Normal distributions and an introduction to statistical inference

Outline for today

Review of the binomial distribution

The normal distribution

Beginning statistical inference (Bayesian inference)

Announcement: class final projects

Final project proposal was due yesterday

If you have not turned it in yet, turn it in ASAP

Review of probability models

Probability is a way of measuring the uncertainty of the outcome of an event

A **random variable (X)** is a random number

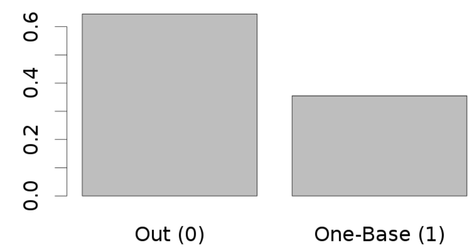
- Notation: Pr[a < X < b] is the probability X is between a and b
- $0 \leq \Pr(X = x) \leq 1$
- $\Sigma \Pr(X = x) = 1$

A **Probability mass functions** tell the probability of each outcome in the *sample space* of a *discrete distribution*

- Definition of a sample space?
- Definition of a discrete distribution?

Bernoulli Distribution

Models the probability of two outcomes:



- Success: X = 1, Failure: X = 0
- Examples?

Model has one parameter π, which is the probability of getting a success

- i.e., Pr(X = 1) = π

Bernoulli probability is given by the mass function:

$$Pr(X = k) = f(k; \pi) = \begin{cases} \pi & \text{if } k = 1 \\ 1 - \pi & \text{if } k = 0 \end{cases}$$

$$f(k; \pi) = \pi^k (1 - \pi)^{1-k} \quad \text{for } k \in \{0, 1\}$$

Binomial distribution

Probability of getting ***k*** successes out of ***n*** trials is:

Parameters: π and n

$$Pr(X = k) = f(k; n, \pi)$$

where $$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

What values can X take on (i.e., what is the sample space)?

- k ranges from 0, 1, …, n

n choose k

N choose k function tells us how many ways there are to order ***k*** items out of ***n*** total

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Q: How many ways are there to choose 3 things out of a total of 8?

A: $(8 \cdot 7 \cdot 6)/ (3 \cdot 2) = 56$

R: choose(n, k)

Binomial distribution example 1

Koji Uehara typically strikes out about 33% of the batters he faces

If he pitched to 5 batters in a game, use the binomial distribution to calculate the probability he would strike out 2 of them?

$$Pr(X = k) = f(k; n, \pi) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Binomial distribution

Koji Uehara typically strikes out about 33% of the batters he faces

If he pitched to 5 batters in a game, use the binomial distribution to calculate the probability he would strike out 2 of them?

What are the parameters?

n = 5, π = .333 k = 2

$$Pr(X = k) = f(k; n, \pi) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}$$

$$Pr(X = 2) = f(k; 5, .333) = \binom{5}{2} .333^2 (1 - .333)^3$$

= .329

R: dbinom(x = 2, size = 5, prob = 1/3)

Binomial distribution

Suppose a Uehara pitched to 20 batters.

How could we calculate the probability that he would get more than 10 strike outs?

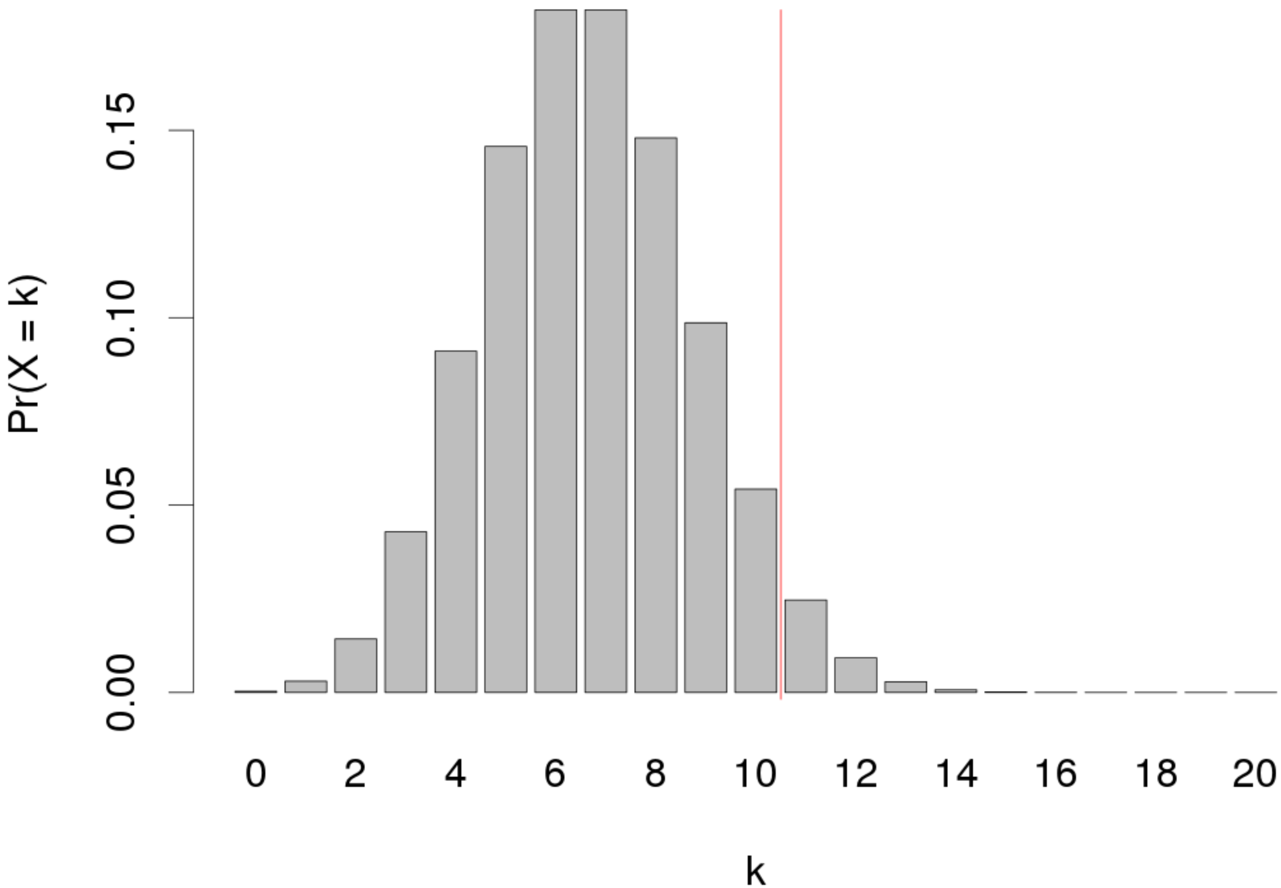
$$Pr(X > 10) = \sum_{k=11}^{20} f(k; n = 20, \pi = .333)$$

$$= \sum_{k=11}^{20} \binom{20}{k} .333^k (1 - .333)^{20-k}$$

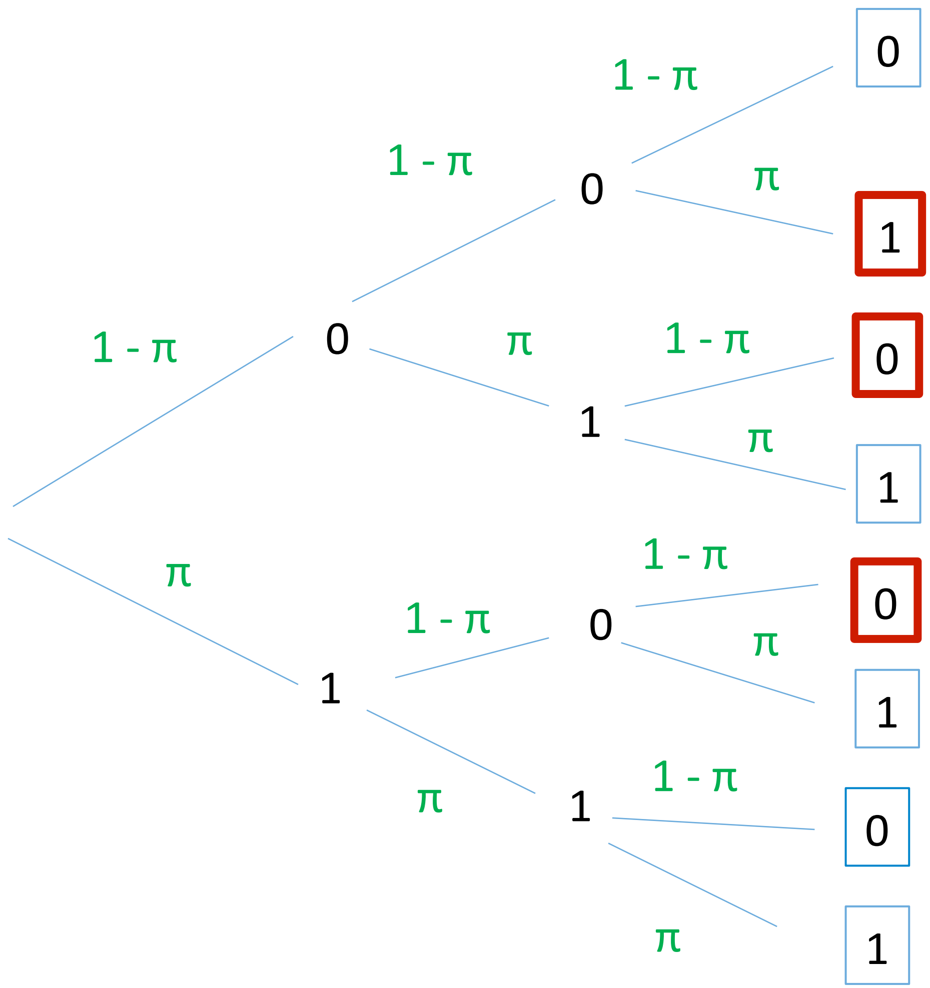
= .0376

R: `sum(dbinom(11:20, size = 20, prob = 1/3))`

Binomial distribution



Verify binomial distribution agrees with tree diagrams



Checking for:

$Pr(X = 1; n = 3, \pi)$

What are the possible sequences for X = 1?

- 0, 0, 1
- 0, 1, 0,
- 1, 0, 0

$$Pr(X = k) = \binom{n}{k}\pi^k(1-\pi)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Is this a good model for on-base events?

<u>Example 2</u>: If a player comes to bat 4 times in a game, what is the probability of getting on-base k times

<u>Ortiz 2014</u>

Number of times on base

Plate appearances

	0	**1**	**2**	**3**	**4**	**5**
1	4	2	0	0	0	0
3	3	4	0	0	0	0
4	14	29	32	2	2	0
5	3	15	13	10	3	0
6	0	0	2	0	1	1
7	0	0	0	1	0	0
8	0	0	1	0	0	0

Is this a good model for on-base events?

If a player comes to bat 4 times in a game, what is the probability of getting on-base k times

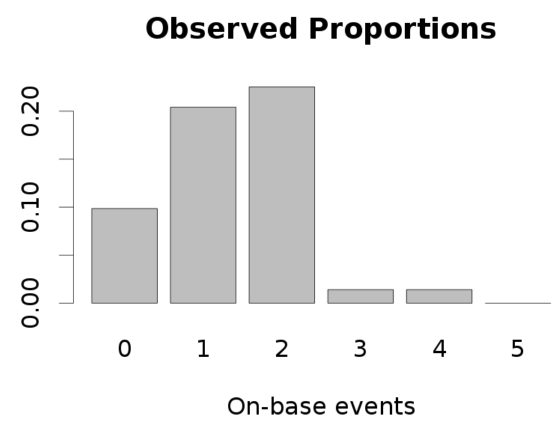
Ortiz 2014
OBP = .355
G with 4 PA = 79

Number of times on base

	0	1	2	3	4
4	14	29	32	2	2

What are n and π in the binomial probability mass function?

$$f(k;n,\pi) = \binom{n}{k}\pi^k(1-\pi)^{n-k}$$



Is this a good model for on-base events?

If a player comes to bat 4 times in a game, what is the probability of getting on-base k times

Ortiz 2014
OBP = .355
G with 4 PA = 79

Number of times on base

	0	1	2	3	4
4	14	29	32	2	2

What are n and p in the binomial probability mass function?

n = 4, π = .355
Compute the binomial distribution!

$$f(k;n,\pi) = \binom{n}{k}\pi^k(1-\pi)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Is this a good model for on-base events?

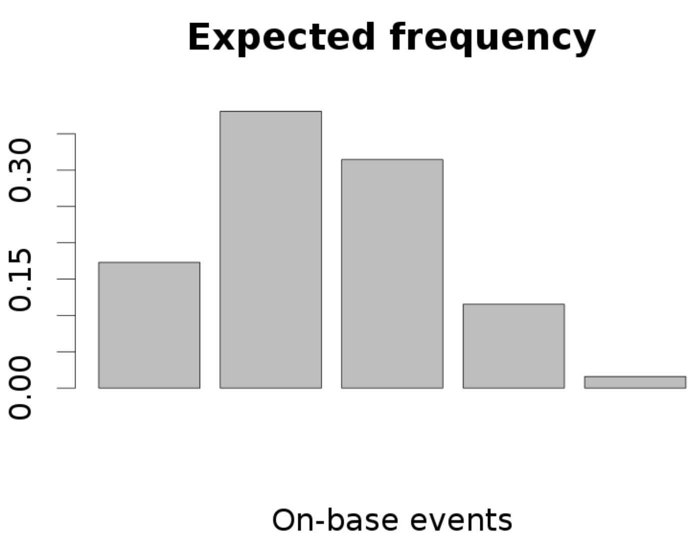
X	Binomial pmf			
0	f(0; 4, .355)			
1	f(1; 4, .355)			
2	f(2; 4, .355)			
3	f(3; 4, .355)			
4	f(4; 4, .355)			

Is this a good model for on-base events?

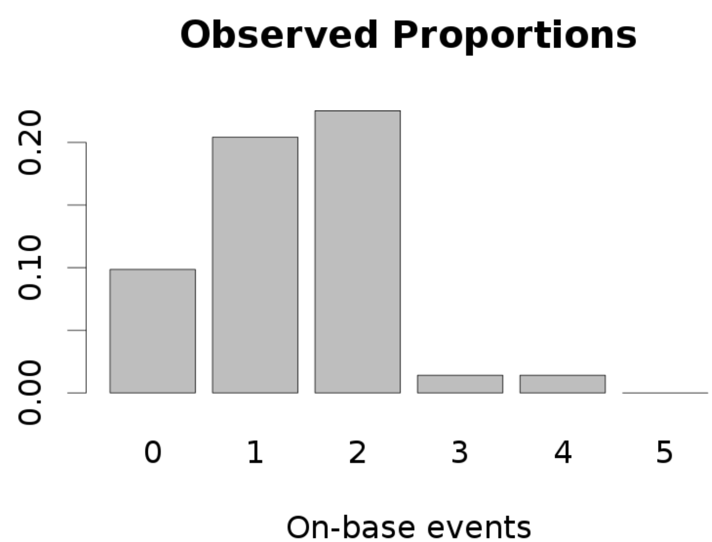
X	Binomial pmf	Binomial pmf value		
0	f(0; 4, .355)	0.173		
1	f(1; 4, .355)	0.381		
2	f(2; 4, .355)	0.315		
3	f(3; 4, .355)	0.115		
4	f(4; 4, .355)	0.016		

Is this a good model for on-base events?

Binomial mass function



Observed data



Is this a good model for on-base events?

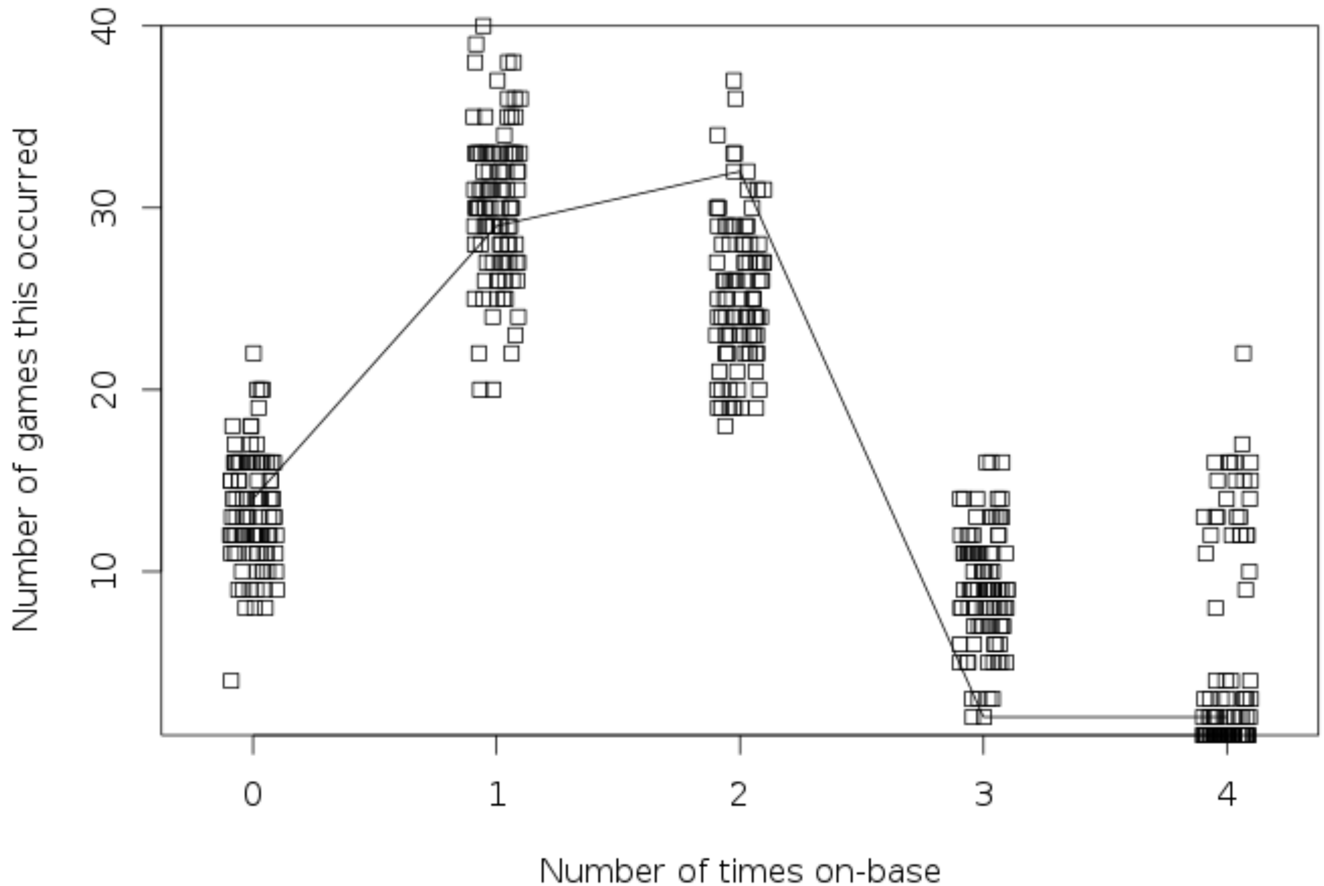
X	Binomial pmf	Binomial pmf value	Expected number	
0	f(0; 4, .355)	0.173	13.667	
1	f(1; 4, .355)	0.381	30.099	
2	f(2; 4, .355)	0.315	24.885	
3	f(3; 4, .355)	0.115	9.085	
4	f(4; 4, .355)	0.016	1.264	

Expected number = N * f(k; 5, .355) = 79 * f(k; 5, .355)

Is this a good model for on-base events?

X	Binomial pmf	Binomial pmf value	Expected number	Observed number
0	f(0; 4, .355)	0.173	13.667	14
1	f(1; 4, .355)	0.381	30.099	29
2	f(2; 4, .355)	0.315	24.885	32
3	f(3; 4, .355)	0.115	9.085	2
4	f(4; 4, .355)	0.016	1.264	2

Generating data from a binomial distribution



Discrete probability distributions

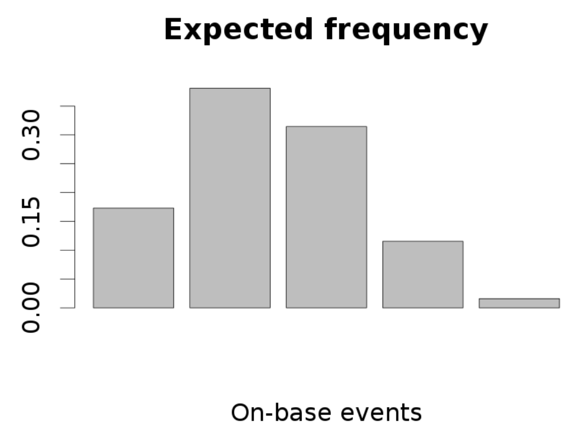
The binomial distribution is a discrete probability distribution

- i.e., the outcomes are discrete numbers (non-negative integers)

Discrete distributions have **probability mass functions** that give the probability of any one event in the sample space

- This can be used to find Pr(a < X < b)

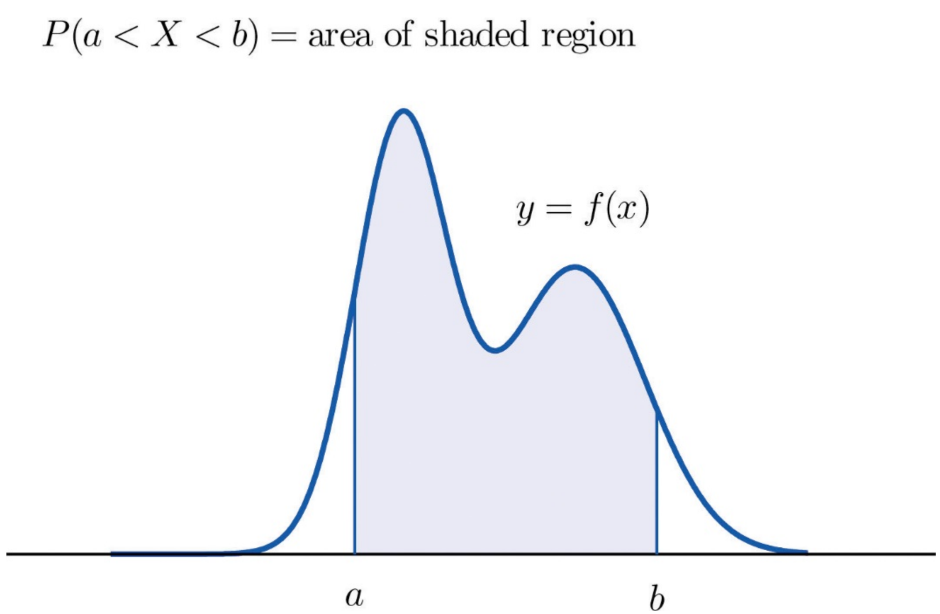
$$Pr(a \leq X \leq b) = \sum_{k=a}^{b} f(k; \ldots)$$



Continuous probability distributions

Continuous distributions have **probability density functions** that can be used to calculate the probability that an even is in some interval [a b]

$$Pr(a \leq X \leq b) = \int_a^b f(x; ...)dx$$



Normal Density Curve

A **normal distribution** follows a bell-shaped curve

- Useful for describing the mean of many random outcomes

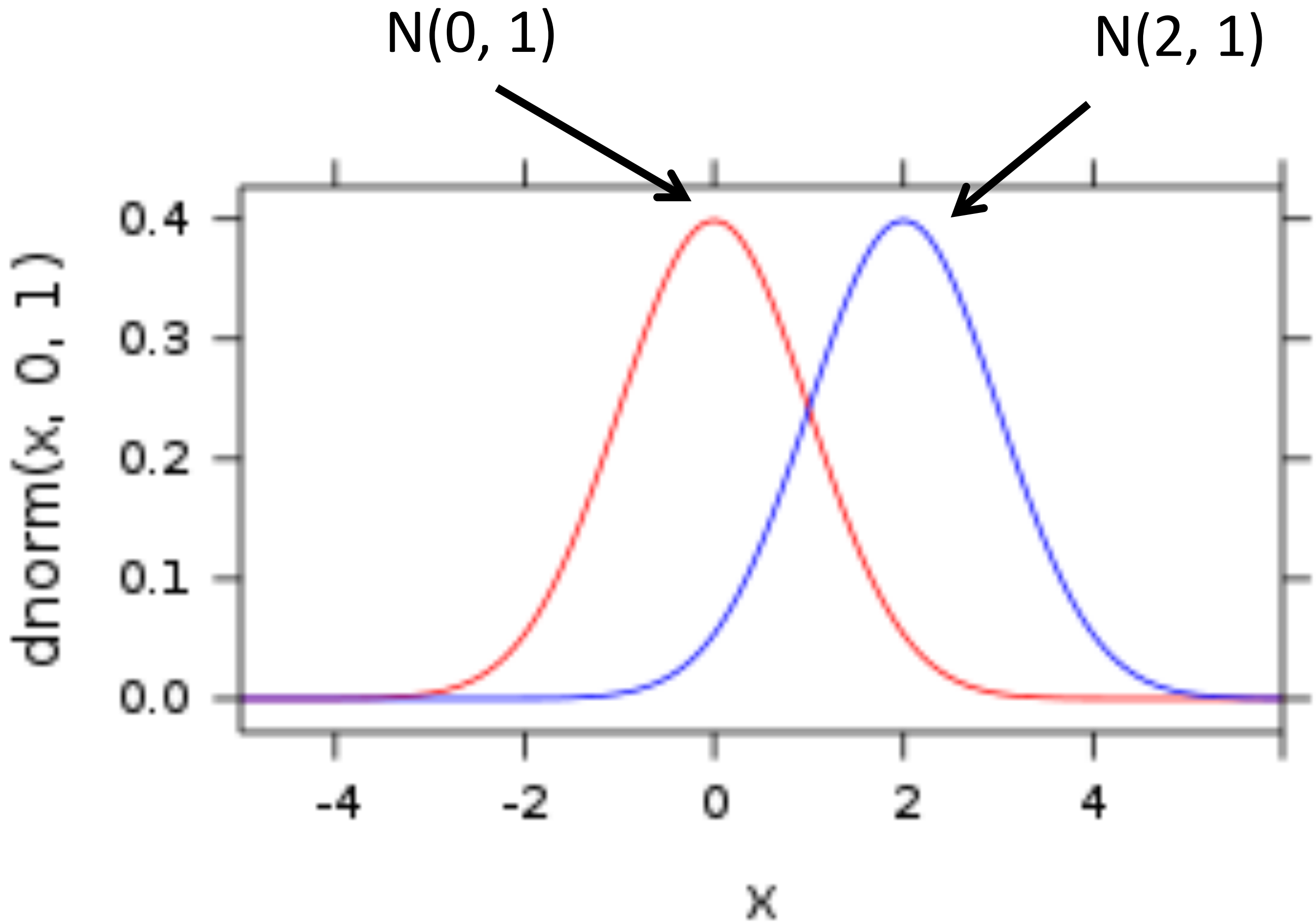
There are two parameters that characterize normal curves, which are:

- The mean: μ
- The standard deviation: σ

Notation: X ~ N(μ, σ)

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Normal curves with different means



Normal curves with different variances