

MLW / KUHeS Statistics and R short course

Session 3 - Practical (solutions)

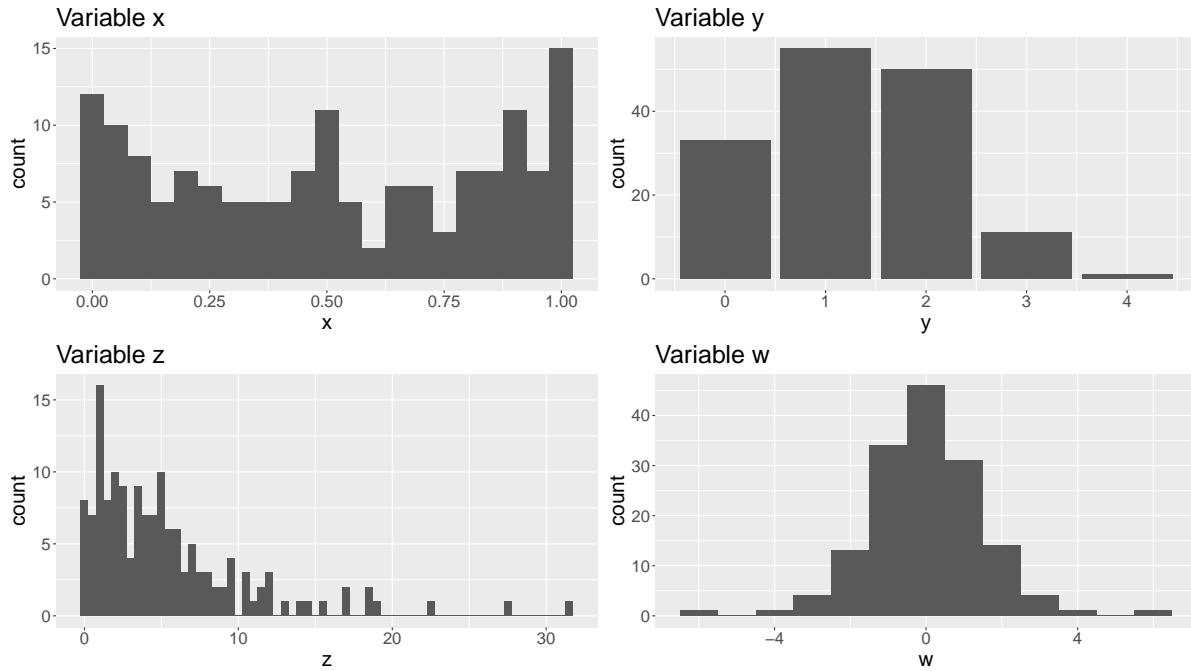
Marc Henrion, Evaristar Kudowa

2026-01-26

Session 3 - Practical

Exercise 1 - Distributions

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



Solution

The distributions used to generate each graph were:

1. $X \sim \beta(1/2, 1/2)$ - beta distribution
2. $Y \sim \text{Bin}(4, 0.3)$ - binomial distribution
3. $Z \sim \text{Exp}(1/5)$ - exponential distribution (obviously gamma, of which the exponential is a special case, would have been an equally valid answer)
4. $W \sim t(4)$ - Student's t distribution (you could also have tried with a Normal distribution)

Exercise 2 - Distributions

1. Generate 20 random values from a Gamma(4,6) distribution.
2. You are recruiting participants to a cohort study of 200 patients. The chance that a given participant drops out of the study before completing the last follow-up visit is 10%. How likely is it that you will end up with complete data for fewer than 180 participants?
3. You have found that the Poisson distribution, with rate $\lambda = 1.12$, is a good fit for the number of cardiac events recorded over a month by a given patient in your ward. What is the probability that a given participant experiences 0 cardiac events?
4. For a normal distribution with mean $\mu = 5$ and standard deviation $\sigma = 1$, what is the probability of observing a value of 8 or larger?

Solution

1. `rgamma(n=20, shape=4, rate=6)`
2. We need to compute, using a binomial distribution, $P(X < 180|n = 200, p = 1 - 0.1) = P(X \leq 179|n = 200, p = 0.9)$. In R: `pbinom(179, size=200, prob=1-0.1)` or `sum(dbinom(1:179, size=200, prob=1-0.1))`.

```
pbinom(179, size=200, prob=1-0.1)
```

[1] 0.4408252

3. We need to compute, using a Poisson distribution, $P(X = 0|\lambda = 1.12)$. In R: `dpois(x=0, lambda=1.12)` which, in this particular case (since the Poisson distribution cannot take values less than 0) is equivalent to `ppois(q=0, lambda=1.12)`.

```
dpois(x=0, lambda=1.12)
```

[1] 0.3262798

4. We need to compute, using a normal distribution, $P(X \geq 8 | \mu = 5, \sigma = 1) = 1 - P(X < 8 | \mu = 5, \sigma = 1)$. In R: `pnorm(8, mean=5, sd=1)`.

```
pnorm(8,mean=5,sd=1)
```

```
[1] 0.9986501
```

Exercise 3 - Central Limit Theorem

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

Solution

We can use the same function we used in the lecture.

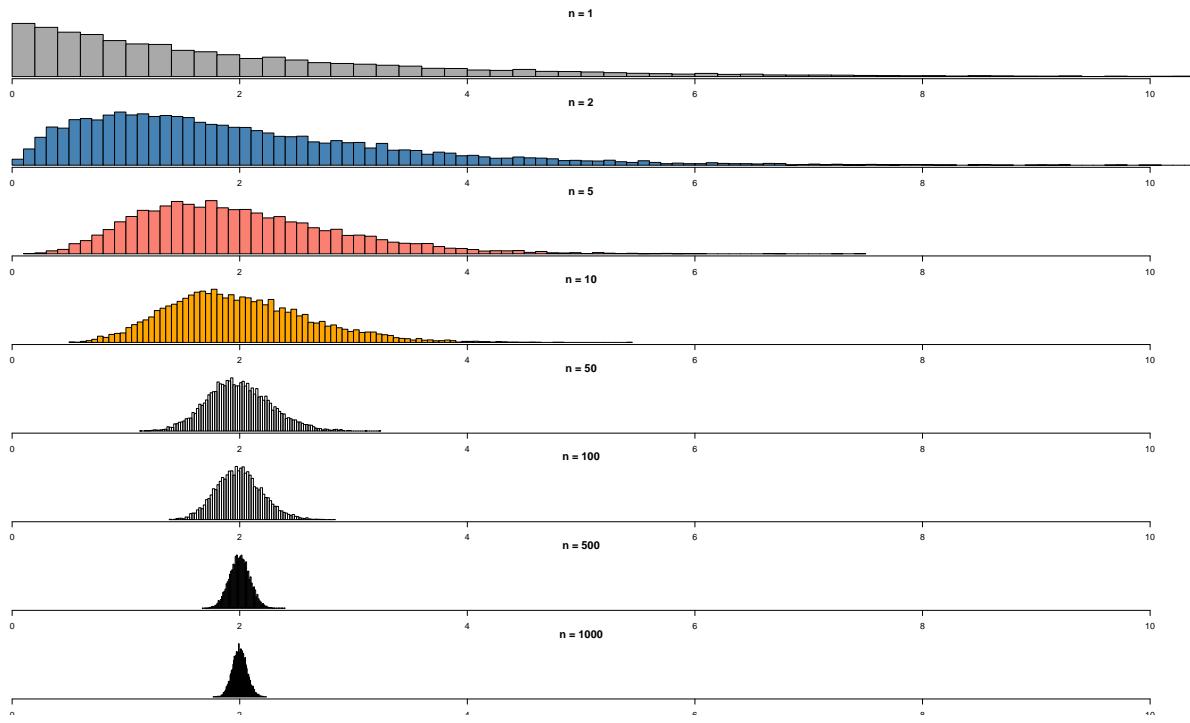
```
cols<-c("darkgrey", "steelblue", "salmon", "orange")

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(1.5,2.5,1.5,0.5),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=cols[j])
    }else{
      hist(breaks=100,sim[,j],main=paste(sep="", "n = ",n[j]),xlim=xlim,yaxt="n",col=cols[j])
    }
  }
}
```

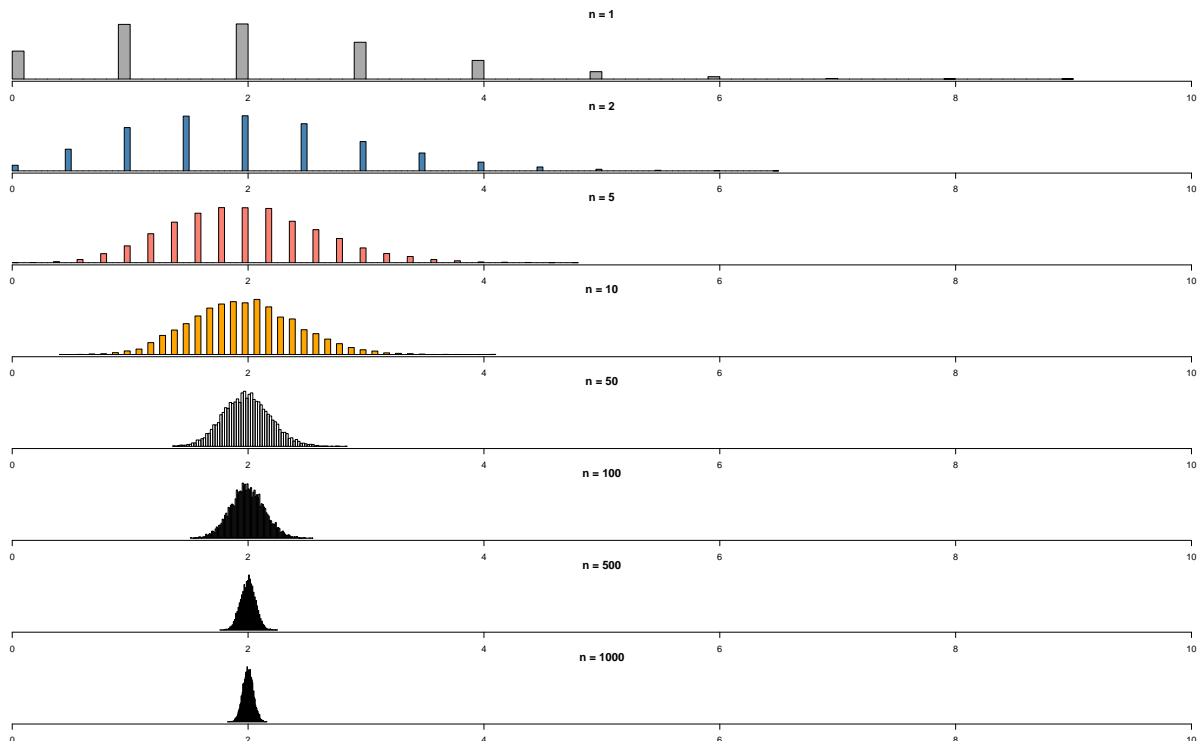
And then we just call it for an exponential distribution (say with $\lambda = 1/2$):

```
generateMeans(rDistFun=rexp,rate=1/2,xlim=c(0,10))
```



And for a Poisson distribution (say with $\lambda = 2$):

```
generateMeans(rDistFun=rpois,lambda=2,xlim=c(0,10))
```



Exercise 4 - 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?

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 question.

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) or unmatched 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 self-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 interventional design, such as a multi-arm multi-stage design, could be useful.