10. Named Continuous Distributions in R

Principles of Data Science with R

Dr. Uma Ravat

PSTAT 10

Announcement

1. Exam 1 is in one week - next Thursday 11/03 during class 8:00am - 9:15am.
2. Format: about 40 questions (multiple choice, short answers) in 60 minutes
   - 30 - 50 % similar/same as questions you've seen before(lecture, worksheet, hw, quiz).
   - All material covered till day before midterm.
3. No collaboration
4. No cheat sheets, formula sheets
5. Seating chart will be released on Canvas on Wednesday 11/02 Nov 2 5pm. (It is likely some students will take the exam at Embarcadero Hall, in which case you need to go there directly. Check the seating chart before heading to the exam location on Thursday.)

How to prepare

6. HW5 due Wednesday 11/02 will be released early on Thursday(today) at noon
7. Quiz 4 will open tomorrow(Friday) at 9am and will close at 9pm
   - *No collaboration*
   - **No posting to sites like Chegg**
   - Good Luck!!
8. Review all material - slides, your turn, worksheet, hw, quiz
9. As you review make a summary of concepts, functions for you to browse as you study for the exam. You will not be allowed any notes during the exam.
10. Extra Practice Problems will be posted on Saturday.

Summary:

- Named discrete distributions
  - Discrete uniform
  - Binomial Distribution

Next:

- Named continuous distributions
  - Uniform Distribution
  - Normal Distribution

Recall

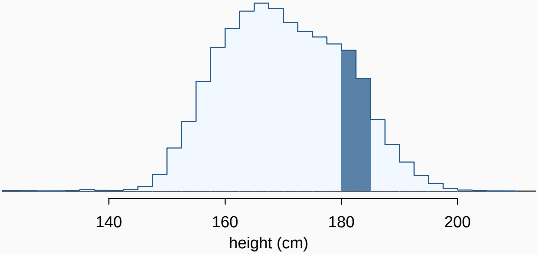
- A random variable is discrete when we can *count* the number of outcomes.
- A random variable is continuous when the outcomes can be *measured*.
    - A continuous rvtakes all values in an interval of real numbers.

Examples of continuous RVs

- Height
- Weight
- Time
- Temperature

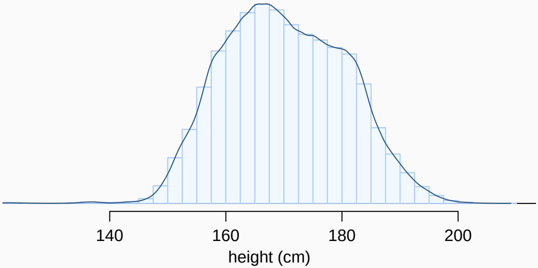
Probability from histogram

- Below is a histogram of the distribution of heights of US adults.
- The proportion of data that falls in the shaded bins gives the probability that a randomly sampled US adult is between 180 cm and 185 cm (about 5'11" to 6'1").



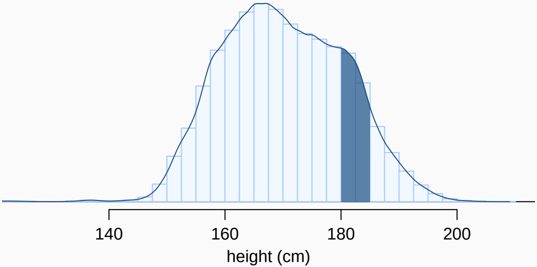
From histograms to continuous distributions

Since height is a continuous numerical variable, its **probability density function** is a smooth curve.



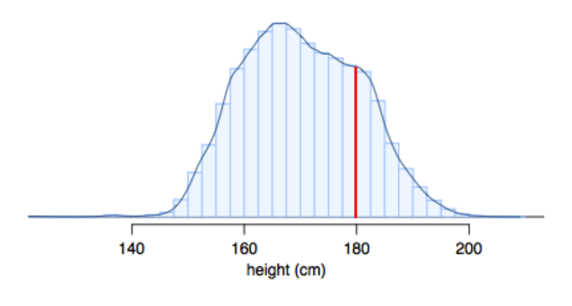
Probabilities from continuous distributions

Therefore, the probability that a randomly sampled US adult is between 180 cm and 185 cm can also be estimated as the shaded area under the curve.



By definition. . .

Since continuous probabilities are estimated as “the area under the curve”, the probability of a person being exactly 180 cm (or any exact value) is defined as 0.



What does this say about $\mathbb{P}(X \leq 180)$ vs. $\mathbb{P}(X < 180)$?

Distribution of a continuous RV

- is specified by its Probability Density Function (p.d.f.)
- The pdf can be represented by a function or its graph called density function or the density curve respectively
- the probabilities are given by the area under the graph between specified values
  - If X is a continuous r.v., then $P(X = x) = 0$ for all values x.
- The total area under a density curve is always equal to 1.

Continuous Probability Distributions

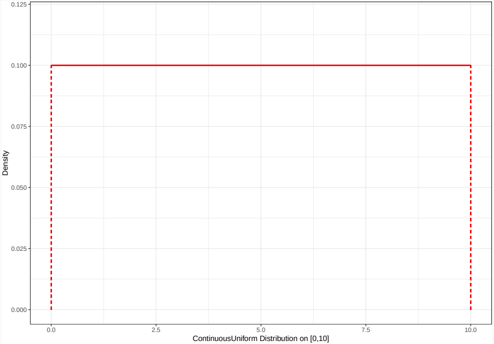
1. The Uniform Distribution
2. The Normal Distribution

The Continuous Uniform Distribution

- All values are equally likely to occur
- the pdf has a uniform shape (looks the same) across the entire range of values.

The Continuous Uniform Distribution

- All values are equally likely to occur
- the pdf has a uniform shape (looks the same) across the entire range of values.



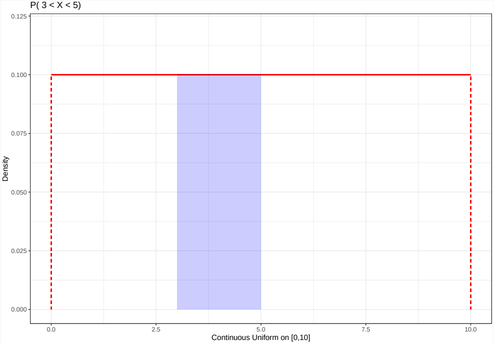
- The mean of the distribution?

The Continuous Uniform Distribution

Probability calculations: By hand

$P(X < 5), P(X \leq 5), P(3 \leq X \leq 5), P(3 < X \leq 5), P(3 \leq X < 5), P(X > 5), P(X \geq 5)$

- Area under the density curve -> Area of rectangles



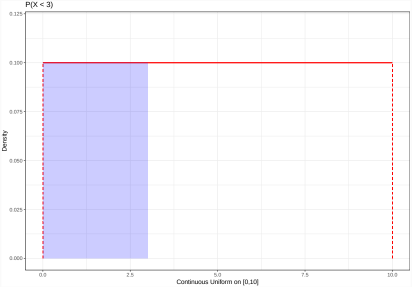
The Continuous Uniform Distribution

Probability calculations: Using R:

$P(X < 5), P(X \leq 5), P(3 \leq X \leq 5), P(3 < X \leq 5), P(3 \leq X < 5)$

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

$P(X \leq q)$ is the area under the density curve to the **left** of q



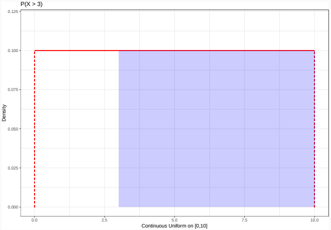
The Continuous Uniform Distribution

Probability calculations: Using R:

$P(X > 5), P(X \geq 5)$

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

$P(X \geq q)$ is the area under the density curve to the **right** of q



The continuous Uniform distribution on [a,b]

$X \sim \mathrm{UNIF}(a, b)$ then

- a and b are the parameters of the distribution, defining the lower and upper limit for the possible values X can take.
- The *probability density function* (p.d.f) is given by $f(x) = \frac{1}{b-a}$, if $a \leq x \leq b$
  - area under the density curve is 1
  - p.d.f in R: `dunif(x, min = a, max = b, log = FALSE)`
- **mean:** $E(X) = \mu = \frac{a+b}{2}$
- **probability calculations**
  - By hand: Area under the density curve -> Area of rectangles
  - By R: `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
```

Example: 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?

- 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 $0 \leq x \leq 10$
- We make the height of the density curve 0.10 so that the total area under the curve is $(0.10)(10) = 1$:

Example: Time Spent Waiting for a Bus

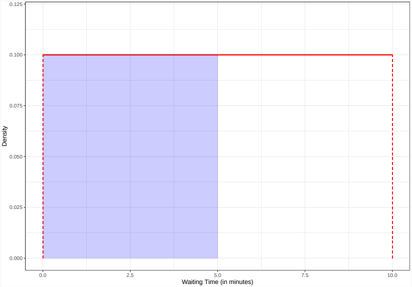
What is the probability the waiting time, X,

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

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?

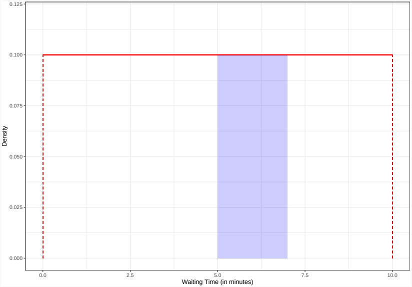


```r
punif(5, min = 0, max = 10)
```

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

Example: Time Spent Waiting for a Bus

2. between 5 and 7 minutes?

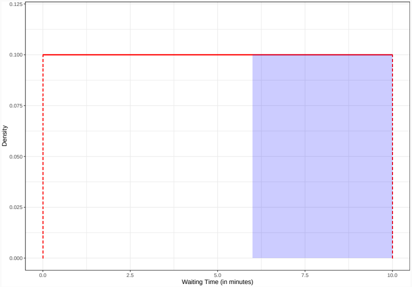


```
punif(7, min = 0, max = 10) - punif(5, min = 0, max = 10)
```

```
## [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)
```

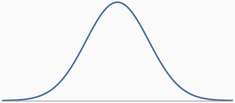
```
## [1] 0.4
```

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

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

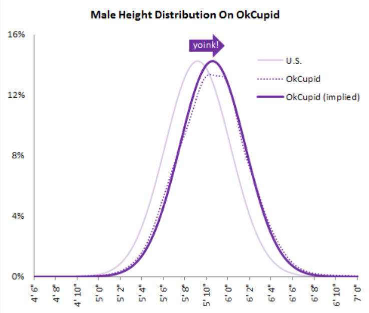
Named Continuous distribution: Normal distribution

- 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

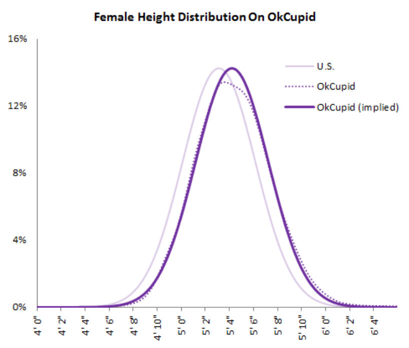


"The male heights on OkCupid very nearly follow the expected normal distribution – except the whole thing is shifted to the right of where it should be. Almost universally guys like to add a couple inches."

"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."

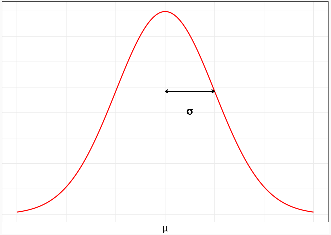
{http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating/}

Normal Distribution

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

- μ and σ are parameters for the normal distribution denoting the mean and standard deviation respectively.
- The *probability density function* (p.d.f) is given by

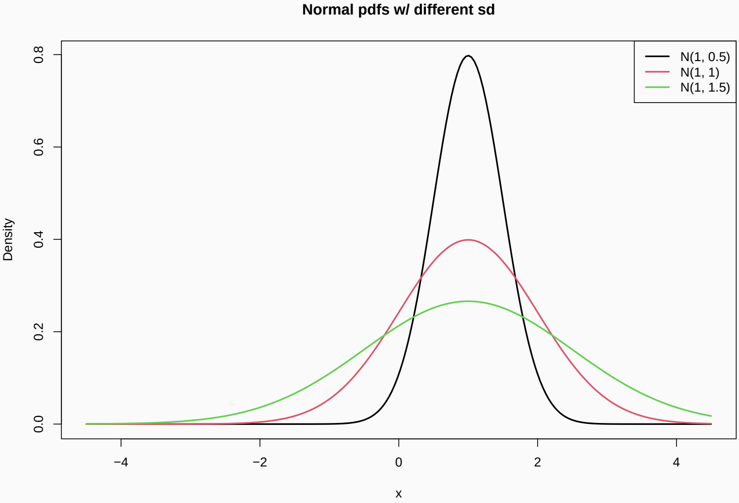
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



- The total area that lies under the curve is 1 or 100%

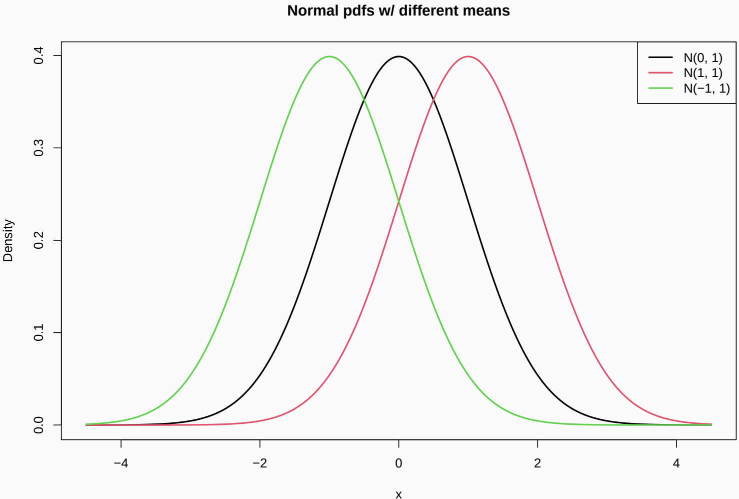
A Family of Density Curves

- If we fix the same mean ($\mu = 1$) and change the *standard deviations*, we obtain the following family of curves:

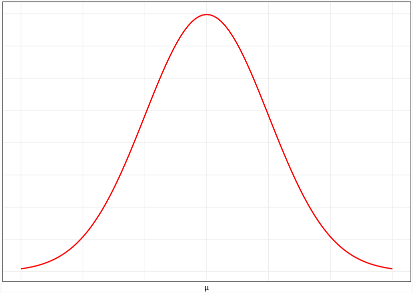


A Family of Density Curves

- If instead we fixed the same standard deviation ($s = 1$) and allow the *means* to vary, we obtain the following family of curves:



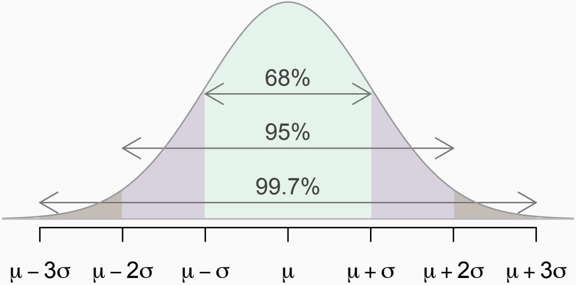
Properties of the Normal Distribution



- 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

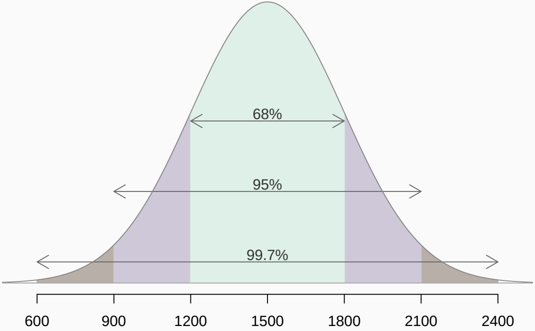
- For nearly normally distributed data,
  - about 68% falls within 1 SD of the mean,
  - about 95% falls within 2 SD of the mean,
  - about 99.7% falls within 3 SD of the mean.
- It is possible for observations to fall 4, 5, or more standard deviations away from the mean, but these occurrences are very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

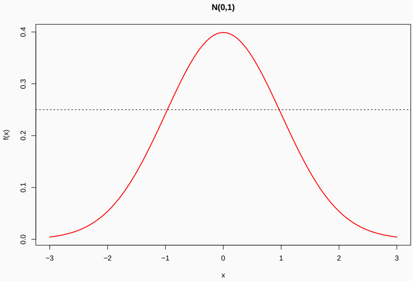


Normal Distribution Functions in `R`

- **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** `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)
```

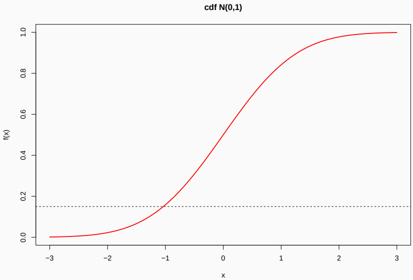


```r
dnorm(1) # pdf at x = 1
```

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

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

```
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)
```

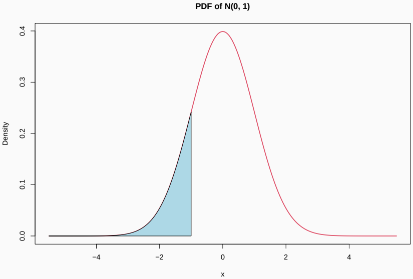


15th percentile: That x such that area to the left of x is 0.15

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

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

Recall : 5 number summary and Box plots

Summarizing numerical data : 5 number summary

- Min, 1st quatile, Median, 3rd quartiles, Max
- `c(min(x), quantile(x,0.25), median(x), quantile(x,0.75), max(x))`
- `summary()`

5 number summary

```r
library(tidyverse)
x <- na.omit(starwars$height)
c(min(x), quantile(x,0.25), median(x),
  quantile(x,0.75), max(x))
```

```
##      25%     75%
##  66 167 180 191 264
```

```r
summary(x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    66.0   167.0   180.0   174.4   191.0   264.0
```

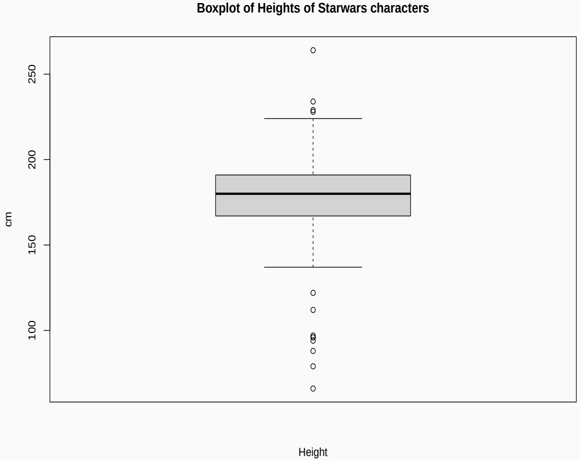
Inter quartile range = Middle 50% of the distribution

Whiskers : 1.5*interquartile range below 1st quartile(25th percentile) and above the 3rd quartile(75th percentile)

```
iqr = quantile(x,0.75) - quantile(x,0.25)
iqr1.5 = 1.5*iqr
lower = quantile(x,0.25) - iqr1.5
upper = quantile(x,0.75) + iqr1.5
boxplot_points = c(lower,quantile(x,0.25), quantile(x,0.50
names(boxplot_points) <- c("lower", "25th percentile", "med
print(boxplot_points)
```

```
##           lower 25th percentile          median 75th per
##             131             167             180
```

Visualizing numerical data : box plot



Back to normal distributions

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

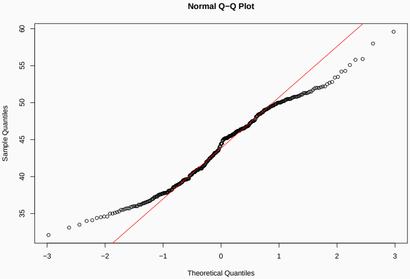
- **Visual check for normality**:
  - The Normal Q Q plot, or quantile quantile plot, is a graphical tool to help us assess if a set of data plausibly came from a normal distribution.

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

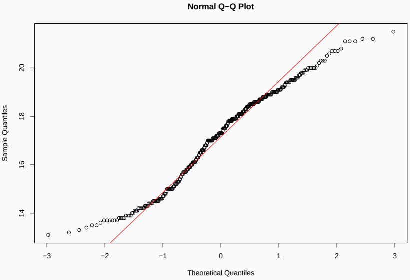
- In R, In R, we create Q Q plots using qqnorm()

Checking for normality: Is our data normally distributed

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

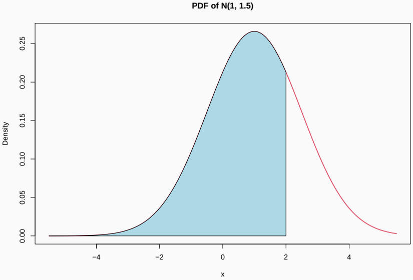


```
library(palmerpenguins)
qqnorm(penguins$bill_depth_mm)
qqline(penguins$bill_depth_mm, col = "red")
```



Probability calculation: Shade the required area

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

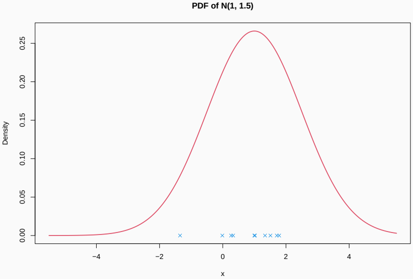


```r
pnorm(2, mean = 1, sd = 1.5)
```

```
## [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.
##  [7]  1.71126107  1.01991149 -0.01380774  0.26431305
```

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)`