Chapter 8.2

Poisson random variable

Review of Last Week

Variance: $\sigma^2 = Var[X] = E\left[(X-\mu)^2\right]$ provides a measure of spread from the expected value μ.

- Easier formula for calculating variance:

$$Var[X] = E\left[X^2\right] - \mu^2$$

Standard deviation: $\sigma = SD[X] = \sqrt{Var[X]}$ is the typical size of the deviation from μ.

Chebyshev's inequality: the probability that a random variable is k or more σ from the mean is no bigger than $\frac{1}{k^2}$.

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

Review of Last Week

Geometric random variable: the number of **failures** before first success in independent trials with probability of success π on each trial.

$$X \sim Geom(\pi)$$

- PMF: $f(x) = (1-\pi)^x \, \pi, \quad x = 0, 1, 2, 3 \ldots$
- For any integer $x \geq 0$ we have the result $P(X \geq x) = (1-\pi)^x$. (example 8.2)
- $E[X] = \frac{1-\pi}{\pi}$ (odds of failure)

Poisson Experiment

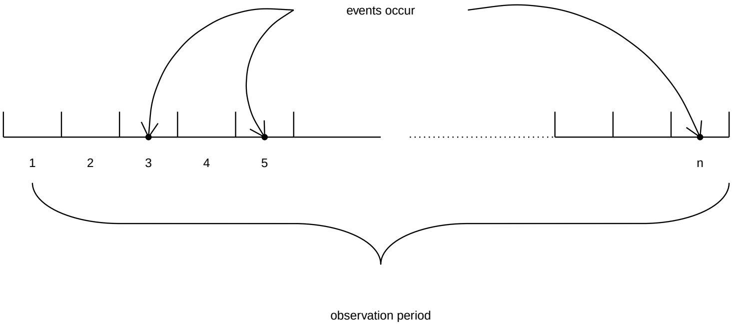
Suppose some event occurs "at random times" over a fixed observation period. Let X be the random variable which counts the number of occurrences of this event over this observation period.

X is called a **Poisson** random variable.

Poisson PMF

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.

- Divide the time into n non-overlapping sub-intervals of equal length.
- 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.
- The number of occurrences in one interval is independent of the number in the other sub-intervals.


events occur
1
2
3
4
5
n
observation period

Poisson PMF

A good approximation for X is

$$X \approx Binom(n, \frac{\lambda}{n})$$

because we have n independent sub-intervals (trials) with probability $\pi = \lambda/n$ of occurrence in each one.

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

Poisson limit to the binomial

The binomial approximation to the Poisson experiment should get better and better as $n \to \infty$. In fact, when n is very large:

$$P(X = x) = \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \to e^{-\lambda} \frac{\lambda^x}{x!}.$$

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

Proof of the Poisson limit to binomial

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

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, for

$$X \sim Binom(n, \pi)$$

if n is large but π is small enough so that $n\pi$ remains constant, then X is called a Poisson random variable with parameter $\lambda = n\pi$.

Poisson PMF

Definition 8.1 The PMF for a **Poisson random variable** with parameter λ (> 0) is

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

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 8.4

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.

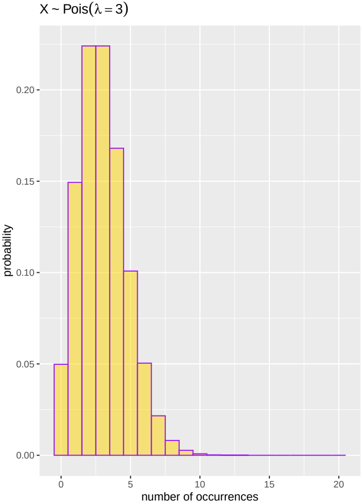
Poisson calculations in R

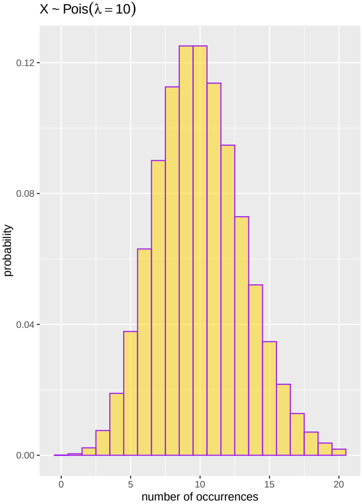
```r
dpois(x = 3, lambda = 3)    #P(X = 3)
## [1] 0.224

ppois(q = 2, lambda = 3)    #P(X <= q)
## [1] 0.423

ppois(q = 2, lambda =3, lower.tail = F)    #P(X > q)
## [1] 0.577
```

Probability histogram





Example 8.4 contd.

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

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

Expectation and variance

Lemma 8.1 Let $X \sim Poisson(\lambda)$. Then

- $E[X] = \lambda$
- $Var[X] = \lambda$

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 8.5

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.

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

Example 8.5

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.

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

Example 8.5

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.

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

Example 8.6

Is the Poisson distribution a good fit for modeling the number of fumbles in NCAA football?

```
#include packages in setup
library(fastR2)                 # for the dataset Fumbles
library(tidyverse)              # for ggplot + dplyr packages
```

Example 8.6

```
#you can type data(fumbles) in Console to load dataset in Environment
#
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~
```

```
#please see STAT 311 course resources "Data verbs" slidedeck, "Data basics" lab
```

Example 8.6

```r
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
```

Example 8.6

```
Fumbles %>% count(week1) #what are the values in this column and how often is each value observed?
##   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(), #n() counts the number of rows
                      xbar = mean(week1), #find mean of values
                      s = sd(week1),      #find SD of values
                      min = min(week1),   #find min of values
                      max = max(week1) )  #find max of values
##     n xbar    s min max
## 1 120 1.75 1.36   0   7
```

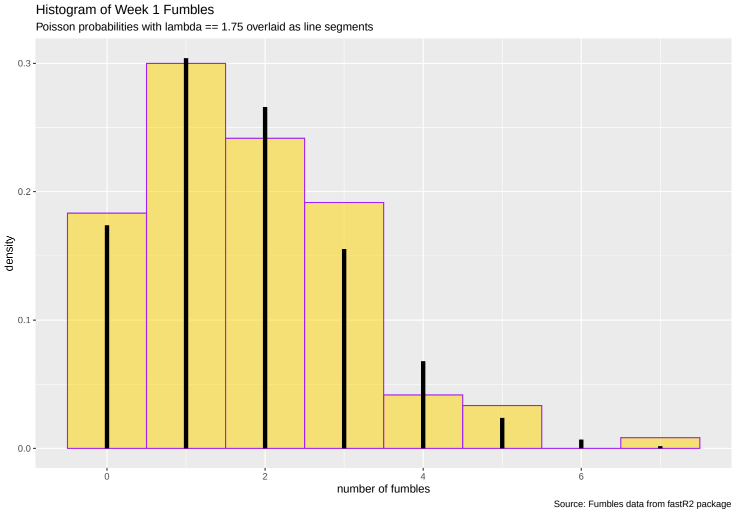
Example 8.6

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 *successes* in a given period of time, but does that automatically mean it has a Poisson distribution? Not necessarily.

Example 8.6


Histogram of Week 1 Fumbles
Poisson probabilities with lambda == 1.75 overlaid as line segments
0.3
0.2
0.1
0.0
density
0
2
4
6
number of fumbles
Source: Fumbles data from fastR2 package

Example 8.6

Code to make histogram with Poisson probabilities overlaid

```
# 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)),
                 fill = "gold",
                 color = "purple",
                 alpha = 0.5,
                 binwidth = 1) +
  geom_segment(data = pois_fit,
               mapping = aes( x = num_fumbles,
                              xend = num_fumbles,
                              y = 0, yend = f),
               linewidth = 2) +
  labs(x = "number of fumbles",
       title="Histogram of Week 1 Fumbles",
       subtitle =paste("Poisson probabilities with", expression(lambda==1.75), "overlaid as line segments"),
       caption="Source: Fumbles data from fastR2 package")
```