STA220H1: The Practice of Statistics I

Elena Tuzhilina

February 14, 2023

Please turn on your videos :)

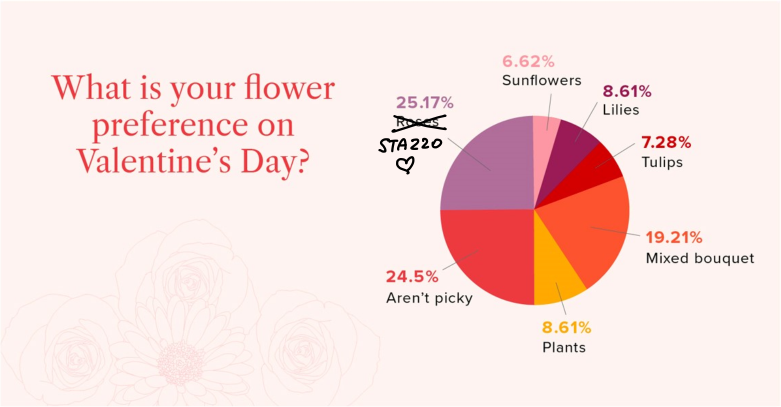


Figure 1: [picture source]

Learning strategy

1. Attend lectures
2. Watch modules at https://sta220.utstat.utoronto.ca
3. Do practice sets, attend TAs office hours if something is not clear
4. Do Quiz, attend my office hours on Monday if something is still not clear
5. Post your questions on Piazza (not my personal email, pls!)

Agenda for today

- Recap: expectation and variance
- Continuous random variables: uniform and normal distribution
- Sample mean distribution: normal sample and CLT

Discrete: expected value

Expected value measures the average value of random variable in long term.

$$E(X) = \sum_x x \cdot P(X = x)$$

Variance and standard deviation measure the spread of the values of a random variable in long term.

$$Var(X) = \sum_x (x - E(X))^2 \cdot P(X = x)$$

$$sd(X) = \sqrt{Var(X)}$$

Exercise

What is the expectation and variance of Bernoulli random variable with $p = 0.1$? What is the formula for general p?

X	1	0
$P(x)$		

Expected value vs. sample mean

- We have a random variable X
- We generate a sample of size n using this random variable $x_1, \ldots, x_n$

What is the relationship between $E(X)$ and $\bar{x} = \frac{x_1 + \ldots + x_n}{n}$?

Important rules

Expectation

- If X is a random variable and a, b are some numbers then

$$E(a \cdot X + b) = a \cdot E(X) + b$$

- If Y is also a random variable then

$$E(X + Y) = E(X) + E(Y)$$

Variance

- If X is a random variable and a, b are some numbers then

$$Var(a \cdot X + b) = a^2 \cdot Var(X)$$

- If Y is also a random variable and it is independent of X then

$$Var(X + Y) = Var(X) + Var(Y)$$

Exercise

(!) If $X_1 \ldots X_n \sim \text{Bernoulli}(p)$ then
$y = X_1 + \ldots + X_n \sim \text{Binomial}(n, p)$

What is the expectation and variance of Binomial random variable with $n = 5$ and $p = 0.1$? What is the formula for general n and p?

Random variable: discrete

Discrete - takes one of a countable list of distinct values

- sex of a baby
- number of heads when tossing ten coins
- number of zeroed in your student ID

Special cases:

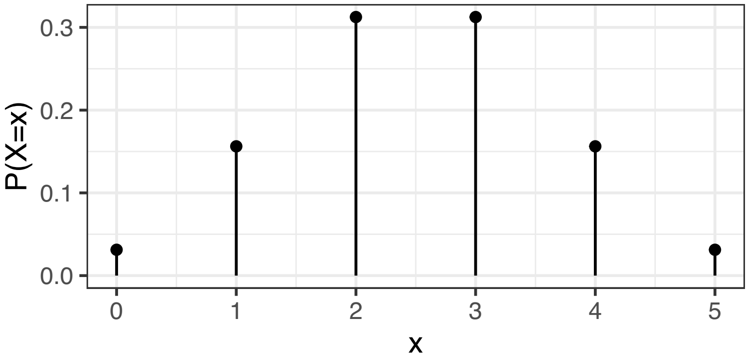
$$X \sim Bernoulli(p)$$

$$X \sim Binomial(n, p)$$

Distribution: discrete

Two ways to present the distribution: a table and a plot

X	0	1	2	3	4	5
$P(X)$	0.03125	0.15625	0.3125	0.3125	0.15625	0.03125



Random variable: continuous

Continuous - takes any value in an interval or collection of intervals

- the wait time for the next bus
- birth weight of a baby
- life expectancy in Canada

Special cases:

$$X \sim Uniform(l, u)$$

$$X \sim Normal(\mu, \sigma^2)$$

Continuous: uniform

You live in a building that has an elevator. Once you push the button to call the elevator, it takes between 0 and 5 seconds (equally likely) for the elevator to arrive.

$$X = \text{wait time (in seconds)} \sim \mathit{Uniform}(0, 5)$$

$$P(0 \leq X \leq 2.5) =$$

$$P(4 \leq X \leq 5) =$$

Continuous: uniform

Since continuous variables can take *so many* possible values, we cannot use distribution tables. Instead we use **probability density functions**.

$P(a \leq X \leq b) =$ area under the curve bounded by a and b vertical lines

Exercise

Find $P(2.5 \leq X \leq 4.5)$ and $P(X \geq 4)$.

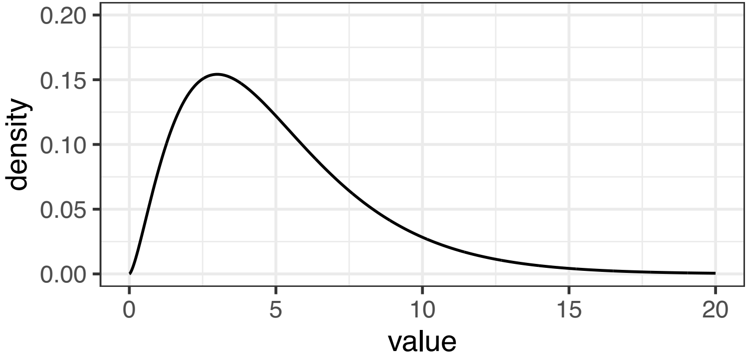
What is $P(X = 4)$? What is $P(X > 4)$?

Continuous: density function

A curve is a valid **density function** if:

- It is greater or equal to 0
- The total area under the curve is 1

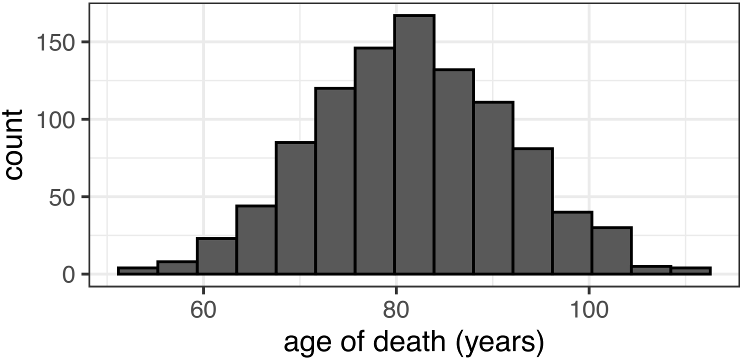
$P(a \leq X \leq b) =$ area under the curve bounded by a and b vertical lines



Histogram vs density curve

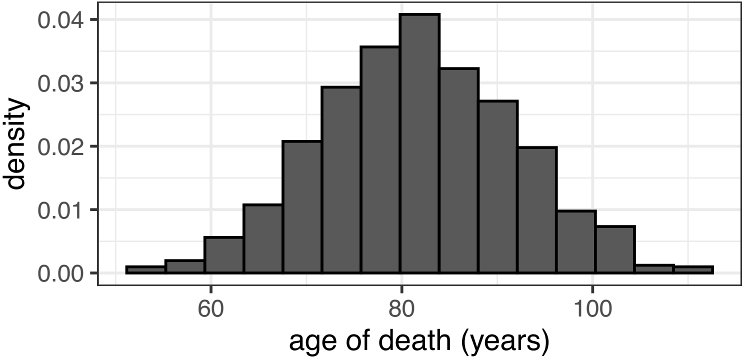
Suppose you are interested in the life expectancy in Canada. You record the age of death for 1000 Canadians.

```
## [1] 72.03085 82.84849 96.87845 69.69624 80.19748 82.32420
```



Histogram vs density curve

You can convert a histogram to an approximate density by changing the scale of y-axis.



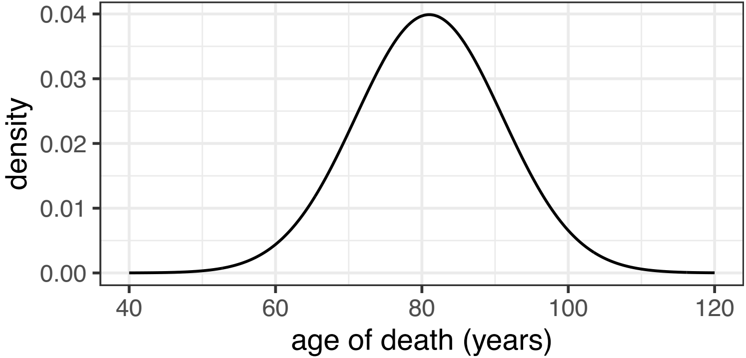
Histogram vs density curve

The “smoothed” version of you histogram is density.

Continuous: normal

Normal random variable (or Gaussian) has symmetric, bell-shaped and unimodal distribution.

$$X = \text{age of death (in years)} \sim \mathit{Normal}(81, 100)$$

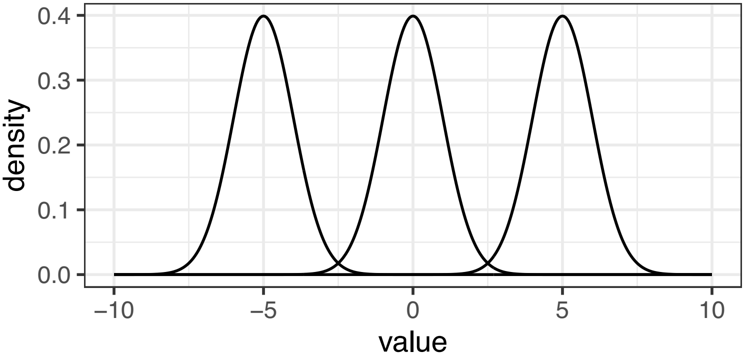


Continuous: normal

Normal distribution $X \sim Normal(\mu, \sigma^2)$ has **two parameters**

$$\mu = E(X) \text{ and } \sigma^2 = Var(X)$$

- μ controls the "center" of the distribution

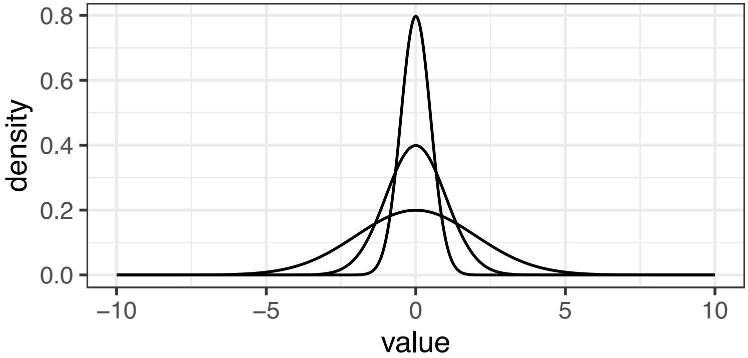


Continuous: normal

Normal distribution $X \sim Normal(\mu, \sigma^2)$ has **two parameters**

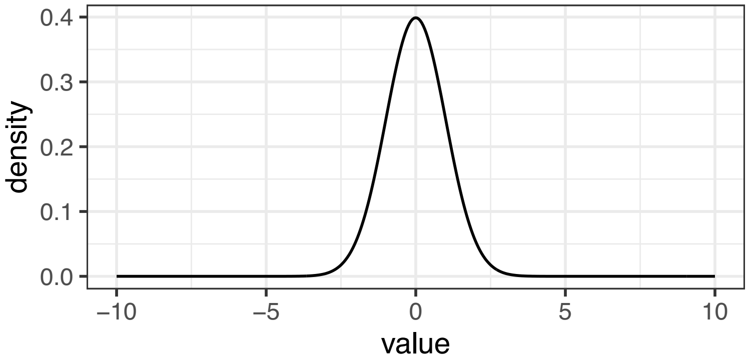
$$\mu = E(X) \text{ and } \sigma^2 = Var(X)$$

- σ^2 controls the "spread" of the distribution



Standard normal

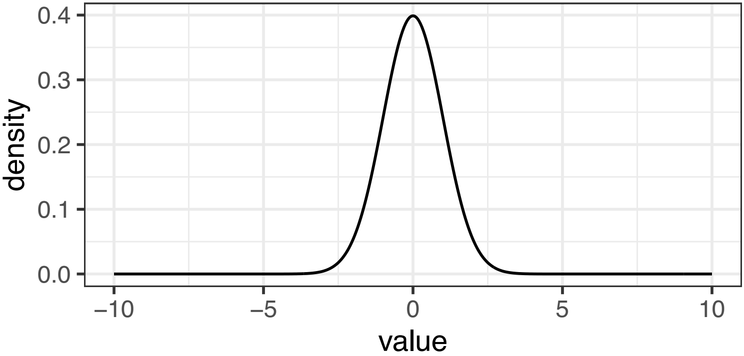
Standard normal distribution has $\mu = 0$ and $\sigma^2 = 1$.



Standard normal

To find the probabilities for standard normal we use the distribution table.

If $X \sim Normal(0, 1)$ what is the probability $P(X \leq -1.25)$?



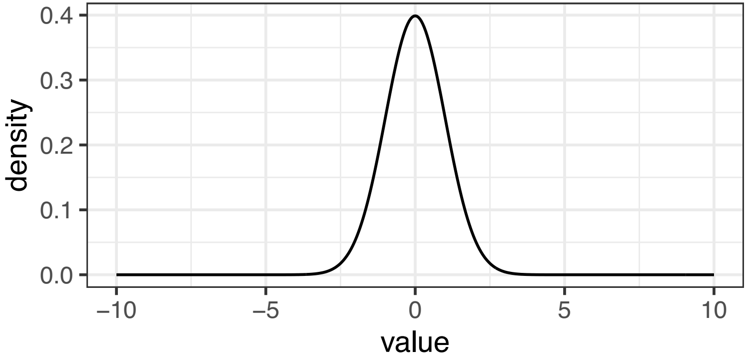

```
pnorm(-1.25)
```

```
## [1] 0.1056498
```

Standard normal

To find the probabilities for standard normal we use the distribution table.

If $X \sim Normal(0, 1)$ what is the probability $P(X \leq 1.25)$?




```
pnorm(1.25)
```

```
## [1] 0.8943502
```

Exercise

If $X \sim Normal(0, 1)$ what is the probability $P(X > 1.25)$?

If $X \sim Normal(0, 1)$ what is the probability $P(X = 1.25)$?

If $X \sim Normal(0, 1)$ what is the probability $P(X \geq 1.25)$?

Normal: properites

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.68$$
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.95$$
$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.997$$

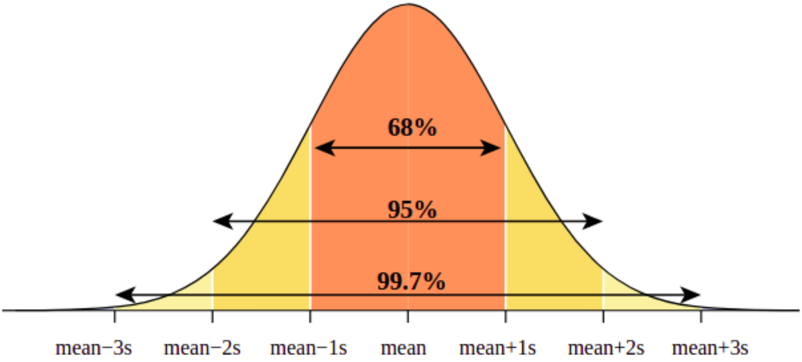


Figure 9: [...]

Normal: properites

- if X is normal then $a \cdot X + b$ is also normal

- X and Y are independent and normal then $X + Y$ is normal

Standard normal

Standardization: if $X \sim Normal(\mu, \sigma^2)$ then $\frac{X-\mu}{\sigma}$ is standard normal.

For example, if life expectancy

$$X \sim Normal(81, 100)$$

then

$$P(X \leq 75) =$$

```
pnorm(75, mean = 81, sd = 10)
```

```
## [1] 0.2742531
```

Exercise

The life expectancy in Canada follows normal distribution with mean 81 and standard deviation 3. What is the probability to live longer than 90 years?

What is the life expectancy corresponding to the first quartile?

Sample mean distribution

The theoretical model of the life expectancy is $X \sim Normal(81, 100)$. We record the ages of death for 10 Canadians and average them.

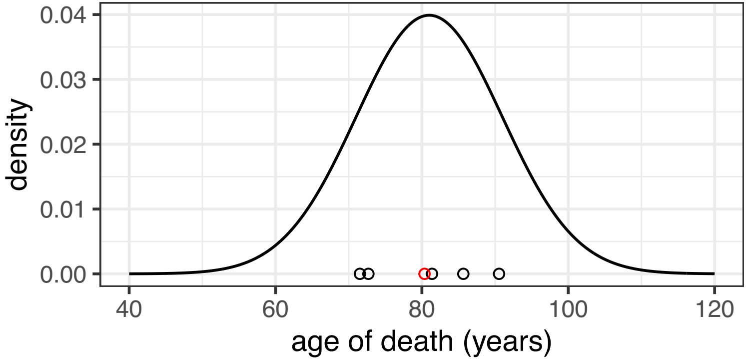
Sample mean distribution

```
ages = rnorm(5, mean = 81, sd = 10)
ages
```

```
## [1] 85.66200 90.54666 71.52794 81.38563 72.69118
```

```
mean(ages)
```

```
## [1] 80.36268
```



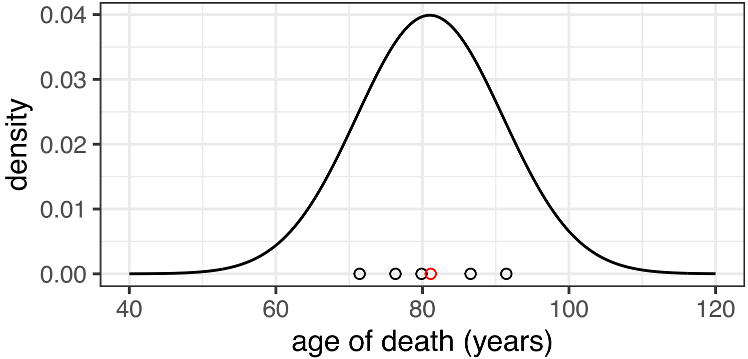
Sample mean distribution

```
ages = rnorm(5, mean = 81, sd = 10)
ages
```

```
## [1] 86.57172 79.87897 71.42362 76.31927 91.44757
```

```
mean(ages)
```

```
## [1] 81.12823
```



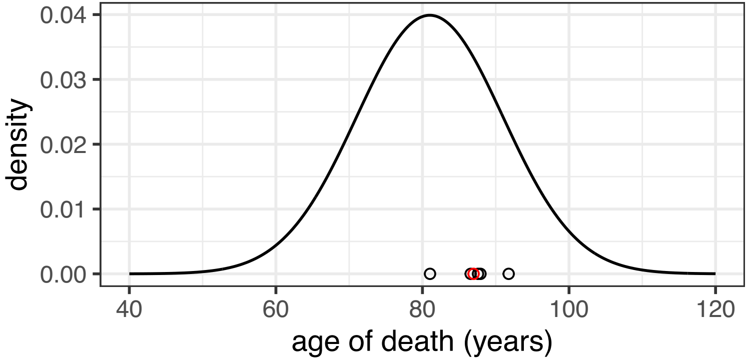
Sample mean distribution

```
ages = rnorm(5, mean = 81, sd = 10)
ages
```

```
## [1] 81.03306 91.76645 86.58143 87.91968 87.57774
```

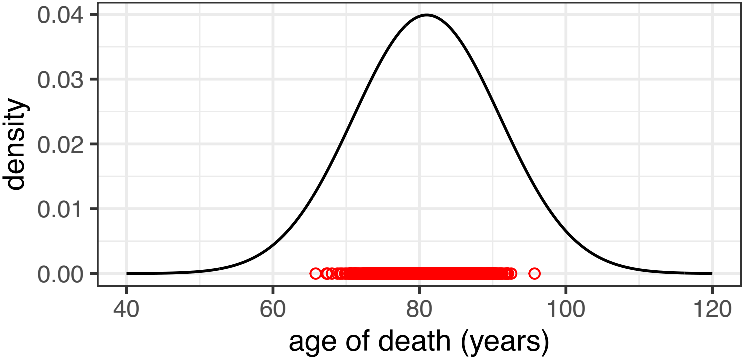
```
mean(ages)
```

```
## [1] 86.97567
```



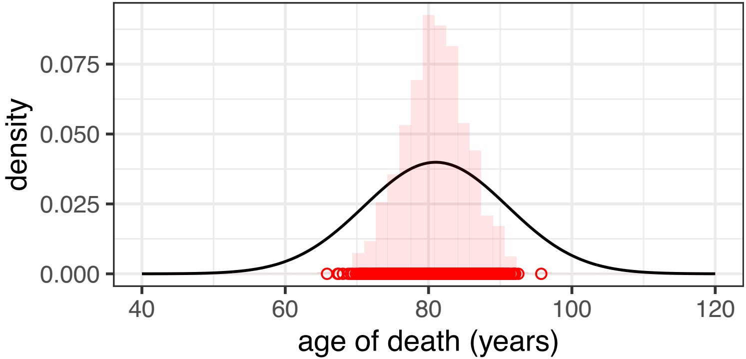
Sample mean distribution

Let's repeat this experiment many times.



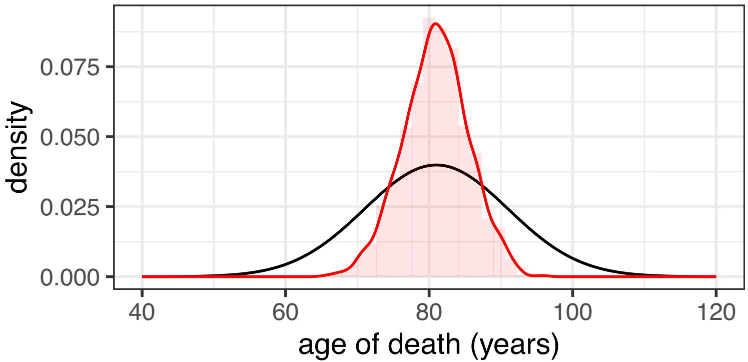
Sample mean distribution

What is the distribution of the sample mean?



Sample mean distribution

What is the distribution of the sample mean?



Exercise

We have 5 observations, each of them were generated from $Normal(81, 100)$.

If $X_1, \ldots, X_5 \sim Normal(81, 100)$ and independent, then

$$\bar{X} = \frac{X_1 + \ldots + X_5}{5} \sim Normal(81, 20).$$

Sample mean distribution

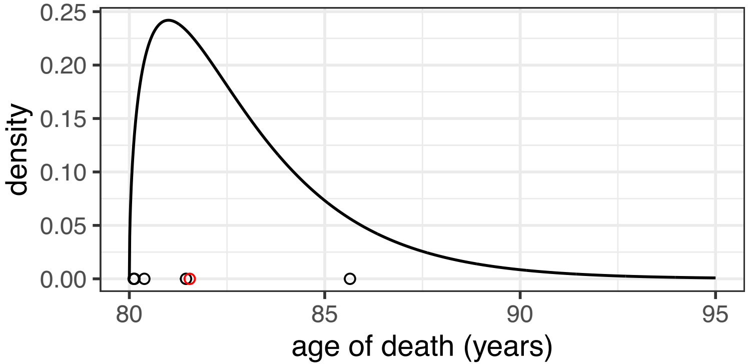
We have n observations, each of them were generated from $Normal(\mu, \sigma^2)$.

If $X_1, \ldots, X_n \sim Normal(\mu, \sigma^2)$ and independent, then

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n} \sim Normal\left(\mu, \frac{\sigma^2}{n}\right).$$

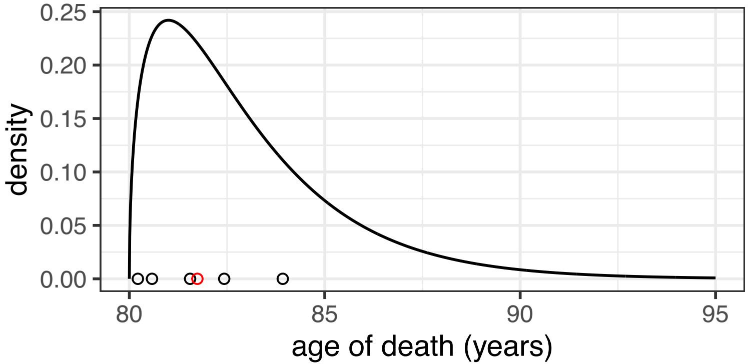
Sample mean distribution

What if the distribution of age is not normal?



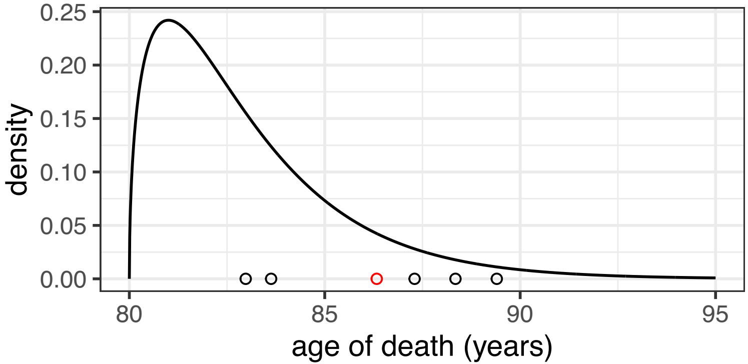
Sample mean distribution

What if the distribution of age is not normal?



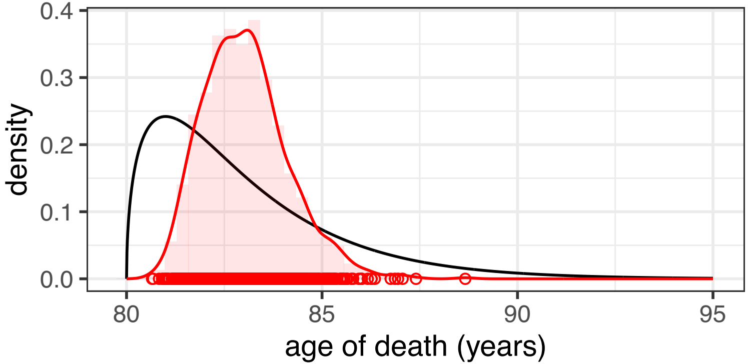
Sample mean distribution

What if the distribution of age is not normal?



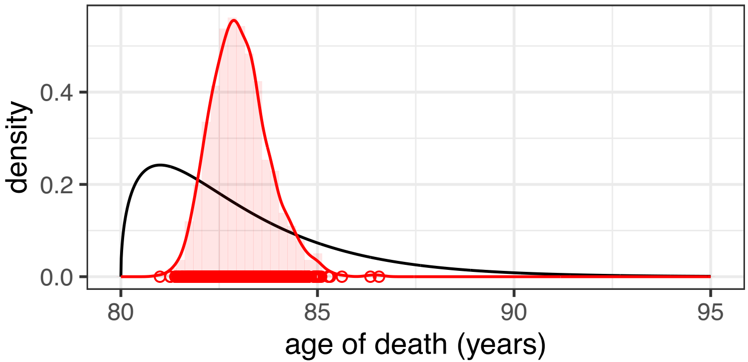
Sample mean distribution

What if the distribution of age is not normal?



Sample mean distribution

Now, let's increase the sample size from 5 to 10.



Sample mean distribution

Now, let's increase the sample size from 10 to 100.

Central limit theorem

If $X_1, \ldots, X_n$ have the same distribution (not necessary normal!) with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$, then for n large enough

- the distribution of $\bar{X} = \frac{X_1 + \ldots + X_n}{n}$ is well approximated by normal
- the expectation of normal distribution is μ
- the variance of normal distribution is $\frac{\sigma^2}{n}$

$$\bar{X} \text{ approximately} \sim Normal\left(\mu, \frac{\sigma^2}{n}\right)$$

Central limit theorem

CLT works even if $X_1, \ldots, X_n$ have discrete distribution.

For example, if $X_i \sim Bernoulli(p)$, then
$Y = X_1 + \ldots + X_n \sim Binomial(n, p)$ total number of successes.

The proportion of successes $\bar{X} = \frac{Y}{n}$ has approximately normal distribution for large n.

Exercise

What are the parameters of this normal distribution?

TO DO

1. Module 2. Probability: Random Variables and Module 3. Sampling Distributions
2. Quiz 6 due Monday (February 20) @ 11:59 PM (EST)
3. Practice Problem Set 6