# Lecture 2: Generative Models for Discrete Data

Jing Ma, Statistics, TAMU

28 August, 2019

# This lecture

- Experiment with the most useful generative models for discrete data: *Poisson*, *binomial*, *multinomial*.
- Use R functions for computing probabilities and counting rare events.
- Generate random numbers from specified distributions.

# Examples of discrete data

- How many reads of DNA match a reference pattern?
- How many CG digrams we observe in a sequence?
- How many binding sites?

# Random variables

- A random variable assigns a number to each outcome of a random circumstance, or, equivalently, a random variable assigns a number to each unit in a population.
- The distribution of a random variable is a model that shows us what values are possible for that particular random variable and how often those values are expected to occur (i.e. their probabilities).
- The model can be expressed as a function or table or picture, depending on the type of variable it is.
- We will consider mainly discrete random variables.

# Discrete random variable

- A discrete random variable, $X$, is a random variable with a finite or countable number of possible outcomes.
- The probability distribution function (pdf) for a discrete random variable $X$ is a table or rule that assigns probabilities to the possible values of the X.

# Discrete random variabl examples

- Bernoulli

| | Success | Failure |
|---|---|---|
| Probability | 0.5 | 0.5 |

- Multinormial

| | A | T | C | G |
|---|---|---|---|---|
| Probability | 0.25 | 0.25 | 0.25 | 0.25 |

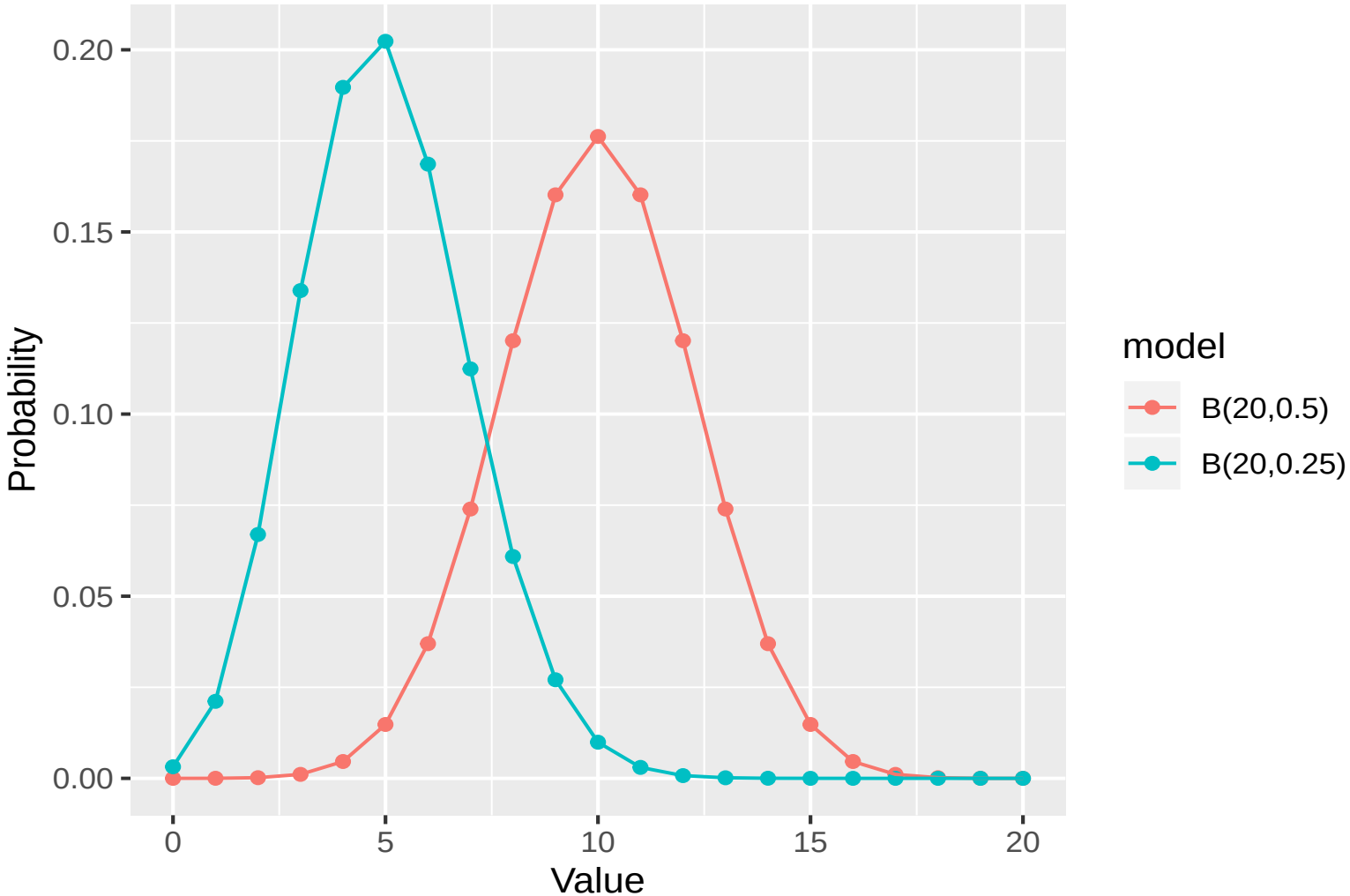
# Binomial random variable $B(n, p)$:

The binomial random variable counts the number of times a certain event occurs out of a particular number of observations or trials of a random experiment.

A binomial experiment is defined by the following conditions:

1. There are $n$ "trials" where $n$ is determined in advance and is not a random value.
2. There are two possible outcomes on each trial, called "success" (S) and "failure" (F).
3. The outcomes are independent from one trial to the next.
4. The probability of a "success" remains the same from one trial to the next, and this probability is denoted by $p$.

# Probability distribution of Binomial



# A special binomial

- The Bernoulli distribution is a special case of the binomial distribution where a single trial is conducted ($n = 1$)

# Poisson random variable: Pois($\lambda$)

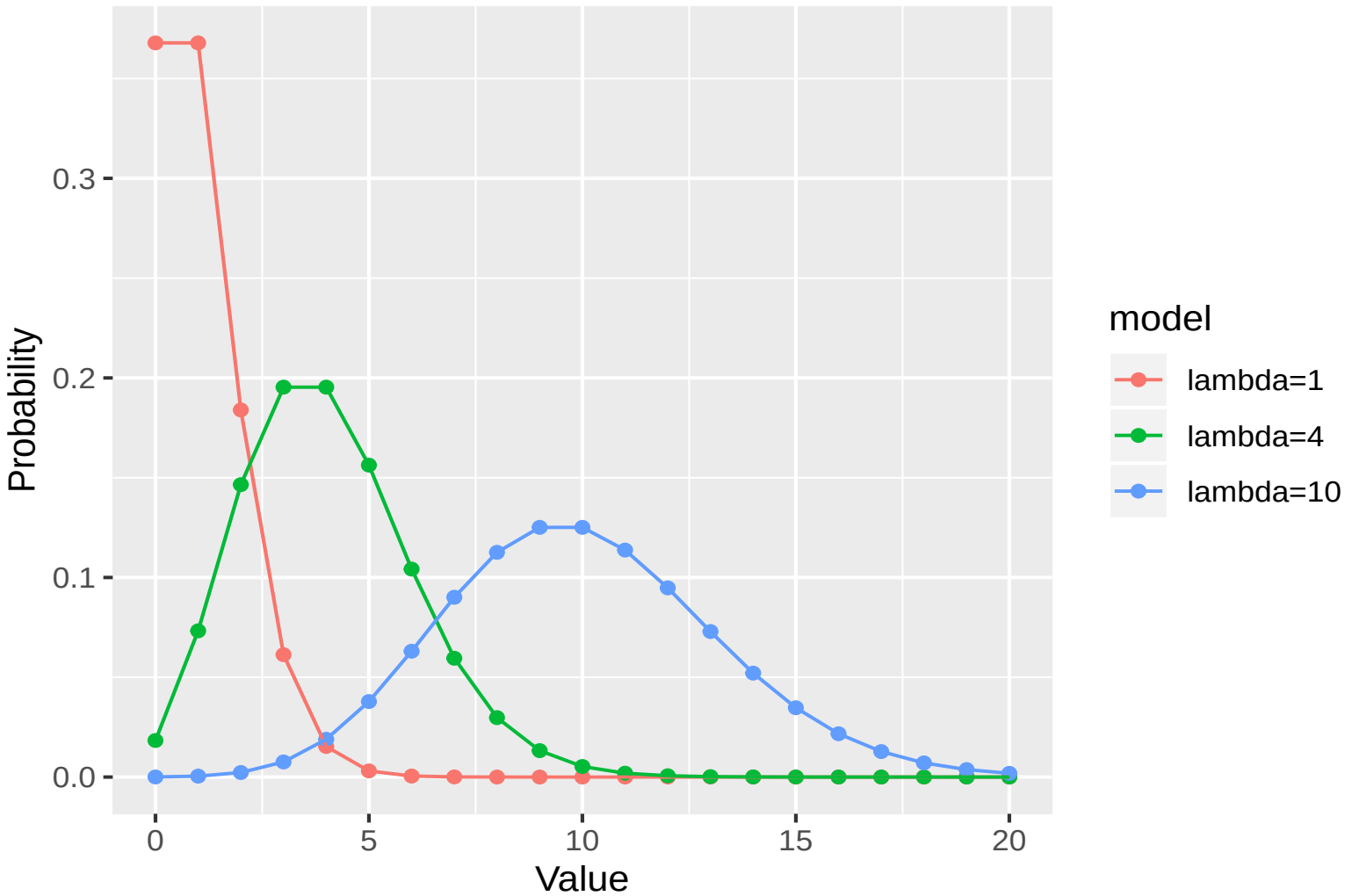
An event can occur $0, 1, 2, \ldots$ times in an interval. The probability of observing $k$ events in an interval is given by

$$P(k \text{ events in interval}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where

- $k = 0, 1, 2, 3, \ldots$
- $\lambda =$ average number of events
- $e =$ Euler's constant $\approx 2.71828$
- $k! = k \times (k-1) \times \ldots \times 2 \times 1$ is the factorial of $k$

# Example probability distribution of Poisson



# Poisson approximation to binomial

- Poisson can be used as an approximation of the binomial distribution if $n$ is sufficiently large and $p$ is sufficiently small.

$$B(n, p) \approx \text{Pois}(\lambda = np)$$

- A rule of thumb: the Poisson distribution is a good approximation of the binomial distribution if $n \geq 20$ and $p \leq 0.05$, and an excellent approximation if $n \geq 100$ and $np \leq 10$.

# Try the code yourself

Verify that Pois(0.5) is a good approximation to $B(5000, 0.01)$.

```
set.seed(5427121)
m50=matrix(rbinom(n=5000,size=1,prob=0.01),ncol=100,nrow=50)
s100=apply(m50,2,sum);table(s100)

t100=rpois(n=100,lambda=0.5);table(t100)
c(mean(s100), mean(t100))
```

# Sums of Poisson-distributed random variables

If $X_i \sim Pois(\lambda_i)$ for $i = 1, \ldots, k$, and $\lambda = \sum_{i=1}^{k} \lambda_i$, then

$$Y = \sum_{i=1}^{k} X_i \sim Pois(\lambda).$$

# The multinomial distribution

The multinomial distribution is a generalization of the binomial distribution. It models the probability of counts of each side for rolling a $k$-sided die $n$ times.

A multinomial experiment has the following properties:

1. The experiment consists of $n$ repeated trials.
2. Each trial has a discrete number of possible outcomes.
3. The trials are independent; that is, the outcome on one trial does not affect the outcome on other trials.
4. In any given trial, the probability that a particular outcome will occur is constant.

# The multinomial distribution

$$P(X_1 = n_1, \ldots, X_k = n_k) = \frac{n!}{n_1! \ldots n_k!} p_1^{n_1} \times \ldots \times p_k^{n_k},$$

where $\sum_{i=1}^{k} n_i = n$.

# Test your understanding

Suppose a card is drawn randomly from an ordinary deck of playing cards, and then put back in the deck.

This exercise is repeated five times.

**Question:** What is the probability of drawing 1 spade, 1 heart, 1 diamond, and 2 clubs?

# Test your understanding

- The experiment consists of 5 trials, so $n = 5$.
- Each trial has four possible outcomes: spade, heart, diamond or club; so $k = 4$.
- In any trial, the probability of drawing a spade, heart, diamond, or club is 0.25, 0.25, 0.25, and 0.25, respectively. Thus, $p_1 = 0.25, p_2 = 0.25, p_3 = 0.25$, and $p_4 = 0.25$.
- The 5 trials produce 1 spade, 1 heart, 1 diamond, and 2 clubs; so $n_1 = 1, n_2 = 1, n_3 = 1, n_4 = 2$.

# Descriptive statistics: Measures of central tendency

Suppose we have data values $x_1, \ldots, x_n$.

# Descriptive statistics: Measures of central tendency

Suppose we have data values $x_1, \ldots, x_n$.

- The sample *mean* is $\bar{x}$ where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + \ldots + x_n).$$

  - Since it's the sum of all data values divided by the sample size, a few extreme data values may largely influence its size.
  - In other words, the mean is not robust against outliers.

# Descriptive statistics: Measures of central tendency

Suppose we have data values $x_1, \ldots, x_n$.

- The sample *mean* is $\bar{x}$ where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + \ldots + x_n).$$

  - Since it's the sum of all data values divided by the sample size, a few extreme data values may largely influence its size.
  - In other words, the mean is not robust against outliers.

- The *median* is defined as the 50th percentile.

  - When the data are symmetrically distributed around the mean, then the mean and the median are equal.
  - Median is robust against outliers.

# Descriptive statistics: Measures of central tendency

Suppose we have data values $x_1, \ldots, x_n$.

- The sample *mean* is $\bar{x}$ where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + \ldots + x_n).$$

  - Since it's the sum of all data values divided by the sample size, a few extreme data values may largely influence its size.
  - In other words, the mean is not robust against outliers.

- The *median* is defined as the 50th percentile.

  - When the data are symmetrically distributed around the mean, then the mean and the median are equal.
  - Median is robust against outliers.

- Robustness is important because biological data are frequently contaminated by extreme or otherwise influential data values.

# Example 1: Measures of central tendency

```
data("golub")
mean(golub[1042, golub.cl==0])
median(golub[1042, golub.cl==0])
```

# Descriptive statistics: Measures of spread

The most important measures of spread are the standard deviation, the interquartile range, and the median absolute deviation.

# Descriptive statistics: Measures of spread

The most important measures of spread are the standard deviation, the interquartile range, and the median absolute deviation.

- The *standard deviation* is the square root of the sample variance, defined as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

# Descriptive statistics: Measures of spread

The most important measures of spread are the standard deviation, the interquartile range, and the median absolute deviation.

- The *standard deviation* is the square root of the sample variance, defined as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

- The *interquartile range* is defined as the difference between the third and the first quartile. See `IQR(x)` in R.

# Descriptive statistics: Measures of spread

The most important measures of spread are the standard deviation, the interquartile range, and the median absolute deviation.

- The *standard deviation* is the square root of the sample variance, defined as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

- The *interquartile range* is defined as the difference between the third and the first quartile. See `IQR(x)` in R.
- The *median absolute deviation* (MAD) is defined as a constant times the median of the absolute deviations of the data from the median.
- The MAD equals the standard deviation in case the data come from a bell-shaped (normal) distribution.

# The Poisson distribution

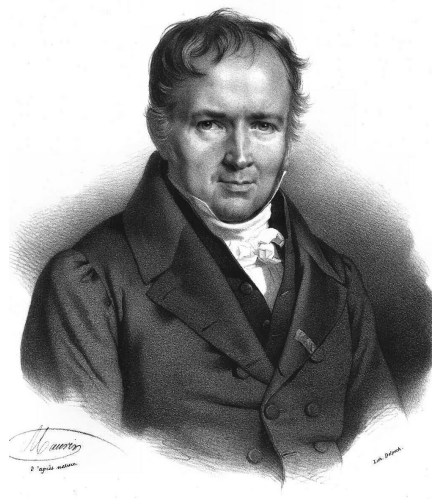


Fig. 1: Simeon Poisson

- When an event is quite rare, like a mutation, the number of times it occurs follows a Poisson distribution. If we only look at one trial, most of the time, we won't see the event occurring.

# The Poisson distribution

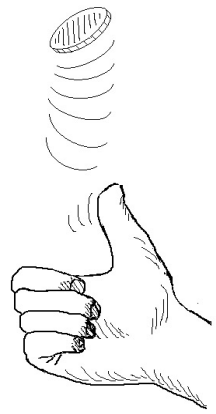
In general, if an event occurs with probability $p$ and and we have $n$ trials, the total number of events is random with a Poisson distribution with mean parameter $\lambda = np$.

```
sum(rpois(1000,lambda=0.02))
```

```
## [1] 19
```

If you do this quite a few times, you'll see the values you get differ, we get a distribution of values centered around 20 or so.

# A real example



Mutations at each position of the RT (reverse transcriptase) gene of HIV follows a Poisson(0.0005) distribution.

- Average number of mutations in the first 10,000 positions?
- Standard error of the above estimate?

# Poisson model for rare events

In the HIV model above, suppose we want to find the probability of seeing 3 mutations in a sequence of length 10,000?

```
dpois(x=3,lambda=5)
```

```
## [1] 0.1403739
```

# Poisson model for rare events

There is a formula:

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

For $\lambda = 5$, we get $P(X = 3) = 5^3 e^{-5}/(3 \times 2 \times 1)$.

```
(5^3*exp(-5))/6
```

```
## [1] 0.1403739
```

# Poisson model for rare events

Ugarte and colleagues report that the average number of goals in a World Cup soccer match is approximately 2.5 and the Poisson model is appropriate.[3]

Because the average event rate is 2.5 goals per match, $\lambda = 2.5$.

$$P(k \text{ goals in a match}) = \frac{2.5^k e^{-2.5}}{k!}$$

$$P(k = 0 \text{ goals in a match}) = \frac{2.5^0 e^{-2.5}}{0!} = \frac{e^{-2.5}}{1} \approx 0.082$$

$$P(k = 1 \text{ goal in a match}) = \frac{2.5^1 e^{-2.5}}{1!} = \frac{2.5 e^{-2.5}}{1} \approx 0.205$$

$$P(k = 2 \text{ goals in a match}) = \frac{2.5^2 e^{-2.5}}{2!} = \frac{6.25 e^{-2.5}}{2} \approx 0.257$$

The table below gives the probability for 0 to 7 goals in a match.

| $k$ | $P(k$ goals in a World Cup soccer match$)$ |
|---|---|
| 0 | 0.082 |
| 1 | 0.205 |
| 2 | 0.257 |
| 3 | 0.213 |
| 4 | 0.133 |
| 5 | 0.067 |
| 6 | 0.028 |
| 7 | 0.010 |

**Once in an interval events: The special case of $\lambda = 1$ and $k = 0$**

Suppose that astronomers estimate that large meteorites (above a certain size) hit the earth on average once every 100 years ($\lambda = 1$ event per 100 years), and that the number of meteorite hits follows a Poisson distribution. What is the probability of $k = 0$ meteorite hits in the next 100 years?

$$P(k = 0 \text{ meteorites hit in next 100 years}) = \frac{1^0 e^{-1}}{0!} = \frac{1}{e} \approx 0.37$$

Under these assumptions, the probability that no large meteorites hit the earth in the next 100 years is roughly 0.37. The remaining 1 – 0.37 = 0.63 is the probability of 1, 2, 3, or more

# Using probabilistic models for epitope detection

- When testing certain pharmaceutical compounds it is important to detect proteins that provoke an allergic reaction, the sites that are responsible for such reactions are called epitopes.
- The technical definition of an epitope: a specific piece of a macromolecular antigen to which an antibody binds.
- An antibody is a type of protein made by certain white blood cells in response to the foreign substance which is called the antigen.
- Each antibody can bind to only a specific antigen.
- The purpose of this binding is to help destroy the antigen.
- Some antibodies destroy antigens directly.
- An epitope (or antigenic determinant), is the part of an antigen that is recognized by the immune system.

# ELISA error model with known parameters

**ELISA** (Enzyme-Linked ImmunoSorbent Assay) detects specific epitopes along proteins.

- The baseline noise level per position is 1% (false positive rate); that is, the probability of declaring a hit (that we have an epitope) when it is not there is 0.01.
- The length of the protein tested is 100 positions.
- We are going to examine a collection of 50 patient samples.

# Results from 50 patients

```
load("e100.RData")
e100
```

```
##   [1] 2 0 1 0 0 0 2 1 0 0 0 0 0 0 0 0 0 1 1 1 0 0 1 0 1 0 0 0 1 0 0 0 0 1 0
##  [36] 1 1 0 1 2 2 7 1 0 2 0 1 0 1 1 1 1 0 1 0 0 0 0 1 2 2 1 0 0 0 0 0 1 0 0
##  [71] 0 1 1 0 0 1 1 0 0 0 0 1 0 0 0 1 0 1 0 0 1 1 0 0 1 1 0 0 1 0
```

Here 1 signifies a reaction or a hit and the zeros signify no reaction at that position. A number greater than 1 indicates a hit appeared in multiple patients.

# Results from 50 patients

- If there are no epitope, the counts follow a Bernoulli distribution with probability 0.01. Each individual position of each patient has a probability of 1 in 100 of being 1.
- For each position, after seeing 50 patients, we expect the sum to have a Poisson distribution with parameter 0.5.

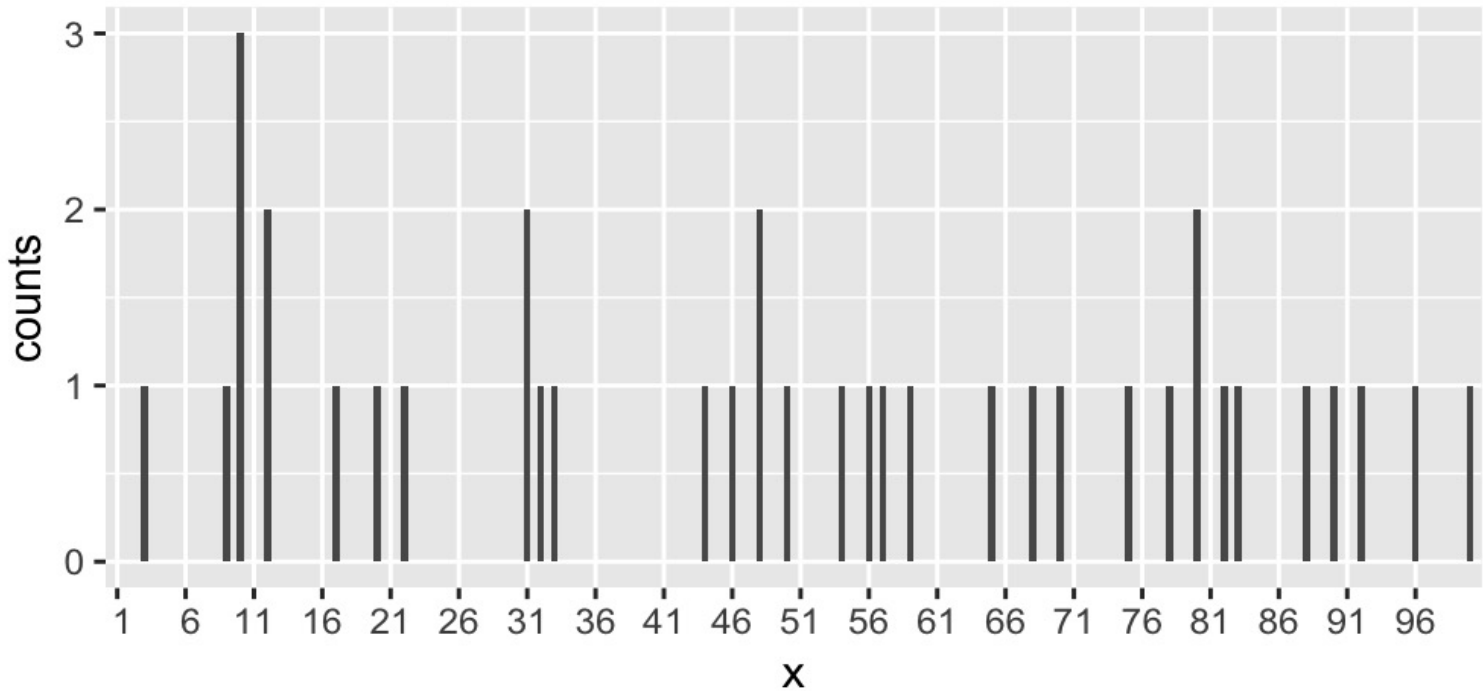
# Question

Run a little simulation experiment to show that looking at the sum of 50 Bernoulli(0.01) variables is a good approximation to one Poisson(0.5) random variable.

```
x50=sum(rpois(50,0.01))
x50

set.seed(5427121)
m50=matrix(rbinom(5000,1,0.01),ncol=100,nrow=50)
s100=apply(m50,2,sum);table(s100)

t100=rpois(100,0.5);table(t100)

c(mean(s100), mean(t100))
```

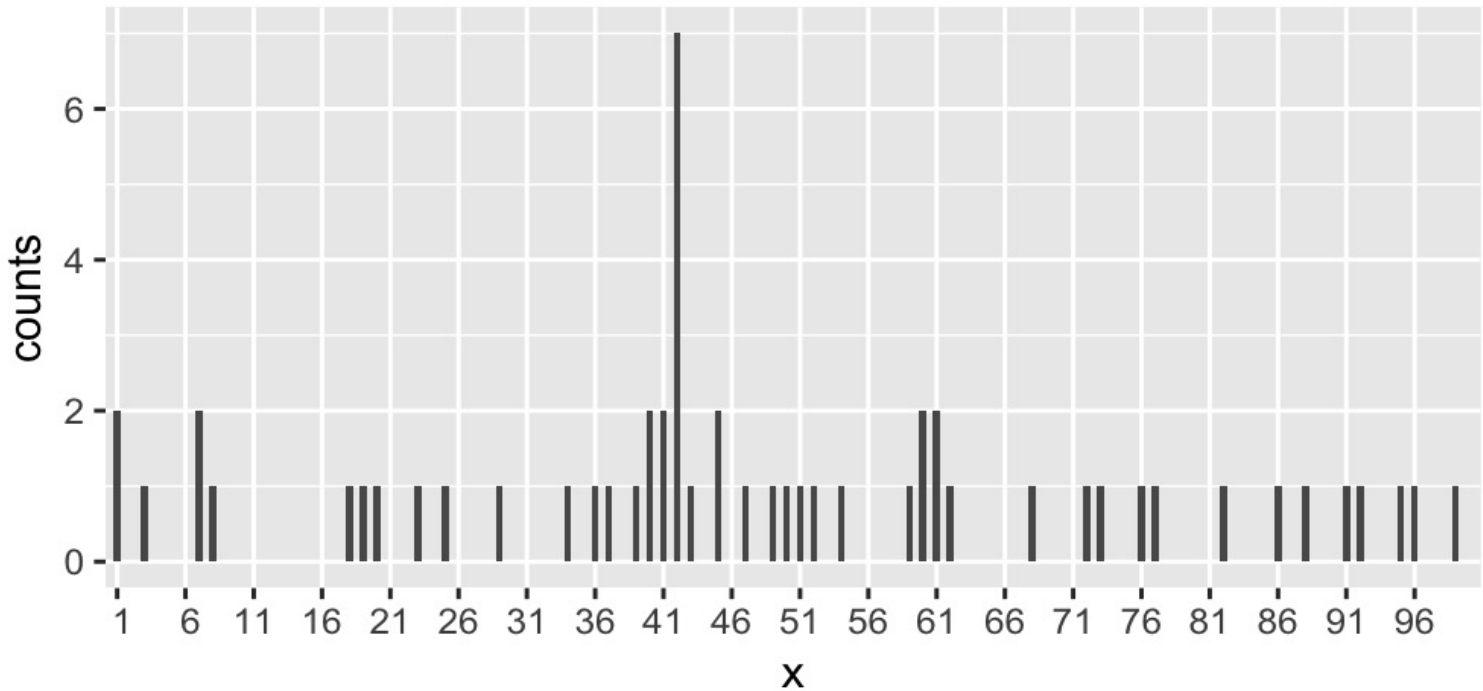
# Simulated data

Plot of 100 draws from a Poisson(0.5)



# Actual data

Suppose that the actual data (the output of Elisa array on 100 positions of 50 patients) we see is




We see a surprising spike.

# What are the chances of seeing a value as large as 7?

The probability of seeing a number at least as large as 7 when considering one Poisson(0.5) random variable is

$$P(X \geq 7) = 1 - P(X \leq 6) = 1 - F_{Pois}(6).$$

Call this number $\epsilon$. How big is it?

```
round(ppois(0:6,0.5),5)
```

```
## [1] 0.60653 0.90980 0.98561 0.99825 0.99983 0.99999 1.00000
```

```
1-ppois(6,0.5)
```

```
## [1] 1.00238e-06
```

# Is this right?

**No.**

- We looked at all 100 positions and 7 was chosen because it was the maximum, so we have to ask ourselves what are the chances of seeing a number as large as 7 **in 100 trials.**
- We use extreme values: order the values $X_1, X_2, \ldots, X_n$ and rename them $X_{(1)}, X_{(2)}, \ldots, X_{(100)}$ so that $X_{(100)}$ denotes the maximum.
- For the largest to be as large as 7 is the **complementary event** (opposite) of having all 100 counts smaller or equal to 6.

# The probability of having all 100 counts smaller or equal to 6

$$P(X_{(100)} \geq 7) = 1 - P(X_{(100)} \leq 6) = 1 - \prod_{i=1}^{100} P(X_i \leq 6)$$

$$= \left( \sum_{k=0}^{6} \frac{e^{-\lambda} \lambda^k}{k!} \right)^{100} = \mathrm{ppois}(6, \lambda)^{100}.$$

# Estimating small numbers & tail probabilities

Do not run this as it takes quite a lot of time and memory (on a 4GB machine this computation took about an hour, whereas the analytic calculation took 2 minutes.)

```
M100=rpois(1000000000,0.5)
table(M100)
Matrix100=matrix(M100,ncol=100)
vecmaxes=apply(Matrix100,1,max)
table(vecmaxes)
```

giving an approximation of $9.48 \times 10^{-5}$ for $P(X_{max} \geq 7)$ and $3 \times 10^{-7}$ for $P(X_{max} \geq 9)$.

# Estimating small numbers & tail probabilities

We can however prove that a probability is smaller than 0.000001 by sampling just $10^6$ times.

```
million=rpois(1000000,0.5)
mat100=matrix(million,ncol=100)
maxes100=apply(mat100,1,max)
table(maxes100)
```

```
## maxes100
##    2    3    4    5    6
## 2310 6070 1454  153   13
```

We can only conclude that $P(X_{max} \geq 7) \leq 10^{-6}$.

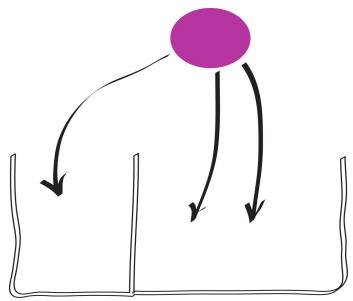
Although it is doable to find small probabilities by simulation (Monte Carlo) it is not always optimal.

# Conclusion for this study

- We postulated the Poisson distribution for the noise and were able to conclude through mathematical deduction.
- Everything we have done up to now uses the probabilistic generative model.
- Now suppose that we knew the number of patients, the length of the proteins, and we observed the data from an unknown distribution, then we would have to use **statistical modeling** which will be developed in the next lecture.

# Binomial success counts

- When we have a binary outcome, e.g. success/failure, CpG/NonCpG, M/F, diseased/healthy, true/false.
- We can model it as a simple random variable with probability of success equal to $p$ (and the probability of failure $1 - p$).
- A sequence of trials SSSSSFSSSSFFFSF is summarized as (#Success=10,#Failures=5).
- A possible generative model: the number of successes follows a binomial distribution with parameters *prob* $= 2/3$ and *size* $= 15$.



# Using R to explore the binomial distribution

Suppose we want to simulate a sequence of fifteen fair coin tosses. We can write

```
rbinom(15,prob=0.5,1)
```

```
##  [1] 0 0 1 1 0 0 1 1 1 0 0 1 0 0 1
```

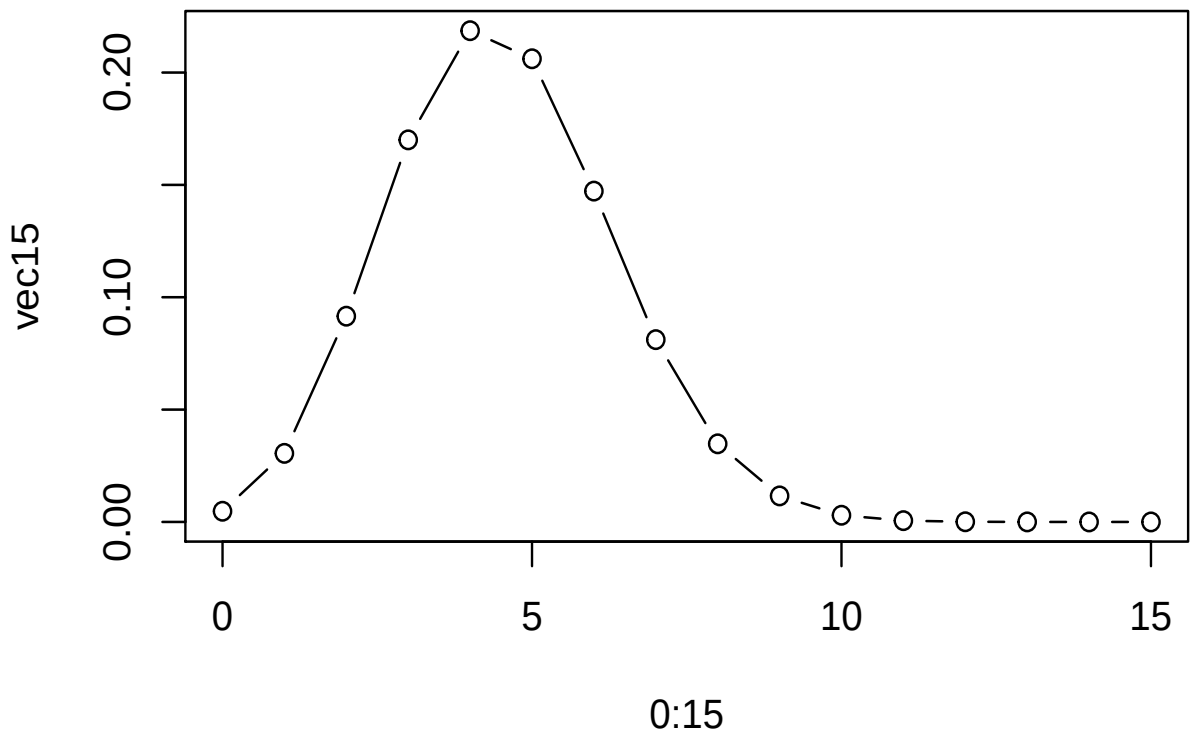
# More examples on `rbinom`

```
set.seed(235569511)
rbinom(1,prob=0.3,15)
rbinom(1,prob=0.3,15)
rbinom(1,prob=0.3,15)
rbinom(1,prob=0.3,15)
rbinom(1,prob=0.3,15)
```

# Using R to explore the binomial distribution

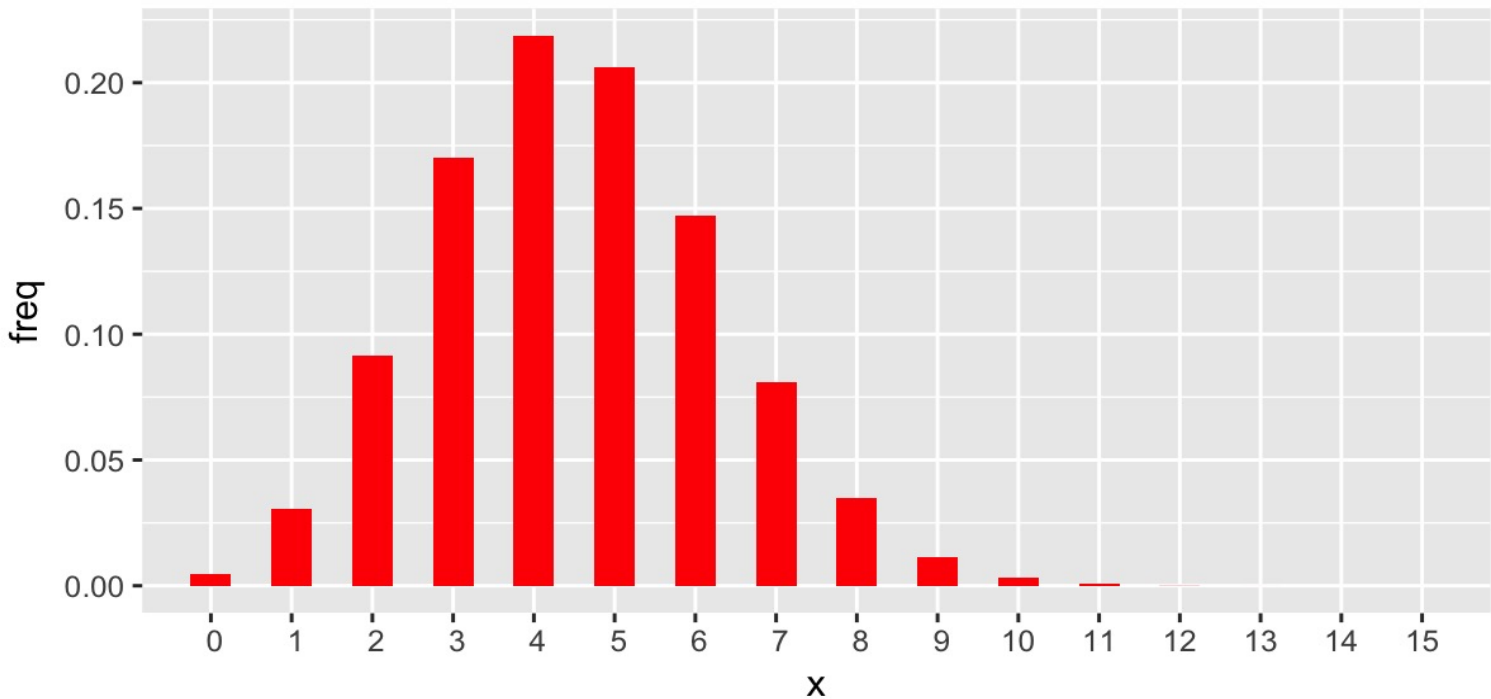
Notice that 5 appears quite often. The theoretical proportion of the occurrences of 5 is the value of the probability of $X = 5$ where $X$ is a binomial with $n = 15, p = 0.3$ and its probability mass function is

```
vec15=dbinom(0:15,prob=0.3,15)
plot(0:15,vec15,type='b')
```



# Using R to explore the binomial distribution

Here's a plot of the theoretical distribution of Binomial(15,0.3)



# Using R to explore the binomial distribution

- The parameter $n$ is the total number of events.
- The parameter $p$ is the probability of a single event being success.

The probability of seeing 5 successes out of $n = 15$ events with $p = 0.3$ is

```
dbinom(5,p=0.3,15)
```

```
## [1] 0.2061304
```

There is a closed form formula:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

# Multinomial events

- Several possibilities for levels: (AA, Aa, aa).
- Number of levels in a categorical variable can be very large.
- If we are only measuring one categorical variable, we usually tally the frequencies of the different levels in a vector of counts.

```
genotype1=c("AA","AO", "BB", "AO", "OO", "AO", "AA", "BO", "BO", "AO",
"BB","AO", "BO", "AB", "OO", "AB", "BB", "AO", "AO")
table(genotype1)
```

```
## genotype1
## AA AB AO BB BO OO
##  2  2  7  3  3  2
```

# Multinomial events

We encode these variables as factor variables

```
genotype=factor(genotype1)
genotype
```

```
##  [1] AA AO BB AO OO AO AA BO BO AO BB AO BO AB OO AB BB AO AO
## Levels: AA AB AO BB BO OO
```

# Multinomial events

Example with 4 categories:

```
set.seed(1)
rmultinom(1,prob=c(3/4,1/12,1/12,1/12),size=1)
```

```
##      [,1]
## [1,]    1
## [2,]    0
## [3,]    0
## [4,]    0
```

# Multinomial events

We could have replaced four draws with one draw of 4:

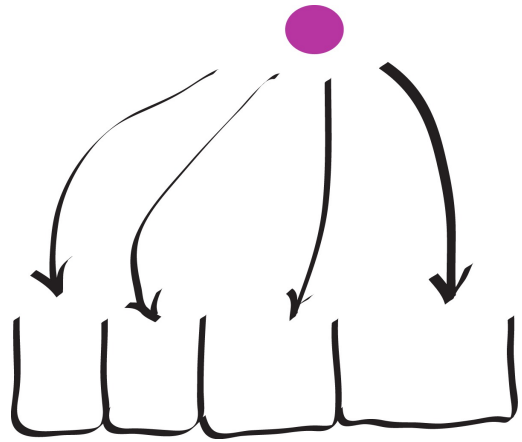
```
rmultinom(1,prob=c(3/4,1/12,1/12,1/12),size=4)
```

```
##      [,1]
## [1,]    3
## [2,]    0
## [3,]    1
## [4,]    0
```

```
rmultinom(4,prob=c(3/4,1/12,1/12,1/12),size=1)
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    0    1    1
## [2,]    0    1    0    0
## [3,]    0    0    0    0
## [4,]    0    0    0    0
```

# Multinomial distributions: the case of DNA



Just as in the binomial case the sum of the probabilities of all possible outcomes is 1:

$$p_A + p_C + p_G + p_T = 1.$$

# Multinomial distribution: the formula

Here is the formula which computes the probability of a multinomial vector of counts $(x_1, \ldots, x_m)$ being observed:

$$P(x_1, x_2, \ldots, x_m \mid p_1, \ldots, p_m) = \frac{n!}{\prod x_i!} \prod p_i^{x_i}$$
$$= \binom{n}{x_1, x_2, \ldots, x_m} p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m}.$$

# Important things about generative models for discrete data

- Several ways to model count data: *Poisson*, *binomial* or *multinomial*.
- Can generate random discrete data using the specialized R functions tailored for each type of distribution.
- The epitope example showed us how to use a probability model to compute the probability of an event under a parametric model when the parameters are known.
- P-values are probabilities under certain assumptions.
- Simulations under known models help design better experiments.

# R notes

- All known distributions can be used to simulate data using the functions `rXXXX` where `XXXX` could be `pois`, `binom`, `multinom`.
- If we need a theoretical computation of a probability under one of these models, we would use the functions `dXXXX`, such as `dpois` (what is the difference between `dpois` and `ppois`?)
- This also works for distributions that are not counts, such the normal distribution (`rnorm`, `dnorm`), the $\chi^2$ distribution (`rchisq`, `dchisq`) or the exponential (`rexp`, `dexp`).

# How to write R functions

R functions are useful for generalizing useful sequences of commands.

```
#Argument by default for the Poisson take 0.5, for protein length take n=100
#Function to compute the probability of having a maximum out of n as big as max
pmax=function(lam=0.5,n=100,max=7){
  epsilon=1-ppois(max-1,lam)
  proba=1-exp(-n*epsilon)
  return(proba)
}
pmax(lam=0.5)
```

```
## [1] 0.0001002329
```

```
pmax(lam=mean(e100))
```

```
## [1] 0.0001870183
```