value, and denoting $q = (1-\pi)$ we have

$$\begin{aligned} E[X] &= \sum_{-\infty}^{\infty} x \cdot f(x), \\ &= \sum_{x=0}^{\infty} x(1-\pi)^x \pi, \\ &= \pi \sum_{x=1}^{\infty} x q^x, \\ &= \pi \left[q + 2q^2 + 3q^3 + 4q^4 + ... \right], \\ &= \pi \frac{q}{(1-q)^2} \end{aligned}$$

where we have used the result in (8.2) with $a = q$ and $r = q$.

Replacing q with $(1-\pi)$ yields the result:

$$E[X] = \pi \cdot \frac{1-\pi}{(1-(1-\pi))^2} = \frac{1-\pi}{\pi}.$$

$\square$

The expected value calculation is intuitive. It emphasizes that our waiting time for the first success depends on the odds of a failure. In the case of the six sided die, we should expect

$$\frac{\frac{5}{6}}{\frac{1}{6}} = 5$$

failures before we roll a "6".

8.1.1 Practice Problems

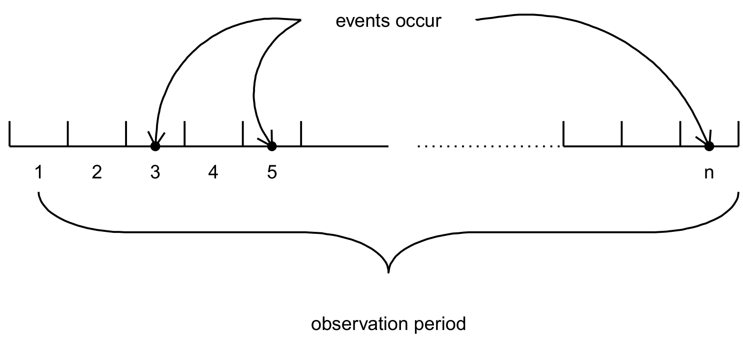
1. Let X be a geometric random variable with $\pi = 1/4$. Calculate:
   a. $P(X \leq 14)$
   b. $P(X > 20)$
   c. $P(X = 25)$
2. Suppose that a basketball player sinks a basket from a certain position on the court with probability 0.35.
   a. What is the probability that the player sinks three baskets in 10 independent throws?
   b. What is the probability that the player gets her first basket in her 10th shot?

8.2 Poisson Distribution

Suppose some event occurs at random times over a fixed observation period. Let X be the random variable which counts the number of occurrences of this event over this observation period. What is the PMF of X?

The derivation of the PMF of X begins by approximating X with something we know, namely the binomial distribution, using the following chain of reasoning.

- We divide the time into n non-overlapping sub-intervals of equal length.
- We assume that the probability that an event occurs during a given sub-interval, π remains constant from sub-interval to sub-interval and is proportional to $\frac{1}{n}$ - let's call this probability λ/n.
- If n is large, the probability of having two occurrences in one sub-interval is very small – we will approximate this with 0.
- We assume the number of occurrences in one interval is independent of the number in the other sub-intervals.



A good approximation for X then is

$$X \approx Binom\left(n, \pi = \frac{\lambda}{n}\right)$$

because we have n independent sub-intervals (trials) with constant probability of occurrence in each one.

The above assumptions imply that

$$\begin{aligned} P(X = x) &\approx P(\text{x of the sub-intervals contain 1 event and the other (n-x) contain 0 events}), \\ &= \binom{n}{x}\left(\frac{\lambda}{n}\right)^{x}\left(1 - \frac{\lambda}{n}\right)^{n-x}. \end{aligned}$$

The binomial approximation to the Poisson experiment should get better and better as n increases. In fact, when $n \to \infty$, we have the result:

$$P(X = x) \quad \to \quad e^{-\lambda}\frac{\lambda^x}{x!}.$$

This is referred to as the **Poisson limit** to the binomial PMF as a nod to Simon Denis Poisson, the French mathematician who discovered it.

The proof of the Poisson limit for the binomial is as follows:

$$
\begin{aligned}
P(X=x) &= \binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}, \\
&= \frac{n!}{x!(n-x)!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}, \\
&= \frac{n\cdot(n-1)\cdot(n-2)\dots(n-x+1)}{n^x}\cdot\frac{\lambda^x}{x!}\cdot\frac{(1-\lambda/n)^n}{(1-\lambda/n)^x}
\end{aligned}
$$

As $n \to \infty$, we have:

$$\left(1-\frac{\lambda}{n}\right)^n \approx e^{-\lambda},\ \frac{n\cdot(n-1)\cdot(n-2)\dots(n-x+1)}{n^x}\approx 1,\ \left(1-\frac{\lambda}{n}\right)^x \approx 1.$$

In other words, if n independent trials, each of which result in a success with probability π are performed, then when n is large but π is small enough so that $n\pi$ remains constant, the number of successes which occur is a Poisson random variable with parameter $\lambda = n\pi$.

Definition 8.2. The PMF for a **Poisson random variable** with parameter λ (> 0) is

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!},\quad x=0,1,2,3,\dots$$

We denote $X \sim Poisson(\lambda)$.

Recall from calculus (Taylor series) that

$$1+\lambda+\frac{\lambda^2}{2!}+\frac{\lambda^3}{3!}+\cdots=\sum_{x=0}^{\infty}\frac{\lambda^x}{x!}=e^{\lambda},$$

and therefore we have defined a legitimate PMF since

$$\sum_{x=0}^{\infty}f(x)=\sum_{x=0}^{\infty}e^{-\lambda}\frac{\lambda^x}{x!}=e^{-\lambda}\cdot\sum_{x=0}^{\infty}\frac{\lambda^x}{x!}=e^{-\lambda}\cdot e^{\lambda}=1.$$

Example 8.3. Suppose that the number of accidents occurring on a highway each day is a Poisson random variable with parameter $\lambda = 3$.

a. Find the probability that 3 or more accidents occur today.

b. Repeat part a under the assumption that at least 1 accident occurs today.

....

We can use `dpois`, `ppois` and `rpois` to calculate probabilities related to the Poisson distribution in R.

```
dpois(x = 3, lambda = 3)    #P(X = 3)
```

```
## [1] 0.2240418
```

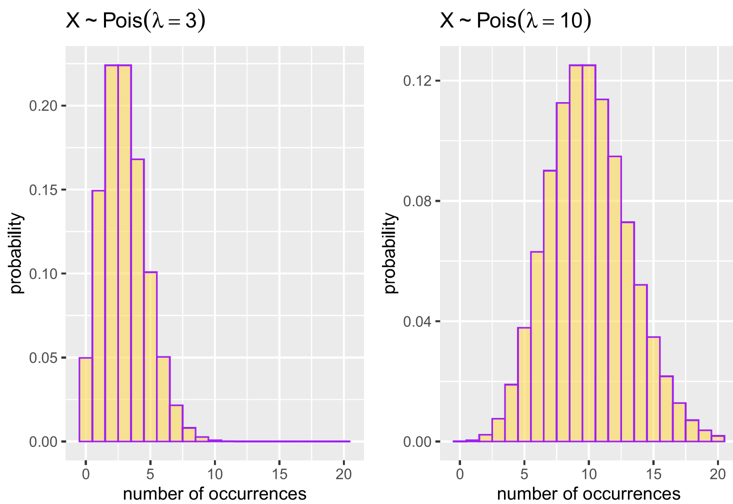
```
ppois(q = 3, lambda = 3)    #P(X <= q)
```

```
## [1] 0.6472319
```

```
ppois(q = 2, lambda =3, lower.tail = F)    #P(X > q)
```

```
## [1] 0.5768099
```

The probability histogram for a Poisson random variable with two different values of λ are shown below. When λ is small,



We will now state, without proof, the expected value and variance of a Poisson random variable[1].

Lemma 8.1. *Suppose $X \sim Pois(\lambda)$. Then*

$$E[X] = \lambda$$

and

$$Var[X] = \lambda.$$

The important take aways here are that if $X \sim Pois(\lambda)$, then

- the mean and variance of X are equal
- the parameter λ is the expected number of occurrences of the event during the observation period and is referred to as the **rate** parameter.

It is often the case that the number of **arrivals** at a server (ATM machine, telephone exchange, wireless network) for some specific length of time t

- can be modeled by a $Pois(\lambda t)$ distribution where λ is the rate per unit time
- and is such that arrivals in non-overlapping intervals are independent.

We call such a model a **Poisson process**

Example 8.4. Customers come to a small business at an average rate of 6 per hour. Let's assume that a Poisson process is a good model for customer arrivals.

[1]Please consult page 93 of the text for a detailed step-by-step derivation

a. Calculate the probability that there are exactly 5 customers in the next 20 minutes?

b. Calculate the probability that there are exactly 5 customers in the next 20 minutes and 5 more customers in the following 10 minutes.

c. Calculate the probability that the next 5 customers will arrive within 15 minutes of each other.

. . . .

Example 8.5. The Poisson distribution has a tremendous range of application as a model for data. The most frequent and obvious application is to model the number of times a certain event occurs during each of a series of units (typically time or space).

Let us now fit the Poisson model to a set of data. The `Fumbles` dataframe from the **fastR2** package gives the total number of fumbles by each NCAA team for the first three weeks in November 2010.

```
library(fastR2)                    # for the dataset Fumbles
library(tidyverse)
```

```
glimpse(Fumbles)
## Rows: 120
## Columns: 7
## $ team  <fct> Air Force, Akron, Alabama, Arizona, Arizona St, Arkansas, Arkans~
## $ rank  <int> 53, 19, 68, 31, 94, 46, 60, 94, 18, 94, 89, 76, 4, 38, 41, 53, 4~
## $ W     <int> 8, 1, 9, 7, 5, 9, 4, 6, 12, 4, 7, 10, 6, 2, 2, 6, 5, 8, 3, 4, 6,~
## $ L     <int> 4, 11, 3, 4, 6, 2, 7, 5, 0, 8, 5, 1, 5, 10, 10, 5, 6, 3, 9, 6, 5~
## $ week1 <int> 4, 2, 0, 1, 2, 0, 0, 3, 1, 2, 5, 3, 0, 1, 2, 1, 3, 3, 5, 2, 1, 0~
## $ week2 <int> 2, 3, 3, 0, 1, 1, 0, 2, 1, 2, 2, 2, 2, 1, 3, 1, 1, 3, 5, 2, 5, 2~
## $ week3 <int> 2, 2, 2, 2, 3, 0, 4, 0, 0, 2, 1, 2, 4, 2, 3, 3, 2, 0, 0, 2, 2, 3~

slice_head(Fumbles, n = 5)      #peek at first five rows
##         team rank W  L week1 week2 week3
## 1  Air Force   53 8  4     4     2     2
## 2      Akron   19 1 11     2     3     2
## 3    Alabama   68 9  3     0     3     2
## 4    Arizona   31 7  4     1     0     2
## 5 Arizona St   94 5  6     2     1     3
```

The frequency distribution of the fumbles for week 1 and some summary statistics are given below.

```
Fumbles %>% count(week1)
##   week1  n
## 1     0 22
## 2     1 36
## 3     2 29
## 4     3 23
## 5     4  5
## 6     5  4
## 7     7  1

Fumbles %>% summarize(n=n(),
                      mean = mean(week1),
                      var = var(week1),
                      min = min(week1),
                      max = max(week1) )
##     n mean      var min max
## 1 120 1.75 1.852941   0   7
```

Let X_i denote the number of fumbles made by team i in week 1. We have observed $x_1 = 4, x_2 = 2, x_3 = 0, x_4 = 1, x_5 = 2$ and so on. What can be said about the distribution of X_i in general? Clearly, X_i is the number of <u>successes</u> in a given period of time, but does that automatically mean it has a Poisson distribution? Not necessarily. As we noted at the beginning of the section, we need to be able to assume that whether or not there is a fumble in one subinterval has no bearing on another sub-interval (independence). We also need to assume that the probability of a fumble is the same for every sub-interval. These are strong assumptions in any realistic setting. Furthermore, even if the X_i individually follow a Poisson distribution, there is no reason to think that the parameter λ will be the same for each team.

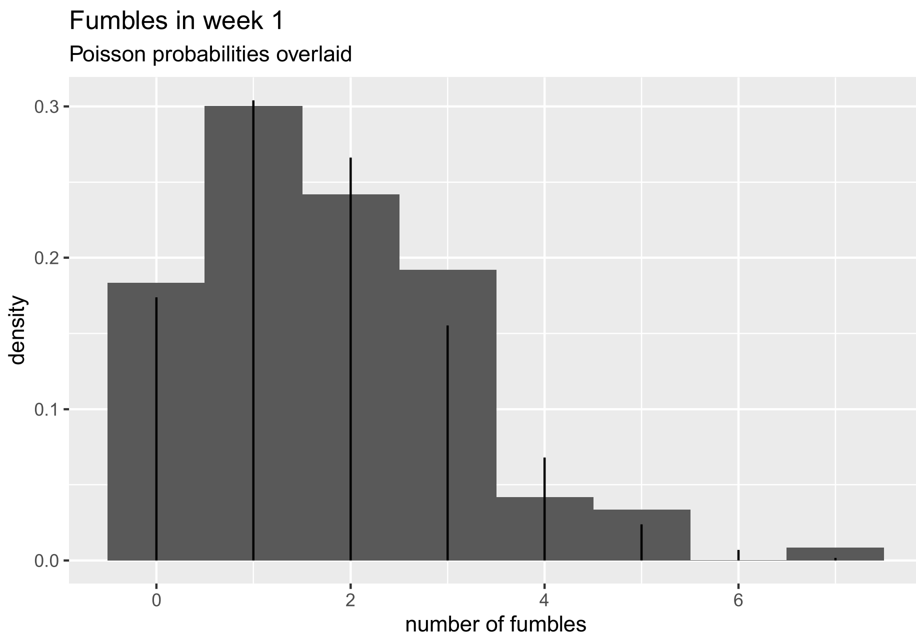
With all these caveats in mind, let's compare a histogram of the fumbles in week 1 to a Poisson distribution with λ equal to the average number of fumbles in week 1. The histogram is drawn on a density[2], rather than frequency, scale since we wish to make comparisons with the Poisson probabilities.

As shown in the code snippet below, the `geom_segment` layer is used to add lines corresponding to the Poisson probabilities.

```
# data frame containing P(X = x) assuming X ~ Pois(lambda = 1.75)
#
pois_fit <- tibble(
                  num_fumbles = 0:7,
                  f = dpois(num_fumbles, lambda = 1.75)
            )

ggplot( ) +
  geom_histogram(data = Fumbles,
                 mapping = aes(x = week1, y = after_stat(density)),
                 binwidth = 1) +
  geom_segment(data = pois_fit,
               mapping = aes( x = num_fumbles,
                              xend = num_fumbles,
                              y = 0, yend = f)) +
  labs(x = "number of fumbles",
       title="Fumbles in week 1",
       subtitle ="Poisson probabilities overlaid")
```

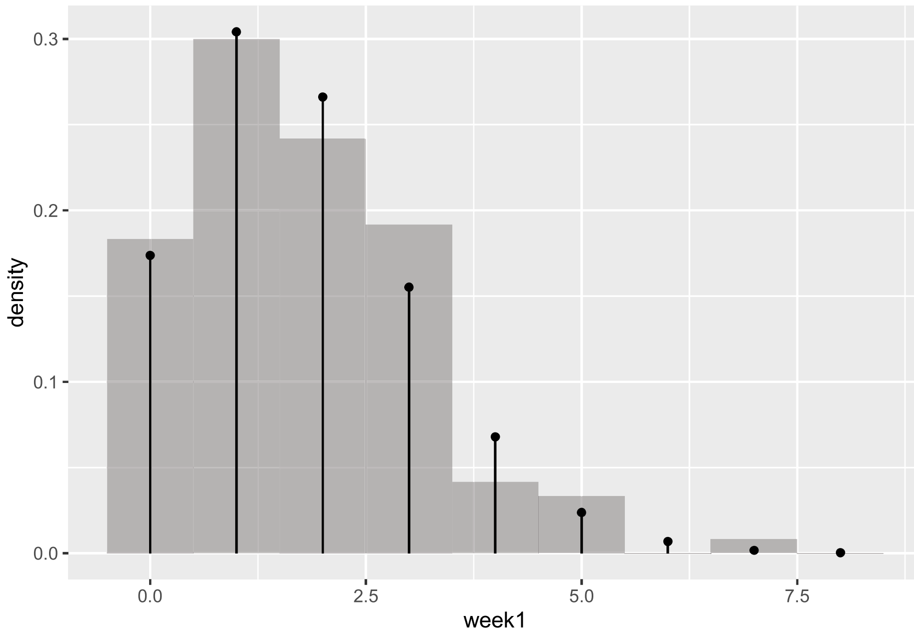
[2]For discrete random variables, the density scale simply means the y axis represents frequencies as percentages rather than as counts.



For an easy alternative, feel free to use the built in functions from the **fastR2** package.

```
library(fastR2(

))
gf_dhistogram(~ week1, data=Fumbles, binwidth=1, alpha=0.3) %>%
  gf_dist("pois", params=list(lambda = mean(~ week1, data=Fumbles) ) )
```



The visualizations show a surprisingly good agreement between the observed data and the Poisson model, especially for values of 0 and 1. The variance of our data is also close to the mean as we would expect for data sampled from a Poisson distribution.

8.2.1 Practice Problems

1. Compare the Poisson approximation with the correct binomial probability for the following cases.

   a. $P(X = 2)$ when $n = 8, \pi = 0.1$.
   b. $P(X = 9)$ when $n = 10, \pi = 0.95$.

2. A computer programmer on the average makes one error in every 500 lines of code. A typical program they write has 500 lines of code. Calculate the probability that they make between 0 and 2 errors (both inclusive).

3. Suppose an urn contains 100 marbles – one of these is black and the remaining 99 are white. 10 marbles are drawn from the urn randomly with replacement. What is the probability that 2 black marbles are drawn? Calculate the probability using the binomial distribution. Repeat using the Poisson approximation.

4. The Content Delivery Network (CDN) on a website fails on average once every 60 days. Assume that a Poisson model is a good model for CDN failures. What is the probability that there are no failures in a week?