Problem Set 3

C. Durso

February 11, 2019

These questions were rendered in R markdown through RStudio (https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf, http://rmarkdown.rstudio.com).

Please complete the following tasks regarding the data in R. Please generate a solution document in R markdown and upload the .Rmd document and a rendered .doc, .docx, or .pdf document. Your work should be based on the data's being in the same folder as the .Rmd file. Please turn in your work on Canvas. Your solution document should have your answers to the questions and should display the requested plots.

1. This question revisits the problem of whether two null hypotheses that we applied to the polio trials are equivalent. Though they both resulted in very low probabilities of the data under the respective null hypotheses, you will show that they aren't exactly the same.

Consider two non-empty multisets $C = \{c_1, c_2, ...c_n\}$ and $D = \{d_1, d_2, ...d_m\}$ of binary outcomes, say "success" and "failure", with k successes in C and l successes in D.

Define a random variable X as the number of successes in a sample of size n drawn without replacement from $\{c_1, c_2, ...c_n, d_1, d_2, ...d_m\}$ in such a way that all samples are equally likely. Recall that the number of samples is $\begin{pmatrix} n+m \\ n \end{pmatrix}$.

Define another random variable Y to be $Binomial\left(k+l, \frac{n}{n+m}\right)$.

(These are abstract versions of the probability models for a couple of the null hypotheses considered in inference_example_polio.Rmd if you take C to be the Vaccinated population, D to be the Placebo population, k to be the number of polio cases in the Vaccinated population, and l to be the number of polio cases in the Placebo population.

Then X corresponds to the null model that the number of polio cases in the two populations is consistent with pooling the two populations, randomly selecting n cases to be designated as Vaccinated and counting the number of polio cases in this sample.

The random variable Y corresponds to the null hypothesis that the $k+l$ polio cases were assigned to the Vaccinated group with probability $\frac{n}{m+n}$ and to the Placebo group with probability $1 - \frac{n}{m+n} = \frac{m}{m+n}$.)

Please calculate the probability that $X = 0$ and the probability that $Y = 0$. Use these to demonstrate that distribution of X is not equal to the distribution of Y when k, l, m, and n are all non-zero. You may use the fact that $0 < r \leq p < q$ implies $\frac{p}{q} > \frac{p-r}{q-r}$. Suggestion: simplify the factorials in the probability that $X = 0$ in such a way that you have $k+l$ terms in the numerator and denominator. (10 points)

The probability for the random variable X : $P(X = a) = \binom{k+l}{a} \frac{\binom{n+m-(k+l)}{n}}{\binom{n+m}{n}}$

The probability for the random variable Y: $P(Y = a) = \binom{k+l}{a}(1 - \frac{n}{n+m})^{k+l} = \binom{k+l}{a}(\frac{m}{n+m})^{k+l}$

The probability that we have X=0 and Y=0 is : $P(Y = 0) = (\frac{m}{n+m})^{k+l}$

$P(X = 0) = \frac{(n+m-(k+l))!}{(m-(k+l))!n!} \frac{n!m!}{(n+m)!} = \frac{m(m-1)...(m-(k+l)+1)}{(n+m)(n+m-1)...(n+m-(k+l)+1)}$

If we group together the terms on the top and bottom of $P(X = 0)$ into $\frac{m}{n+m} \frac{m-1...}{n+m-1...} \frac{m-(k+l)+1}{n+m-(k+l)+1}$ we notice that the first term is $\frac{n}{n+m}$ is the same as the expression of $P(Y = 0)$. However, the following terms of $P(X = 0)$ are smaller than $\frac{n}{n+m}$, i.e. $\frac{n-1}{n+m-1}$ and so on. Therefore we can conclude that $P(Y)$ is bigger and

from the setup of the problem of $P(Y)$ being the probability of contracting polio in the placebo group we can conclude that the two groups do not have equal probability of contracting polio.

Another possible route to have solved this problem would have been to plug in small values (i.e. $m = 20, n = 22, k = 3, l = 4$) into the expressions : $P(Y = 0) = (\frac{m}{n+m})^{k+l}$ and $P(X = 0) = \frac{(n+m-(k+l))!}{(m-(k+l))!n!} \frac{n!m!}{(n+m)!}$ and solved numerically. This would have given you two disctinct values showing that the probabilities were different.

Note : The reason we do not plug in the values of the populations (~200k people for each) is because representing such large numbers and their division would be difficult to represent numerically in a computer due to an inherent limit in numerical representation of very large or very small numbers in computers.

2. The parts of this problem lead to a proof that that if X is a Poisson random variable with $\lambda = \lambda_x$ and Y is a Poisson random variable with $\lambda = \lambda_y$ and X and Y are independent, then the random variable $W = X + Y$ is a Poisson random variable with $\lambda = \lambda_x + \lambda_y$. Please complete each part rather that providing a proof of the result directly. Recall that the probability that $X = k$ is $\frac{(\lambda_x)^k \exp(-\lambda_x)}{k!}$, and likewise for Y. (2 points each)

   a. Consider the probability space Q with outcomes in $\mathbb{N} \times \mathbb{N}$ such that the first component has the same distribution as X, the second component has the same distribution Y, and the two components are independent. What is the probability of the outcome $(1, 2)$?

Knowing expression : $P(X = k) = \frac{\lambda_x^k e^{-\lambda_x}}{k!}$ and $P(Y = l) = \frac{\lambda_y^l e^{-\lambda_y}}{l!}$, we plug in the values $(1, 2)$ into $P(X)$ and $P(Y)$ respectively and calculate the probability keeping in mind that these two probabilities are independent, therefore must be multiplied together. We obtain : $P(X = 1)P(Y = 2) = \frac{\lambda_x e^{-\lambda_x}}{1!} \frac{\lambda_y^2 e^{-\lambda_y}}{2!}$. Simplifying this expression further is optional. ($P(X = 1)P(Y = 2) = \frac{\lambda_x \lambda_y^2 e^{-(\lambda_x + \lambda_y)}}{2}$ =)

   b. The probability that $W = n$ is equal to the sum of the probabilities of the outcomes (k, l) in Q for which $k + l = n$: $P(W = n) = \sum_{(k,l) \text{ such that } k+l=n} P(X = k)P(Y = l)$. Please rewrite this using a summation of the form $\sum_{k=0}^{n} \ldots$.

$P(W = n) = \sum_{k=0}^{n} P(X = k)P(Y = n - k)$

   c. Rewrite the summation from 2.b using the Poisson formulas for the probabilities.

$P(W = n) = \sum_{k=0}^{n} \frac{\lambda_x^k e^{-\lambda_x}}{k!} \frac{\lambda_y^{(n-k)} e^{-\lambda_y}}{(n-k)!}$

   d. Recall the binomial formula: $(a + b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$. Multiply your summation from 2.c. by $\frac{n!}{n!}$. Explain how rearranging terms verifies the claim that W is a Poisson random variable with $\lambda = \lambda_x + \lambda_y$.

$P(W = n) = \sum_{k=0}^{n} \frac{\lambda_x^k e^{-\lambda_x}}{k!} \frac{\lambda_y^{(n-k)} e^{-\lambda_y}}{(n-k)!}$

$= e^{-\lambda_x} e^{-\lambda_y} \sum_{k=0}^{n} \frac{\lambda_x^k}{k!} \frac{\lambda_y^{(n-k)}}{(n-k)!} \frac{n!}{n!}$

$= \frac{e^{-\lambda_x} e^{-\lambda_y}}{n!} \sum_{k=0}^{n} \frac{n!}{(n-k)!} \frac{\lambda_x^k \lambda_y^{(n-k)}}{k!}$

$= \frac{e^{-\lambda_x} e^{-\lambda_y}}{n!} \sum_{k=0}^{n} \binom{n}{k} \frac{\lambda_x^k \lambda_y^{(n-k)}}{k!}$

$= \frac{e^{-\lambda_x \lambda_y}}{n!} (\lambda_x + \lambda_y)^n$ which gives us $W \sim Pois(\lambda_x + \lambda_y)$

3. Normal Distributions shift and scale while remaining Normal. This greatly simplifies working with Normal distributions. This property is not generally true of parametrized families of distributions. (5 points each)

   a. Prove that if X is a random variable with the $Normal\left(\mu, \sigma^2\right)$ distribution, then $X + a$ is a random variable with the $Normal\left(\mu + a, \sigma^2\right)$ distribution. Please do this by showing that the cumulative

distribution for $X + a$ defined in terms of the cumulative distribution for X equals the cumulative distribution for a random variable with the $Normal\left(\mu + a, \sigma^2\right)$ distribution.

$P(X + c \leq t) = P(X \leq t + c) = \int_{-\infty}^{t-c} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$

We perform a change of variable of the form : $u = x + c \rightarrow x = u - c, dx = du$. We get:

$= \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(u-(c+\mu))^2}{2\sigma^2}} du$

Which can be interpreted as $P(Y \leq t)$, Y~$Normal(\mu + c, \sigma^2)$

b. Prove that if X is a random variable with the $Normal\left(\mu, \sigma^2\right)$ distribution, then cX is a random variable with the $Normal\left(c\mu, c^2\sigma^2\right)$ distribution. Please do this by showing that the cumulative distribution for cX equals the cumulative distribution for a random variable with the $Normal\left(c\mu, c^2\sigma^2\right)$ distribution.

$P(Xc \leq t) = P(X \leq \frac{t}{c}) = \int_{-\infty}^{\frac{t}{c}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$

We perform a change of variable of the form : $u = xc, dx = \frac{1}{c}du$. We get:

$= \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{c} e^{-\frac{(\frac{u}{c}-\mu))^2}{2\sigma^2}} du$

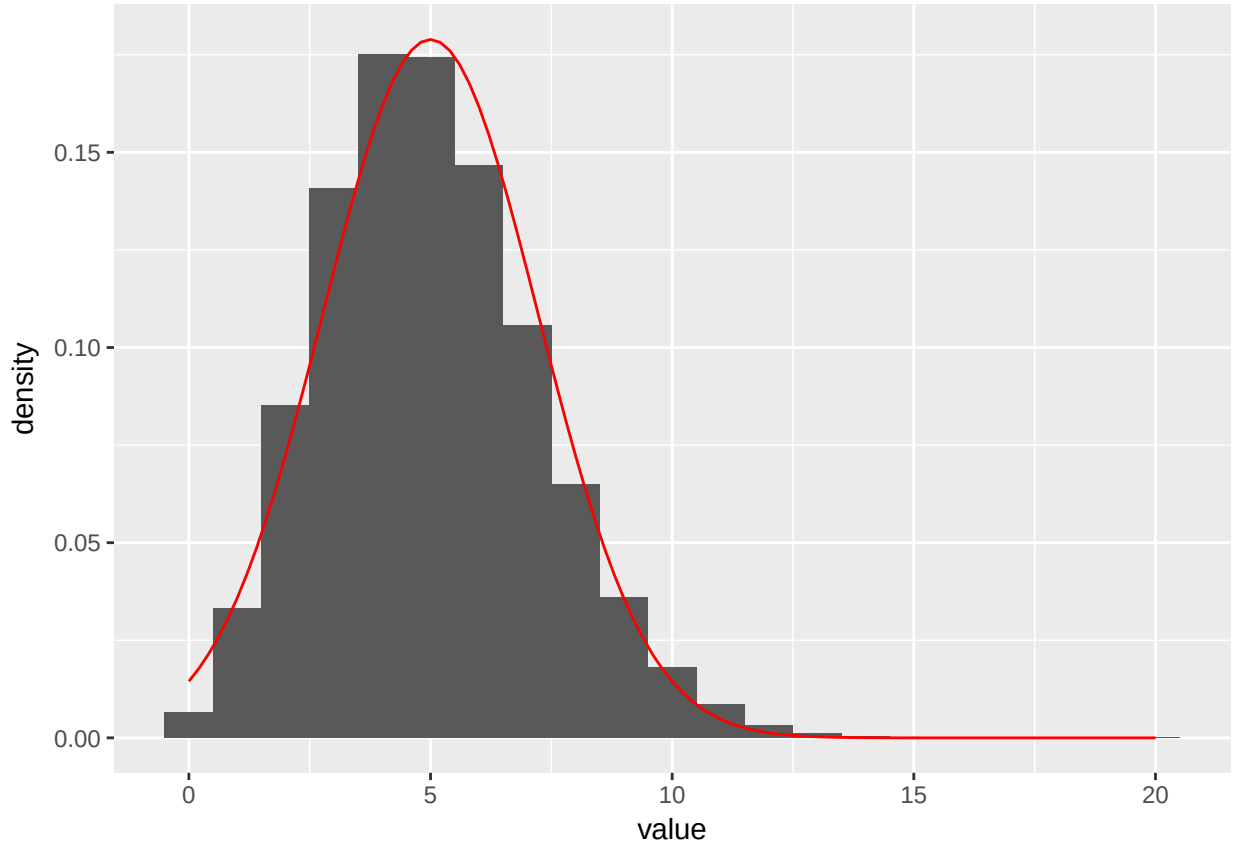
$= \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi(c\sigma)^2}} e^{-\frac{(u-c\mu)^2}{2(c\sigma)^2}} du$

Which can be interpreted as $P(Y \leq t)$, Y~$Normal(\mu + c, \sigma^2)$

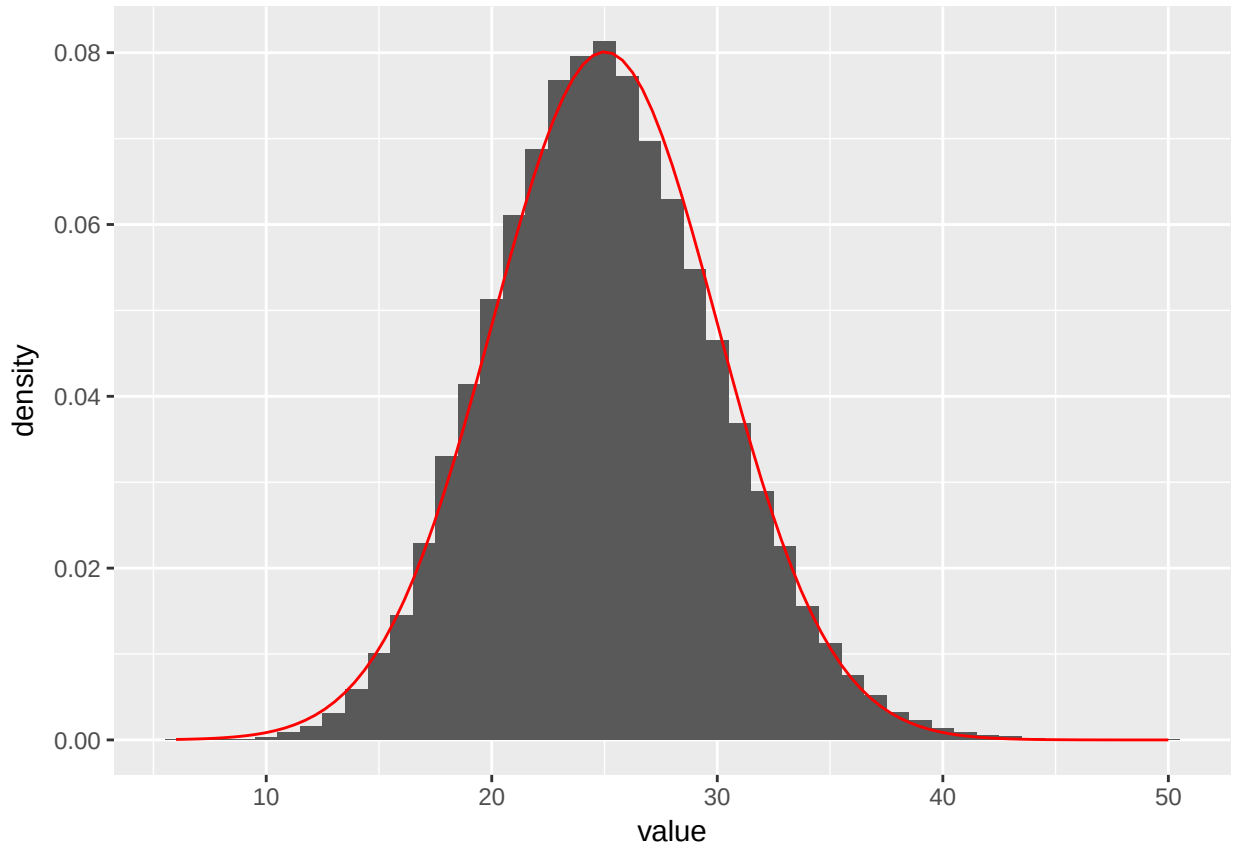
4. The code below draws samples of size 100,000 from the Poisson distributions with $\lambda = 5, 25, 50, 100$ and puts them in a matrix. For each sample, calculate the maximum likelihood values of μ and σ^2 to approximate the sample as a sample from a Normal distribution. For each λ, plot the density histogram of the sample and the density curve of the Normal distribution with the maximum likelihood values of μ and σ^2 for the sample on the same graph. (7 points)

```
set.seed(34567)
mat<-matrix(c(rpois(100000,5),rpois(100000,25),rpois(100000,50),rpois(100000,100)),ncol=4)

#We apply the formula derived in class:
mu<-apply(mat,2,mean)
sigmasq.ml<-function(x){
  return (var(x)*(length(x)-1)/length(x))
  }
sigmasq<-apply(mat,2,sigmasq.ml)

# Plot of the density histogram of the poisson and normal for lambda=5
dat.5<-data.frame(value=mat[,1])
g.5<-ggplot(dat.5, aes(x=value)) + geom_histogram(aes(y =..density..), binwidth =1) + stat_function(fun=

# Plot of the density histogram of the poisson and normal for lambda=25
dat.25<-data.frame(value=mat[,2])
g.25<-ggplot(dat.25, aes(x=value)) + geom_histogram(aes(y =..density..), binwidth =1) + stat_function(f

# Plot of the density histogram of the poisson and normal for lambda=50
dat.50<-data.frame(value=mat[,3])
g.50<-ggplot(dat.50, aes(x=value)) + geom_histogram(aes(y =..density..), binwidth =1) + stat_function(f

# Plot of the density histogram of the poisson and normal for lambda=50
dat.100<-data.frame(value=mat[,4])
g.100<-ggplot(dat.100, aes(x=value)) + geom_histogram(aes(y =..density..), binwidth =1) + stat_function
```

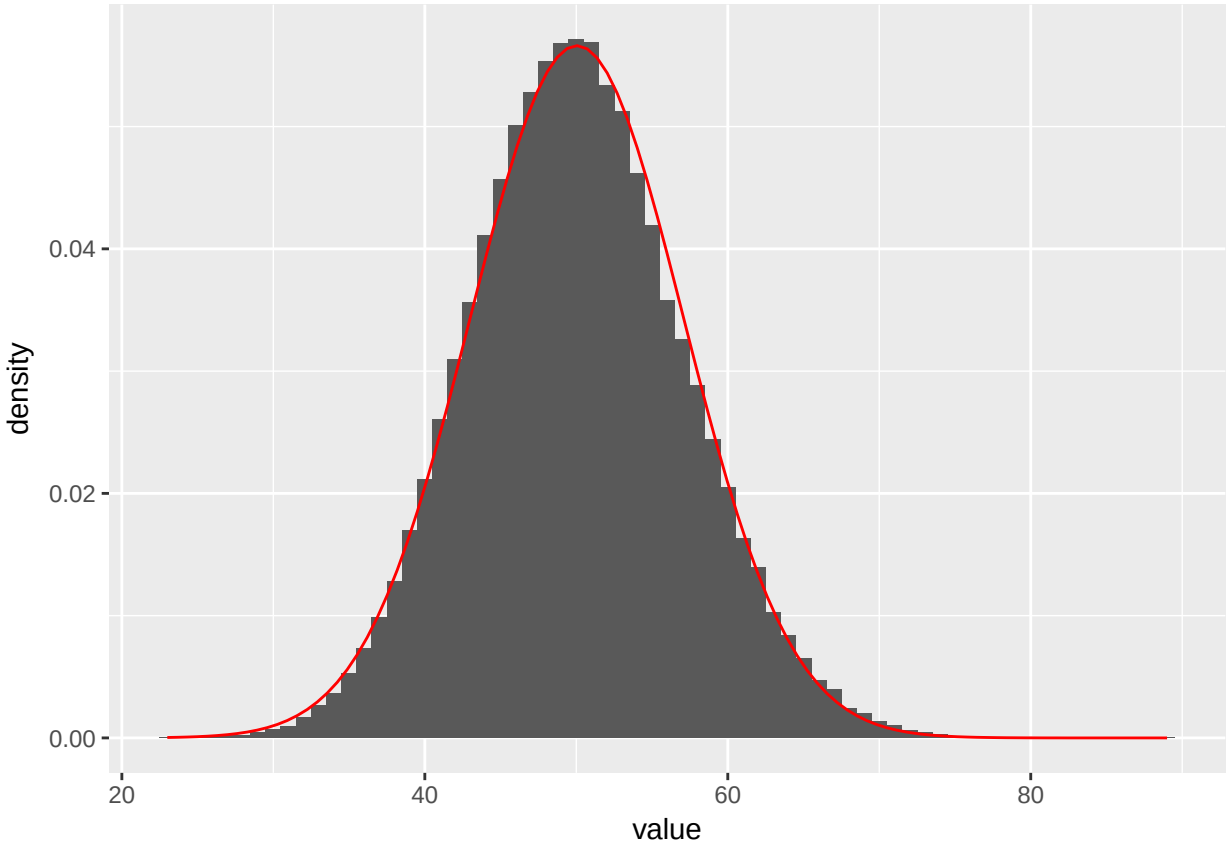
```
g.5
```



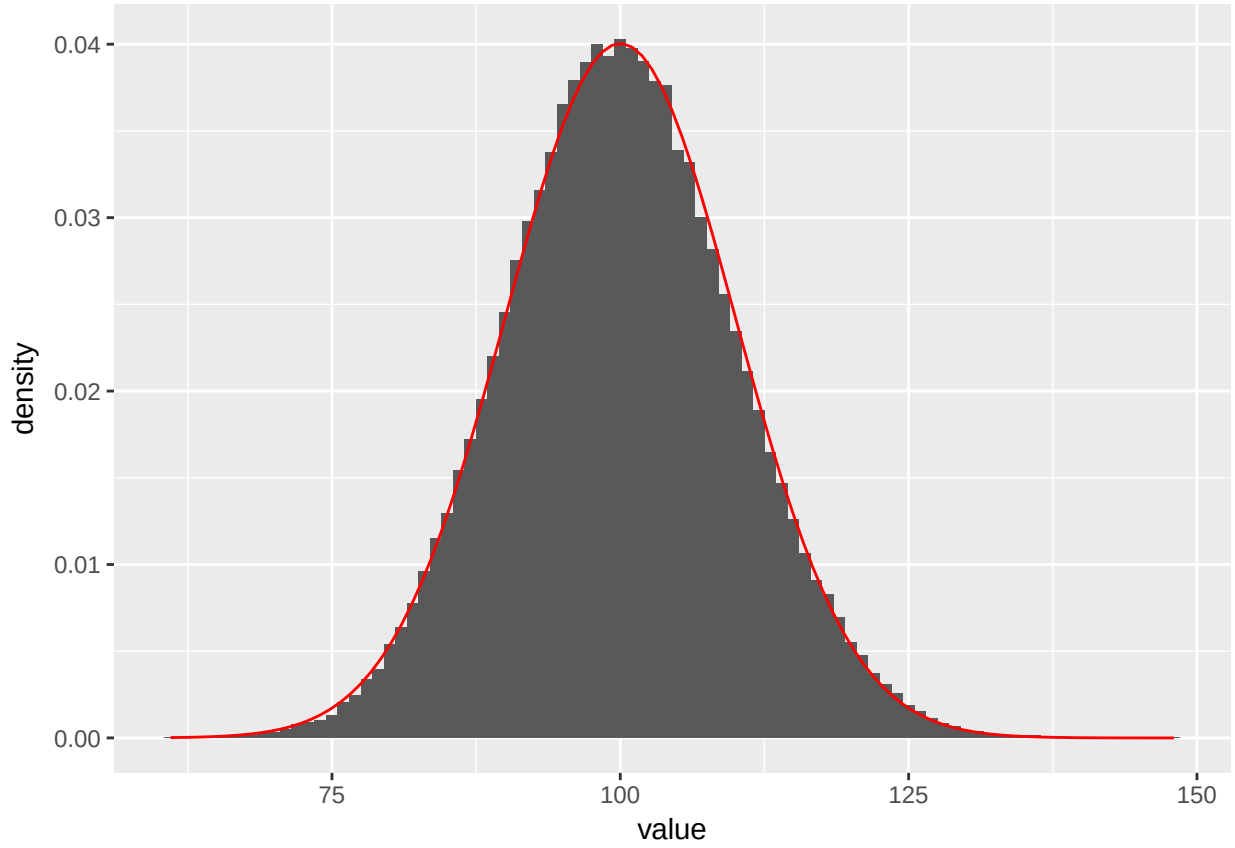

```
g.25
```




```
g.50
```




```
g.100
```



5. Consider approximating the probability $P(X = k)$ for a Poisson random variable X with parameter λ by the probability that a Normal random variable with parameters $\mu = \sigma^2 = \lambda$ lies in $(k - 0.5, k + 0.5)$. These should be close to the values you found by simulation. (Remember the parameter "sd" is σ, not σ^2.) Using R, calculate the sum of the absolute values of the differences between the probabilities for $k = 0, 1, ...1000$ for the values $\lambda = 5, 25, 50, 100$. Please base the Poisson probabilities on the probability density function rather than on simulation.

This is a measure of how well the Normal distribution approximates the Poisson distribution: it's the sum over all reasonably probable ks of the absolute value of errors from using the approximation.

What do you observe about this difference as λ increases? Note this is a computation based on the theoretical probabilities, not on simulations. (8 points)

```
ks<-0:1000
pois.vs.norm<-function(x){
  prob.pois<-dpois(ks, x)
  prob.norm<-pnorm(ks+.5, x, sqrt(x))-pnorm(ks-.5, x, sqrt(x))
  return(sum(abs(prob.pois-prob.norm)))
}

#Calculate the difference with different lambda's
pois.vs.norm(5)
```

```
## [1] 0.1071244
```

```
pois.vs.norm(25)
```

```
## [1] 0.05051787
```

```
pois.vs.norm(50)
```

```
## [1] 0.03562743
```

```
pois.vs.norm(100)
```

```
## [1] 0.02518573
```

We notice that the sum of the absolute values of the differences between the probabilities decreases as the value of λ increases.

6. Repeat exercise 5 using the Binomial distributions with probability of success=0.7 and the size parameter equal to $5, 25, 50, 100$. This time take $\mu = 0.7n$ and $\sigma^2 = n(0.7)(1 - 0.7)$. These are close to the values you would find by simulation. For each n, you can limit the summation to $k = 0, 1, ...n$.(4 points)

```
prob.error<-function(n,p=.7){
  prob.binom<-dbinom(0:n, n, p)
  prob.norm<-pnorm(0:n+.5,n*p,sqrt(n*p*(1-p)))-
    pnorm(0:n-.5,n*p,sqrt(n*p*(1-p)))
  return(sum(abs(prob.binom-prob.norm)))
}

prob.error(5)
```

```
## [1] 0.09256642
```

```
prob.error(25)
```

```
## [1] 0.04546888
```

```
prob.error(50)
```

```
## [1] 0.03135613
```

```
prob.error(100)
```

```
## [1] 0.02192309
```

7. View the results of question 5 as giving measures of the quality of a Normal approximation for sums of 5,25,50,and 100 independent Poisson random variables with $\lambda = 1$ and the results of question 6 as giving measures of the quality of a Normal approximation for sums of independent random variables that equal 1 with probability 0.7 and that equal 0 with probability 0.3. What do you observe about the quality of the approximation as the number of terms in the sum increases? (3 points)

We notice that the sum of the absolute values of the differences between the probabilities gets smaller and smaller as the number of terms increase.