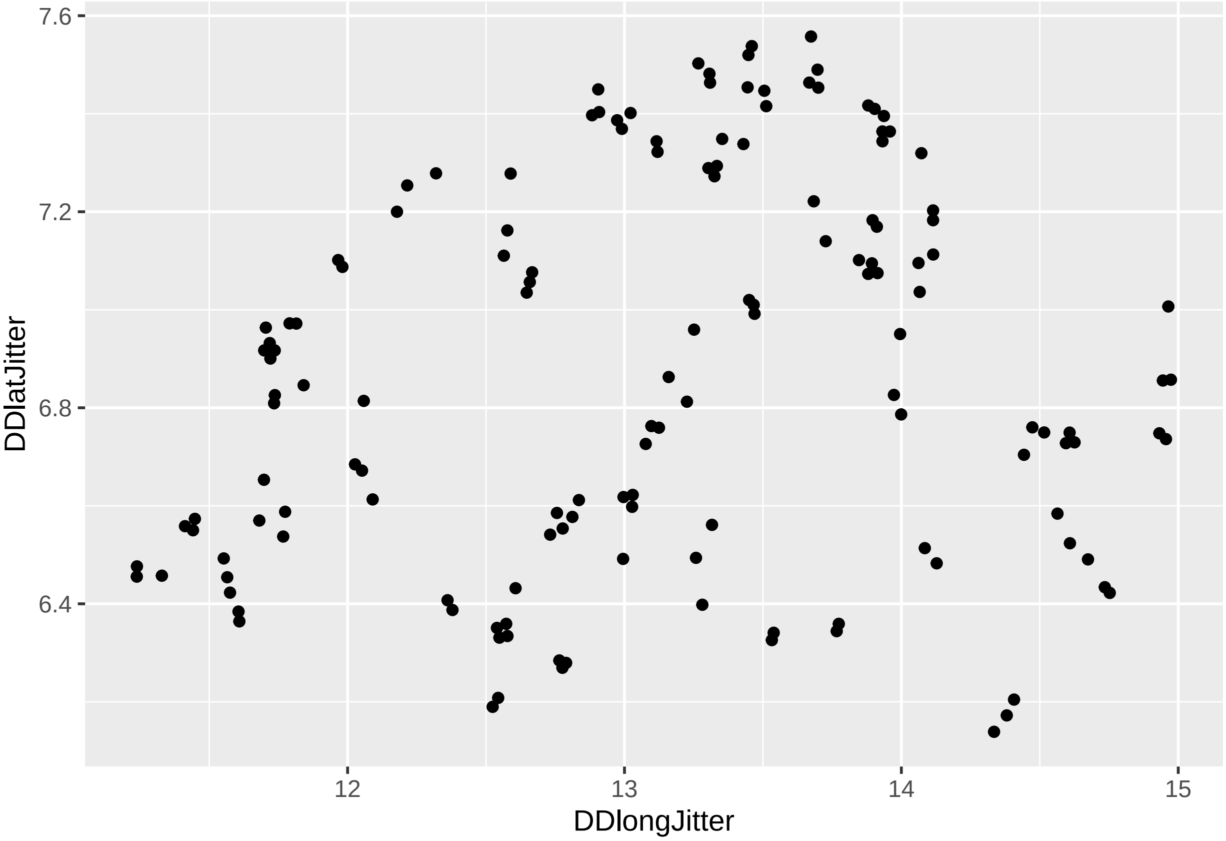
```r
skim(dat)
```

```r
view(dat)
```

Are the herds where you think they should be? I have added in colour by whether they were FMD positive herd.

```
ggplot(data=dat, aes(x=DDlongJitter, y=DDlatJitter) ) +
  geom_point()
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
```



What does the distribution of positive herds look like?

```
ggplot(data=dat, aes(x=DDlongJitter, y=DDlatJitter, color = disltyr)) +
  geom_point()
```

For categorical variables it is useful to make tables and spot where there are cells with very few counts etc.

```
dat %>%
  tabyl(Div) %>%
  gt() %>%
  fmt_number(columns = vars(percent), decimals = 3)
```

```
## Warning: Since gt v0.3.0, ‘columns = vars(...)‘ has been deprecated.
## * Please use ‘columns = c(...)‘ instead.
```

So when we look through the data are there anythings that stand out as an issue?

Div	n	percent	valid_percent
DJEREM	30	0.204	0.2054795
FARO_ET_DEO	15	0.102	0.1027397
MAYO_BANYO	28	0.190	0.1917808
MBERE	25	0.170	0.1712329
VINA	48	0.327	0.3287671
NA	1	0.007	NA

Div	0	1
DJEREM	18	12
FARO_ET_DEO	4	11
MAYO_BANYO	15	13
MBERE	11	14
VINA	11	37
NA	1	0

```r
dat %>%
  tabyl(Div, disltyr) %>%
  gt()
```

So it looks like there is a observation missing a lot of information. One option is to delete this if you can not find out the information. However, we do have the missing information and so can add it.

```r
dat <- dat %>%
  dplyr::select(-c('...1')) %>%
  mutate(DDlatJitter = case_when(hcode == 118 & is.na(DDlatJitter) ~ 6.43197, TRUE ~ DDlatJitter),
         DDlongJitter = case_when(hcode == 118 & is.na(DDlongJitter) ~ 12.4228, TRUE ~ DDlongJitter),
         Ccode = case_when(hcode == 118 & is.na(Ccode) ~ "DAM", TRUE ~ Ccode),
         Div = case_when(hcode == 118 & is.na(Div) ~ "DJEREM", TRUE ~ Div),
         cluster = case_when(hcode == 118 & is.na(cluster) ~ 42, TRUE ~ cluster),
         vc.herds = case_when(hcode == 118 & is.na(vc.herds) ~ 203, TRUE ~ vc.herds),
         weight.h = case_when(hcode == 118 & is.na(weight.h) ~ 101.50000, TRUE ~ weight.h))
```

Now lets check the data again and make sure all the corrections have been made.

```r
ggplot(data=dat, aes(x=DDlongJitter, y=DDlatJitter)) +
  geom_point(color = "blue", size = 3) +
  geom_label_repel(aes(label = Ccode),
                   box.padding   = 0.35,
                   point.padding = 0.5,
                   segment.color = 'grey50')
```

Day 2.2

MEASURING DISEASE FREQUENCY

Prevalence: This is a measure of existing cases or exposure.

The proportion of individuals in a defined population that have the outcome under study at a defined instant (a point in time).

$$prevalence = \frac{the\ number\ of\ cases\ at\ a\ given\ point\ in\ time}{number\ of\ animals\ at\ that\ time}$$

Eg. What is the prevalence of FMD positive herds in the Adamawa Province of Cameroon in 2000?

What we need to do is look at how many animals had Lepto antibodies in our sample. One way to do this would be to make a table and count up the number of animals in each group.

```
dat %>%
  group_by(disltyr) %>%
  tally()
```

What do you get?

We can do all this within R.

```
ci(dat$disltyr)
ci.binomial(dat$disltyr)
```

There are many other ways you can get to the same result in R. But many of these function rely on first creating a table or getting the number tallied before entering and does not make for very robist coding. Fine for quick check.

```
tab <- dat %>%
     group_by(disltyr) %>%
     tally()
ci.binomial(as.numeric(tab[2,2]),
            as.numeric(tab[2,2], tab[1,2]))
```

```
##   events total probability se exact.lower95ci exact.upper95ci
##       87    87           1  0       0.9583908               1
```

```
ci.binomial(87, 147)
```

```
##   events total probability         se exact.lower95ci exact.upper95ci
##       87   147   0.5918367 0.04053771       0.5077959       0.6721224
```

If the prevalence is very low you need to be aware that some of the estimators (they formulas for estimating the standard errors start to be not very good and you need to use exact methods)

Lots of functions will give results using approximate methods for estimating the CI. The asymptotic one is the one based on a normal approximation that most stats packages will give as a default. Most of the time this is not a major issue but when the prevalence is low it can cause major errors.

Try replacing with 1 out of 147 and rerunning estimation…

```
## think about why the second version might be safer.
# an even simpler way
binom.exact(1, 147)
```

```
##   x   n  proportion        lower      upper conf.level
## 1 1 147 0.006802721 0.0001722152 0.03731818       0.95
```

```r
binom.approx(1, 147)
```

```
##   x   n  proportion        lower      upper conf.level
## 1 1 147 0.006802721 -0.006484939 0.02009038       0.95
```

```r
ci.binomial(1, 147)
```

```
##  events total probability          se exact.lower95ci exact.upper95ci
##       1   147 0.006802721 0.006779543    0.0001721088      0.03731837
```

The exact (binomial) is more reliable and the one used in general for low (or high) prevalence situations.

Incidence: There are a few variations on this but the basic idea is the number of new cases of disease in a given period for a certain population eg. 23 cases of mastitis per 100 cow years at risk.

The **cumulative incidence** is usually given as a proportion like the the prevalence but will always have a given period of risk underlying it.

$$cumulative\ incidence = \frac{the\ number\ of\ new\ cases\ during\ a\ given\ period}{number\ of\ subjects\ at\ risk\ at\ the\ start\ of\ the\ study}$$

We will use a different dataset that has some time component we can work with (slightly fixed as you will see but just pretend).

READ IN THE DATA

```r
dat <- read_csv(here("data", "cattle_data.csv"))
dim(dat)
class(dat)
head(dat, 14)
tail(dat, 12)
summary(dat)
```

Data Dictionary

Variable	Definition
DamID	The unique identifier for the Dam
Lepto	Binary positive/negative result of a serological ELISA for *L. hardjo*
BrucellaLFA	Binary positive/negative result of a LFD for *B. abortus*
BrucellaRBG	Binary positive/negative result of a Rose Bengal for *B. abortus*
GirthDam	The dam girth measurement in cm
CSDam	Dam condition score on a scale 1-10
SL	Sublocation in the study site - number is just an identifier and not to be used as an integer
East	Digital degrees east of Greenwich
North	Digital degrees north of the equator
Elevat	Metres above sea level
CalfSex	Sex of the last calf the dam gave birth to
DAge	Dam age (years) at last calving
DCalving	Number of calvings including current
TDFarm	Number of breeding females on the farm

Variable	Definition
DAge2	Imputed ages randomly drawn from uniform distribution for missing ages in DAge

The diagnostic tests are binary not numeric variables. The '0' stands for negative and '1' for positive.

```
dat %>%
  group_by(Lepto) %>%
     tally()
```

```
## # A tibble: 2 x 2
##   Lepto     n
##   <dbl> <int>
## 1     0   187
## 2     1    12
```

Using our Kenyan data if we assumed that we observed the animals for a year and we had 12 cases we could estimate the cumulative incidence as 12/199 ~6%.

Now it is of course not ever that simple as in the real world animals are bought and sold, some die, etc. Particularly if you want to consider incidence over a long period. Then it becomes really hard to just add up the individuals to see how many were at risk for the year. There are various ways to deal with this: (1) you can say that if there was relatively low turn over you use the number at the end or the start of the period of interest; (2) you can take an average population size by taking the mean of the start and end population; or (3) you can try and work out how much time each animal actually contributed and total this up and then estimate the incidence rate.

Incidence Rate doesn't relate to the number of individuals initially at risk but rather to the amount of time they spend at risk. It won't necessarily lie between 0-1. The measure allows the population at risk to change. If a new individual joins, you add in their 'time at risk' and similarly if an individual leaves (eg. a cow is sold from a herd or dies from another cause) they stop contributing further 'time at risk'.

$$incidence\ rate = \frac{the\ number\ of\ new\ cases\ during\ a\ given\ period}{subject\ time\ at\ risk}$$

Eg. What is the incidence rate for leptospirosis in this Kenyan population of cows? We have to make a few assumptions for convenience here as we don't have the real data but if we assume that the DAge2 column was the age at which they contracted disease and for the negatives it is the period in years they have been observed we can work out the total 'time at risk'. We can also calculate the number of cases using the 'Lepto' column.

```
time_at_risk <- sum(dat$DAge2)
time_at_risk
```

```
## [1] 1434.5
```

```
cases <- sum(dat$Lepto)
cases
```

```
## [1] 12
```

```
cases / time_at_risk ## cases per year at risk
```

```
## [1] 0.008365284
```

```
cases / time_at_risk * 1000 ## cases per 1000 cows per year at risk
```

```
## [1] 8.365284
```

We should get something like 0.008 cases per cow year at risk or 8 cases per 1000 cows per year.

However we are usually estimating these measures from a sample and also working with data in a database and we can do all this more easily with various R packages. One is epiR so we run the library function to call it (assuming it is already installed on your machine).

```
ci.poisson(cases, time_at_risk, alpha=0.05)
```

```
##   events person.time   incidence         se exact.lower95ci exact.upper95ci
##       12      1434.5 0.008365284 0.00241485     0.004316487      0.01461415
```

Other useful measures that you may come across are the **mortality rate** which is the number of deaths in a time period for a defined population (it is the incidence of death). The other commonly used term is the **case fatality rate** which is actually a proportion of the cases of a specific disease that then go on to die. eg. HPAI in people has only caused disease in a small number of people (300) but of these about 50% died so there is a very high case fatality rate.

NB. Some measures were named before risk and rate were defined. For example, **survival rate** (proportion of persons in a group, usually patients with a disease, who survived in a specified period of time) is not a rate, it's a proportion. This can get confusing but just pay attention to the outcome given. Eg. 10 cases per 1000 animals in one year is a risk. 10 cases per 100 cow-years is a rate.

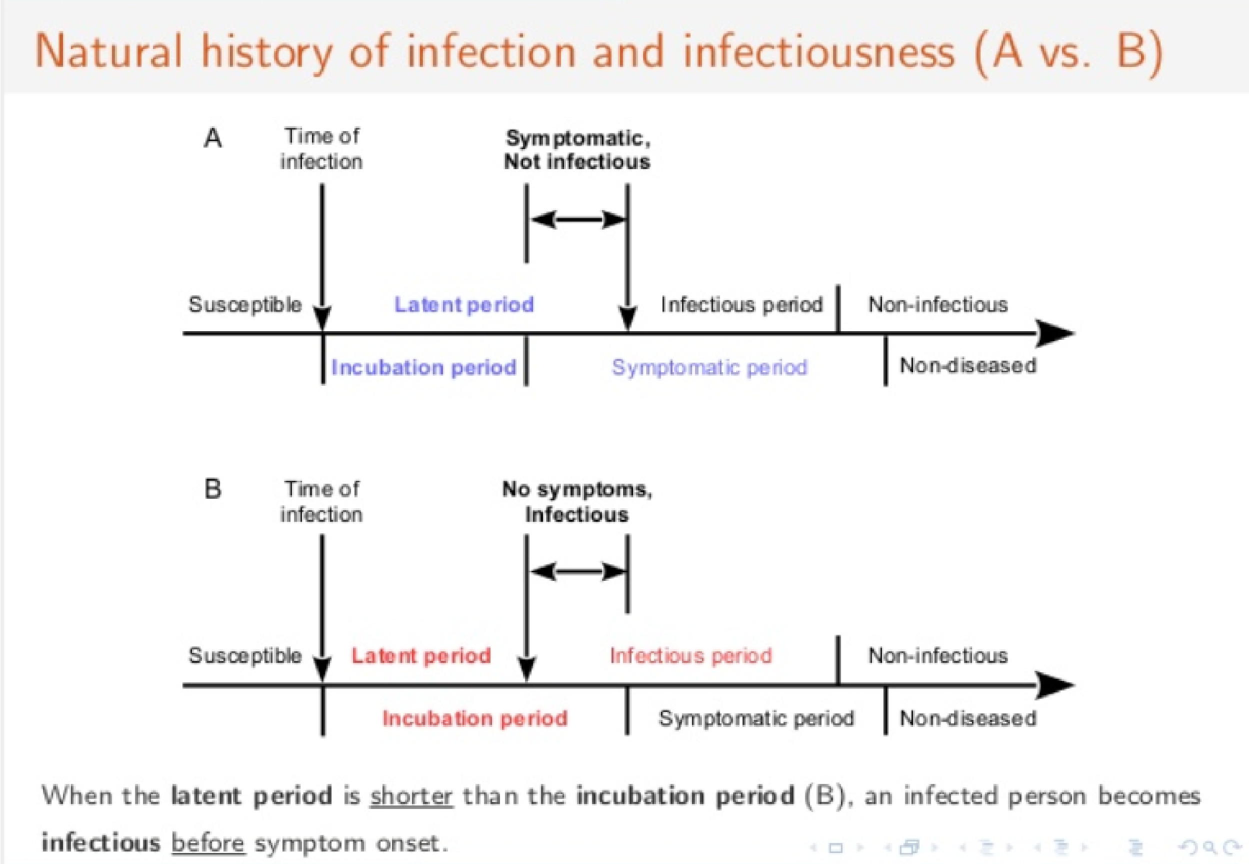
Case definition: What outcome are you investigating? What is your case definition? Ideally you would use the same definition as has been used in other studies because then the results of your study can be sensibly compared with other literature. However, if you can't use an accepted definition (or if there isn't an accepted definition) then it is important to give a clear definition of what you consider as a 'case'. For example, you might define a case as "someone who reports that they have had flu" or "someone who reports that they have had flu and who has had an elevated temperature for two or more days" or you might perform tests and your case definition would be "someone who reports that they have had flu and for whom it was possible to isolate influenza virus type X using technique Y". This sounds obvious at some level but you will be amazed at how complex it can become in many clinical settings and when trying to do case control studies and looking back through patient records trying to decide who was a case.

This can be complicated when considering infectious disease. An individual my be infected but not show signs of disease. Your definition would then depend on what was important. If the disease was such that an infected individual could transmit the infection without showing signs of disease, it would be worth testing individuals so that you knew who had the potential to spread the disease (case definition might be serological). If those without symptoms could not spread the disease then it might be more efficient to focus simply on those with clinical disease (case definition based on symptoms). The current outbreak of Ebola is a really good example of the difficulties of case definition. Individuals might present with fever and lethargy. These symptoms can be indicative of a number of diseases. One might base the decision to isolate on a case definition of "fever and lethargy". E.g. Suspected case defined as individual presenting with "fever and lethargy" who should then be tested for Ebola. If an individual in the areas affected by Ebola died following fever, lethargy, diarrhoea etc. then this might be considered as a confirmed case without testing because, in a situation with limited resources, it is more important to test the living than the dead. Procedures for safe and respectful burial in a way that minimised the chances of ongoing transmission from the body would then be applied. The case definition might also be partially based on exposure (below).

Population at risk: In the same way that you have to carefully define your outcome of interest, it's also important to define your population at risk. For example, students on this MSc course. This may be complicated by vaccinations or prior infections which might remove individuals from an at-risk population.

Exposure: Exposure means that an individual has been exposed to a risk factor for a disease be that a pathogen, a toxin or a particular environment. In infectious disease epidemiology we typically look at exposures using serological screening tests that can detect antibodies to a particular pathogen. Exposure does not guarantee disease and an animal may or may not still be infected or infectious. With respect to case definitions, knowledge that individuals have been exposed to a pathogen or environment might form part of a case definition. For example, in the Ebola case definition above, if the individual presents with fever and lethargy and they report that they had been treating a family member who recently died of Ebola, one might consider that as a case. The case definition would be extended to "individual from whom virus has been isolated or who has symptoms "fever and lethargy" and who reports close contact with previous case".

INFECTIOUS DISEASE



Natural history of infection and infectiousness (A vs. B)

When the **latent period** is shorter than the **incubation period** (B), an infected person becomes **infectious** before symptom onset.

Figure 1: When the **latent period** is *shorter* than the **incubation period** (B), an infected person becomes **infectious** *before* symptom onset

Infected: an animal may become infected following exposure to a pathogen.

Latent period: There is then usually a delay or latent period between infection and the individual being infectious themselves. This latent period can be days (eg. FMD) or years (eg. BSE).

Infectious: An individual is said to be infectious if they are capable of transmitting/releasing a pathogen to infect another individual.

Incubation period: The period between being infected and developing clinical disease is the incubation period. For most diseases this is slightly longer than the latent period and so gives the pathogen a chance to transmit itself.

The period between becoming infectious and the development of clinical signs is a critical period (no specific name yet) but this is often the major transmission period and can have profound impact on our ability to control a disease outbreak. eg. FMD animals become infectious 1-2 days (up to 4 days in sheep) prior to the development of clinical signs alerting us that there is a problem. By comparison in SARS people did not become infectious until just after they developed clinical signs. This proved critical in its control as quarantining people as soon as they developed suspected symptoms stopped the spread in Asia in particular. Similarly, Ebola patients do not become infectious until they develop clinical signs of illness which is why it has previously been possible to control outbreaks.

Exercises

Ex 1.

Using the cattle_data.csv check all the variables in terms of:

- what class they are
- if they have missing values or any unreasonable entries
  - for categorical variables use **table(variablename, useNA="always")** (SL and CalfSex)
  - for numerical variables use **summary(variablename)**
  - perhaps use skim in the skimr package to examine the data

Ex 2.

Correct any errors you think might be typos (hint. check how we changed -888 to NA)

Ex 3.

Work out the binomial confidence interval for *Brucella* prevalence using both *Brucella* tests. Why are they different? Make a 2x2 table to find out where the test disagree. Hint: Try `with(dat, table(BrucellaRBG, BrucellaLFA))`

Ex 4.

Estimate the incidence rate of brucellosis using variable *BrucellaRBG* and assuming that DAge2 column was the age at which they contracted disease and for the negatives it is the period in years they have been observed we can work out the incidence.