Probability Density Functions and the Normal Distribution

Quantitative Understanding in Biology, 1.2

1. Discrete Probability Distributions

1.1. The Binomial Distribution

Question:

You’ve decided to flip a coin. What’s the probability the coin will come up heads? Tails? What about heads 10 times in a row? What about heads, then, tails, then head again?

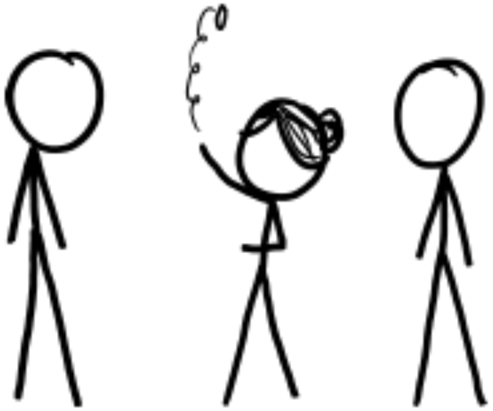
Proposition:

You don't need to flip any coins. If your coin is fair, coin flips follow the binomial distribution.

A probability distribution function is a function that relates an event to the probability of that event.

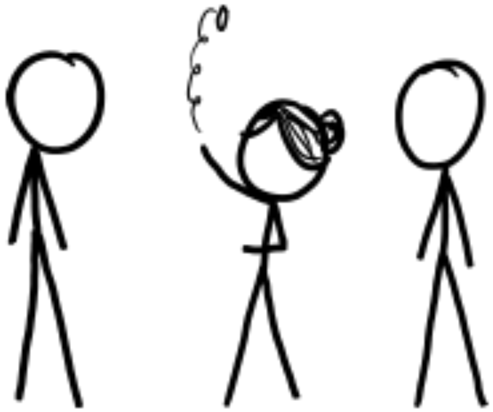
If the events are discrete (i.e. they correspond to a set of specific numbers or specific “states”), we describe it with a probability mass function.

$$p(x) = \begin{cases} x = x_1 & p_1 \\ x = x_2 & p_2 \\ \vdots & \vdots \\ x = x_n & p_n \end{cases}$$



$$p(x) \geq 0$$

$$\sum_{i=1}^{n} p_i = 1$$



[R Example]

To be added to notes. See code on next slide.

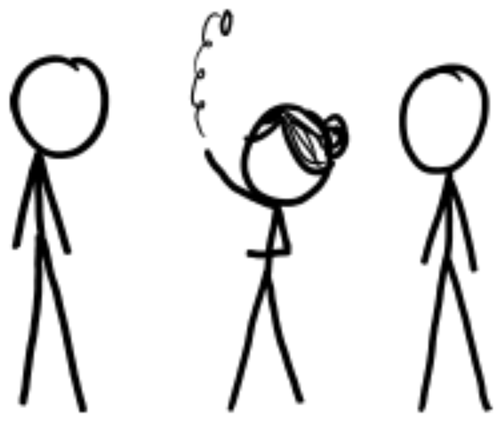
```
# ROLL A FAIR 6-SIDED DIE
sample(1:6, 4, replace=TRUE)
# ROLL AN UNFAIR 5-SIDED DIE
sample(1:5, 4, replace=TRUE, prob=c(0.1, 0.3, 0.4, 0.05, 0.05))
# SAMPLE FROM A SET OF COLORS
sample(c("red", "blue", "green", "white", "black"), 4,
replace=TRUE, prob=c(0.1, 0.3, 0.4, 0.05, 0.05))
```

Proposition:

You don’t need to flip any coins. If your coin is fair, coin flips follow the binomial distribution.

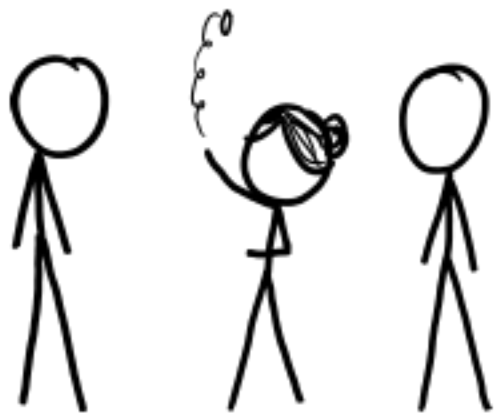
The Binomial Distribution

$$p(x) = \begin{cases} x = success & p_{success} \\ x = failure & 1 - p_{success} \end{cases}$$



The Fair Coin

$$p(x)=\begin{cases} x=heads & 0.5 \\ x=tails & 0.5 \end{cases}$$



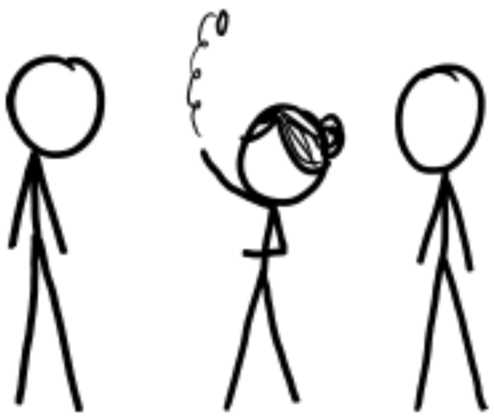
Question:

What about multiple coin flips?

If you are fairly flipping a fair coin, each coin flip is *independent and identically distributed*, also known as iid.

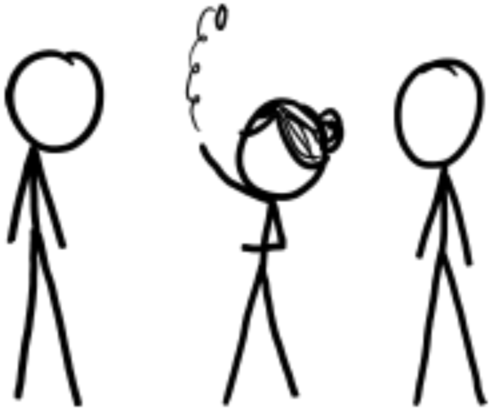
Independent Fair Coin Flips

$$p(heads, heads, tails) = p(heads)\,p(heads)\,p(tails)$$



Independent Fair Coin Flips

$$p(\#\, flips = n) = 0.5^n$$



When you sample a binomial distribution multiple times, you are performing a Bernoulli trial.

We can perform Bernoulli trials in R

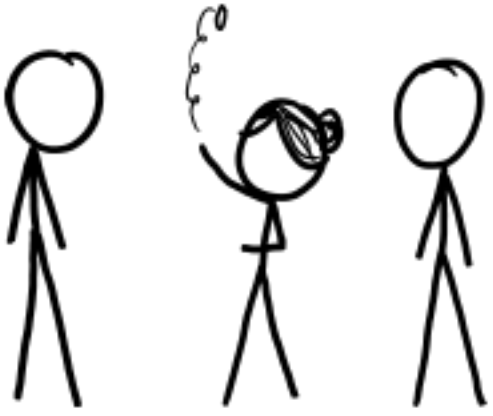
[R Example]

See notes.

Often it is useful to calculate cumulative probabilities; for example, the probability of 7 or more heads when you flip 10 coins.

$$P(\leq \# \, successes, \# \, trials) = \sum_{i=0}^{\# \, successes} p(successes = i, trials = \# \, trials)$$

$$P(\geq \# \, successes, \# \, trials) = 1 - P(\leq \# \, successes, \# \, trials)$$



[R Example]

See notes.

1.2. The Poisson Distribution

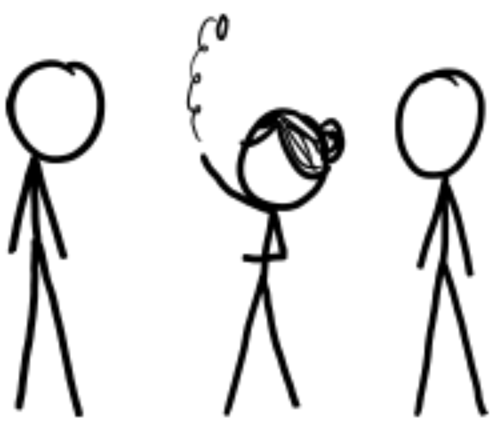
[R Example]

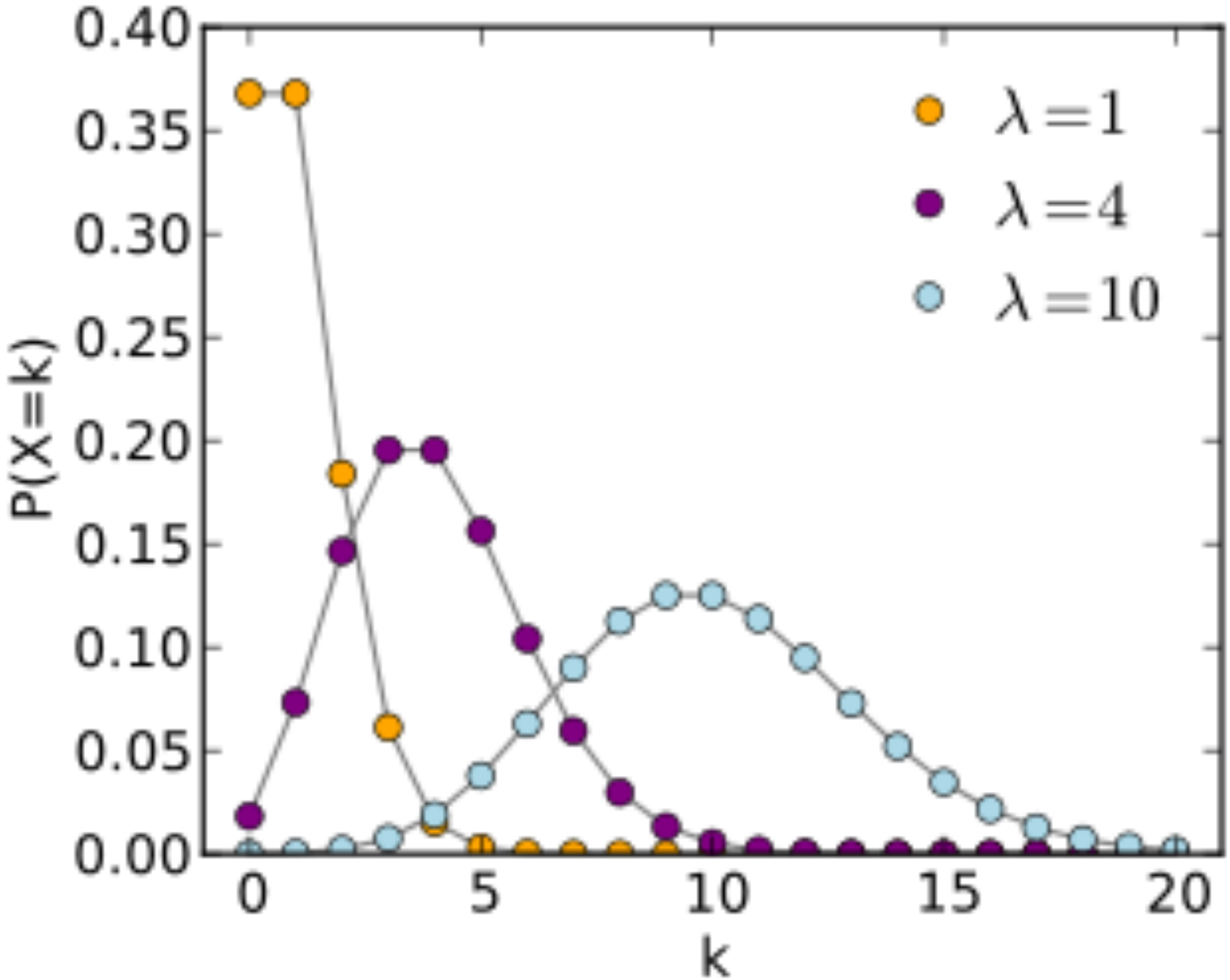
See notes.

The Poisson distribution describes the probability of a certain number of events occurring within a given time interval.

The Poisson Distribution

$$p(\#\,events = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$





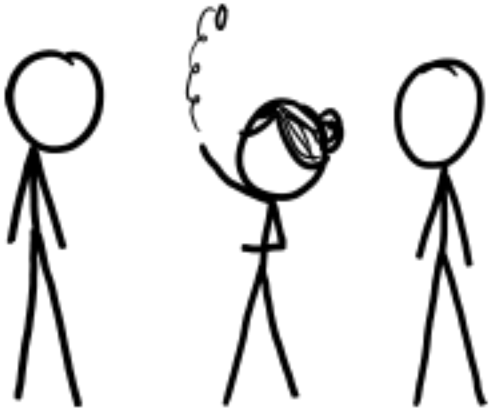
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
P(X=k)
0
5
10
15
20
k
$\lambda = 1$
$\lambda = 4$
$\lambda = 10$

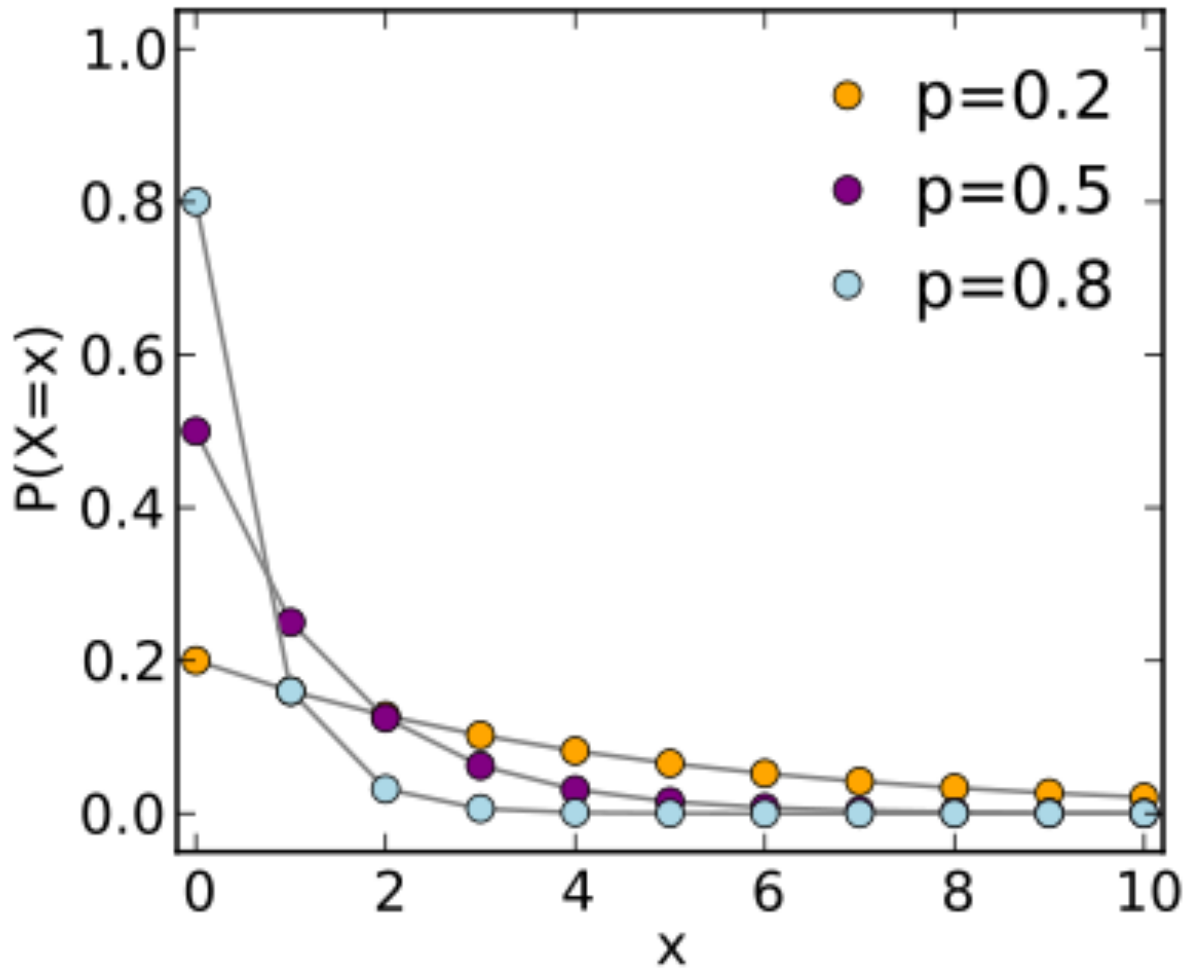
1.3. The Geometric Distribution

You flipped heads. How many tails will you flip before you flip heads again? The Geometric distribution describes the probability of a given “waiting time” between successes.

The Geometric Distribution

$$p(wait = k) = p_{success}\left(1 - p_{success}\right)^k$$




p=0.2
p=0.5
p=0.8
P(X=x)
x
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10

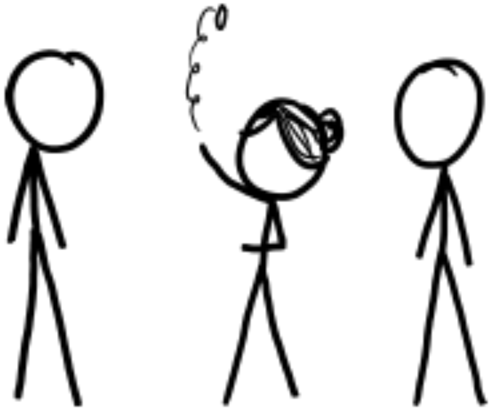
[R Example]

See notes.

1.3. Uniform Distributions

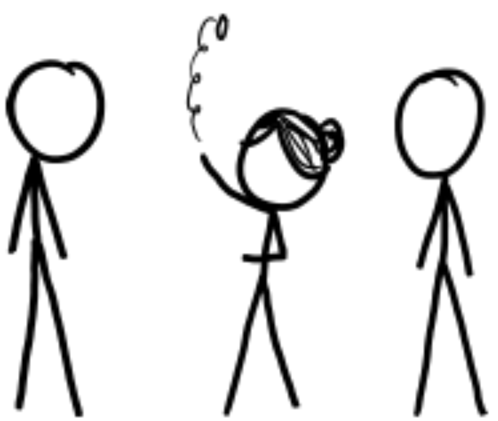
The Uniform Distribution

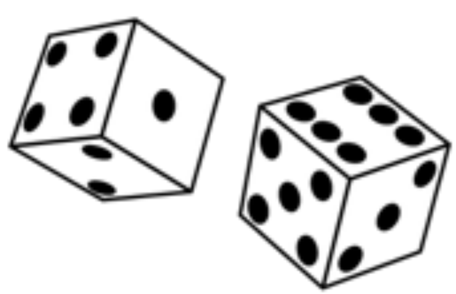
$$p(x) = \frac{1}{n}$$



The Uniform Distribution

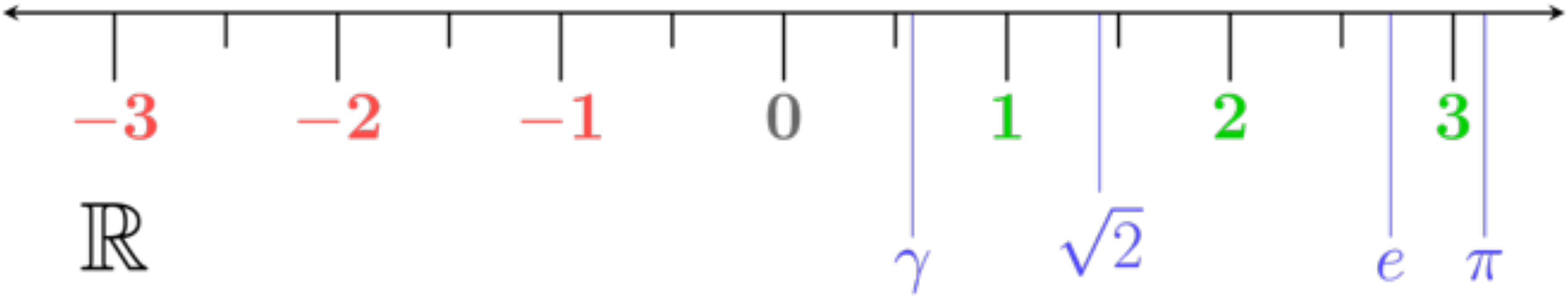
$$p(x) = \frac{1}{n}$$





2. Continuous Probability Distributions

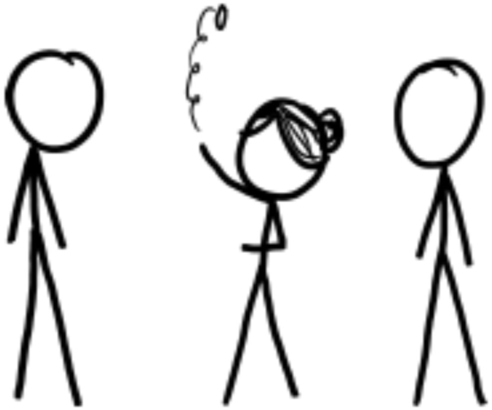
If the events correspond to *real numbers*, we describe it with a probability density function.



−3
−2
−1
0
1
2
3
$\mathbb{R}$
γ
$\sqrt{2}$
e
π

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

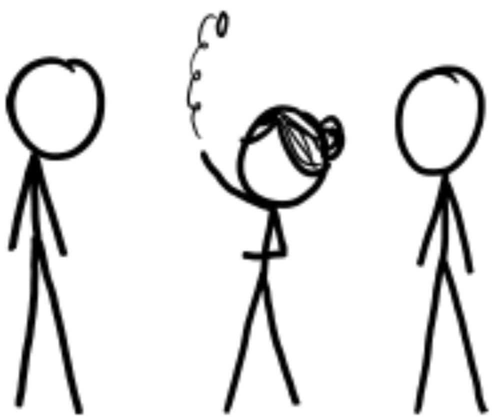


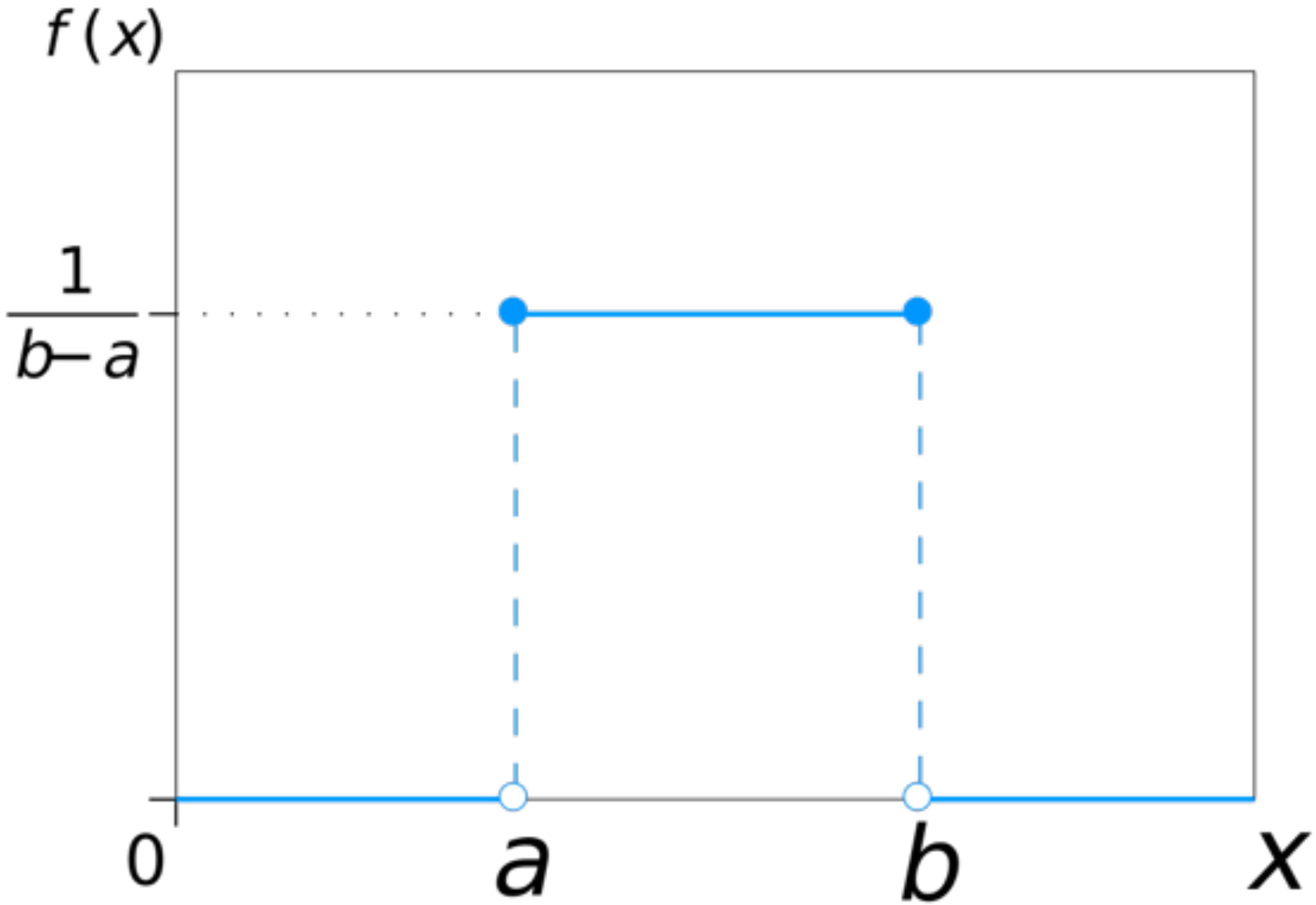
2.1. The Uniform Distribution

The Uniform distribution defines an interval in which the probability density is uniform. Outside of this interval, the probability is 0.

The Uniform Distribution

$$f(x) = \begin{cases} 0 & x < a \\ \dfrac{1}{b-a} & a \le x \le b \\ 0 & x > b \end{cases}$$





f(x)
1
b−a
0
a
b
x

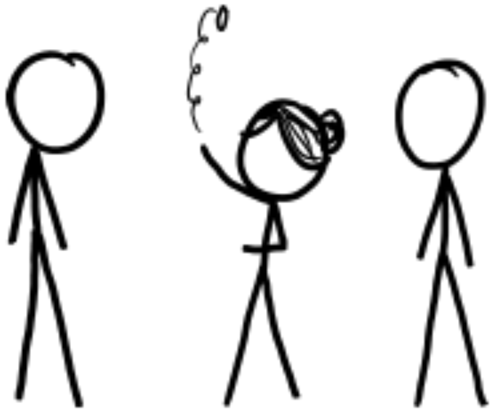
[R Example]

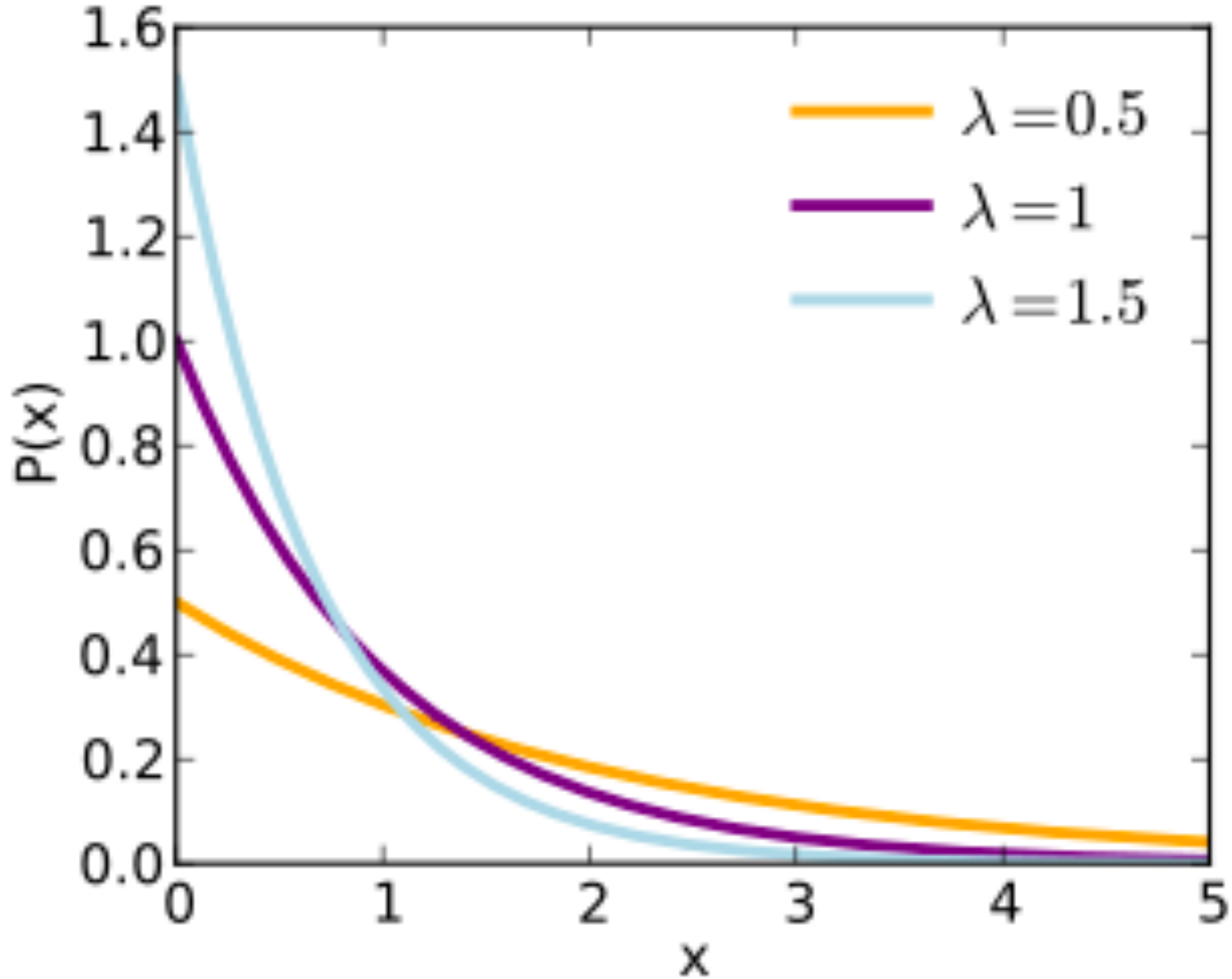
See notes.

2.2. The Exponential Distribution

The Exponential Distribution

$$f(x) = \lambda e^{-\lambda x}$$



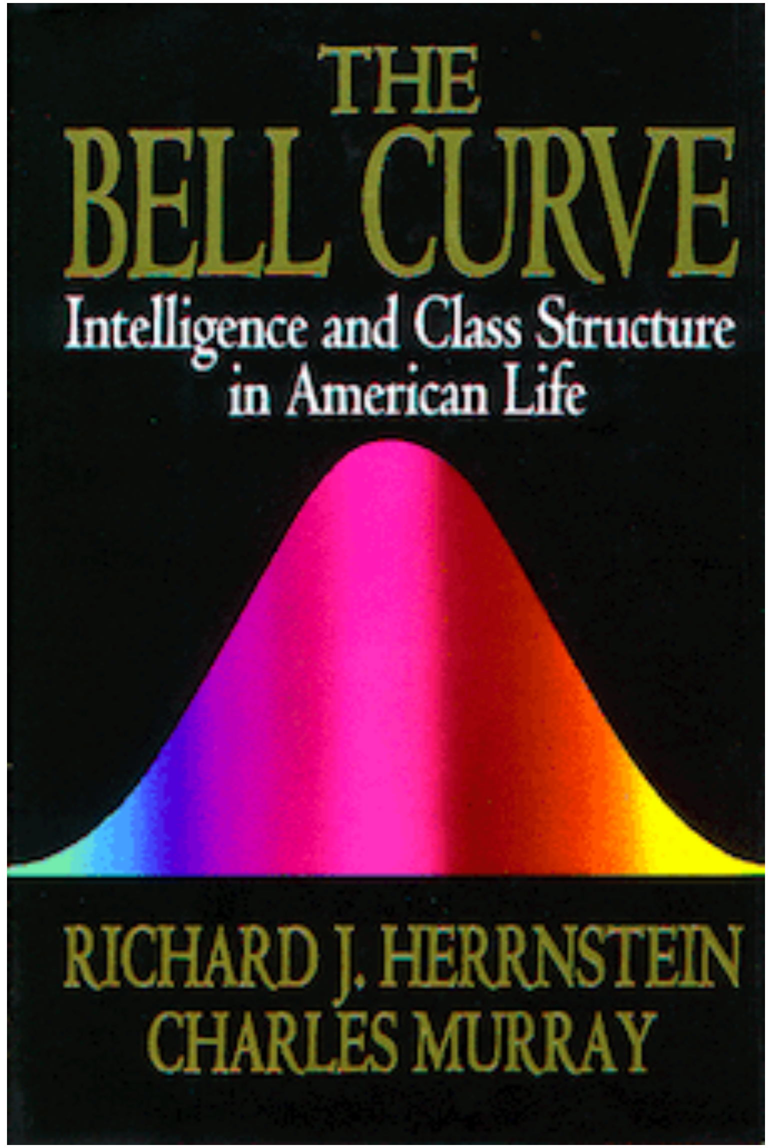


P(x)
x
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
0
1
2
3
4
5
$\lambda=0.5$
$\lambda=1$
$\lambda=1.5$

[R Example]

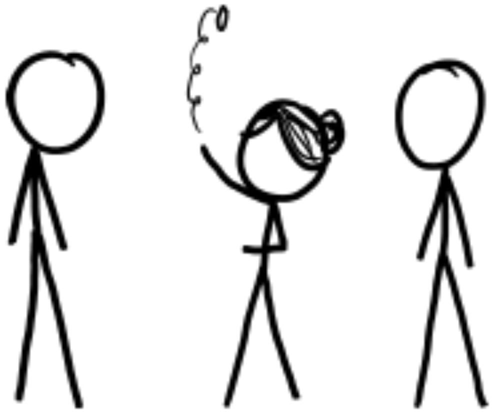
See notes.

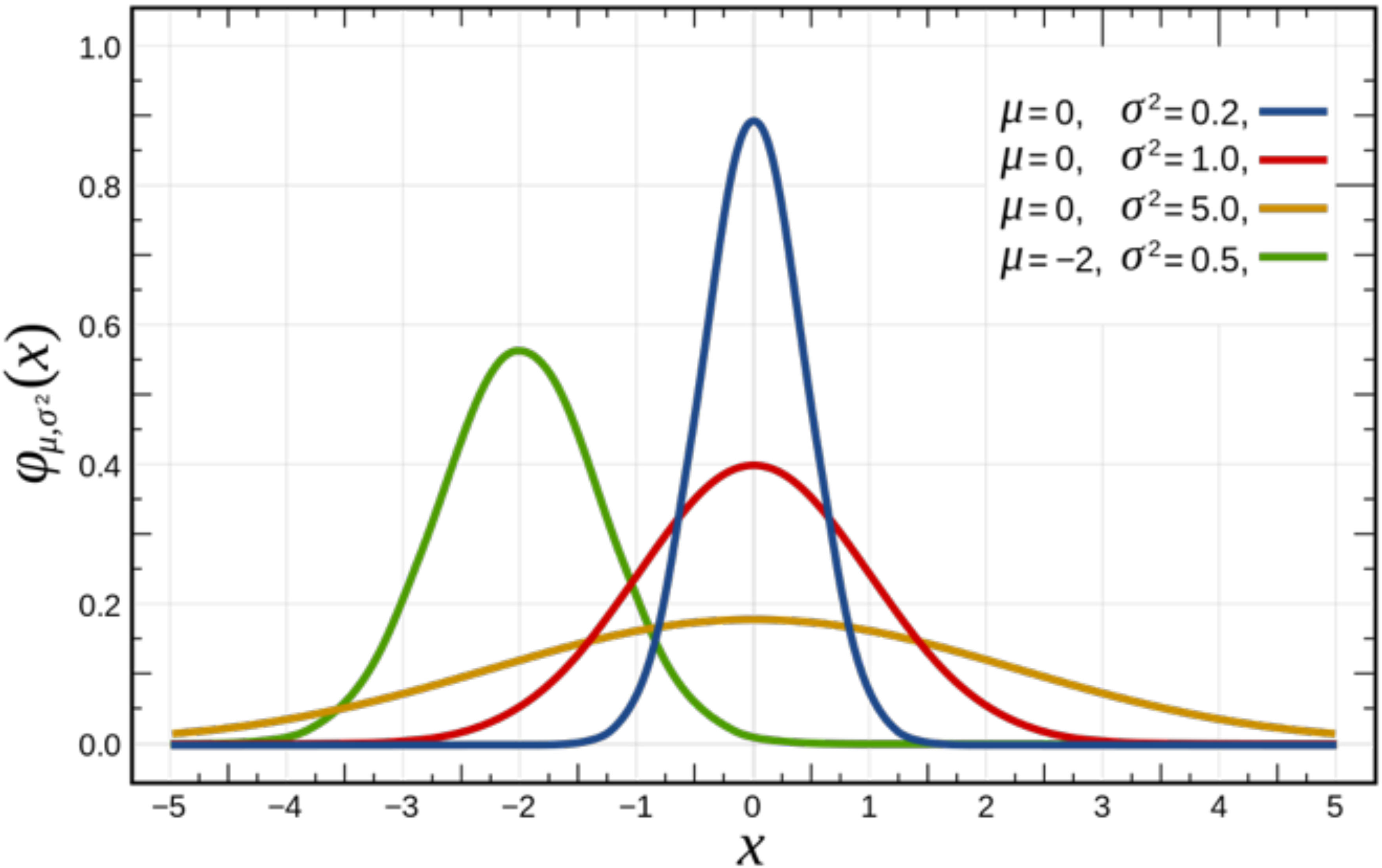
2.3. The Normal Distribution


THE
BELL CURVE
Intelligence and Class Structure
in American Life
RICHARD J. HERRNSTEIN
CHARLES MURRAY

The Normal Distribution

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





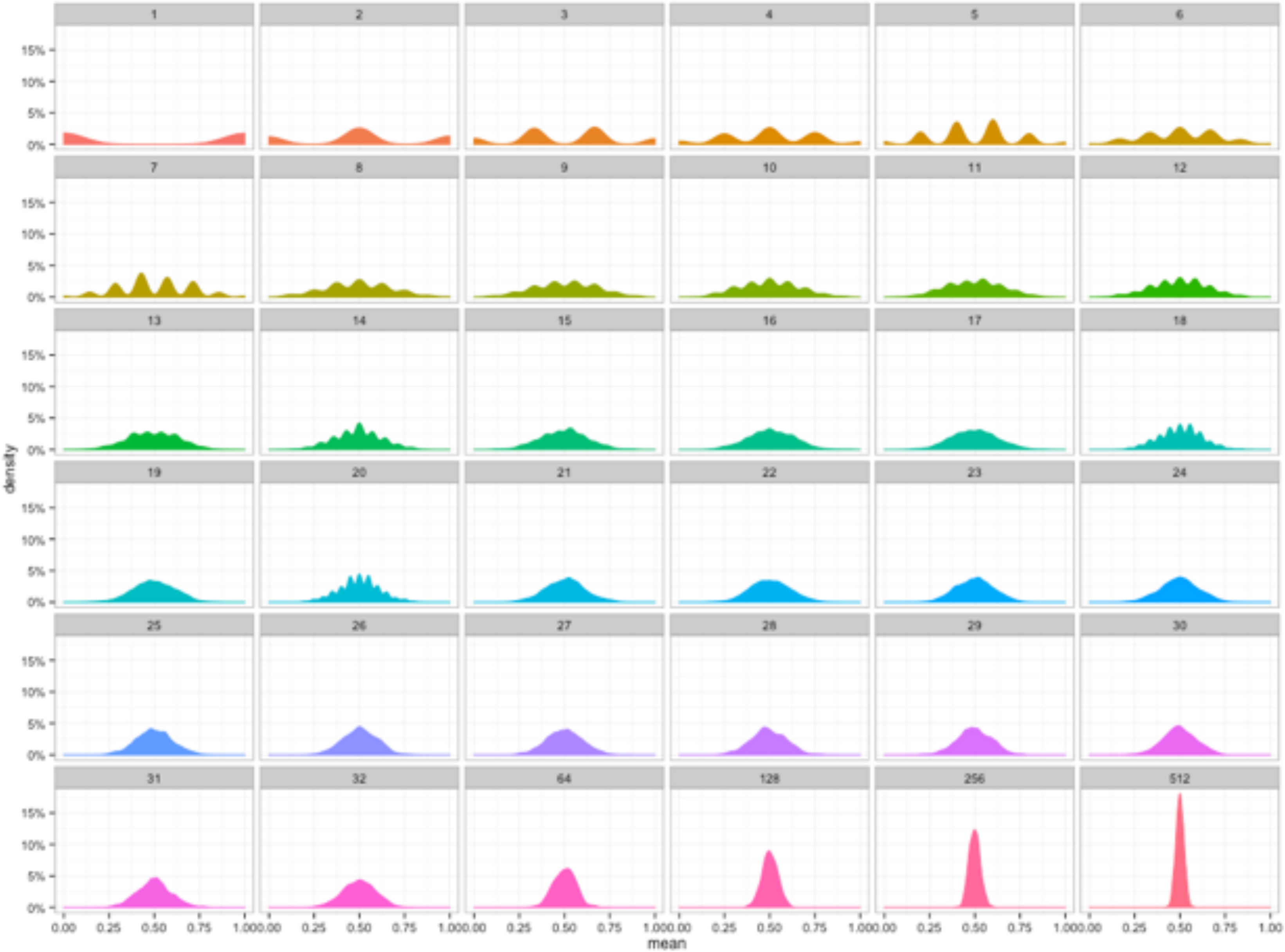
1.0
0.8
0.6
0.4
0.2
0.0
−5
−4
−3
−2
−1
0
1
2
3
4
5
x
$\varphi_{\mu,\sigma^2}(x)$
$\mu=0, \quad \sigma^2=0.2,$
$\mu=0, \quad \sigma^2=1.0,$
$\mu=0, \quad \sigma^2=5.0,$
$\mu=-2, \quad \sigma^2=0.5,$

[R Example]

See notes.

3. Central Limit Theorem

If you measure the mean of a distribution many, many times, those means will tend to become normally distributed.



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
64
128
256
512
15%
10%
5%
0%
density
0.00
0.25
0.50
0.75
1.00
mean

Central Limit Theorem implies:

1. If you sample enough, your estimate of the mean will converge to the true mean.
2. If your measurement is the average of many independent processes, it will tend to be normally distributed.

Central Limit Theorem implies:

1. **If you sample enough, your estimate of the mean will converge to the true mean.**
2. **If your measurement is the average of many independent processes, it will tend to be normally distributed.**

Central Limit Theorem implies:

1. **If you sample enough, your estimate of the mean will converge to the true mean.**
2. **If your measurement is the average of many independent processes, it will tend to be normally distributed.**

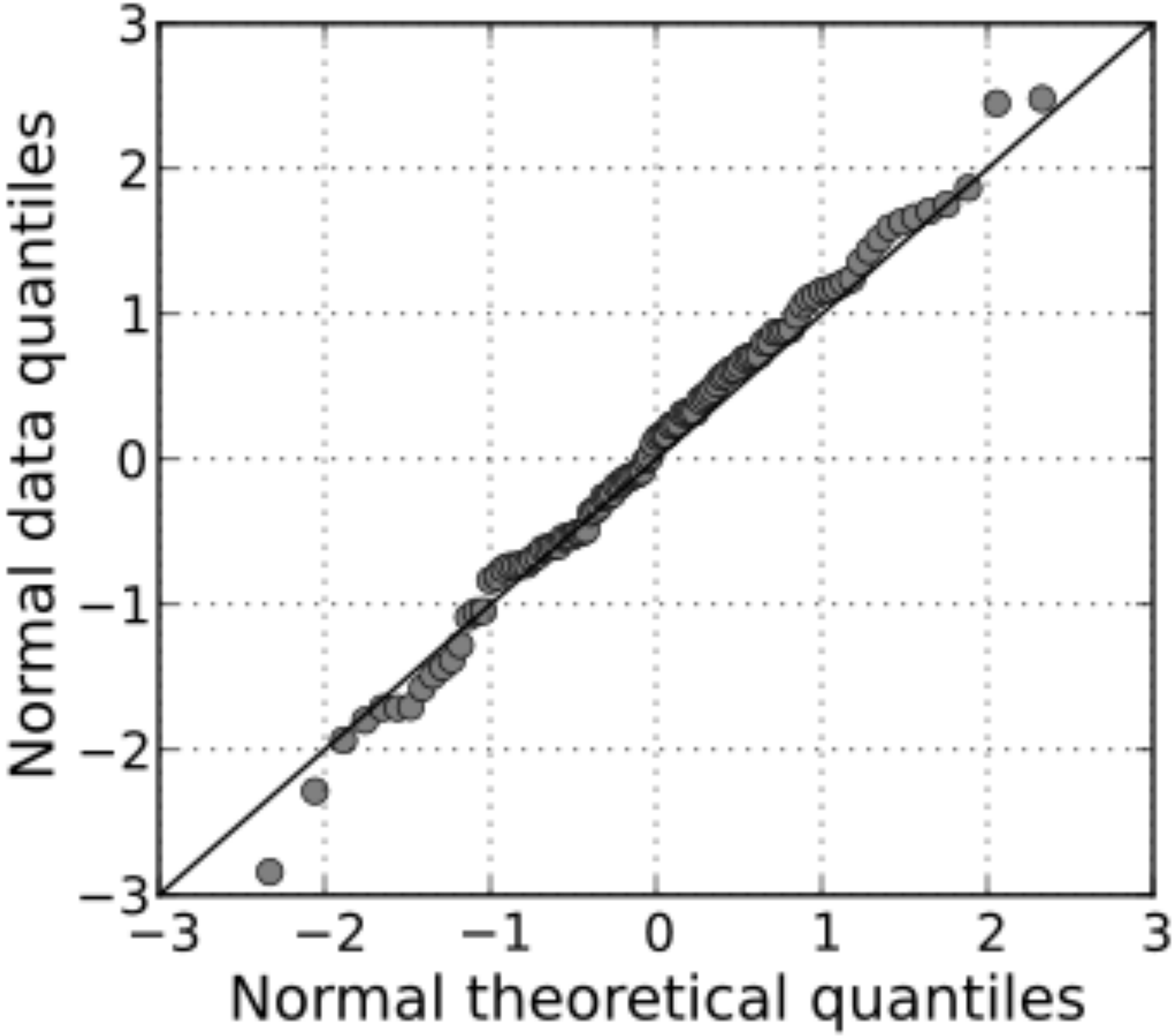
Question:

So you have some data. What should I do with it?

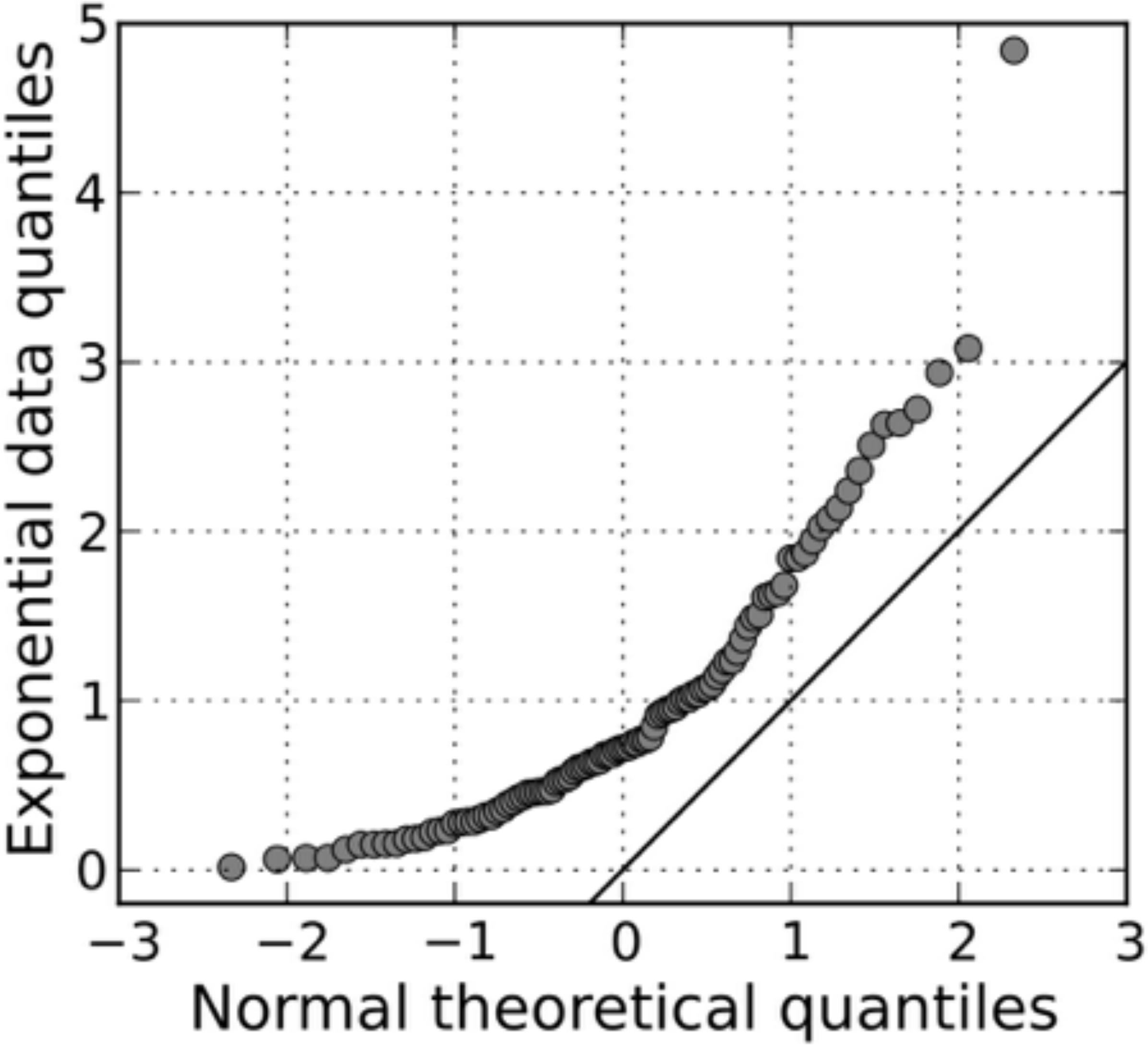
Step 1. Test for normality.

Step 1.1. Look at your data.

One of the easiest ways to “look at your data” is to make a Q-Q plot.



Normal data quantiles
Normal theoretical quantiles
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3



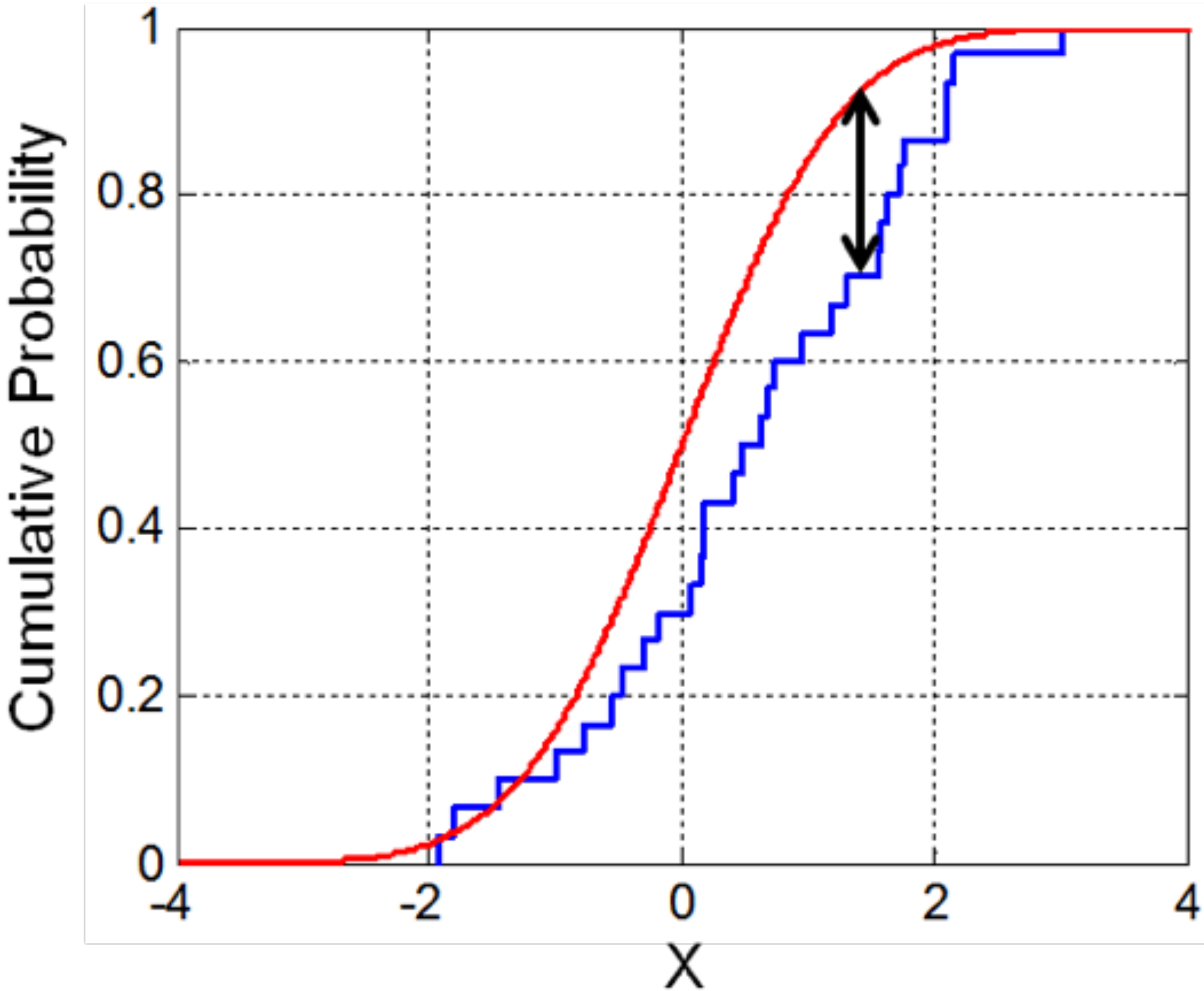
Exponential data quantiles
5
4
3
2
1
0
−3
−2
−1
0
1
2
3
Normal theoretical quantiles

[R Example]

See notes.

Step 1.2. Do a statistical test.

In the Kolmogorov-Smirnov test, you estimate the probability that your data comes from the normal distribution from the distance between the two cumulative distribution functions.



Cumulative Probability
1
0.8
0.6
0.4
0.2
0
-4
-2
0
2
4
X

[R Example #8]

See notes.

Homework: Generate Q-Q plots