Statistics and R short course

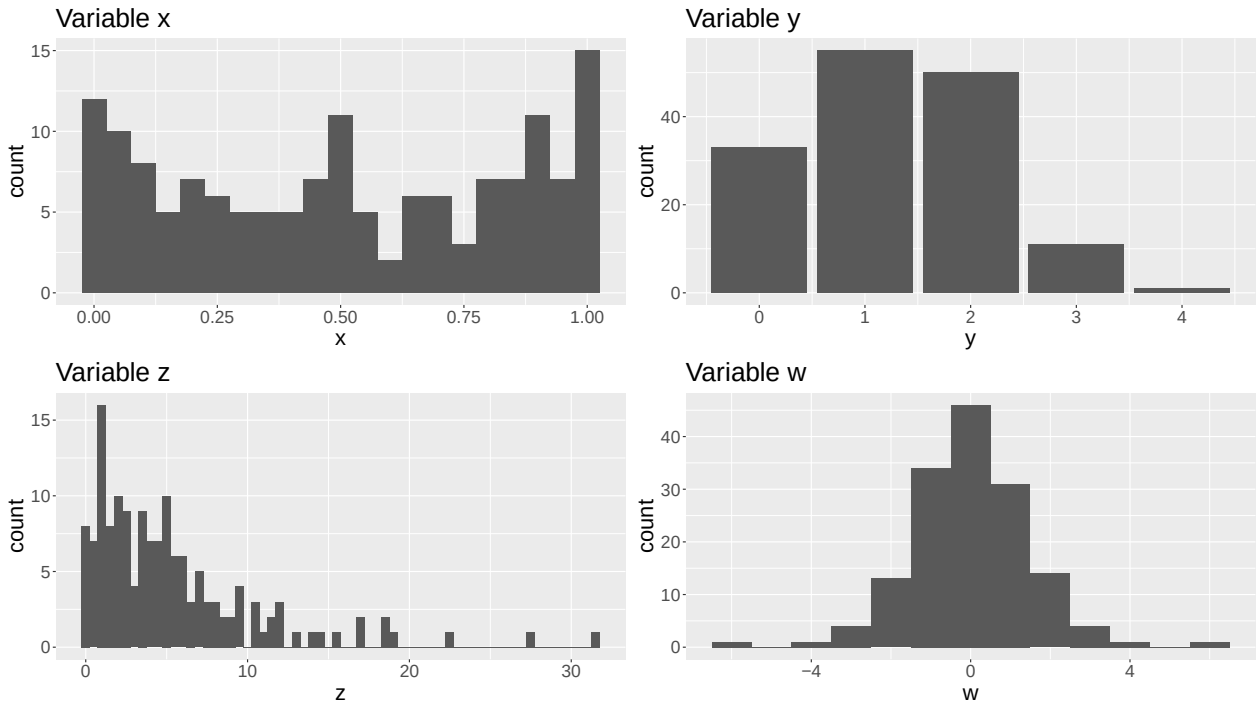
Augustine Choko, Marc Henrion, James Chirombo

01 December 2020

Session 3 - Practical (Solutions)

Exercise 1 - CLT

Which distributions do you think gave rise to each of the variables displayed below?



Exercise 1 (Solution)

This is how the data were generated:

```
set.seed(20201201)

dat<-data.frame(
  x=rbeta(150,shape1=0.5,shape2=0.5),
  y=rbinom(150,size=4,prob=0.3),
  z=rexp(150,rate=1/5),
  w=rt(150,df=4)
)
```

So $X \sim$ Beta$(0.5, 0.5)$, $Y \sim$ Binom(n=4,p=0.3), $Z \sim$ Exp(0.2), $W \sim$ t_4.

Exercise 2 - Central Limit Theorem

In the lecture we saw how to empirically prove the CLT for the normal, beta and negative binomial distributions. Do the same now for the exponential and the Poisson distributions.

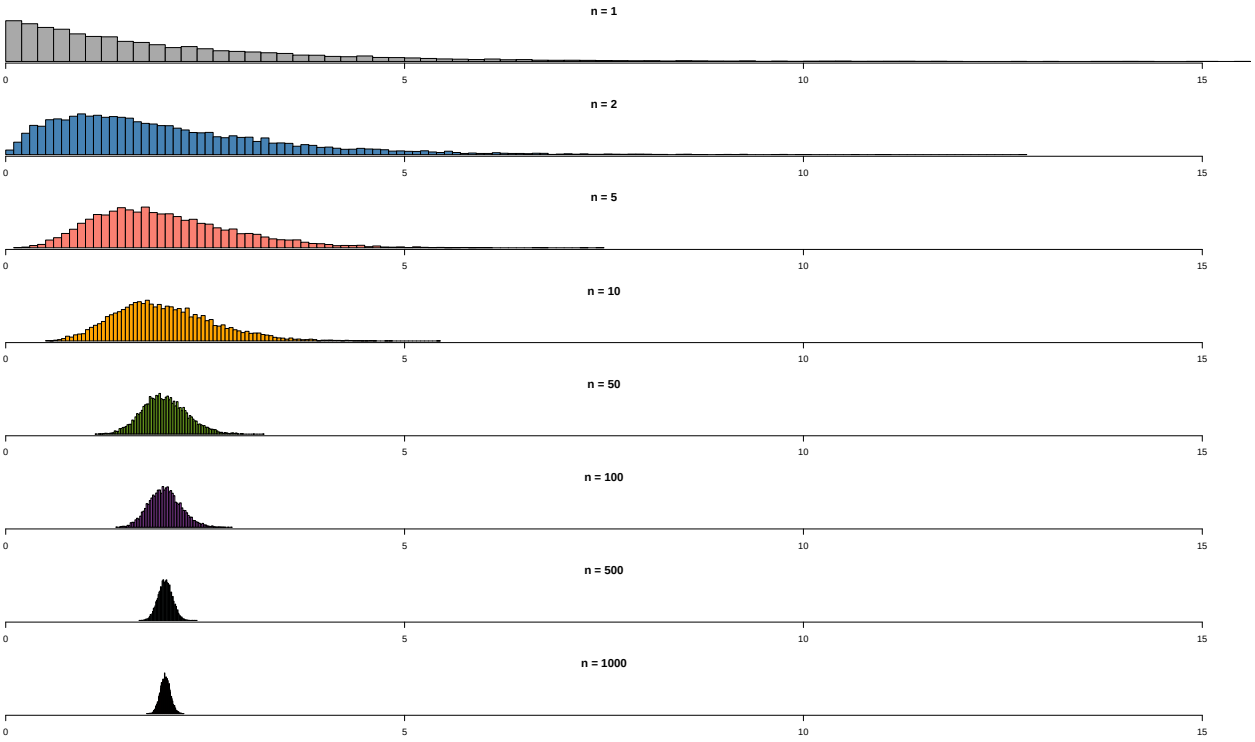
Exercise 2 (Solution)

We will use the same functiom from the lecture:

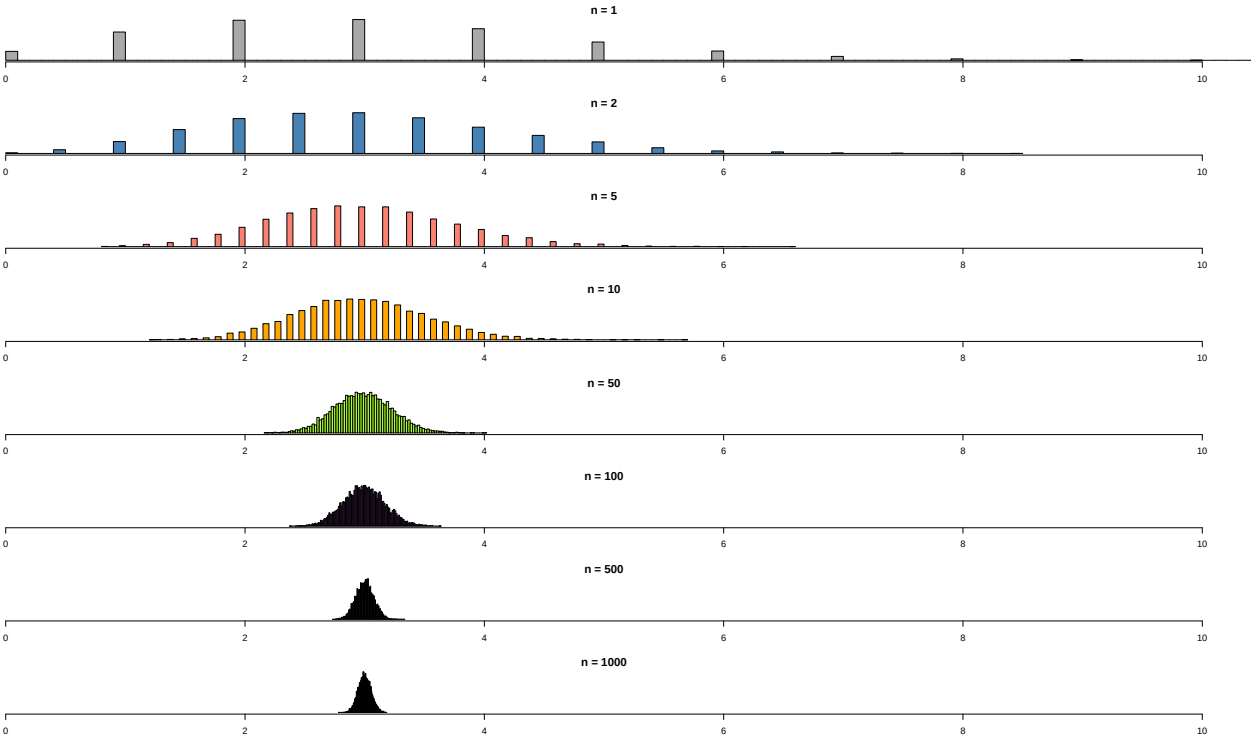
```
cols<-c("darkgrey","steelblue","salmon","orange","greenyellow","mediumorchid","lightcyan","brown")

generateMeans<-function(n=c(1,2,5,10,50,100,500,1000),col=cols,N=1e4,rDistFun,xlim=NULL,...){
  sim<-matrix(nrow=N,ncol=length(n))
  for(j in 1:length(n)){for(i in 1:N){
    sim[i,j]<-mean(rDistFun(n[j],...))
  }}

  par(mar=c(3,2.5,1.5,0.5),mfrow=c(length(n),1))
  for(j in 1:length(n)){
    if(length(xlim)<2){
      hist(breaks=100,sim[,j],main=paste(sep="","n = ",n[j]),yaxt="n",col=col[j])
    }else{
      hist(breaks=100,sim[,j],main=paste(sep="","n = ",n[j]),xlim=xlim,yaxt="n",col=col[j])
    }
  }
}
```

And then just input `rexp` and `rpois` for the argument `rDistFun`.

```
generateMeans(rDistFun=rexp,rate=0.5,xlim=c(0,15))
```




```
generateMeans(rDistFun=rpois,lambda=3,xlim=c(0,10))
```



Exercise 3 - study design

Decide what design could be used to answer the following research questions:

1. What is the prevalence of HIV in urban Blantyre in 2018?
2. Do men experience higher mortality compared to women once they start ART?
3. Does smoking increase the chance of having lung cancer?
4. What is the effect of providing oral HIV self-test kits on the uptake of HIV testing?
5. What interventions may improve linkage to ART following community based HIV testing?

Exercise 3 (Solution)

The answers below may not be the only valid answers - there may be several alternative designs for a given question.

1. If today was 2018, then taking a cross-sectional, random sample from the Blantyre population in 2018 will allow you to answer the question. Given that 2018 is in the past now, a retrospective design will need to be used.

2. A longitudinal design where a cohort of equal numbers of men and women, recruited at ART initiation, are followed over time will be appropriate for this questions.

3. You could again recruit a cohort for a longitudinal study. However you will need a big budget and a lot of time: lung cancer is rare and to develop lung cancer takes years. So here a case-control study may be more efficient: recruit lung cancer patients from a hospital, then recruit matched (by age, sex and other known factors to impact the risk of lung cancer) controls. Then by comparing smoking habits between controls and cases, you may be able to answer the research question (somewhat - the causality implied by the question will be tricky to resolve).

4. The appropriate design depends on the practical circumstances. If there is a government programme distributing self-test kits, then a pragmatic before-after study design will need to be used. However if no such programmes exist, then an intervention study, specifically a randomised controlled trial, where

participants (or more likely health centres where these kits would be distributed, making this a cluster design) are randomised to either receiving HIV sel-test kits or not, will be an appropriate design.

5. The question implies that there are a number of potential interventions and the idea is to both identify effective interventions and evaluate their effect. This suggests an adaptive interentional design, such as a multi-arm multi-stage design, could be useful.

Exercise 4 - sample size calculation

The difference of the mean birth weights between Babies born at Queen Elizabeth Central hospital (QECH) and Kamuzu Central Hospital (KCH) will be determined. Suppose at QECH the mean is 3000 g with a standard deviation of 500 g (from previous/pilot studies). At KCH the mean is 3300 with a s.d of 500 grams. The difference in mean birth weight to be detected is therefore 300 g. The significance level for the test will be 0.05. What would be the sample size required to detect this difference at 80% power?

Exercise 4 (Solution)

```
library(pwr)

sampsize<-pwr.t.test(d=(3300-3000)/500,sig.level=0.05,power=0.8)
print(sampsize)
```

```
##
##      Two-sample t test power calculation
##
##               n = 44.58577
##               d = 0.6
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
n<-ceiling(sampsize$n)
print(n)
```

```
## [1] 45
```

Per group 45 participants will be needed, i.e. 90 participants in total.