10. Named Continuous Distributions in R

Principles of Data Science with R

Dr. Uma Ravat

PSTAT 10

Summary:

- Discrete Random Variables and p.m.f
- Named discrete distributions
  - Discrete uniform
  - Binomial Distribution

Next we will see...

- Continuous Random Variables and Distributions
- Named Continuous Distributions and Probability calculations
  - Uniform
  - Normal

Recall: Types of random variables

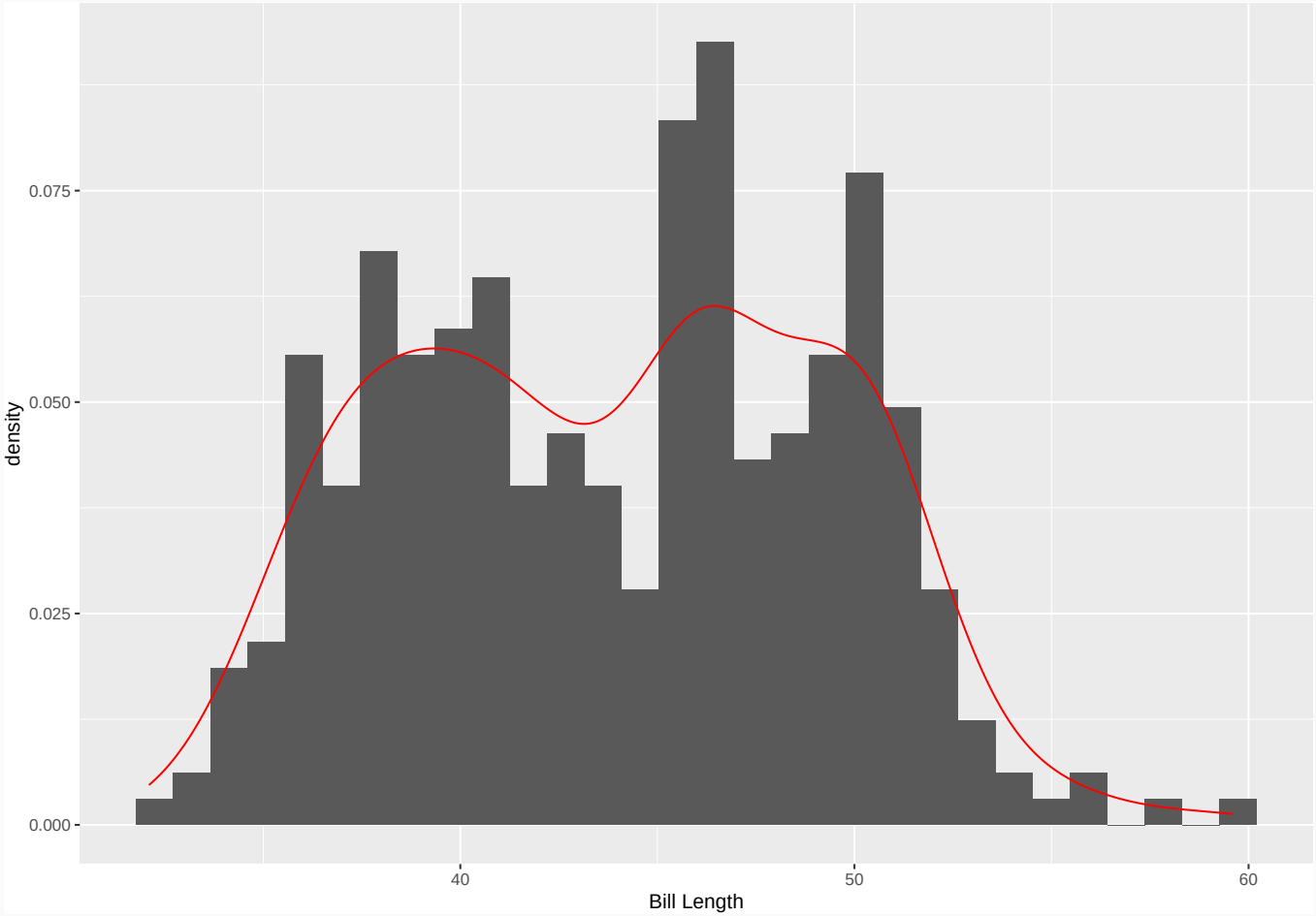
- A random variable is **discrete** when we can *count* the number of outcomes.
  - has a finite or countably infinite number of possible outcomes.
- A random variable is **continuous** when the outcomes can be *measured*.
  - takes all values in an interval of real numbers.

Examples of continuous RVs

- Height
- Weight
- Time
- Temperature

From histograms to continuous distributions

`penguins` dataset bill lengths

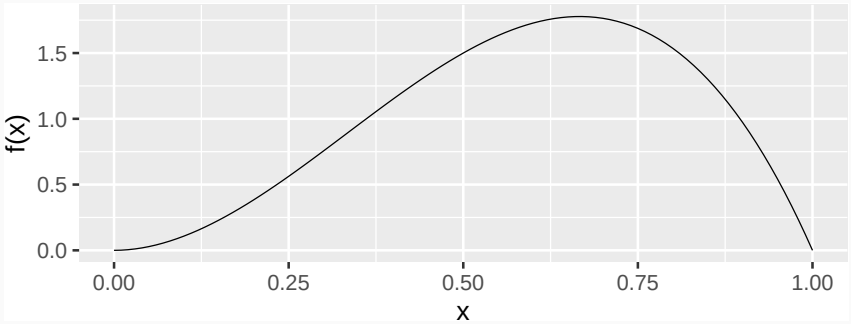


probability density function is a smooth curve.

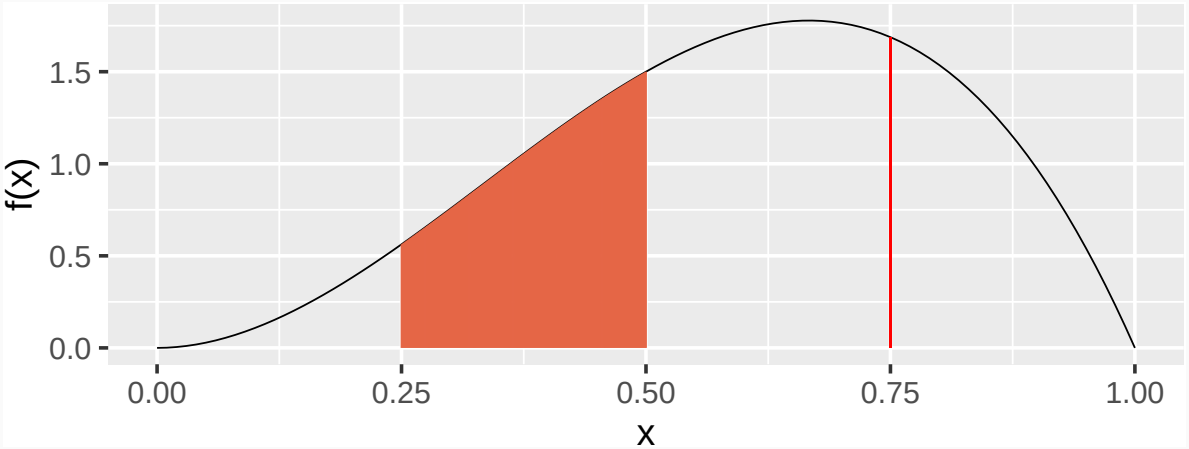
For a continuous RV

- Distribution is specified by its Probability Density Function (p.d.f.)
- The pdf can be represented by
  - a function $f(x)$, the density function or
  - its graph, the density curve

$$f(x) = \begin{cases} 12\ x^2\ (1-x) & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

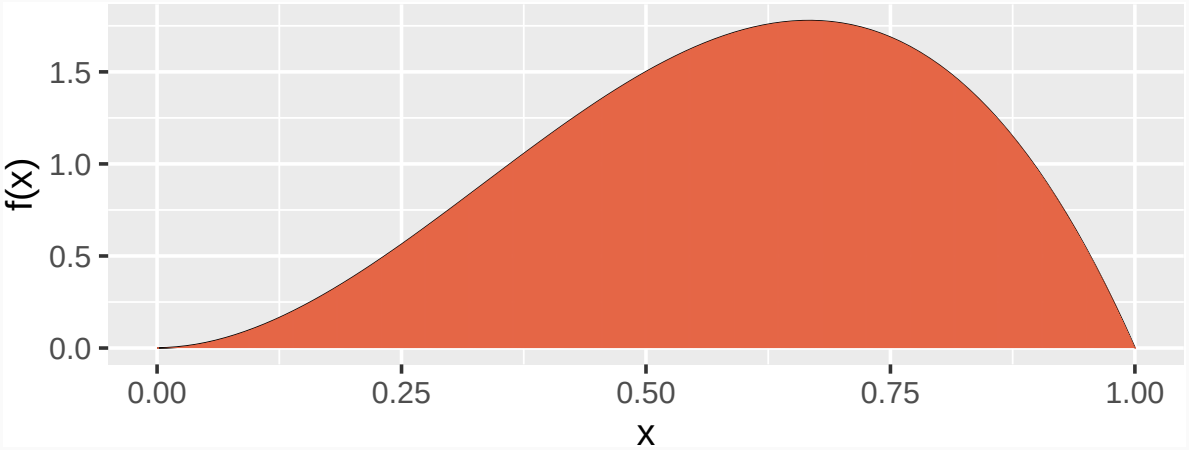


Probability is Area Under the Curve



- $P(0.25 \leq X \leq 0.50)$
- $P(X = 0.75)$

Area under the pdf curve = 1

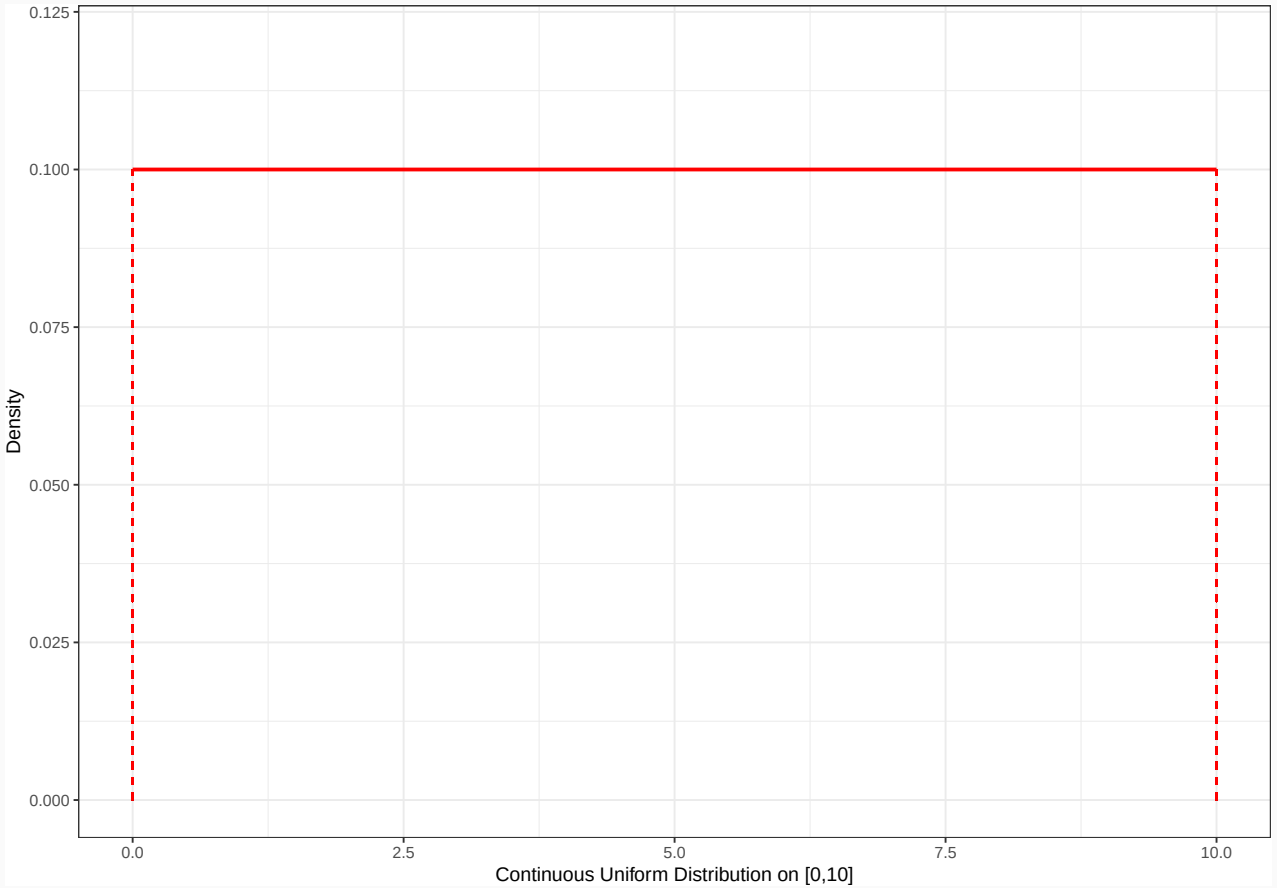


Named Continuous Probability Distributions

1. The Uniform Distribution

2. The Normal Distribution

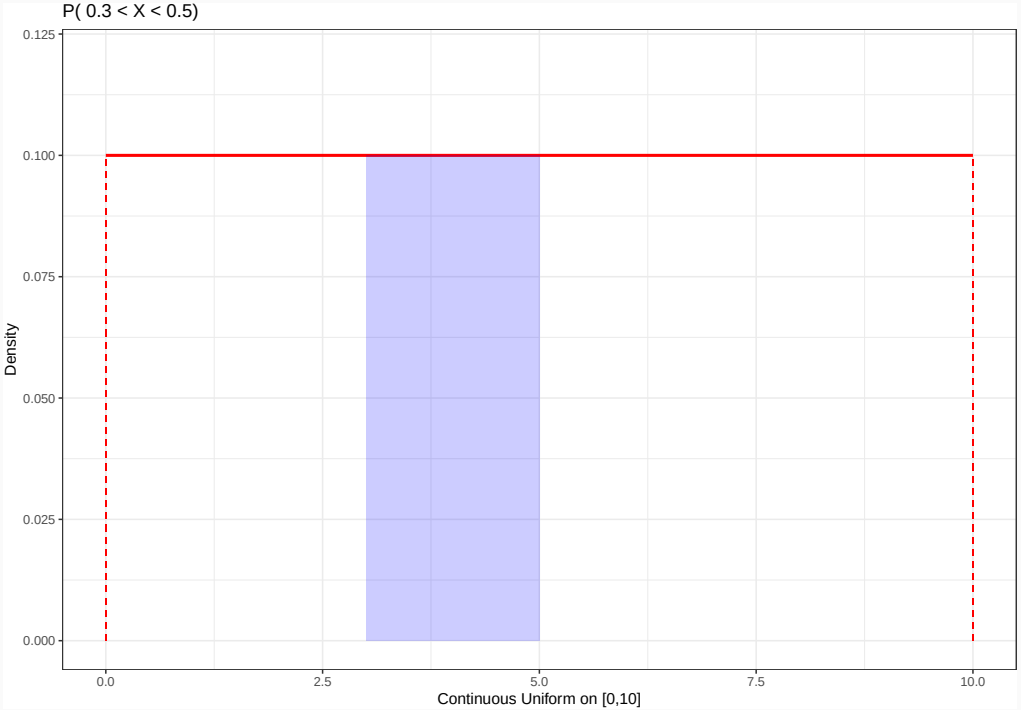
The Continuous Uniform Distribution



- All values are equally likely to occur

Uniform Probability calculations - By hand

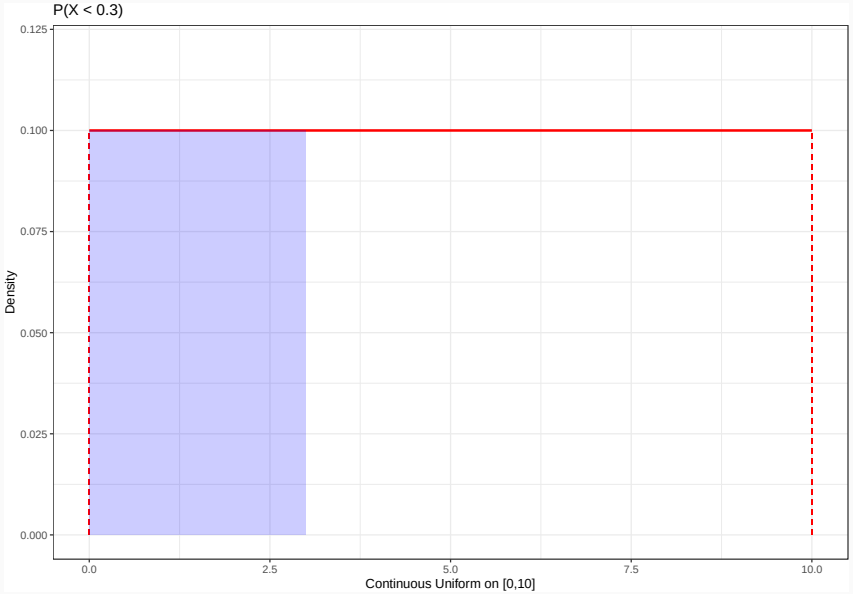
- $P(X < 0.5), P(X \leq 0.5),$
- $P(0.3 \leq X \leq 0.5), P(0.3 < X \leq 0.5), P(0.3 \leq X < 0.5),$
- $P(X > 0.5), P(X \geq 0.5)$



$$P(0.3 < X < 0.5) = (0.5 - 0.3) * 1 = 0.2$$

Uniform Probability calculations - Using R

- $P(X < 0.5), P(X \leq 0.5)$
  - Recall: $P(X \leq q)$ is cdf(q) - the area under the density curve to the **left** of q
- **Syntax:** `punif(q, min = 0, max = 1, lower.tail = TRUE, log.p = FALSE)`

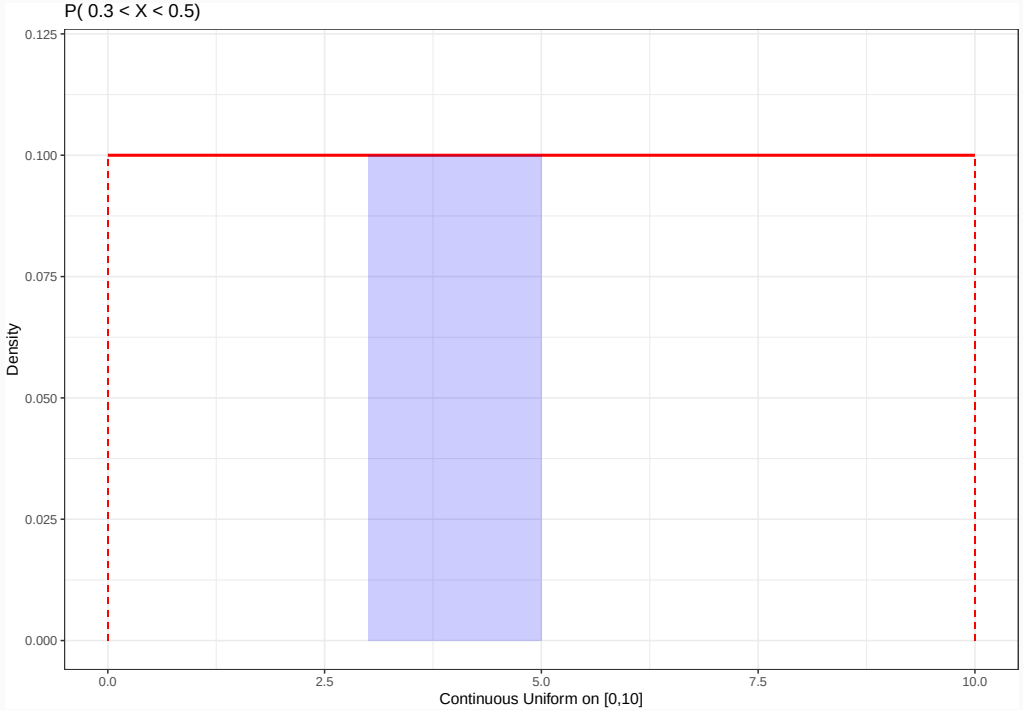



```
punif(0.3, min = 0, max = 1, lower.tail = TRUE, log.p = FALSE)
```

```
## [1] 0.3
```

Uniform Probability calculations - Using R

- $P(0.3 \leq X \leq 0.5), P(0.3 < X \leq 0.5), P(0.3 \leq X < 0.5)$
  - Recall trick: difference of two cdfs

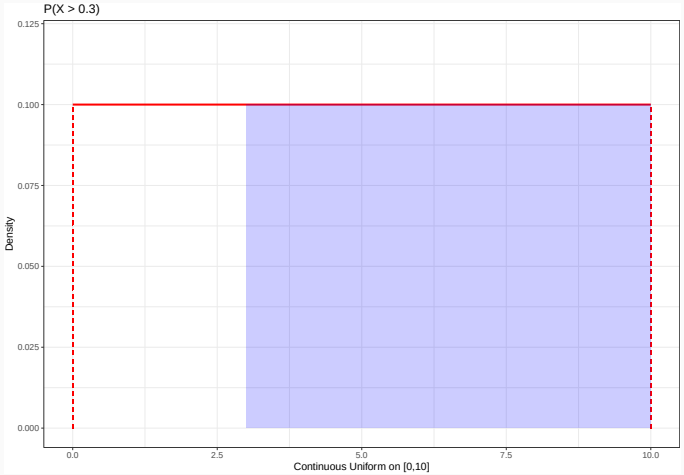



```
punif(0.5, min = 0, max = 1) - punif(0.3, min = 0, max = 1)
```

```
## [1] 0.2
```

Uniform Probability calculations: Using R:

- $P(X > 0.5), P(X \geq 0.5)$
  - Recall: $P(X \geq q)$ is the area under the density curve to the **right** of q
- **Syntax:** `punif(q, min = 0, max = 1, lower.tail = FALSE, log.p = FALSE)`
- `1 - punif(q, min = 0, max = 1)`




```
punif(0.3, min = 0, max = 1, lower.tail = FALSE)
```

```
## [1] 0.7
```

The continuous Uniform distribution on [a,b]

$X \sim \text{UNIF}(a, b)$ then the *probability density function* (p.d.f) is given by

- $f(x) = \frac{1}{b-a}$, if $a \leq x \leq b$
  - **mean:** $E(X) = \mu = \frac{a+b}{2}$
- **p.d.f in R:** `dunif(x, min = a, max = b, log = FALSE)`
- **probability calculations**
  - By hand: Area under the density curve -> Area of rectangles
  - **c.d.f in R: Syntax:** `punif(q, min = a, max = b, lower.tail = TRUE, log = FALSE)`
- generating samples from uniform distribution: `runif`

```
runif(5) # default is a = 0, b= 1
```

```
## [1] 0.79324533 0.50666721 0.56121438 0.75066492 0.09648169
```

```
mean(runif(40, min = 4, max = 10))
```

```
## [1] 6.985599
```

Your Turn: Uniform Distribution. Time Spent Waiting for a Bus

A bus arrives at a stop every 10 minutes. A student is equally likely to arrive at the stop at any time. How long will the student have to wait?

What is the probability the waiting time, X,

1. 5 minutes or less?
2. between 5 and 7 minutes?
3. more than 6 minutes?

Time Spent Waiting for a Bus: 1. Formulate the problem in Math notation

- Let X denote the waiting time until the next bus arrives.
- X is a continuous uniform random variable, measured from 0 to 10 minutes.
- p.d.f is $f(x) = \frac{1}{10}$, if $0 \leq x \leq 10$

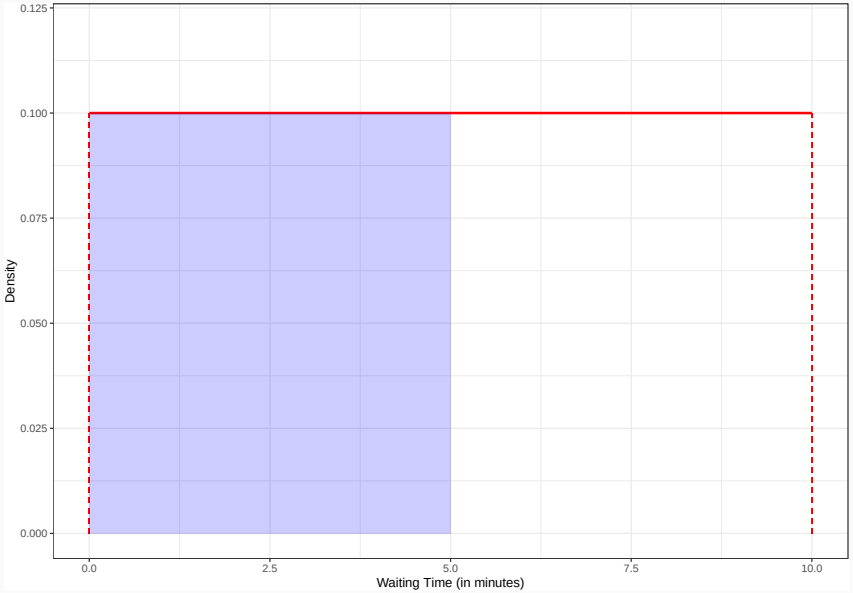
Probability of waiting

1. 5 minutes or less? $P(X \leq 5)$
2. between 5 and 7 minutes? $P(5 \leq X \leq 7)$
3. more than 6 minutes? $P(X \geq 6)$

It is always helpful(and mistakes are avoided) to **draw a picture** of the density and **shade the desired area** under the curve while doing probability calculations.

Example: Time Spent Waiting for a Bus

1. 5 minutes or less?

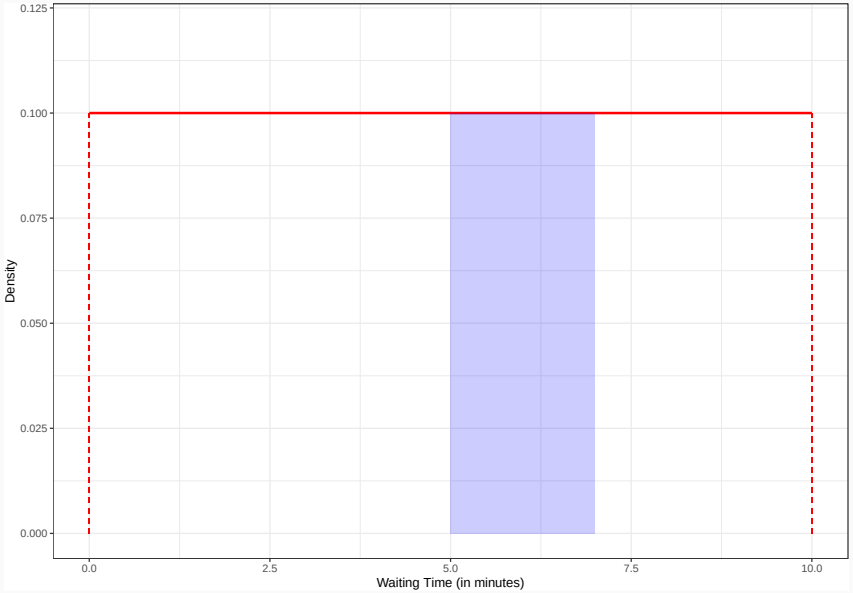



```
punif(5, min = 0, max = 10)
```

```
## [1] 0.5
```

Example: Time Spent Waiting for a Bus

2. between 5 and 7 minutes?

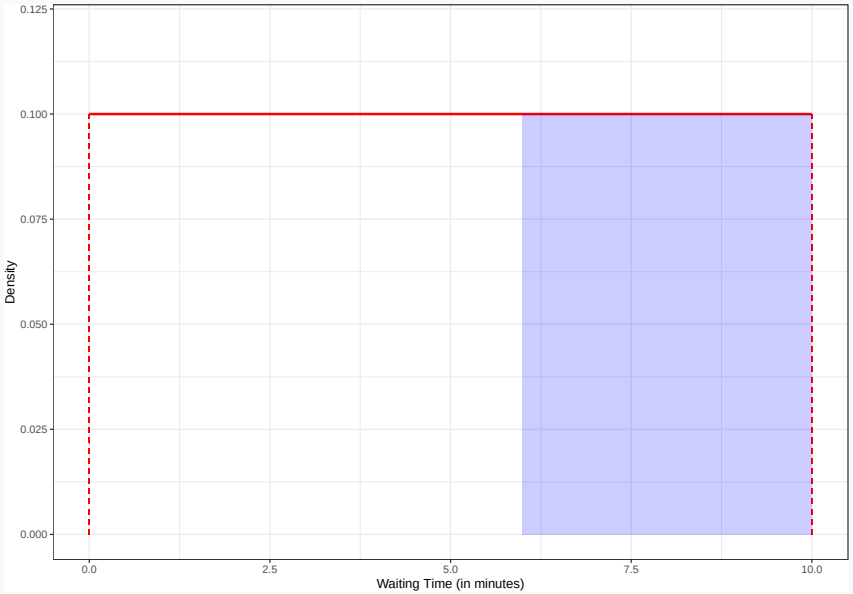



```
punif(7, min = 0, max = 10) - punif(5, min = 0, max = 10)
```

```
## [1] 0.2
```

Example: Time Spent Waiting for a Bus

3. more than 6 minutes?




```
punif(10, min = 0, max = 10) - punif(6, min = 0, max = 10)
```

```
## [1] 0.4
```

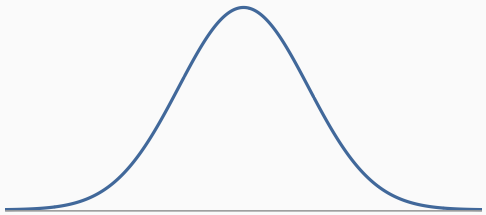
```
# or
punif(6, min = 0, max = 10, lower.tail = FALSE)
```

```
## [1] 0.4
```

Named Continuous distribution: Normal distribution

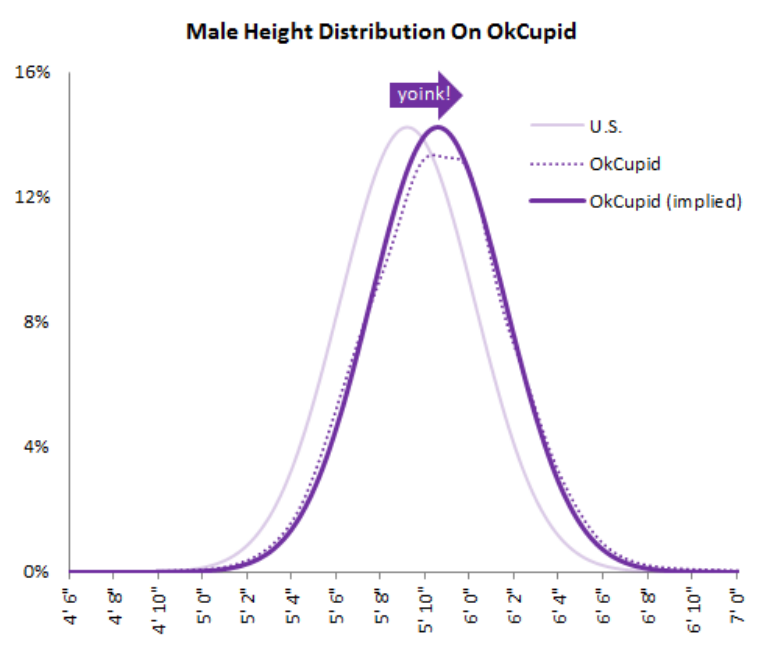
Named Continuous distribution: Normal distribution

- Uni modal and symmetric, bell shaped curve
- Many variables are nearly normal, but none are exactly normal
- Denoted as $\mathbb{N}(\mu, \sigma) \rightarrow$ Normal with mean μ and standard deviation σ



- For example;
  - the heights of people,
  - the weights of similar animals,
  - measurements on machine produced items

Heights of males

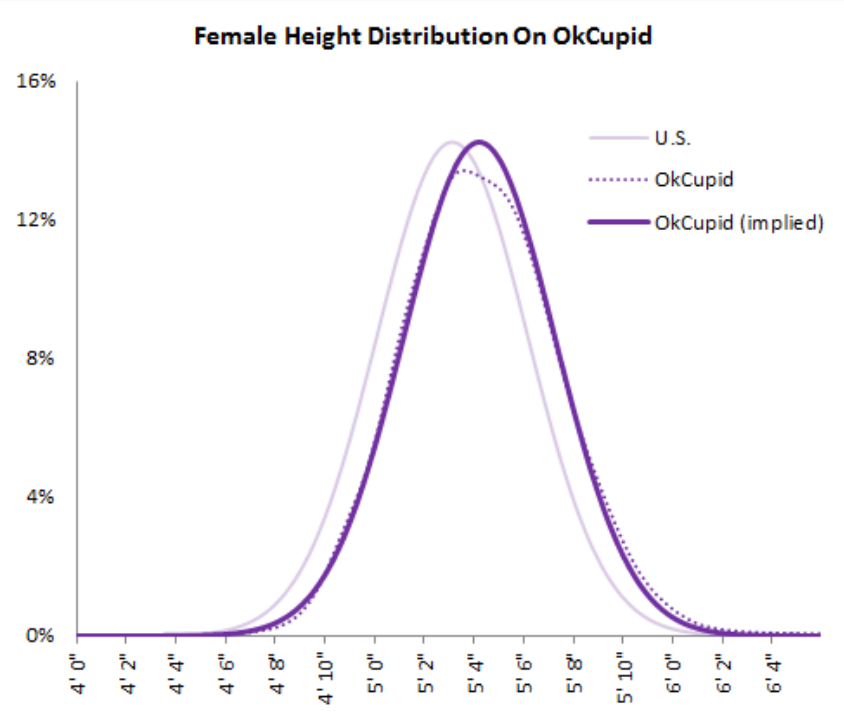


"The male heights on OkCupid very nearly follow the expected normal distribution – except the whole thing is shifted to the right of where it should be. Almost universally guys like to add a couple inches."

"You can also see a more subtle vanity at work: starting at roughly 5' 8", the top of the dotted curve tilts even further rightward. This means that guys as they get closer to six feet round up a bit more than usual, stretching for that coveted psychological benchmark."

{http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating/}

Heights of females



"When we looked into the data for women, we were surprised to see height exaggeration was just as widespread, though without the lurch towards a benchmark height."

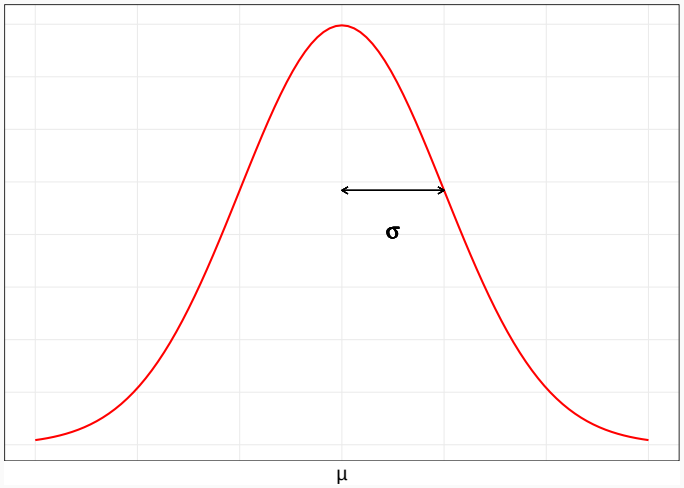
{http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating/}

Normal Distribution

If $X \sim \mathbb{N}(\mu, \sigma)$

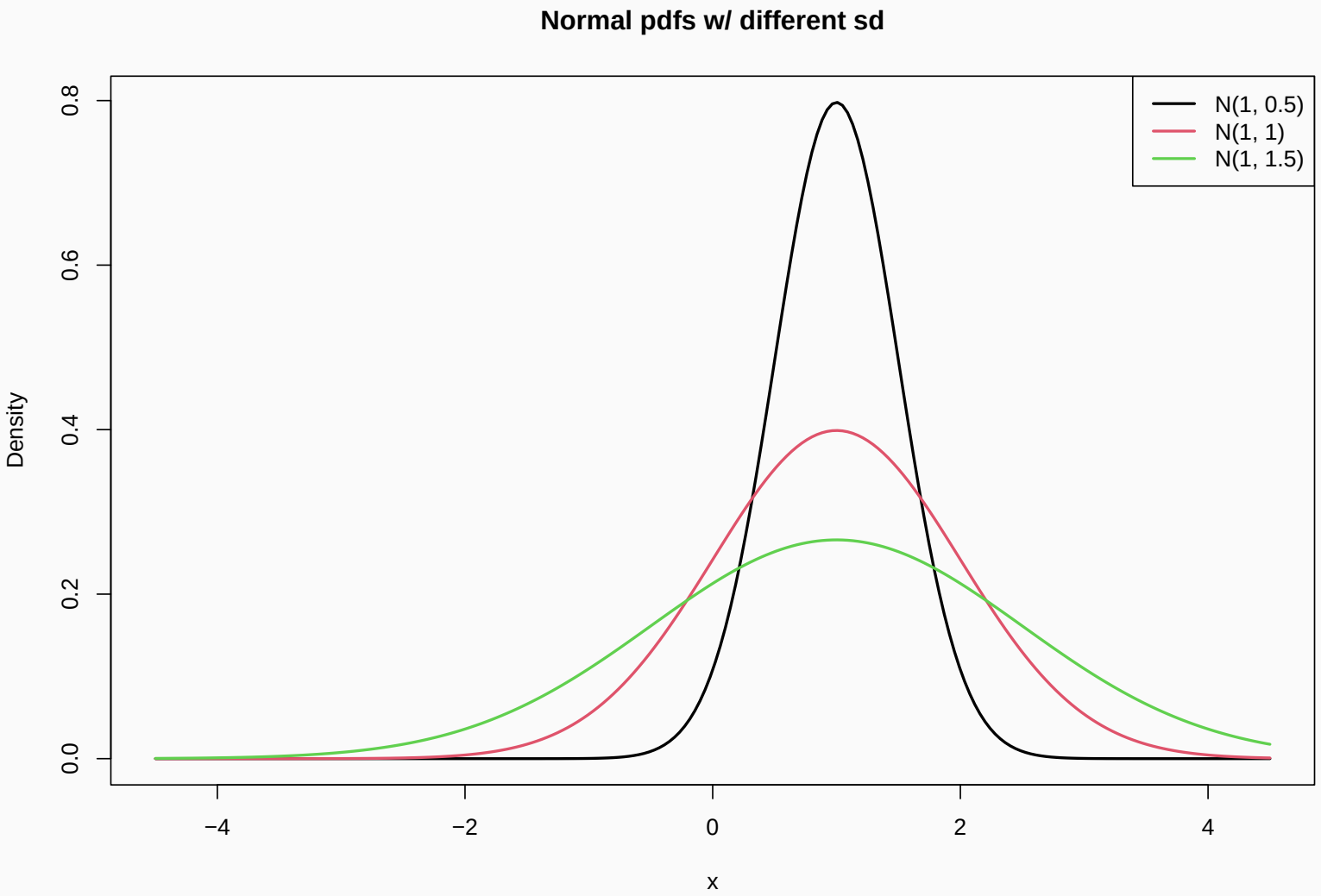
- parameters of the normal distribution - mean μ and standard deviationσ
- The *probability density function* (p.d.f) is given by

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

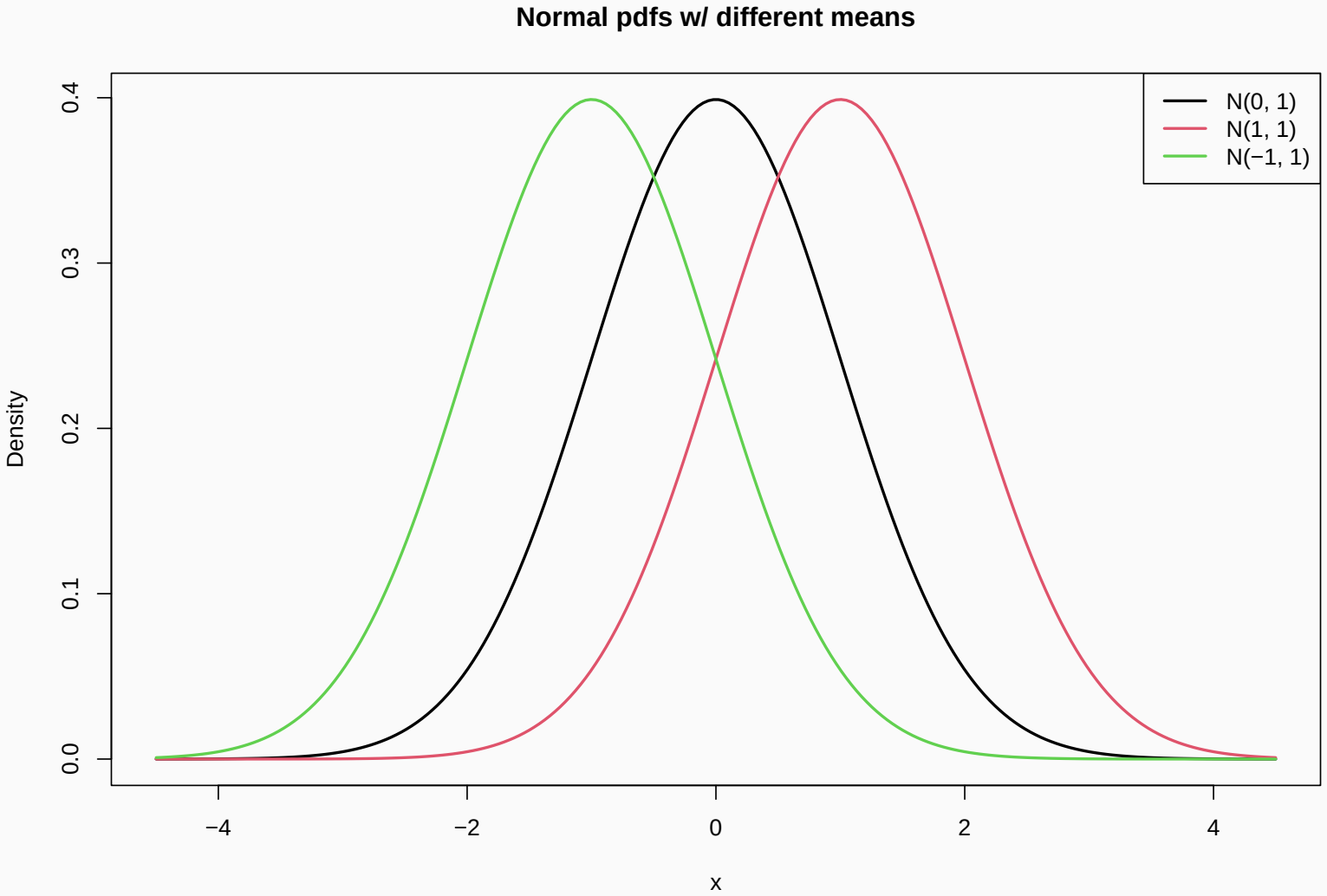


- The total area that lies under the curve is 1 or 100%
- $\mathbb{N}(\mu = 0, \sigma = 1)$ is called the standard normal distribution

A Family of Density Curves with same mean ($\mu = 1$)



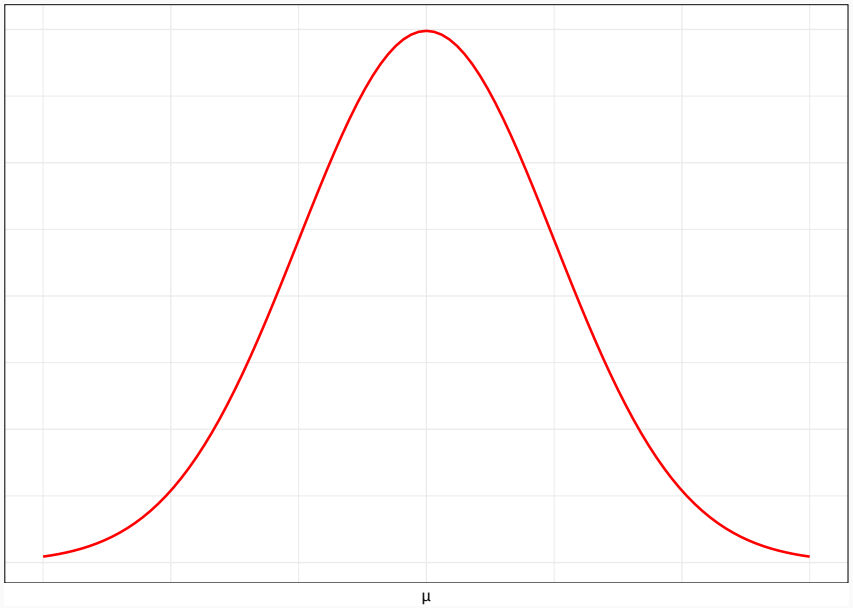
A Family of Density Curves the same standard deviation ($\sigma = 1$)



—>

—>

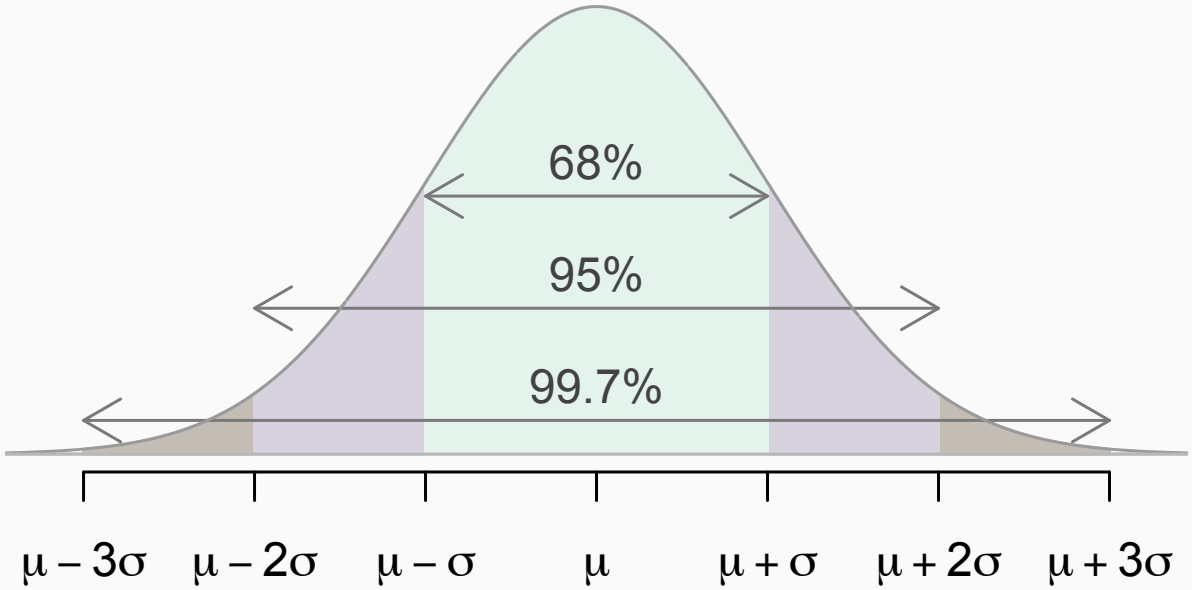
Properties of the Normal Distribution



- The mean, median, and mode are equal
- Bell shaped and is symmetric about the mean
- The total area that lies under the curve is 1 or 100%
- Probabilities are calculated as area under the curve between specific values, generally using the c.d.f
- As the curve extends farther and farther away from the mean, it gets closer and closer to the x axis but never touches it.
- The curve is approximately 6 standard deviations across.

68-95-99.7 (1-2-3 SD) Rule for Normal distribution

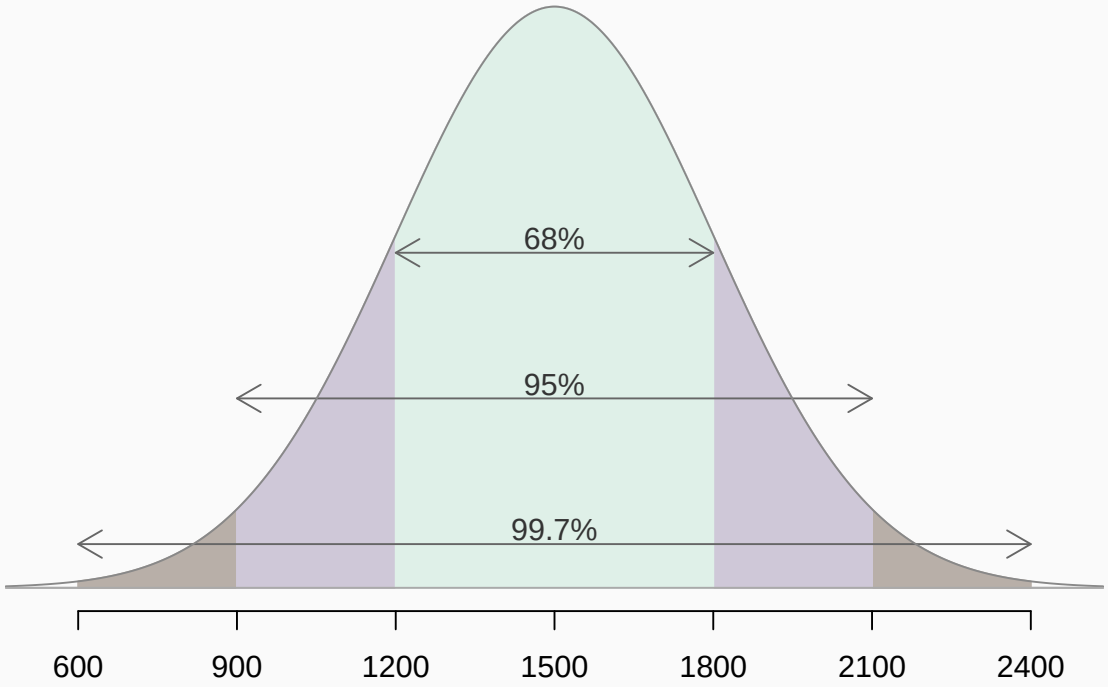
- For nearly normally distributed data,
  - about 68% falls within 1 SD of the mean,
  - about 95% falls within 2 SD of the mean,
  - about 99.7% falls within 3 SD of the mean.
- It is possible for observations to fall 4, 5, or more standard deviations away from the mean, but these occurrences are very rare if the data are nearly normal.



Describing variability using the 68-95-99.7 Rule

SAT scores are distributed nearly normally with mean 1500 and standard deviation 300.

- $\sim 68\%$ of students score between 1200 and 1800 on the SAT.
- $\sim 95\%$ of students score between 900 and 2100 on the SAT.
- $\sim 99.7\%$ of students score between 600 and 2400 on the SAT.

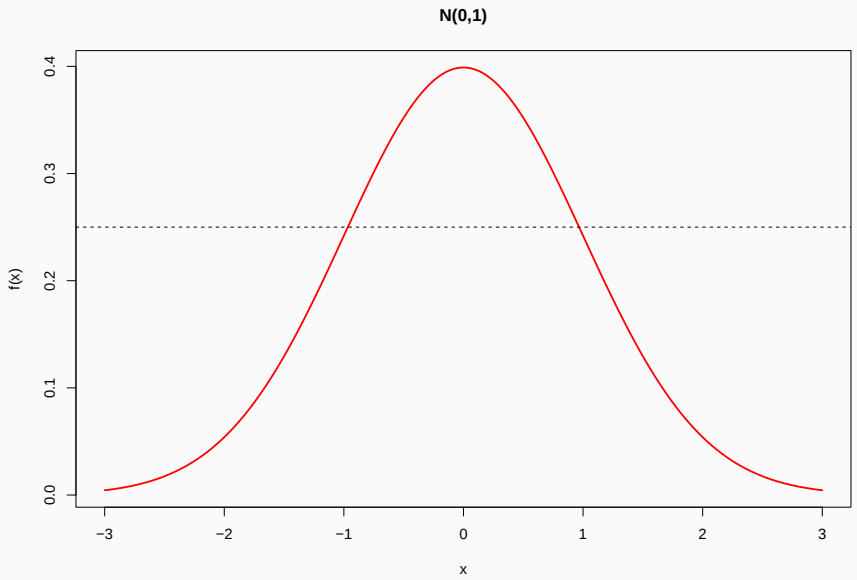


Normal Distribution Functions in R

- **p.d.f:** `dnorm(x, mean, sd)`
- **c.d.f ($\mathbb{P}(X \leq q)$):** `pnorm(q, mean, sd)`
- **Quantile:** `qnorm(p, mean, sd)`
- **simulation/sample generation:** `rnorm(n, mean, sd)`

Plot the standard normal density N(0,1)

```r
x <- seq(-3, 3, by = 0.01)
y <- dnorm(x) # default mean = 0 , sd = 1
plot(x, y, type = "l", col = "red", lwd=2,
     xlab = "x", ylab = "f(x)", main = "N(0,1)")
abline(h = 0.25, lty=2)
```

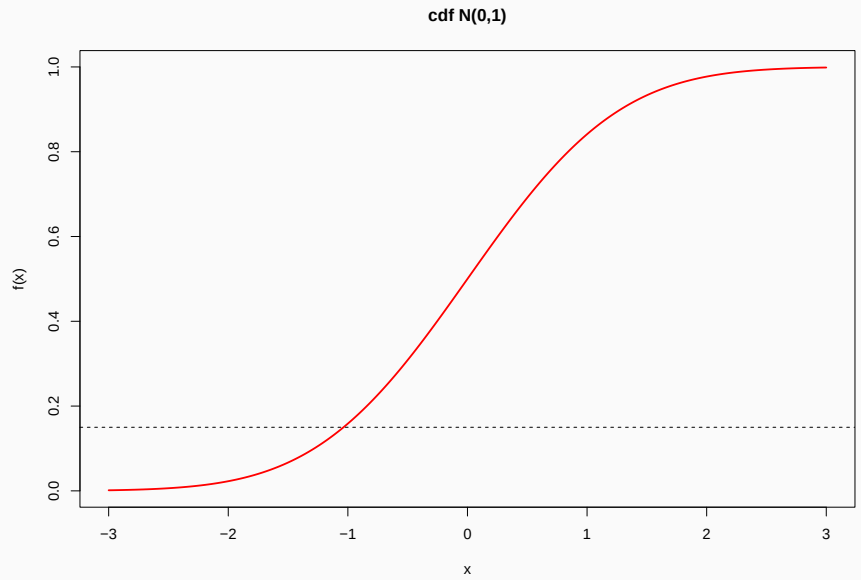



```r
dnorm(1) # pdf at x = 1
```

```
## [1] 0.2419707
```

Plot the cdf of the standard normal RV N(0,1)

```r
x <- seq(-3, 3, by = 0.01)
y <- pnorm(x) # default mean = 0 , sd = 1
plot(x, y, type = "l", col = "red", lwd=2,
     xlab = "x", ylab = "f(x)", main = "cdf N(0,1)")
abline(h = 0.15, lty=2)
```

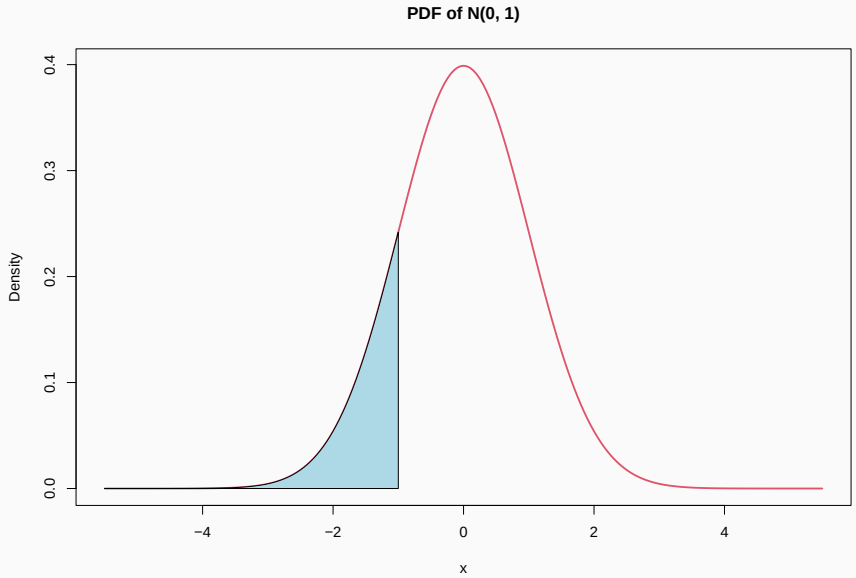


15th percentile: That x such that c.d.f at x $= 0.15$ or $P(X \leq x) = 0.15$ or area to the left of x is 0.15

Percentiles, Quantiles

15th percentile: That x such that area to the left of x is 0.15

$P(X \leq x) = 0.15$ c.d.f at x $= 0.15$




```
pnorm(-1) # cdf(-1) ~ 0.15 or P(X < -1) ~ 0.15
```

```
## [1] 0.1586553
```

```
round(qnorm(0.1586555),2) #  15% percentile is -1(inverse cdf)
```

```
## [1] -1
```

Visual check for normality using Normal QQ plots

Is the Data Normal? `qqnorm()` and `qqline()`

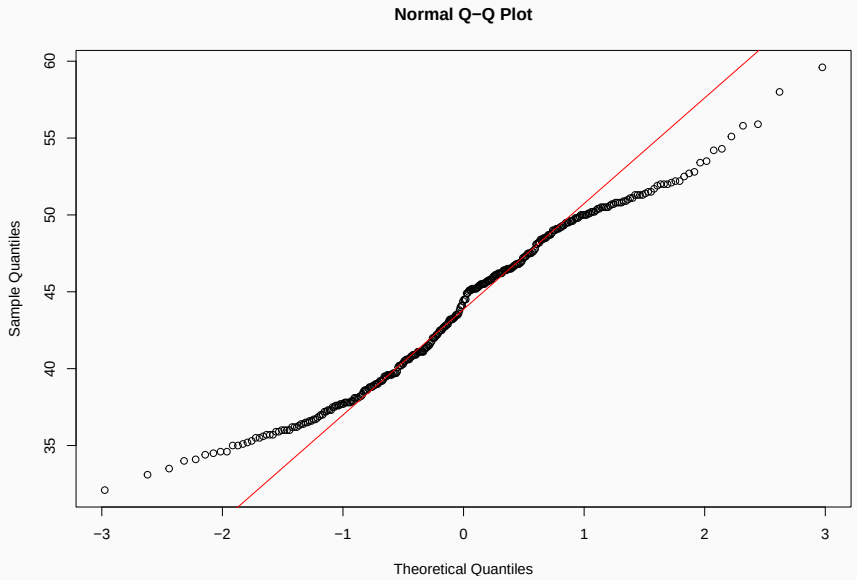
- The Normal QQ plot, or quantile quantile plot, is a graphical tool to help us assess if a set of data plausibly came from a normal distribution.

- **Normal Q-Q plots:**
  - Quantiles taken from sample data, plotted against quantiles calculated from a theoretical distribution.
  - If the points fall on approximately a straight line, we can assume normality

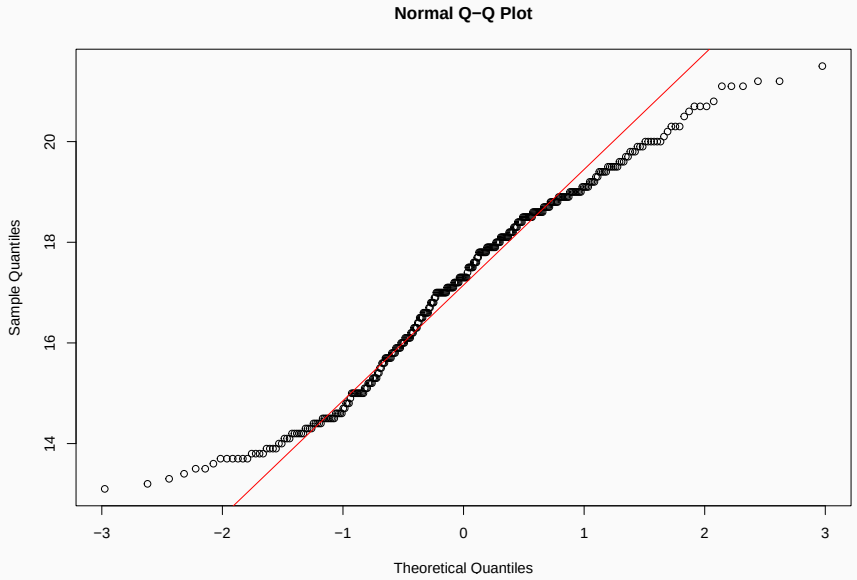
- In R, In R, we create Q Q plots using `qqnorm()`

Checking for normality: Is our data normally distributed

```
library(palmerpenguins)
qqnorm(penguins$bill_length_mm)
qqline(penguins$bill_length_mm, col = "red")
```

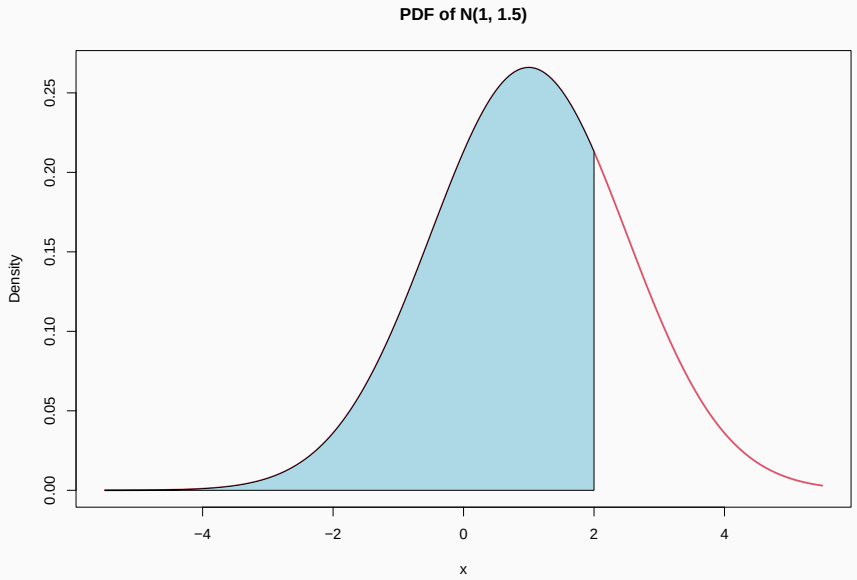


```r
library(palmerpenguins)
qqnorm(penguins$bill_depth_mm)
qqline(penguins$bill_depth_mm, col = "red")
```



Probability calculation: Shade the required area

$X \sim N(1, 1.5)$, what is $P(X \leq 2)$

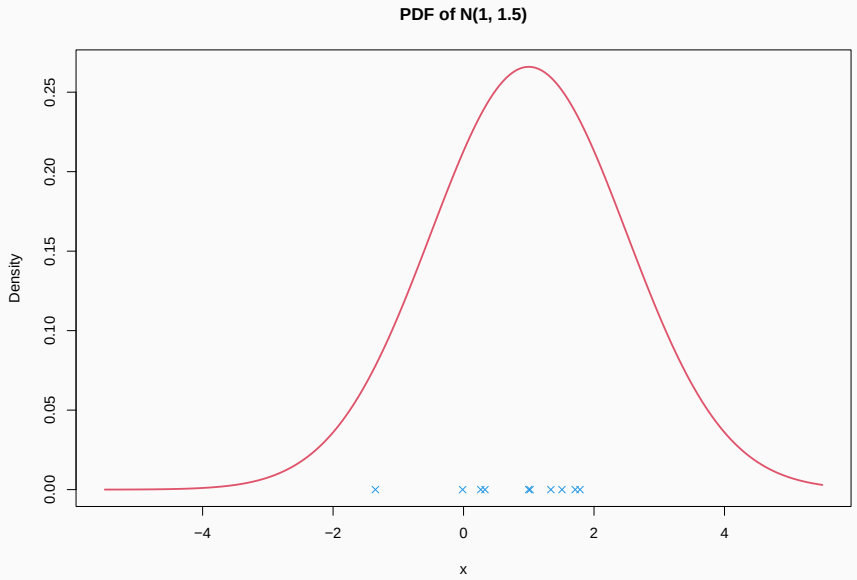



```
pnorm(2, mean = 1, sd = 1.5)
```

```
## [1] 0.7475075
```

Simulating normal variates (observations)

Generate a sample of size 10 from N(1,1.5)




```
set.seed(10262022)
rnorm(10, mean = 1, sd = 1.5)
```

```
##  [1]  0.32833073  1.33860079 -1.35139136  1.78558872  0.99898
##  [7]  1.71126107  1.01991149 -0.01380774  0.26431305
```

questions you should be able to answer

- Identify distributions and calculate probabilities
- Know different R functions for probability calculations
  - d : density
  - p : cdf
  - q : quantile
  - r : random sample/simulation of a sample.

We did:

PDF, plotting pdf, cdf, probability calculations by hand and using R for Continuous uniform, normal distributions.

binomial distribution $\mathrm{Binom}(\text{size}, \text{prob})$

- `dbinom(x, size, prob)`
- `pbinom(q, size, prob)`
- `rbinom(n, size, prob)`

uniform distribution $\mathrm{Unif}(\text{min}, \text{max})$

- `dunif(x, min, max)`
- `punif(q, min, max)`
- `runif(n, min, max)`

normal distribution $N(\text{mean}, \text{sd})$

- `dnorm(x, mean, sd)`
- `pnorm(q, mean, sd)`
- `rnorm(n, mean, sd)`