USC Department Of Translational Genomics
Keck Medicine of USC

TRGN 527: Applied Data Science and Bioinformatics

UNIT I. Introduction and Basic Data Science

Week 3 - Lecture 2

Enrique I. Velazquez Villarreal, M.D., Ph.D., M.P.H., M.S. | **Assistant Professor**

Dept. of Translational Genomics

USC | Keck School of Medicine | Norris Comprehensive Cancer Center

Leader of the USC Bioinformatics Core – *USC CaRE2 Health Equity Center*

David W. Craig, Ph.D. | **Professor and Vice Chair**

Dept. of Translational Genomics

USC | Keck School of Medicine | Norris Comprehensive Cancer Center

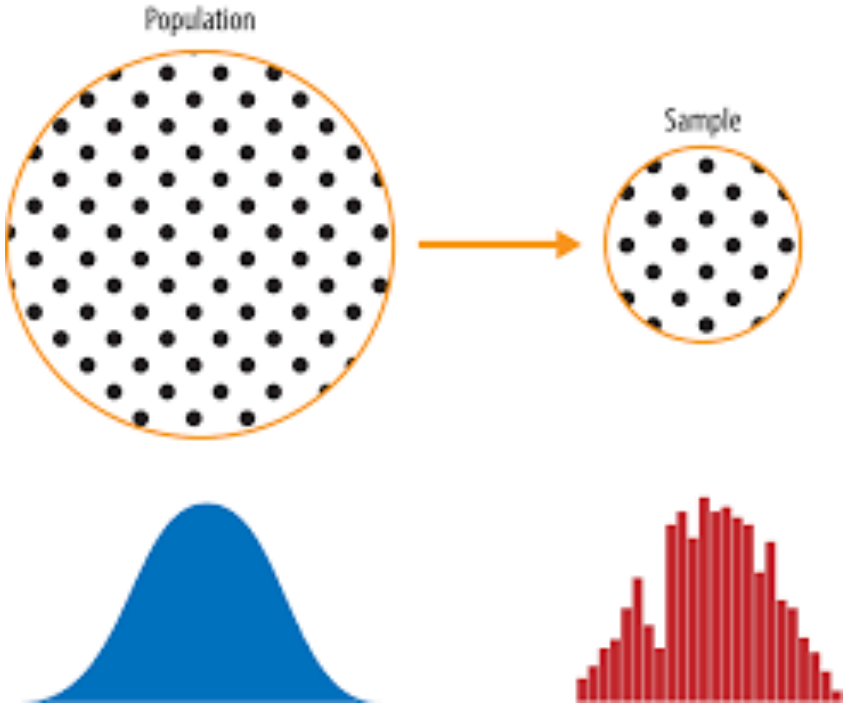
Co-Director, Institute of Translational Genomics

Topics

- Distributions, sampling distributions

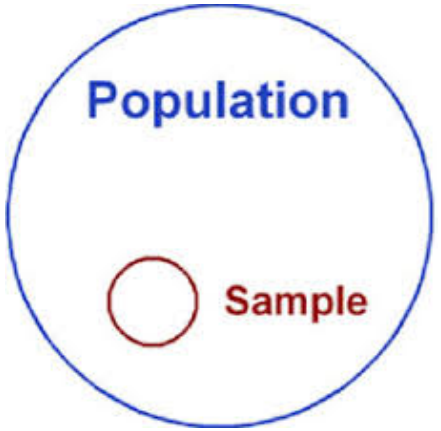
Data distributions

- In general data scientist need not worry about the theoretical nature of the Population, and instead should focus on the sampling procedures and the data at hand.



Key Terms for Random Sampling

- Sample
  - A subset from a larger data set.
- Population
  - The larger data set or idea of a data set.
- N(n)
  - The size of the population (sample).
- Random sampling
  - Drawing elements into a sample at random



Key Terms for Random Sampling

- Stratified sampling
  - Dividing the population into strata and randomly sampling from each strata.
- Simple random sample
  - The sample that results from random sampling without stratifying the population.
- Sample bias
  - A sample that misrepresents the population.
- Bias
  - Systematic error.

Key Terms for Random Sampling

- Example of Random Variables:
  - Genotypes of a bi-allelic gene.
    - Sample space, S = {*AA*, *Aa*, *aa*}
    - Various events, such as the homozygous event H M = {*AA*, *aa*}
  - Probability distributions
    - Random variable X assigns a numerical value to each possible outcome (and event) of a random phenomenon.
    - Below X can be defined based on possible genotypes of a bi-allelic gene **A** as follows:

$$X = \begin{cases} 0 & \text{for genotype } AA, \\ 1 & \text{for genotype } Aa, \\ 2 & \text{for genotype } aa. \end{cases}$$

    - Above, the random variable assigns 0 to the outcome *AA*, 1 to the outcome *Aa*, and 2 to the outcome *aa*.

Key Terms for Random Sampling

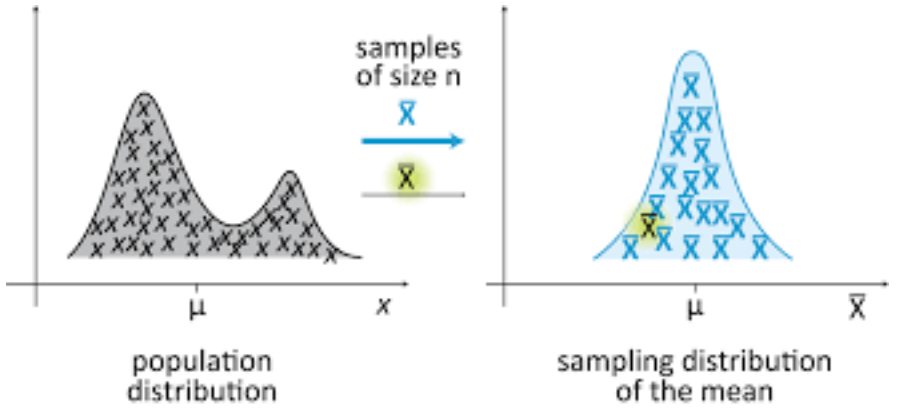
- Example of Random Variables:
  - For the previous example, we can define another random variable Y as follows:

$$Y = \begin{cases} 0 & \text{for genotypes } AA \text{ and } aa, \\ 1 & \text{for genotype } Aa. \end{cases}$$

  - Above, Y assigns 0 to the homozygous event and assigns 1 to the heterozygous event.

Sampling Distribution of a Statistic

- Refers to the distribution of some sample statistic, over many samples drawn from the same population.
- Much of classical statistics is concerned with making inference from (small) samples to (very large) populations.



Probability Distribution of a random variable

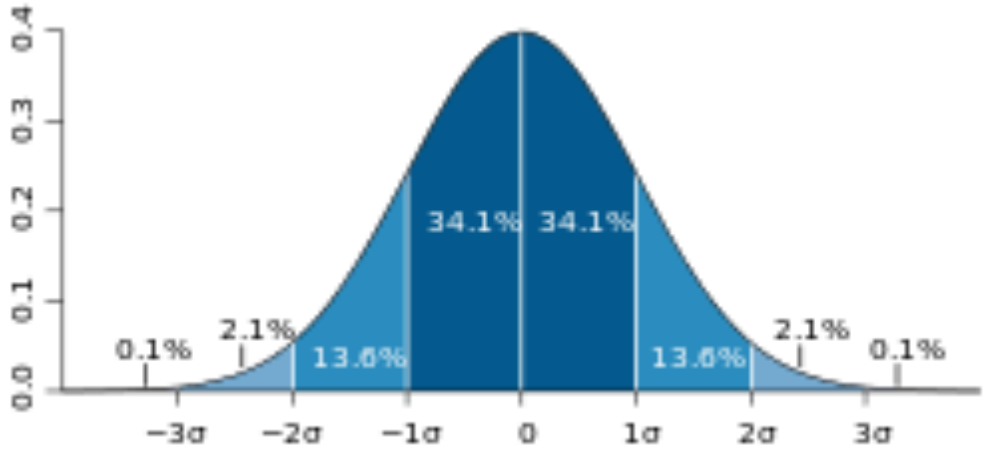
- Provides the required information to find the probability of its possible values.
- Recall that the total probability of all possible values is equal to 1.
- Concern about all the possible values a random variable can take and their corresponding probabilities (i.e., the chance of observing those values) as opposed to a sample of observations.

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N},$$

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N},$$

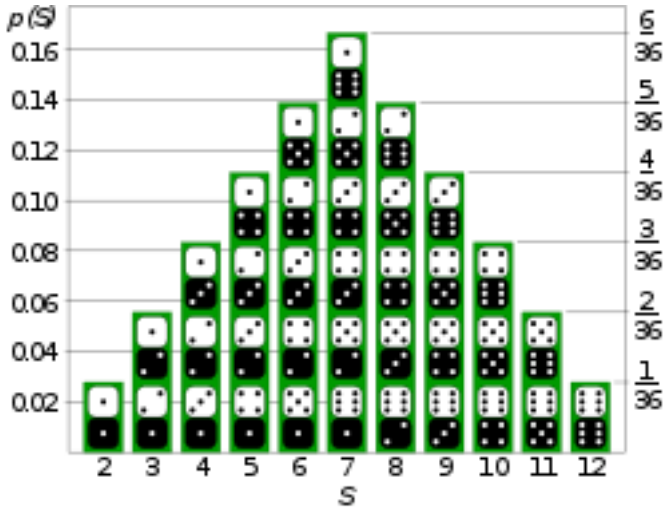
Probability Distribution of a random variable

- It is easier to discuss this concept with a population in mind.
- The mean and variance here refer to the population mean and population variance (theoretically).
- The probability distribution of a random variable specifies its possible values (i.e., its range) and their corresponding probabilities.



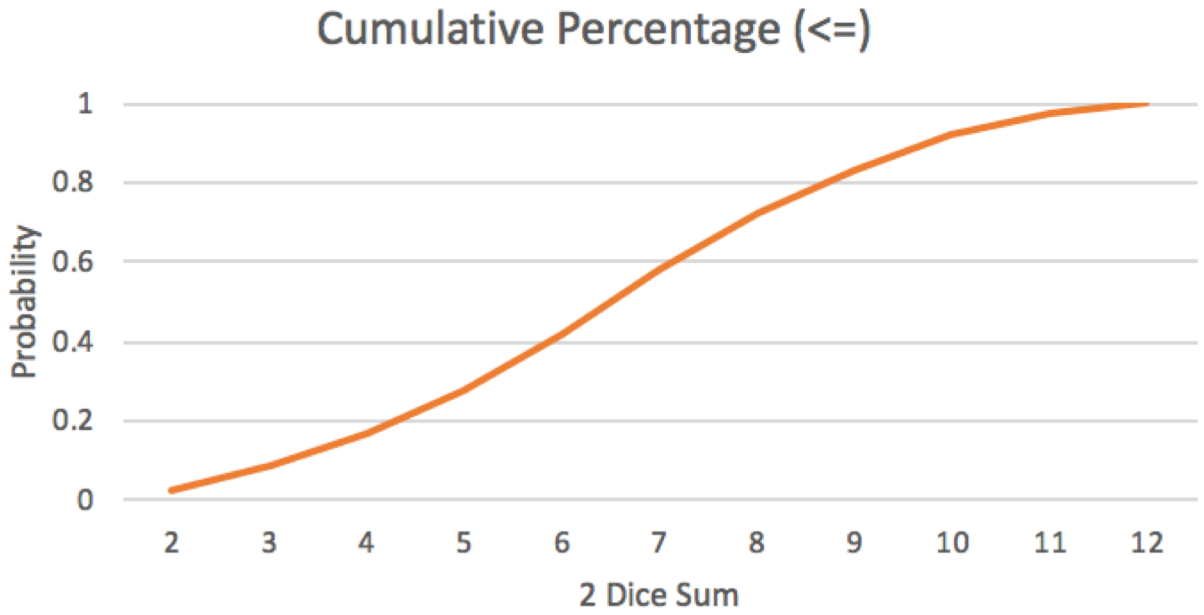
Probability Distribution of a random variable

- A joint probability distribution is a description of all the possibilities and the probability that they will occur.
- If I role a six sided dice, I have an equal probability (16.6%) of getting a 1, as is the case for a 2,3,4,5,6.
- Its often characterized by a **Probability Mass Function (PMF)** which provides the probability of each event.
- If I role a six sided dice, what is the probability that I get a 7?



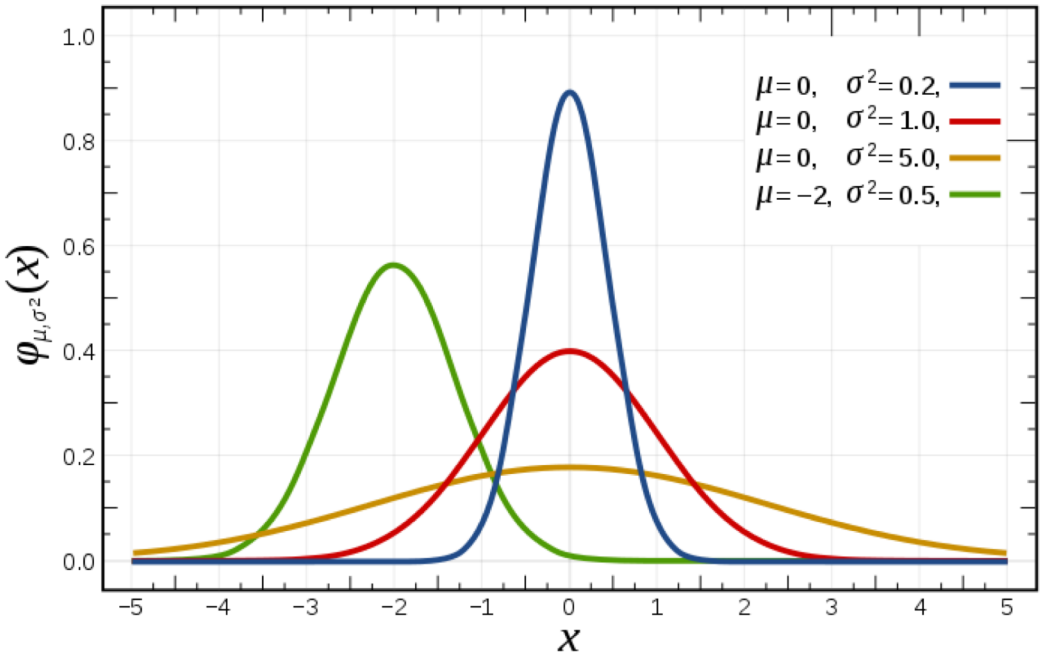
Probability Distribution of a random variable

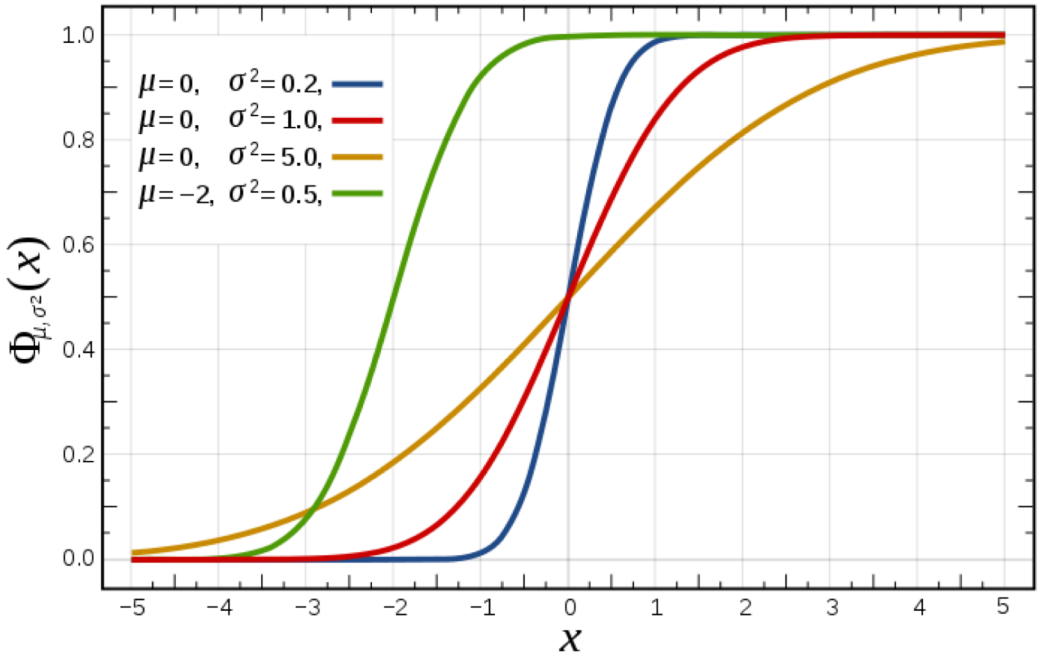
- Another key term:
  - Cumulative Probability Function (CMF), which is the sum that event or lower will have occurred.



Probability Distribution of a random variable

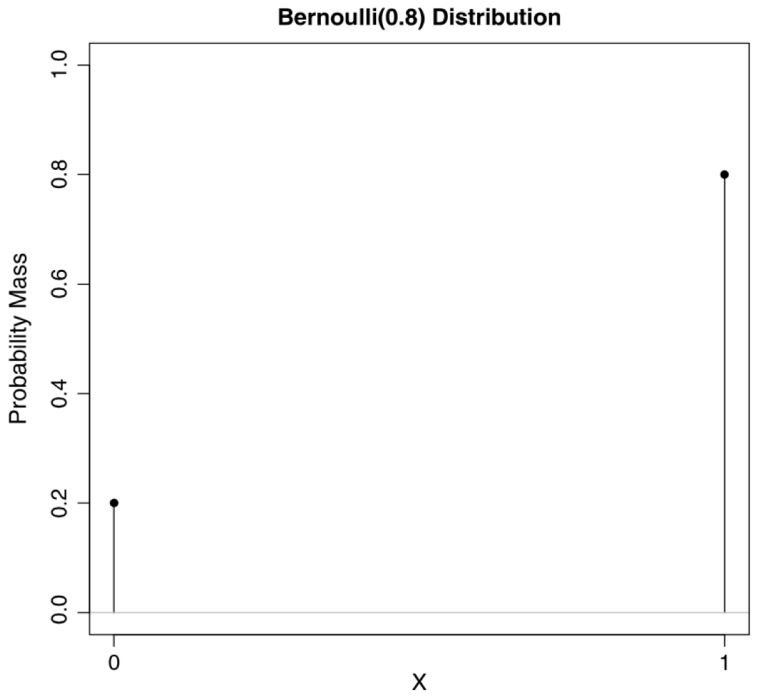
- There are a few probability distributions we care a lot about in bioinformatics.
  - The normal distribution (or Gaussian) and binomial distribution (binary events, such as a coin toss).
  - Normal distributions are common in biology – but cannot always be presumed.
    - A normal distribution is famously represented as a bell curve.





Probability Distribution of a random variable

- Bernoulli Distribution:
  - Binary random variables are abundant in scientific studies.
  - Examples include disease status (healthy and diseased), gender (male and female), survival status (dead, survived), and a gene with two possible alleles (A and a).
  - We usually regard one of the values as the outcome of interest and denote it as X = 1.
  - The other outcome is denoted as X = 0.
  - The probabilities for all possible values sum to one: P (X = 0) + P (X = 1) = 1.



$$X \sim \text{Bernoulli}(0.8).$$

$$P(X = x) = \begin{cases} 0.2 & \text{for } x = 0, \\ 0.8 & \text{for } x = 1. \end{cases}$$

Probability Distribution of a random variable

- Example:
  - Assume that we want to examine 10 people for a disease that has probability of 0.2 in the population of interest.
  - The number of people (out of 10) who are affected, denoted as Y, has Binomial(10, 0.2) distribution. Let us first simulate five random samples from this distribution (i.e., examine five groups each with 10 people)
    - The first argument to the rbinom() functions specifies the number of random samples.
    - The size option is the number of Bernoulli trials (here, n = 10), and the prob option is the probability for the outcome of interest.
    - Each randomly generated number represents the number of people affected by the disease out of 10 people.

````
10 ## Distributions:
11 # Binomial Distribution
12 ```{R}
13 rbinom(5, size = 10, prob = 0.2)
14 ```
````

Probability Distribution of a random variable

- Example:
  - From last example:
    - If we set size=1:
      - it will be simulating random samples from the corresponding Bernoulli distribution.
        - For example, it can simulate the disease status for a group of 10 people:

````
13 ```{R}
14 rbinom(10, size = 1, prob = 0.2)
15 ```

     [1] 0 0 0 1 0 0 0 0 1 0

16
````

Probability Distribution of a random variable

- Example:
  - Now suppose that it is required to know the probability of observing 3 out of 10 people affected by the disease: P (X = 3).
  - Then it needs the probability mass function **dbinom( )**, which returns the probability of a specific value:

```
18 ```{R}
19 dbinom(3, size = 10, prob = 0.2)
20 ```
```

```
[1] 0.2013266
```

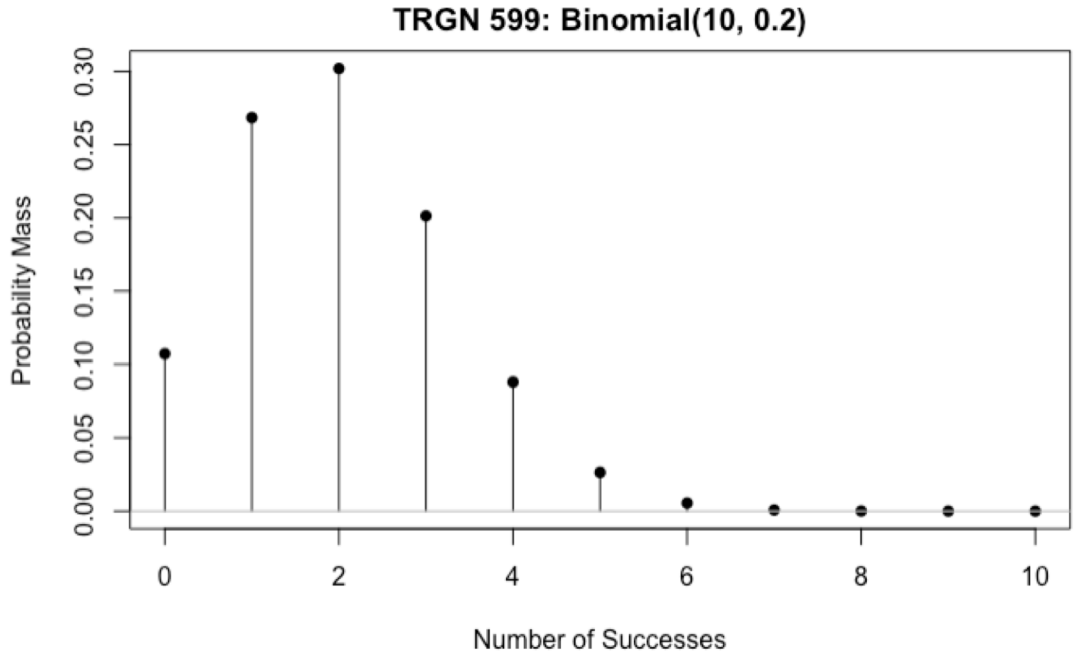
  - Along with the value of the random variable, 3, the other arguments of the dbinom() function are the number of Bernoulli trials (size=10) and the probability (prob=0.2) for the event of interest.

Probability Distribution of a random variable

- Example:
  - We can also create a vector x of the possible values of X and then use this vector as input to dbinom() function:

```
21
22 ```{R}
23 x <- 0:10
24 Px <- dbinom(x, size = 10, prob = 0.2)
25 round(Px, 2)
26
27 plot(x, Px, type = "h", xlab = "Number of Successes",
28      ylab = "Probability Mass",
29      main = "TRGN 599: Binomial(10, 0.2)")
30 points(x, Px, pch = 16)
31 abline(h = 0, col = "gray")
32 ```
```


```
[1] 0.11 0.27 0.30 0.20 0.09 0.03 0.01 0.00 0.00 0.00 0.00
```



Probability Distribution of a random variable

- Now suppose that we are interested in the probability of observing three or fewer affected people in a group of 10.
- We could of course sum the values of pmf:
  - $P(Y \leq 3)=P(Y=0)+P(Y=1)+P(Y=2)+P(Y=3)$.
- However, it is easier to use the *cumulative distribution function* for a binomial random variable, pbinom(), to obtain the lower tail probability:

````
33 ```{R}
34 pbinom(3, size = 10, prob = 0.2, lower.tail = TRUE)
35 ```

[1] 0.8791261
````

- the arguments size=10 and prob=0.2 specify the parameters of the binomial distribution.
- The option lower.tail=TRUE tells R to find the lower tail probability.
- By changing the lower.tail option to false (FALSE), we can find the upper tail probability P (Y > 3).

Probability Distribution of a random variable

- On the other hand, to obtain the 0.879 quantile,
  - we use the qbinom() function:

```{R}
qbinom(0.879, size = 10, prob = 0.2, lower.tail = TRUE)
```

```
[1] 3
```

Discrete Probability Distributions

- For discrete random variables:
  - The probability distribution is fully defined by the **probability mass function (pmf)**.
    - This is a function that specifies the probability of each possible value within range of random variable.
    - Example: using genotypes, the pmf of the random variable X is:

$$P(X = x) = \begin{cases} 0.49 & \text{for } x = 0, \\ 0.42 & \text{for } x = 1, \\ 0.09 & \text{for } x = 2. \end{cases}$$

$$X = \begin{cases} 0 & \text{for genotype } AA, \\ 1 & \text{for genotype } Aa, \\ 2 & \text{for genotype } aa. \end{cases}$$

    - The probabilities for all possible values of the random variable sum to one.

Discrete Probability Distributions

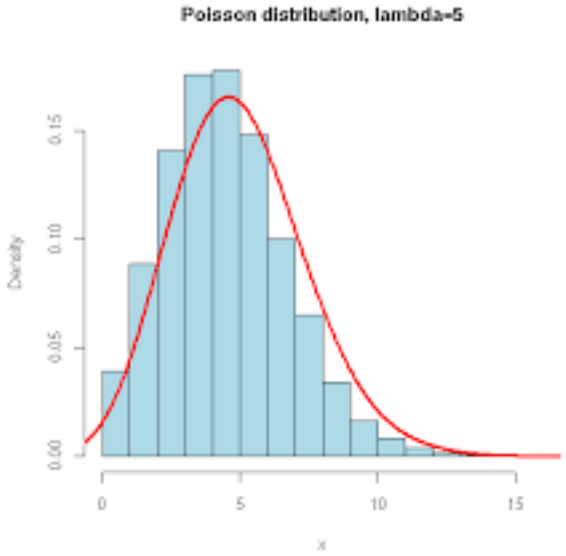
- Example:
  - Suppose Y is a random variable that is equal to 1 when a newborn baby has low birthweight, and is equal to 0 otherwise.
  - Below, Y is a *binary*

$$P(Y = y) = \begin{cases} 0.7 & \text{for } y = 0, \\ 0.3 & \text{for } y = 1. \end{cases}$$

$$Y = \begin{cases} 0 & \text{for genotypes } AA \text{ and } aa, \\ 1 & \text{for genotype } Aa. \end{cases}$$

Poisson Distribution

- From historical data we can estimate the average number of events per unit of time or space, but we might also want to know how different this might be from one unit of time/space to another.
- The Poisson distribution tells us the distribution of events per unit of time or space when we sample many such units.
- It is useful when addressing queuing questions like “How much capacity do we need to be 95% sure of fully processing the internet traffic that arrives on a server in any 5-second period?”

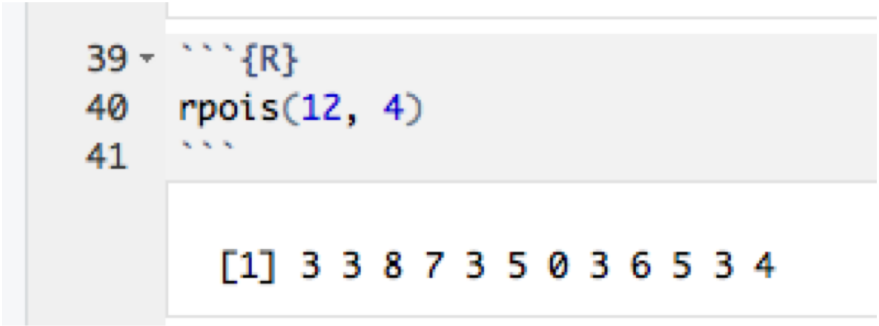


Poison Distributions

- Random variables representing counts within temporal and/or spatial limits but without pre-specified upper limits are often assumed to have **Poisson** distributions.
- The range of these variables is the set of all nonnegative integers (i.e., the lower limit is zero, but there is no upper limit).
- A Poisson distribution is specified by a parameter λ, which is interpreted as the rate of occurrence within a time period or space limit.
- We show this as $X \sim \text{Poisson}(\lambda)$, where λ is a positive real number ($\lambda > 0$).
- The mean and variance of a random variable with Poisson(λ) distribution are the same and equal to λ.
- That is, $\mu = \lambda$ and $\sigma_2 = \lambda$.

Poison Distributions

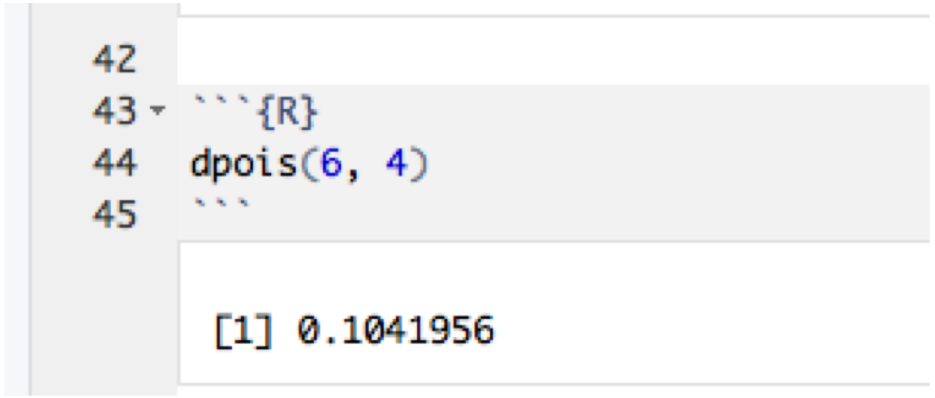
- *Example:*
  - *Poisson Distribution*
    - Suppose that on average 4 people visit the hospital each hour.
    - Then we can represent the hourly number of hospital visitation as X ~ Poisson(4) and simulate 12 samples from this distribution:



    - These randomly generated numbers can be regarded as the number of people visiting the hospital at different hours.
    - Similar to the rbinom() function, the first parameter to the rpois() function is the number of samples, and the remaining argument specifies the distribution parameter.

Poison Distributions

- *Example:*
  - *Poisson Distribution*
    - Suppose that we want to know the probability that six people visit the hospital in an hour.
    - Then we would use the probability mass function dpois():



```
42
43 ```{R}
44 dpois(6, 4)
45 ```

[1] 0.1041956
```

    - Here, 6 is the specific value of the random variable, and 4 is the distribution parameter.

Poison Distributions

- *Example:*
  - *Poisson Distribution*
    - We can create a plot of the pmf by first creating a vector of possible values and finding their corresponding densities.
    - To find the probability of six or fewer people visiting the hospital (as opposed to the probability that exactly six people visit), we need to find the lower tail probability of x = 6.
    - For this, we use the ppois() function:

```
47
48 ```{R}
49 ppois(6, 4)
50 ```
```

```
[1] 0.889326
```

    - The 0.889 quantile of the distribution is:

```
52
53 ```{R}
54 qpois(0.889, 4)
55 ```
```

```
[1] 6
```

Key Terms for Poisson Distribution and Related Distributions

- Lambda
  - The rate (per unit of time or space) at which events occur.
- Poisson distribution
  - The frequency distribution of the number of events in sampled units of time or space.
- Exponential distribution
  - The frequency distribution of the time or distance from one event to the next event.
- Weibull distribution
  - A generalized version of the exponential, in which the event rate is allowed to shift over time.

Binomial Distributions

- A sequence of binary random variables $X_1, X_2, \ldots, X_n$ is called **Bernoulli trials** if they all have the same Bernoulli distribution (i.e., the same probability θ for the outcome of interest) and are independent (i.e., not affecting each other's probabilities).
- For example,
  - Suppose that we plan to recruit a group of 50 patients with breast cancer and study their survival within five years from diagnosis.
  - We represent the survival status for these patient by a set of Bernoulli random variables $X_1, \ldots, X_{50}$.
    - For each patient, the outcome is either 0 or 1.
    - Assuming that all patients have the same survival probability, $\theta = 0.8$, and the survival status of one patient does not affect the probability of survival for another patient, $X_1,\ldots,X_{50}$ form a set of 50 Bernoulli trials.

Binomial Distributions

- Now we can create a new random variable Y representing the number of patients out of 50 who survive for five years.
- The number of survivals is the number of 1s in the set of Bernoulli trials. This is the same as the sum of Bernoulli trials, whose values are either 0 or 1:

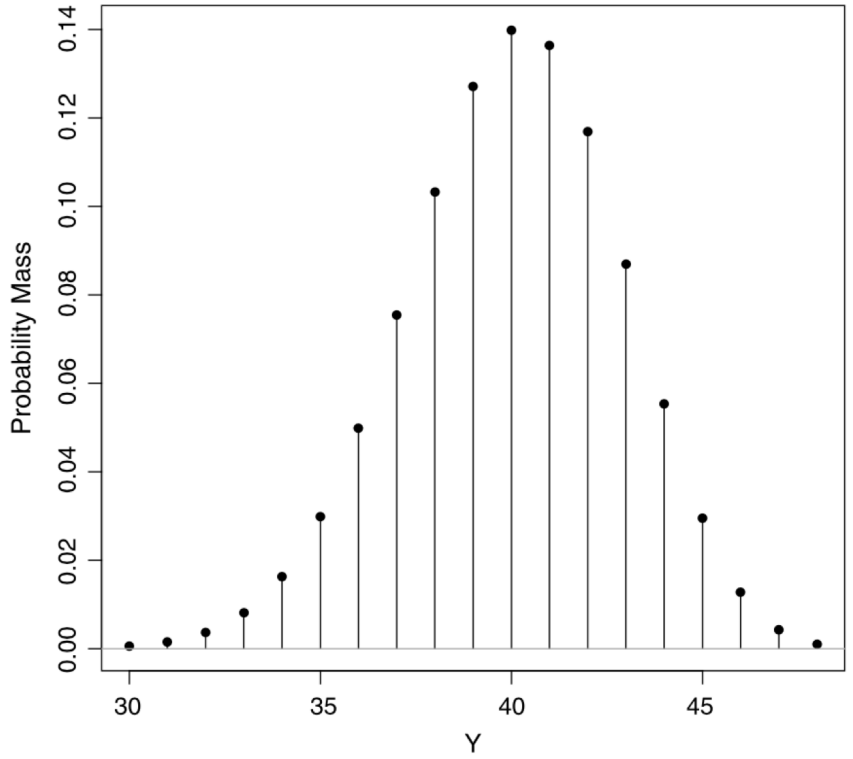
$$Y = \sum_{i}^{n} X_i,$$

  - where Xi = 1 if the ith patient survive and Xi = 0 otherwise.
- Y can be any integer number from 0 (no one survives) through 50 (everyone survives)
  - Y range is {0, 1, . . . , 50} a countable set, discrete.
  - The distribution of Y is a **binomial** distribution, shown as: $Y \sim \text{Binomial}(50, 0.8).$

Binomial Distributions

- The pmf of a binomial(n, θ) specifies the probability of each possible value (integers from 0 through n) of the random variable.
- For example, the pmf of Binomial(50, 0.8) distribution specifies the probability of 0 through 50 survivals:

$$Y \sim \text{Binomial}(50, 0.8).$$



Binomial Distribution

- Binomial outcomes are important to model, since they represent, among other things, fundamental decision (Yes/No).

- A binomial trial is an experiment with two possible outcomes: one with probability p and the other with probability 1-p.

- With large n, and provided p is not too close to 0 or 1, the binomial distribution can be approximated by the normal distribution.

Binomial Distribution

- Yes/No (binomial) outcomes lie at the heart of analytics since they are often the culmination of a disicion or other process.
  - buy / don’t buy
  - Click / don’t click
  - Survive / die

- For example:
  - Flipping a coin 10 times is a binomial experiment with 10 trials
    - Each trial having two possible outcomes (heads or tails)

Key Terms of Distribution of a Statistic

- Sample statistic
  - A metric calculated for a sample of data drawn from a larger population.
- Data distribution
  - The frequency distribution of individual values in a data set.
- Sampling Distribution
  - The frequency distribution of a sample statistic over many samples or resamples.