Ghana Complex Surveys Day 2 part 2

```r
#load packages
library(tidyverse)
library(survey)
library(srvyr)
library(knitr)
library(here)
library(ggrepel)
```

Reference

Good online tool for understanding sampling https://shiny.rit.albany.edu/stat/

Website with details on surveys https://www.jstatsoft.org/article/view/v009i08

https://stats.oarc.ucla.edu/r/seminars/survey-data-analysis-with-r/

https://tidy-survey-r.github.io/tidy-survey-book/c01-intro.html#survey-analysis-in-r

Sampling

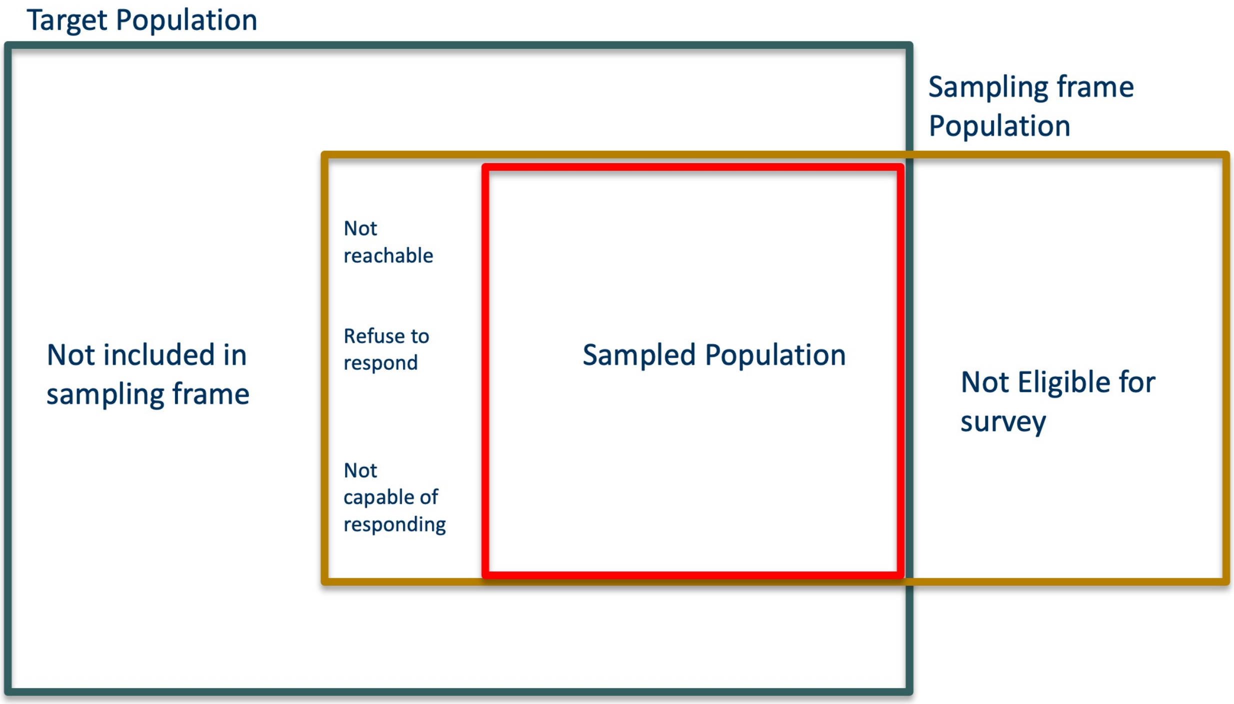
Epidemiology and statistics in general is all about estimating parameters or meansurements about groups of animals or people or farms.

In a cost free world we would just go and sample every individual in the population of interest, measure the thing we are intersted in and get the mean. For example if we wanted to know the milk yield of dairy cows in Ghana we would go and sample visit ecery dairy cow in Ghana and get it milk yield and add then and dividie by the sample size and we have our mean. We could also calculate the standard deviation or variance to know how much cows idffer etc. This is basically a census and many countries do try and do human censuses every 10 years.

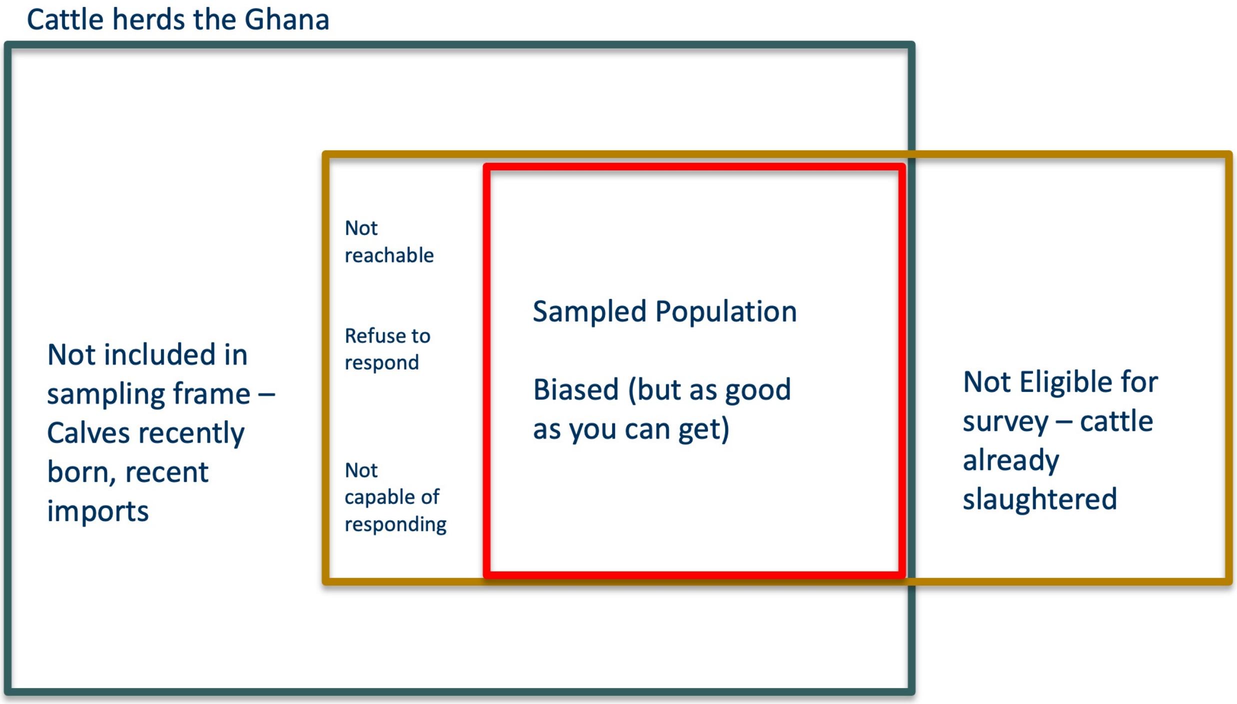
However, we do not live in a cost free world. Field sampling is, in fact, extremely expensive and time consuming and requires teams to head out into the country and find households or farms and gather the information. Therefore to make it a cost effective exercise and managable in every way we take samples.

The challange is how do we identify the individual units to sample? If we just go and sample farms say near where we live we may get a very biased sample and thus an incorrect estimate. For example, sampling dairy farms near a major city like Kumasi may give a very biased estimate of milk production compared to a sample from a more remote area. Farmers may have different access to genetics, veterianry care etc.

So you might start off with your target population of dairy cattle in Ghana and you go out and get a list of all the farms and what animals they have which is your sampling frame. But as you are doing this you will find that some of the farms no longer keep dairy cattle any more, just beef cattle and so are in reality no longer eligible. Some new dairy farms will have been extablished but were not yet listed in the government database. Then when you have selected your sample and you go to sample them you may find that you can not find the farmer for some reason. they may not have a mobile or be off site and some may refuse.



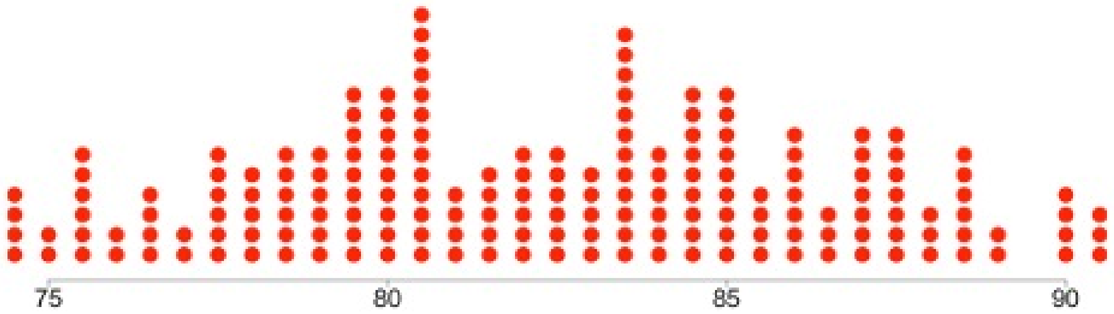
What you end up with is a sample that may not be quite as representative of the target population as you had hoped. However, if you do it all propery and systematically you will at least have an idea about how many refusal or dropout farms etc you have and in practice when you do this you try and capture something about tere frms to know if you are missing specific groups like the small holders or the commercial farms or women etc.


Cattle herds the Ghana
Not included in sampling frame – Calves recently born, recent imports
Not reachable
Refuse to respond
Not capable of responding
Sampled Population
Biased (but as good as you can get)
Not Eligible for survey – cattle already slaughtered

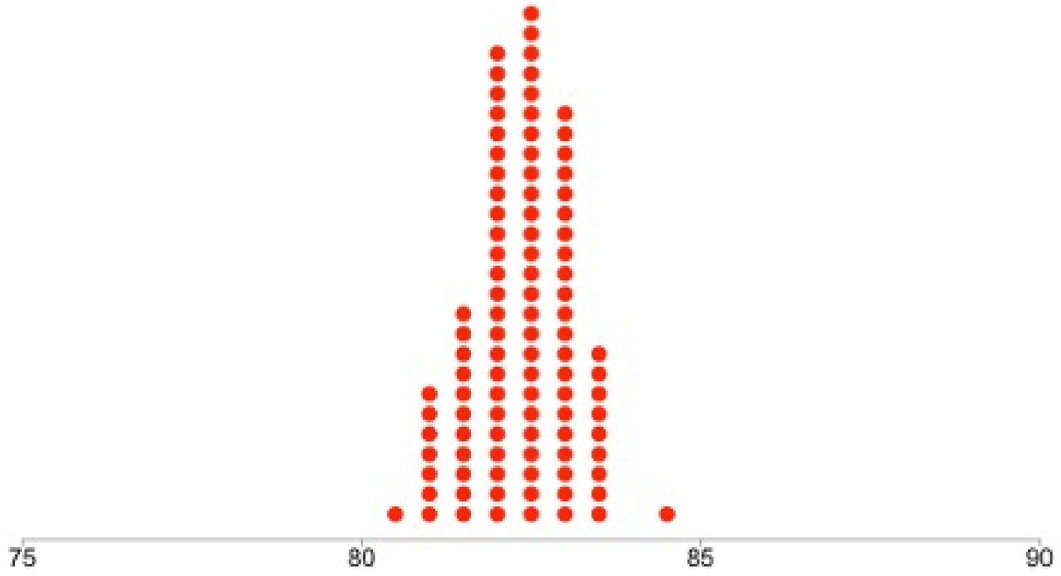
Sampling variation

So say we were wanting to know the mean weight of 6 month old dairy calves in Volta Province. But we have no money so we go and sample 10 animals and get an estimate. We calculate our mean and get 82.1 Kg. But what might the estimate have been if we had taken a different random sample?

We can keep doing this and then get a distribution of mean weights for 6 month old calves from samples of size 10.



What you see is that you could get an estimate as low as ~ 74Kg or as high as 92Kg. However, most of the samples you take will give you an estimate. closer to the true estimate somewhere around 85Kg. But we can see that there is a reasonably large degree of uncertainty.



If instead of 10 we took a sample of 500 and did the same thing again and looked at the distribution we see a different pattern. Now the means range from ~81 to 84Kg. This is much small than with the sample of 10. So by increaseing the sample size (ASSUMING IT IS A RANDOM SAMPLE) increasing the sample size gives us more precision and reduces the uncertainty about the estimate.

```
x <- c(100, 77, 56, 84, 68, 63, 83, 81, 93, 78)
mean(x)
```

```
## [1] 78.3
```

```
sd(x)
```

```
## [1] 13.28366
```

```
# manually calculate the 95% CI using normal approximation
mean(x) - 1.96*sd(x)/sqrt(length(x))
```

```
## [1] 70.0667
```

```
mean(x) + 1.96*sd(x)/sqrt(length(x))
```

```
## [1] 86.5333
```

```
# manually calculate the 95% CI using t distribution for small sample sizes
t.test(x)
```

```
##
##  One Sample t-test
##
## data:  x
## t = 18.64, df = 9, p-value = 1.688e-08
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  68.79744 87.80256
## sample estimates:
## mean of x
##      78.3
```

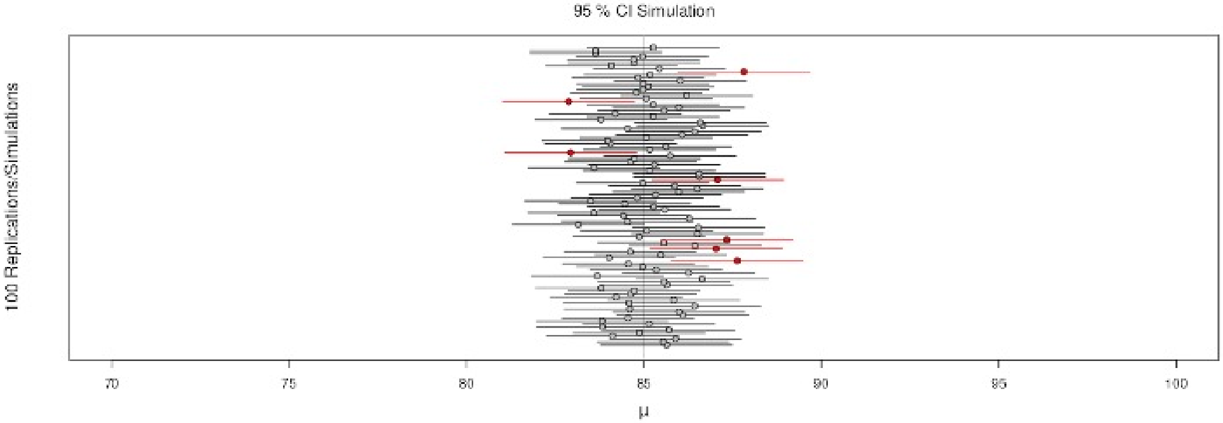
```
mean(x) - 2.262*sd(x)/sqrt(length(x))
```

```
## [1] 68.7981
```

```
mean(x) + 2.262*sd(x)/sqrt(length(x))
```

```
## [1] 87.8019
```

If we were to repeated draw our samples as before but also construct a 95% CI we would get something like this.



Interpretation of a CI

A 95% confidence interval is a range of values, calculated from the sample data, that is expected to contain the true population parameter 95% of the time in repeated sampling.

In more detail:

Interpretation in Repeated Sampling: If we were to take many random samples from the population and calculate a 95% confidence interval for each sample, approximately 95% of those intervals would contain the true population parameter. Note: This interpretation emphasizes the process of repeated sampling and the long-run frequency of intervals capturing the true parameter. #### What It Does Not Mean: It does not mean there is a 95% probability that the true parameter lies within the specific interval calculated from your data. The true parameter is a fixed value; the confidence interval is random and varies with the sample. It also does not guarantee that the true value lies within any one particular interval; it either does or does not. #### Assumptions: The calculation of the confidence interval relies on specific assumptions, such as the sample being randomly drawn and the data meeting the assumptions of the statistical method used (e.g., normality, independence).

Binary outcome CI

The same principles apply for sampling for disease. But here the outcome is binary, diseased or not diseased. So when you take a sample of animals and test for a disease/exposure you get one estimate but if you took a different sample you would get a different estimate. But at the end of the day you have to take the best sample you can (least biased) and be aware that it is an estimate and there is some uncertainty and add confidence intervals to reflect the uncertainty.

So for example if we take a sample of 10 animals from a population with an underlying disease prevalence of 10% we get differet numbers of positive animals in each sample and this will result in different estimates of the prevalence.

```
rbinom( 1 , 10, 0.1)
```

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

```
rbinom( 1 , 10, 0.1)
```

```
## [1] 3
```

```
rbinom( 1 , 10, 0.1)/10*100
```

```
## [1] 10
```

```
rbinom( 1 , 10, 0.1)/10*100
```

```
## [1] 10
```

We can get the confidence interval from our sample asa before.

```
ci.binomial(1,10)
ci.binomial(3,10)
```

If we do this a 100 times and now also add the confidence interval to a datarfame we get this… you can change the seed number to allow it to draw a differnt sample.

```
set.seed(35)
num_pos <- rbinom(100 , 30, 0.1)
estimate <- num_pos/30*100
n <- rep(30,100)
order <- 1:100
bin <- as.data.frame(cbind(order, num_pos, estimate, n))
for (i in 1:100){
bin$lower[i] <- binom.exact(num_pos[i], n[i])$conf.int[1]*100
bin$upper[i] <- binom.exact(num_pos[i], n[i])$conf.int[2]*100
}
threshold <- as.numeric(lower>10)

ggplot(data=bin, aes(y=estimate, x= as.factor(order))) +
  geom_point() +
  geom_errorbar(data = bin, aes(ymin = lower, ymax=upper, colour = as.factor(threshold))) +
  scale_color_manual(values = c("black", "red")) +
  ylab("Estimated Prevalence") +
  theme_bw() +
  theme(axis.title.y=element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank()) +
  coord_flip() +
  guides(color = guide_legend(title = "CI that do not\n include true\n prevalence"))
```

Complex sampling designs

Next lets look at how to think about designing more complex sampling strategies and analysiing these. This is just a very basic intro to get you to start thinking about the issues you need to consider when designing a survey. You need to make sure you consult and statistician and/or epidemiologists comfortable with complex designs when doing your own studies.

Sampling strategies

There are of course many different ways to apporach sampling. At the very core is the idea that any sample of size x has the same probability of being selected as any other sample of size x. This is the important because most statistical tests, models and estimators have the assumption that the sample is independent and random and therefore likely to be representative. Violating the assumptions means any conclusions you draw may be incorrect.

Random samples are more likely to be representative of the population; therefore you can be more confident with your statistical inferences with a random sample. Independence means the value of one observation does not influence or affect the value of other observations. Independent data items are not connected with one another in any way (unless you account for it in your model). This includes the observations in both the “between” and “within” groups in your sample. Non-independent observations introduce bias and can make your statistical test give too many false positives.

So how does this relate to sampling strategies?

Well there are a number of ways you can select your sample. *1) Convenient sample* - here you just go and find some animals or farms that are easy to get to and you know the foarmer will allow you to sample. But this sample is very difficult to extrapolate to the whole population and so very hard to interpret and make sense of. It will still cost you lots to get the samples and test them but the quality of the interomation will be very poor.

2) Simple random sample - here you get you sampling frame and you select a random sample. This is on the surface ideal from a statistical point of view but may in fact nt be the most efficent. Going to lots of farms and just sampling 1 cow for example would be very expensive.

3) Geographical random sample - in situations where there is no sampling frame at all and the population of interst is reasonably evenly spread you could use the computer to select random coordinates within a given boundary and then try to get to the location and select the nearest farm or animal etc. Again practically can be challanging.

4) Systematic random sampling - a good compromise in some settings. For example sampling cows coming through the milking parlour. You say have 20 stalls and you need to randomly select a starting stall and then say take every 7th cow that comes through. This way you avoid just taking the health ones that came inn first.

5) Complex sampling design - in reality things are often clustered or unevenly distributed. For example children re clustered into classes within schools or animals are clustered within herds on farms. The statistical “issue” with this is that because animals on farms are likely to managed in a similar way they are more likley to have similar disease status or milk yields etc. Therefore they are not truely independent.

Similarly you might want to make sure that you have a statistically robust sample from different regions or countries that have quite different numbers of farms and cows.

Why does this matter?

Lets look at the table below:

If we were to sample 200 cows in Scotland and 300 from England and Wales and test them for somethng like bTB (amde up numbers). If we work out the prevalence based on these numebrs we can see that the prevalence in Scotland is 10% compared to 5% in England and Wales. If we just now total all this we get an estimate of 7% overall in the sample.

Is this correct?

Of course NO IT IS NOT. We have failed to account for the fact that as a proportion we have sampled far more animals in Scotland compared to England and Wales and we need to adjust for this. One way is to work out how many animals should be positive in the population of our estimate were correct (Pop. positive). If we now add these to get the true estimate of total positive and divide by the overall population we get an answer of 5.5% instead of 7%.

Another way to think about this is in terms of weighting. The Scottish sample coms from about 9% of the populations compared to the sample from England and Wales that represents 91% of the overall population. Therefore we can take 9% of the 10% estimate for Scotland and 91% of the 5% estimate for England and Wales and we should get the same result.

(0.091 x 0.1) + (0.909 x 0.05) = 0.05455 ~ 0.055 or 5.5%

Country	Scotland	England/Wales	Overall
Sampled	200	300	500
Test positive	20	15	35

Country	Scotland	England/Wales	Overall
Prevalence	10%	5%	7%
Population	1,000,000	10,000,000	11,000,000
Pop. positive	100,000	500,000	600,000
weighting	0.091	0.909	
Pop prev	10%	5%	5.5%

You may wonder why you would use clusters to sample given that they are not independent. Well the main advantage is around cost as in general one of your main costs in field sampling is getting to a farm or unit and then the cost of individual samples within that is usually orders of magnitude cheaper. But if you do use cluster sampling you then need to adjust for it in your estimation.

Pros and cons of stratification and cluster sampling.

Stratified Sampling Ensures specific sub-populations are included Improves precision

Cluster Sampling Improves efficiency Loss of precision Animals within clusters tend to be more similar If ignored risk of under estimating the variance

A worked Example

Import the FMD data again

First lets import the data we used earlier for FMD in Cameroon.

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

The dataset & sampling strategy

A cross sectional study to estimate prevalence of FMD in cattle herds in the Adamawa Region of Cameroon. The region has 5 administrative divisions and 88 veterinary centres.

- 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.
- Up to 3 herds were randomly sampled per vet centre.

Data dictionary

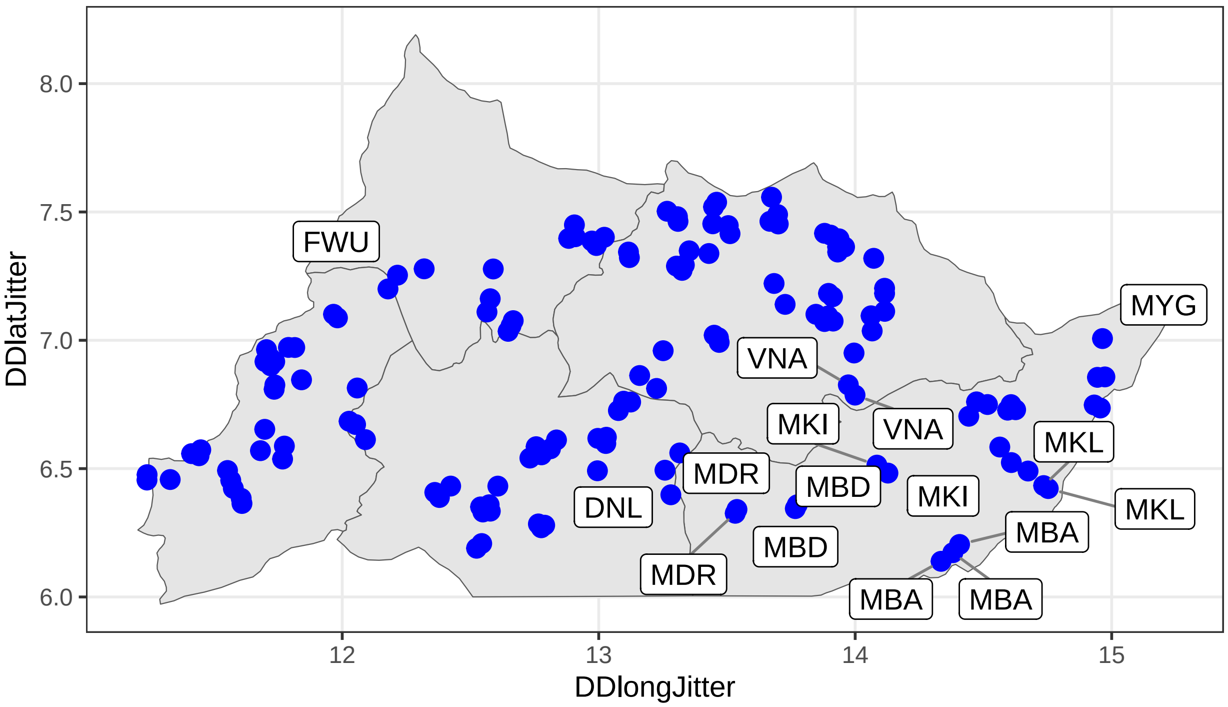
Variable	Description
Ccode	Vet centre code (clusters)
hcode	unique herd ID
Div	Administrative Division
cluster	which cluster group herd belongs to
vc.herds	Number of herds in each vet centre
weight.h	weighing
entries	Number of herds sampled in each vet centre
disltyr	FMD reported in herd by herdsman 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
drkmix	Does the herd mix with otehr herds at the water point
DDlongJitter	longitude (with added noise)
DDlatJitter	latitude (with added noise)

First we clean the data again and sort out the missing row…

```
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))
```

Then we plot it just to check it looks like we were expecting. Normally you would also check summaries etc to make sure all variables are right.

```
adm <- sf::read_sf(here("data", "adamawa.shp"))

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



Create weight variable

This should be the number of herds in each vet centre divided by the number of herds sampled in each centre. We will call this variable **Weight**

```
dat <- dat %>%
  mutate(weight = vc.herds/entries)
```

Syntax for creating a survey design object

```
survey_design <- dat %>%
  as_survey_design(ids = Ccode,
                   # cluster id
                   weights = weight.h,
                   # weights
                   strata = Div
                   # strata id
                   )
```

Estimate the prevalence of FMD

Simple random sample - no adjustments

The assumption here is we have a simple random sample and we are ignoring any design issues.

This then creates the survey object

```
survey_design_simple <- dat %>%
  as_survey_design(ids = 1, # 1 for no cluster ids
                   weights = NULL, # No weight added
                   strata = NULL # sampling was simple
                   #(no strata)
                  )
```

Then we use this to generate the estimate and confidence intervals..

Simple Random Sampling

```
prop_srs <- survey_design_simple %>%
  group_by(disltyr) %>%
  summarize(fmd_prop = survey_prop(vartype = c("ci"))) %>%
  filter(disltyr == 1) %>%
  mutate(type = "SRS")

prop_srs %>% kable()
```

disltyr	fmd_prop	fmd_prop_low	fmd_prop_upp	type
1	0.5918367	0.5096925	0.669152	SRS

Next lets adjust for the stratification effect. In this case we actually designed it to produce a asmple of herds proportional to the number in each strata.

The design object looks like this…

Adjustment for stratified sampling

```
survey_design_strata <- dat %>%
  as_survey_design(ids = 1,
                   weights = NULL,
                   strata = Div
                  )
```

And the estimate and confidence interval are …

Stratified

```
prop_strata <- survey_design_strata %>%
  group_by(disltyr) %>%
  summarize(fmd_prop = survey_prop(vartype = c("ci"))) %>%
  filter(disltyr == 1) %>%
  mutate(type = "STR")

prop_strata %>% kable()
```

disltyr	fmd_prop	fmd_prop_low	fmd_prop_upp	type
1	0.5918367	0.5128897	0.6663131	STR

Then if we ignore stratification but include clustering…

Adjustment for cluster sampling

```
survey_design_cluster <- dat %>%
  as_survey_design(ids = Ccode,
                   weights = NULL,
                   strata = NULL
                  )
```

This gives the new estimates below ## Clustered

```
prop_clus <- survey_design_cluster %>%
  group_by(disltyr) %>%
  summarize(fmd_prop = survey_prop(vartype = c("ci"))) %>%
  filter(disltyr == 1) %>%
  mutate(type = "CLUS")

prop_clus %>% kable()
```

disltyr	fmd_prop	fmd_prop_low	fmd_prop_upp	type
1	0.5918367	0.4877149	0.6883201	CLUS

Now we add in both stratification and clustering…

Adjustement for stratified cluster sampling

```
survey_design_strata_cluster <- dat %>%
  as_survey_design(ids = Ccode,
                   weights = NULL,
                   strata = Div)
```

This produces the following estimates… ## Stratified & Clustered

```
prop_str_clus <- survey_design_strata_cluster %>%
  group_by(disltyr) %>%
  summarize(fmd_prop = survey_prop(vartype = c("ci"))) %>%
  filter(disltyr == 1) %>%
  mutate(type = "STR_CLUS")

prop_str_clus %>% kable()
```

disltyr	fmd_prop	fmd_prop_low	fmd_prop_upp	type
1	0.5918367	0.4977081	0.6796785	STR_CLUS

Finally for this example we put it all together with a weighing...

Adjustment for stratified cluster and weighted sampling

```
survey_design_strata_cluster_weight <- dat %>%
  as_survey_design(ids = Ccode,
                   weights = weight.h,
                   strata = Div)
```

Stratified, Clustered & Weighted (The correct one!)

```
prop_str_clus_weight <- survey_design_strata_cluster_weight %>%
  group_by(disltyr) %>%
  summarize(fmd_prop = survey_prop(vartype = c("ci"))) %>%
  filter(disltyr == 1) %>%
  mutate(type = "STR_CLUS_W")

prop_str_clus_weight %>% kable()
```

disltyr	fmd_prop	fmd_prop_low	fmd_prop_upp	type
1	0.622729	0.4944018	0.735885	STR_CLUS_W

Combine results into single dataframe

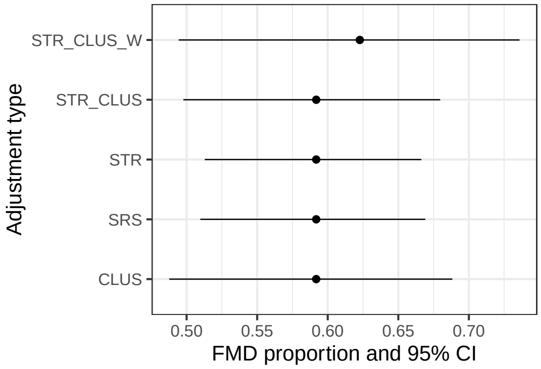
```
df_all <- bind_rows(
  prop_srs,
  prop_strata,
  prop_clus,
  prop_str_clus,
  prop_str_clus_weight
)

df_all %>% kable(digits = 3)
```

disltyr	fmd_prop	fmd_prop_low	fmd_prop_upp	type
1	0.592	0.510	0.669	SRS

disltyr	fmd_prop	fmd_prop_low	fmd_prop_upp	type
1	0.592	0.513	0.666	STR
1	0.592	0.488	0.688	CLUS
1	0.592	0.498	0.680	STR_CLUS
1	0.623	0.494	0.736	STR_CLUS_W

Plot it!

```r
ggplot(data=df_all,
       aes(x=type, y=fmd_prop, ymin=fmd_prop_low, ymax=fmd_prop_upp)) +
        geom_pointrange() +
  coord_flip() +
  labs(x = "Adjustment type", y = "FMD proportion and 95% CI") +
  theme_bw(base_size = 18)
```



So when we put all the diferent estimates together what do we see?

Firstly all those that did not adjust for sample weighting are the same point estimate but slightly power estimate than with the weighting. Then looking at the CIs we can see the narrowest CI are for the SRS. This is what you would expect and highlights the risk of drawing significant conclusions as ignoring the design gives a over confidence effectively. It treats every observation as independent so you end up thinking you have more information than you do in reality. Stratification helps improve the precision while accountng for the clustering accounts for the lack of independence and so the effective sample size is smaller than the number of obersations and so the CIs are wider. Combining stratification and clustering improves on clustering alone but adding in the weighting moves the point estimate to hopefully a more accurate estimate but there is increased uncertainty and the CI get wider.

Exercises on complex design

1) Compare the estimate and CIs for the proportion of herdsmen reporting went on transhumance (trans1yr) assuming we have a SRS compared to the actual design which was a weighted stratified cluster design?

2) Set up a design object just with weighting and stratification and compare to the fully adjusted result from Ex1.

3) Plot the 3 sets of results as we did for FMD above.