Quiz Week 4 - distributions - Solutions

FIT5197 teaching team

Question 1

Given the following sample of weights $w \in \{66, 54, 69, 77, 55, 97, 79, 52, 56, 60, 64, 102\}$ and assuming the random variable defining the weights is normally distributed, i.e. $W \sim N(\mu, \sigma^2)$

(a). estimate the mean μ using the sample mean.

(b). estimate the variance σ^2 using the sample variance.

(c). using the estimate of the mean and variance from (a) and (b) to determine $P(36 < w < 100)$

Answer 1

(a). The estimate of the mean is

$$\hat{\mu} = \frac{\sum_{i=1}^{n} W_i}{n} = \frac{\sum_{i=1}^{12} W_i}{12} = 69.25.$$

(b). The estimate of the variance is

$$\begin{aligned}\hat{\sigma^2} &= \frac{1}{n-1}\sum_{i=1}^{n}(W_i - \overline{W})^2 \\ &= \frac{1}{12-1}\sum_{i=1}^{n}(W_i - 69.25)^2 \\ &= \frac{1}{11}[(66-69.25)^2 + (54-69.25)^2 + \ldots] \\ &= 275.48\end{aligned}$$

Calculating this manually is slow, so you can do it this way instead if you want. Remember how $V[W] = E[W^2] - E[W]^2$. The sample variance analogue is

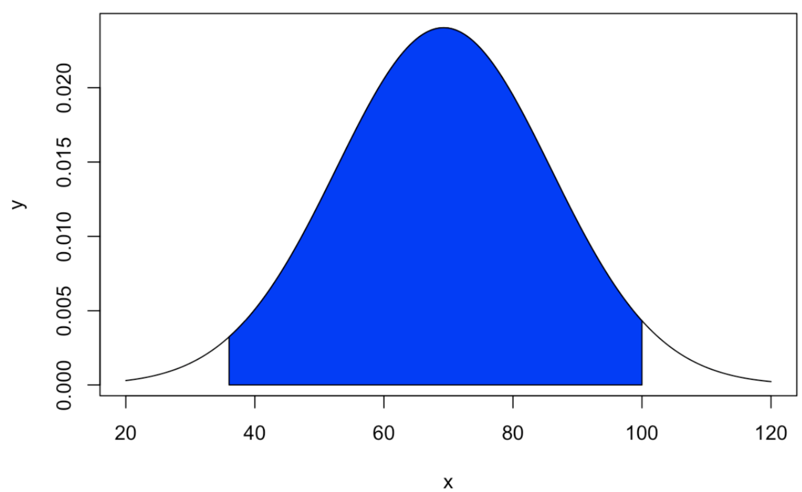
$$\hat{\sigma}^2 = \frac{n}{n-1}(\overline{W^2} - \overline{W}^2)$$
$$= \frac{n}{n-1}(\frac{1}{n}\sum_{i=1}^{n} W_i^2 - (\frac{1}{n}\sum_{i=1}^{n} W_i)^2).$$

(Exercise - Derive this version from the sample variance definition given above.)

Using the 2nd version to compute the estimate of sample variance,

$$\hat{\sigma}^2 = \frac{n}{n-1}(\overline{W^2} - \overline{W}^2)$$
$$= \frac{n}{n-1}(\frac{1}{n}\sum_{i=1}^{n} W_i^2 - (\frac{1}{n}\sum_{i=1}^{n} W_i)^2)$$
$$= \frac{12}{11}(5048.08 - (69.25)^2)$$
$$= 275.48$$

(c). Given the variables are normally distributed, we have $W \sim N(69.25, 275.48)$ since our best guess for the parameters μ, σ^2 are $\hat{\mu}$ and $\hat{\sigma}^2$. We can visualise the distribution below,

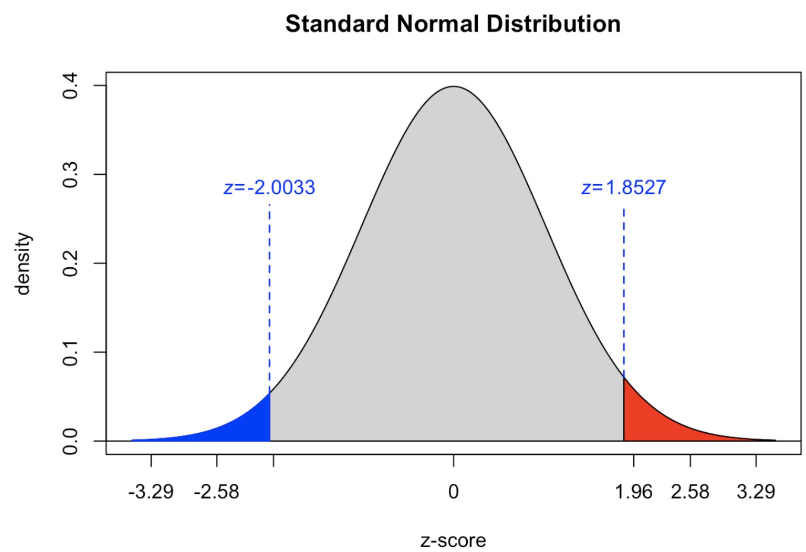


Now $P(36 < w < 100)$ is just the area coloured in blue (note x in the figure above corresponds to w). It is impossible to integrate the Gaussian pdf analytically so it needs to be done numerically using Z-table. We can note that $P(36 < w < 100) = P(w < 100) - P(w < 36)$ by considering the area under the curve, meanwhile $P(w < 100)$ is the area from negative infinity to 100 and $P(w < 36)$ is the area from negative infinity to 36. The blue area is calculated by subtracting $P(w < 36)$ from $P(w < 100)$.

We can get these areas by transforming our normal pdf to standard normal pdf and then apply Z-table.

For $w = 36$, $z = \frac{36-69.25}{\sqrt{275.48}} = -2.0033$;

For $w = 100$, $z = \frac{100-69.25}{\sqrt{275.48}} = 1.8527$.



So,

$$\begin{aligned} P(36 < w < 100) &= P(w < 100) - P(w < 36) \\ &= P(z < 1.8527) - P(z < -2.0033) \end{aligned}$$

Now we need to use a Z-table to help us find the solution. The Z-table we use in this unit is found in the [formula sheet (https://lms.monash.edu/pluginfile.php/11064620/mod_resource/content/7/FIT5197_2020_S2_Formula_Sheet.pdf)](https://lms.monash.edu/pluginfile.php/11064620/mod_resource/content/7/FIT5197_2020_S2_Formula_Sheet.pdf) on the Moodle Unit Information page . A Z-table gives you the area under the standard normal pdf from $-\infty$ to z (i.e. $p(-\infty < Z < z)$), or in other words the value of the standard normal cdf, $p = F(z)$ for a specific z value. Looking up the Z-table to find z-value, we see the closest z-value to $z = -2.0033$ is $z = -2$ which corresponds to $p = 0.0228$, so we approximate $P(z < -2.0033) \approx P(z < -2) = 0.0228$.

Then, we see the closest z-value to $z = 1.8527$ is $z = 1.85$ which corresponds to $p = 0.9678$, do we approximate $P(z < -1.8527) \approx P(z < -1.85) = 0.9678$. Therefore, we have

$$\begin{aligned} P(36 < w < 100) &= P(w < 100) - P(w < 36) \\ &= P(z < 1.8527) - P(z < -2.0033) \\ &\approx 0.9678 - 0.0228 \\ &= 0.945 \end{aligned}$$

P.S. If we are going to use Chebyshev's inequality, the estimation would be less accurate. Recap Chebyshev's inequality means the probability of any number differing greater than or equal to k standard deviation from the mean is $\frac{1}{k^2}$.

$$P\left(\frac{|X-\mu|}{\sigma} \geq k\right) \leq \frac{1}{k^2}.$$

Replace the symbols with $\hat{\mu}$ and $\hat{\sigma^2}$, since they are the best guess so far, and W. We have

$$P\left(\frac{|W-69.25|}{\sqrt{275.48}} \geq k\right) \leq \frac{1}{k^2}.$$

Now we plug two boundaries, 36 and 100 in the formula above.

For $w = 36$, $\frac{|36-69.25|}{\sqrt{275.48}} \geq 2.0033$. Reconstruct the formula

$$P\left(\frac{|W-\mu|}{\sigma} \geq 2.0033\right) \leq \frac{1}{2.0033^2}.$$

and we have the probability of any number differing greater than or equal to 2.0033 standard deviation is at most $\frac{1}{2.0033^2}$ which is the area $P(w < 36) + P(w > 102.5)$ (36 and 102.5 are the values of w for which $\frac{|W-\mu|}{\sigma} = 2.0033$). Given we know W follows a Gaussian distribution, we can divide the probability by 2 due to symmetry. Thus, we get the area $P(w < 36) = \frac{1}{2.0033^2}/2 \approx \frac{1}{8}$, given $P(w < 36)$ and $P(w > 102.5)$ are symmetric about mean $\hat{\mu} = 69.25$.

For $w = 100$, $\frac{|100-69.25|}{\sqrt{275.48}} \geq 1.8527$, Reconstruct the formula

$$P\left(\frac{|W-\mu|}{\sigma} \geq 1.8527\right) \leq \frac{1}{1.8527^2}.$$

and we have the probability of any number differing greater than or equal to 1.8527 standard deviation is at most $\frac{1}{1.8527^2}$ which is the area $P(w < 38.5) + P(w > 100)$ (38.5 and 100 are the values of w for which $\frac{|W-\mu|}{\sigma} = 1.8527$). Given we know W follows a Gaussian distribution, we can divide the probability by 2 due to symmetry. Then, We get the area $P(w > 100) = \frac{1}{1.8527^2}/2 \approx 0.1456$, given $P(w < 38.5)$ and $P(w > 100)$ are symmetric about mean $\hat{\mu} = 69.25$.

When we don't know the shape of the distribution, other versions of one sided chebyshev's are needed like Cantelli's inequality. You may find the relevant information here https://en.wikipedia.org/wiki/Cantelli%27s_inequality (https://en.wikipedia.org/wiki/Cantelli%27s_inequality). Take it easy, we are not assessing you about this.

Hence,

$$\begin{aligned} P(36 < w < 100) &= 1 - P(w > 100) - P(w < 36) \\ &\approx 1 - \frac{1}{8} - 0.1456 \\ &= 1 - 0.125 - 0.1456 \\ &= 0.7294 \end{aligned}$$

Note that this result is a weak inference and less accurate than the result that we got using Z-table.

Question 2

You toss an unfair coin 11 times where the chance of tossing a head is 0.6. What is the combined probability of getting exactly 5 or exactly 6 heads?

Answer 2

(Recall that toss a coin is a Bernoulli trial: because there's only two possibility of trial outcome and repeated Bernoulli trials would under binomial distribution)

Firstly we will use the binomial distribution with probability mass function(PMF):\newline $p(m|n, \theta) = \binom{n}{m}\theta^m(1-\theta)^{n-m} = \frac{n!}{m!(n-m)!}\theta^m(1-\theta)^{n-m}$

In which we should notice that

n is the number of trials,

m is the number of success (of getting head)

θ is the probability of success (of getting head)\newline

Then write out the formula for getting exactly 5 or exactly 6 heads:

$$
\begin{aligned}
p(m = 5 \; or \; m = 6 | n = 11, \theta = 0.6) &= p(m = 5 | n = 11, \theta = 0.6) + p(m = 6 | n = 11, \theta = 0.6) \\
&= \binom{11}{5} 0.6^5 (1 - 0.6)^{11-5} + \binom{11}{6} 0.6^6 (1 - 0.6)^{11-6} \\
&= \frac{11!}{5!(11-5)!} 0.6^5 (1 - 0.6)^{11-5} + \frac{11!}{6!(11-5)!} 0.6^6 (1 - 0.6)^{11-6} \\
&= 0.368
\end{aligned}
$$

Question 3

Two cafes are side by side and are open 7 days a week.

(a) The first cafe sells cappucinos at $5 each and has 30 customers a day. Half the customers buy a cappucino, the rest don't. What is the distribution of dollar sales made by the first cafe in a week?

(b) The second cafe sells cafe lattes at $6 each. Customers arrive at a rate of 60 a day and all the customers buy a cafe latte. What is the distribution of dollar sales made by the second cafe in a week?

(c) Now the Poisson distribution is approximately Gaussian when the rate is larger, and this applies moderately well when the rate is 15 or more. Give Gaussian approximations to the distributions of weekly dollar sales for each of the two cafes.

Answer 3

(a) The number of sales per week is Poisson with rate 105. The dollar sales is 5 times this, i.e. distribution is Poisson with rate parameter of $525 per week.

(b) The number of sales per week is Poisson with rate 420. The dollar sales is 6 times this, i.e. distribution is Poisson with rate parameter of $2520 per week.

(c) The question isn't very clear about when to apply the Gaussian approximation. One possible answer is to apply the normal approximation to the Poisson distribution of sales per week. In this case the sales rates respectively are $N(105, 105)$, and $N(420, 420)$. Now we scale to get dollars. The first is $N(525, 105.5.5) = N(525, 2625)$, and the second is $N(2520, 420.6.6) = N(2520, 15120)$. Note since we already applied the Gaussian approximation to the sales rates, we have to square the dollars as the units for the variance σ^2 are sales^2, so to get variance of total dollars we compute (sales times dollars per sale)^2 =(sales)^2 (dollars per sale)^2

It turns out the above approximation is not as accurate as directly applying the normal approximation to the Poisson distributions of weekly dollar sales in (a) and (b) to obtain $N(525, 525)$ and $N(2520, 2520)$, respectively.

Question 4

The continuous version of the uniform distribution has the following probability density function:

$$p(x|a,b) = \begin{cases} 0, \ x < a \\ \frac{1}{b-a}, \ a \leq x \leq b \\ 0, \ x > b \end{cases}$$

Show that

$$E\left[X\right] = \frac{a+b}{2}$$

and that

$$V\left[X\right] = \frac{(b-a)^2}{12}$$

Answer 4

Question asks for derivation of expected value and variance for continuous uniform distribution.

The formula of expected value is below. p(x) is the PDF function of X.

$$E[X] = \int_{x \in X} x \, p(x) \, dx$$

$$E[X] = \int_a^b x \, \frac{1}{b-a} \, dx$$

$$E[X] = \frac{1}{b-a} \int_a^b x \, dx$$

$$E[X] = \frac{1}{b-a} \Big|_a^b \frac{x^2}{2}$$

$$E[X] = \frac{1}{b-a} \left(\frac{b^2 - a^2}{2}\right)$$

$$E[X] = \frac{(b-a)(b+a)}{2(b-a)} = \frac{b+a}{2}$$

For calculating variance we will use the formula below;

$$V[X] = E[X^2] - E[X]^2$$

We already calculated the value of $E[X]$. We need to calculate $E[X^2]$.

$$E[X^2] = \int_a^b x^2 \ \frac{1}{b-a} \, dx$$

$$E[X^2] = \frac{1}{b-a} \int_a^b x^2 \, dx$$

$$E[X^2] = \frac{1}{b-a} \Big|_a^b \frac{x^3}{3}$$

$$E[X^2] = \frac{1}{b-a} (\frac{b^3 - a^3}{3})$$

$$E[X^2] = \frac{(b-a)(b^2 + ab + a^2)}{3(b-a)}$$

$$E[X^2] = E[X^2] = \frac{(b^2 + ab + a^2)}{3}$$

Therefore;

$$V[X] = E[X^2] - E[X]^2$$

$$V[X] = \frac{b^2 + ab + a^2}{3} - \left(\frac{b+a}{2}\right)^2$$

$$V[X] = \frac{b^2 + ab + a^2}{3} - \frac{b^2 + 2ab + a^2}{4}$$

$$V[X] = \frac{b^2 - 2ab + a^2}{12} = \frac{(b-a)^2}{12}$$

Question 5

Suppose that a binary message—either 0 or 1—must be transmitted by electromagnetic transmission from location A to location B. However, the data sent over the transmission channel are subject to a channel noise disturbance and so to reduce the possibility of error, the value 2 is sent over the channel when the message is 1 and the value -2 is sent when the message is 0. If x is the value sent at location A then R, the value received at location B, is given by $R = x + N$, where N is the channel noise disturbance. When the message is received at location B, the receiver decodes it according to the following rule: \ if $R \geq .5$, then 1 is concluded \ if $R < .5$, then 0 is concluded \ (a) Assuming the channel noise N follows a standard normal distribution, determine the error probabilities of incorrectly determining a 0 if a 1 was sent and incorrectly determining a 1 if a 0 was sent, i.e. $P(\text{error}|1 \text{ was sent})$ and $P(\text{error}|0 \text{ was sent})$.

(b) Now consider that an additional Gaussian noise source, M, with mean $\mu = 0$ and standard deviation $\sigma = 2$ is influencing the channel such that $R = x + N + M$. What are the error probabilities, $P(\text{error}|1 \text{ was sent})$ and $P(\text{error}|0 \text{ was sent})$, under these conditions? \

Answer 5

(a) $P(\text{error}|1 \text{ was sent}) = P(2 + N < .5) = P(N < -1.5)$. Z-value is $(-1.5 - 0)/1 = -1.5$ and corresponding z-value in the table is -1.5 which has p-value of 0.0668. So $P(\text{error}|1 \text{ was sent}) = 0.0668$.

$P(\text{error}|0 \text{ was sent}) = P(-2 + N > .5) = P(N > 2.5)$. Z-value is $(2.5 - 0)/1 = 2.5$ and corresponding z-value in the table is 2.5 which has p-value of 0.9938. So $P(\text{error}|0 \text{ was sent}) = 1 - 0.9938 = 0.0062$.

(b) When combining 2 normal distributions mean and variance add so we are dealing with distribution with variance $1 + 4 = 5$. So standard deviation is 2.236.

$P(\text{error}|1 \text{ was sent}) = P(2 + N < .5) = P(N < -1.5)$. Z-value is $(-1.5 - 0)/2.236 = -0.671$ and corresponding z-value in the table is -0.67 which has p-value of 0.2514. So $P(\text{error}|1 \text{ was sent}) = 0.2514$.

$P(\text{error}|0 \text{ was sent}) = P(-2 + N > .5) = P(N > 2.5)$. Z-value is $(2.5 - 0)/2.236 = 1.118$ and corresponding z-value in the table is 1.12 which has p-value of 0.8686. So $P(\text{error}|0 \text{ was sent}) = 1 - 0.8686 = 0.131$.

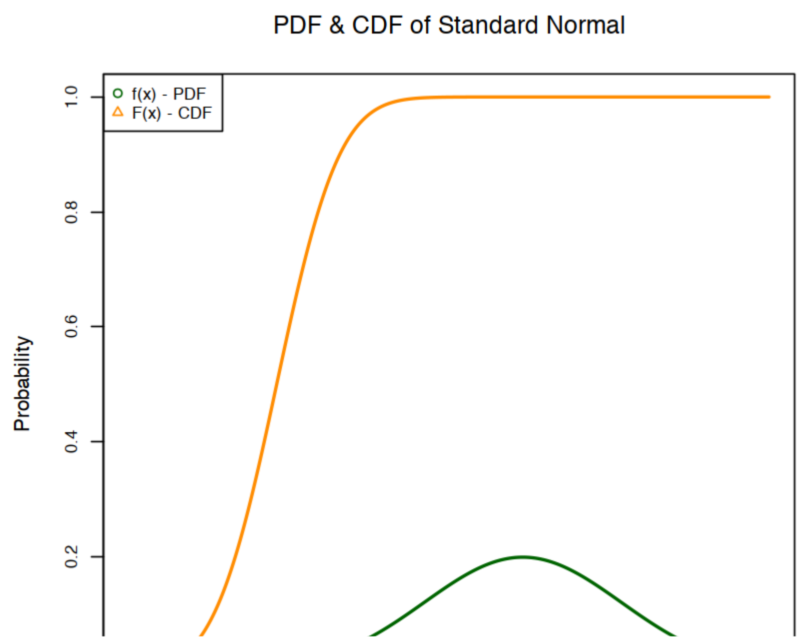
R questions

Create an R program to visualise with clear labelling the following distributions using the given R functions:

1. Gaussian Distribution with mean equal to 5 and standard deviation equal to 2 => Compute and display the PDF using dnorm and the CDF using pnorm
2. Bernoulli Distribution with probability of success equal to 0.5 => Compute and display the approximate PMF using rbinom to simulate drawing m=100,000 samples from the Bernoulli distribution
3. Binomial Distribution with n=20 trials and probability of success equal to 0.3 => Compute and display the approximate PMF using rbinom to simulate drawing m=100,000 samples from the Binomial distribution and the exact PMF using dbinom.
4. Poisson Distribution with rate parameter equal to 5 => Compute and display the approximate PMF using rpois to simulate drawing m=100,000 samples from the Poisson distribution and the exact PMF using dpois.

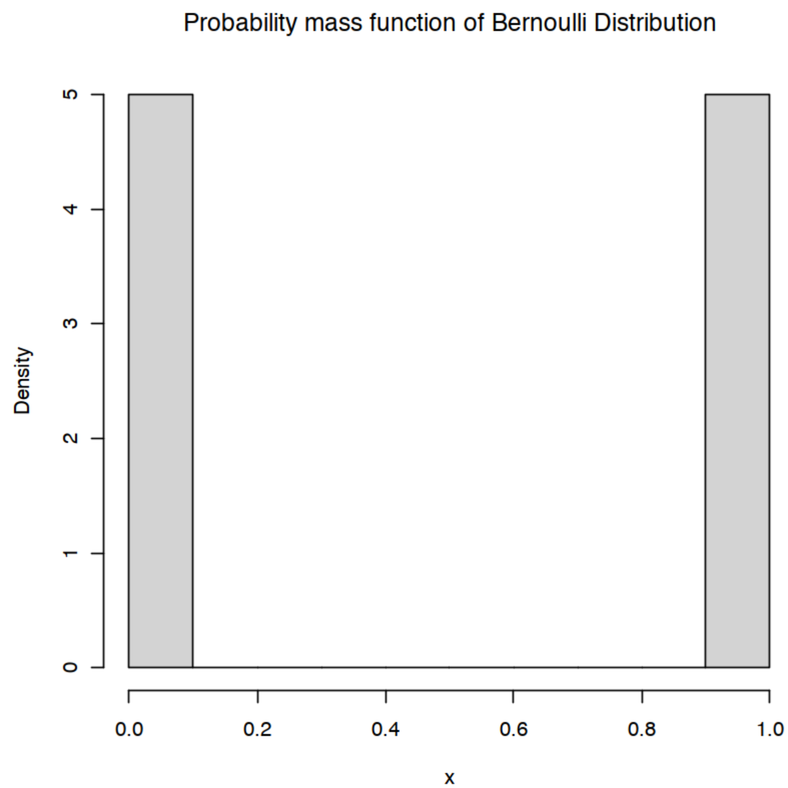
1. Gaussian Distribution

In [1]:
```
xseq<-seq(-3,10,.01)
densities<-dnorm(xseq,5,2)
cumulative<-pnorm(xseq)
plot(xseq, cumulative, col="darkorange", xlab="x", ylab="Probability",type="l",lwd=2, cex=2, main="PDF & CDF of S
lines(xseq, densities,col="darkgreen",lwd=2, cex=2)
legend("topleft",c("f(x) - PDF","F(x) - CDF"),cex=.8,col=c("darkgreen","darkorange"),pch=c(1,2))
```



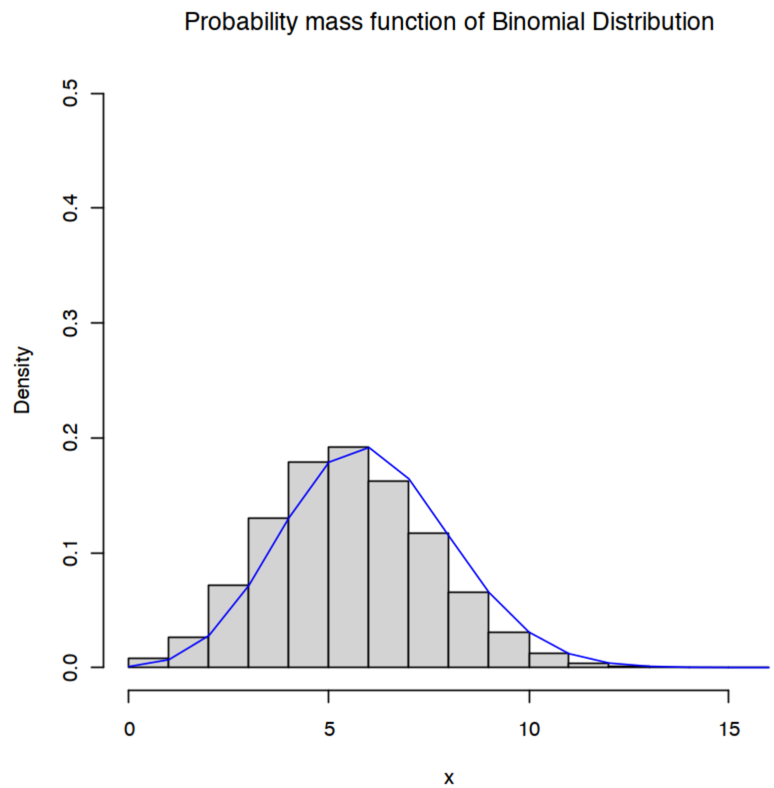
2. Bernoulli Distribution

In [2]:
```
xseq<-c(rep(0,50),rep(1,50))
hist(xseq, main="Probability mass function of Bernoulli Distribution",  xlab="x", freq=F)
```



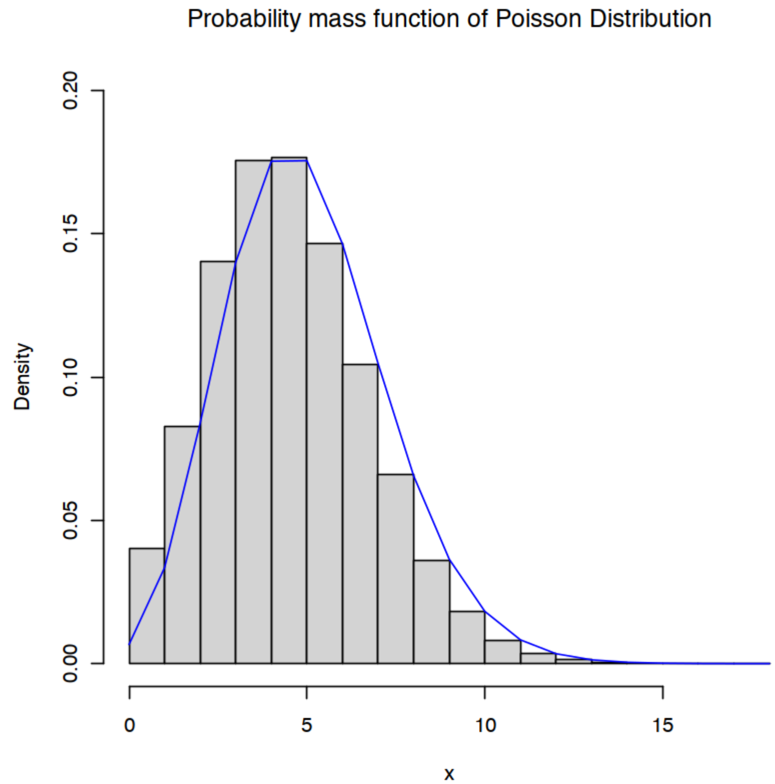
3. Binomial Distribution

In [3]:
```
vec <- rbinom(100000,20,0.3)
hist(vec, prob=TRUE, ylim = c(0, .50), xlab="x", main="Probability mass function of Binomial Distribution")
lines(0:max(vec), dbinom(min(vec):max(vec), 20, 0.3), col = 'blue')
```



4. Poisson Distribution

In [4]:
```
vec <- rpois(100000, 5)
hist(vec, prob=TRUE, ylim = c(0, .20), xlab="x", main="Probability mass function of Poisson Distribution")
lines(0:max(vec), dpois(0:max(vec), mean(vec)), col = 'blue')
```



Suppose that IQ scores have a bell-shaped distribution with a mean of 100 and a standard deviation of 15.

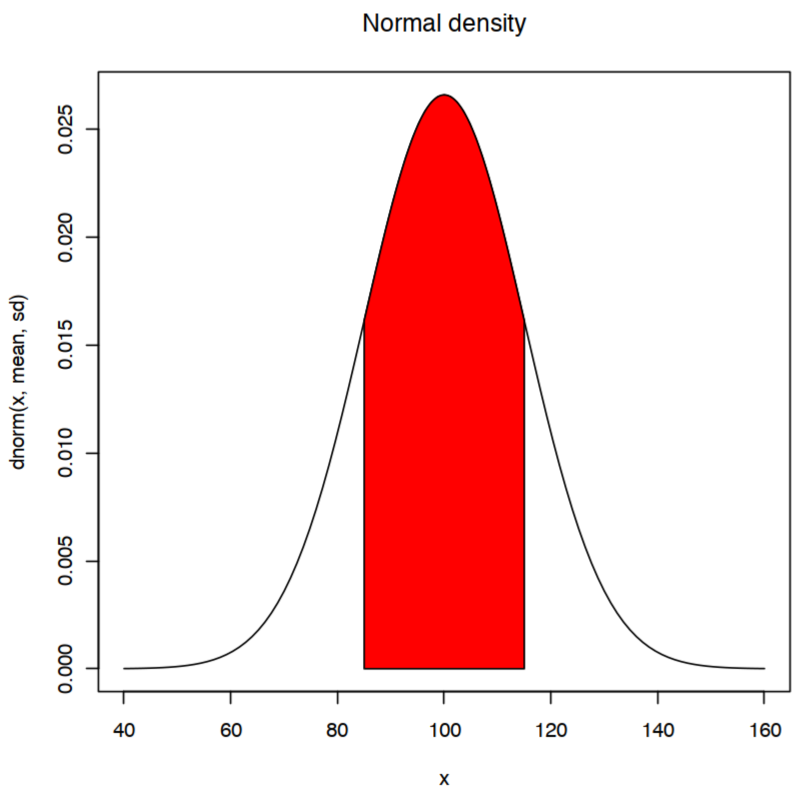
1. What percentage of people should have an IQ score between 85 and 115?
2. What percentage of people should have an IQ score between 70 and 130?
3. What percentage of people should have an IQ score of more than 130?
4. A person with an IQ score greater than 145 is considered a genius. Does the empirical rule support this statement? Explain

In [5]:
```
mean <- 100
sd <- 15

shade_normal <- function(mean, sd, from, to){
    curve(dnorm(x,mean,sd), xlim=c(mean-sd*4,mean+sd*4), main="Normal density")

    S.x  <- c(from, seq(from, to, 0.01), to)
    S.y  <- c(0, dnorm(seq(from, to, 0.01), mean, sd), 0)
    polygon(S.x,S.y, col="red")
}

prob_calc <- function(mean, sd, from, to){
    z.from <- (from-mean)/sd
    z.to <- (to-mean)/sd
    prob <- pnorm(z.to) - pnorm(z.from)
    return(prob);
}
```

What percentage of people should have an IQ score between 85 and 115?

In [6]:
```
from <- 85
to <- 115
shade_normal(mean, sd, from, to)
prob_calc(mean, sd, from, to)
```

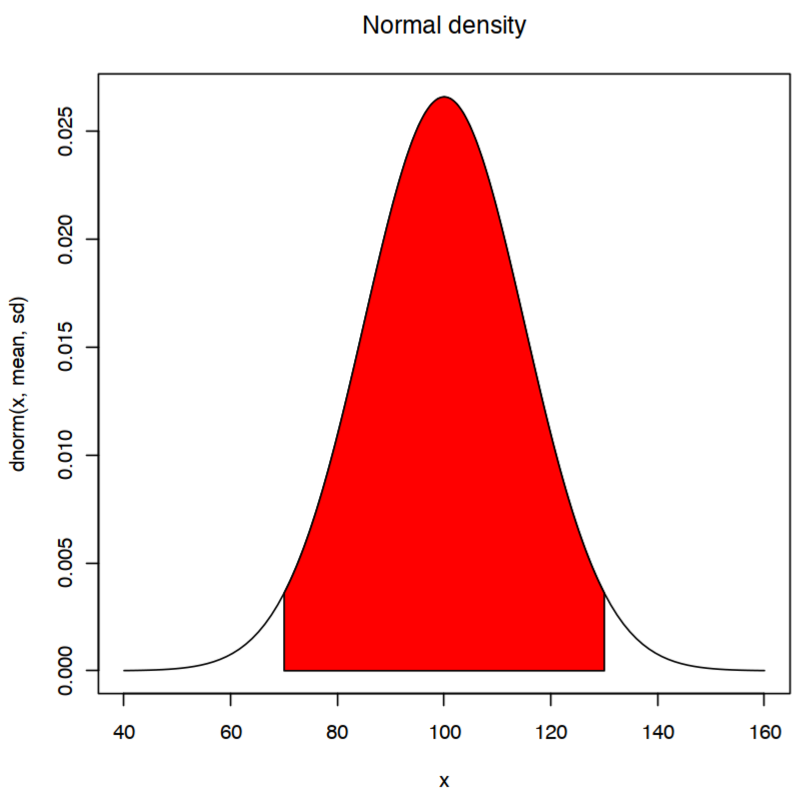
0.682689492137086



What percentage of people should have an IQ score between 70 and 130?

In [7]:
```
from <- 70
to <- 130
shade_normal(mean, sd, from, to)
prob_calc(mean, sd, from, to)
```

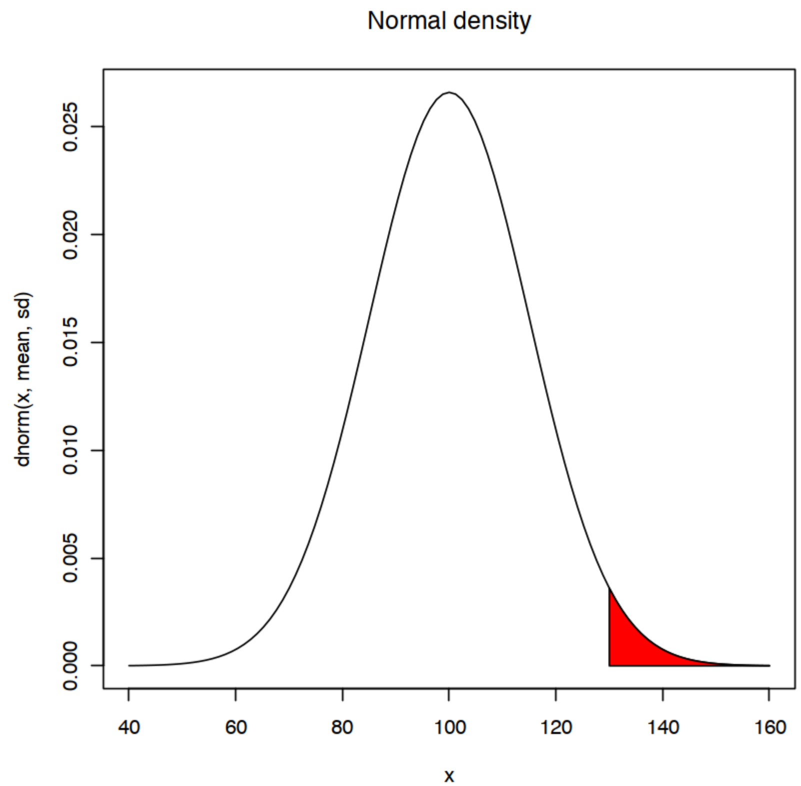
0.954499736103642



What percentage of people should have an IQ score more than 130?

In [8]:
```
from <- 130
to <- 160
shade_normal(mean, sd, from, to)
prob_calc(mean, sd, from, to)
```

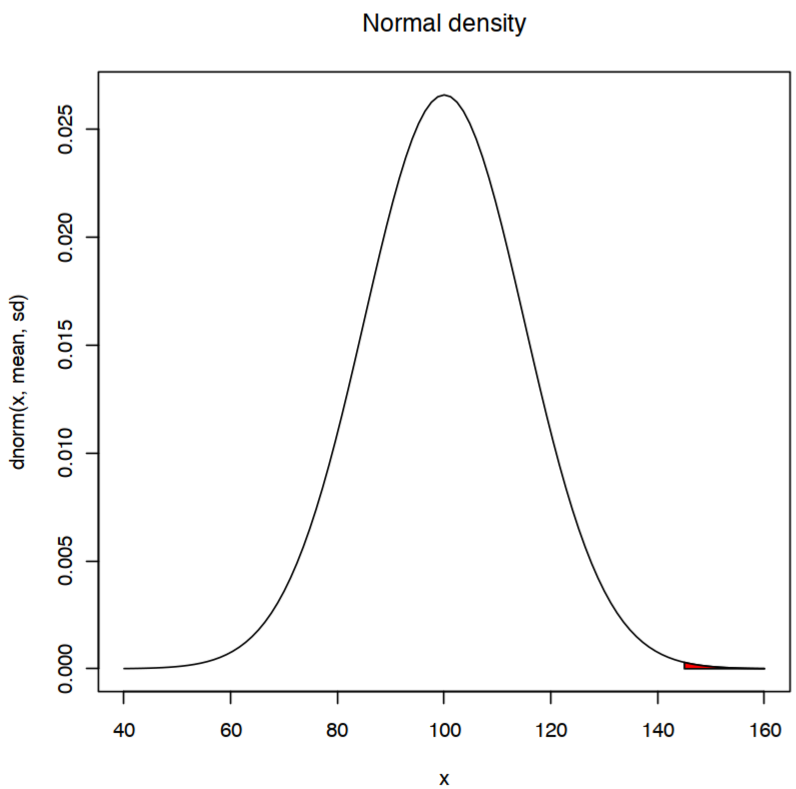
0.0227184607063461



Suppose that IQ scores have a normal distribution with a mean of 100 and a standard deviation of 15. Does the empirical rule support this statement? Explain

In [9]:
```
from <- 145
to <- 160
shade_normal(mean, sd, from, to)
prob_calc(mean, sd, from, to)
```

0.00131822678979698



The calculated probability(and shaded region proportion) is very small which is significant and not ordinary. So, it can be said that person is a genius.