Survival Analysis

Session 1: Probabilities of Survival

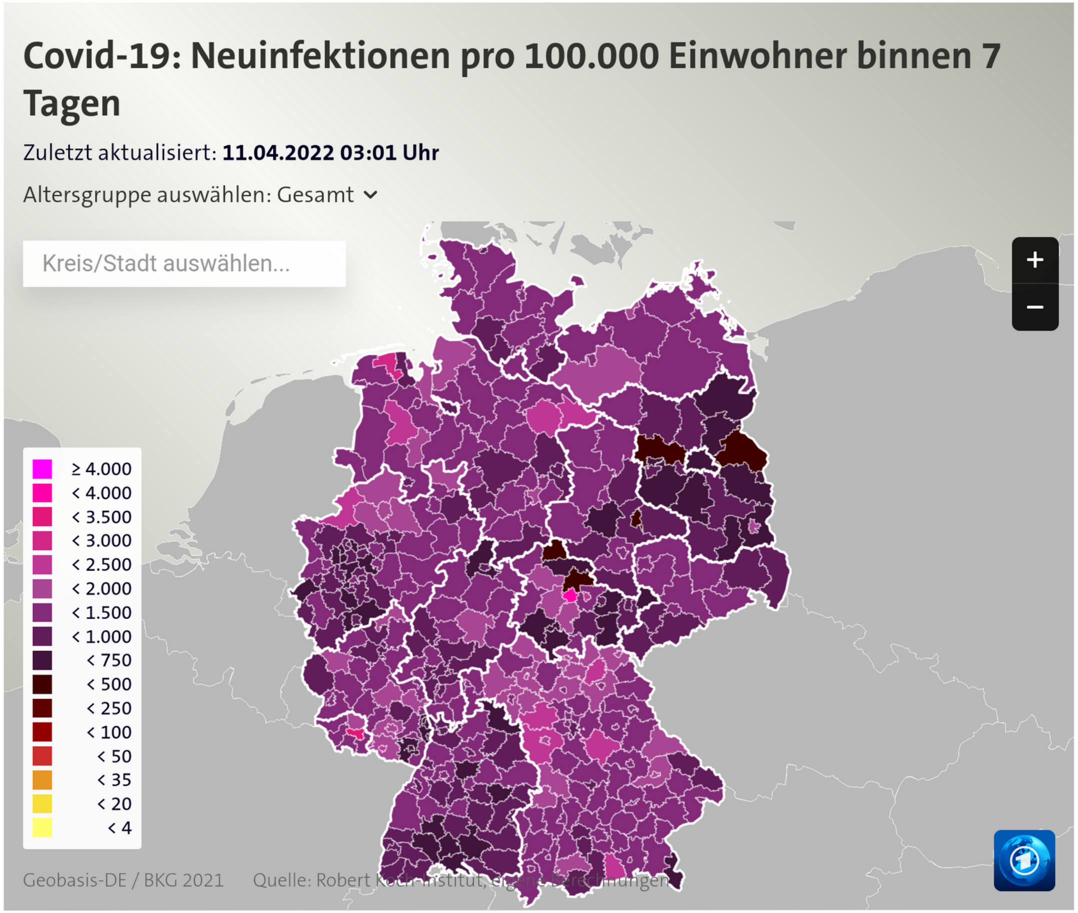
Jonas Schöley

@jschoeley

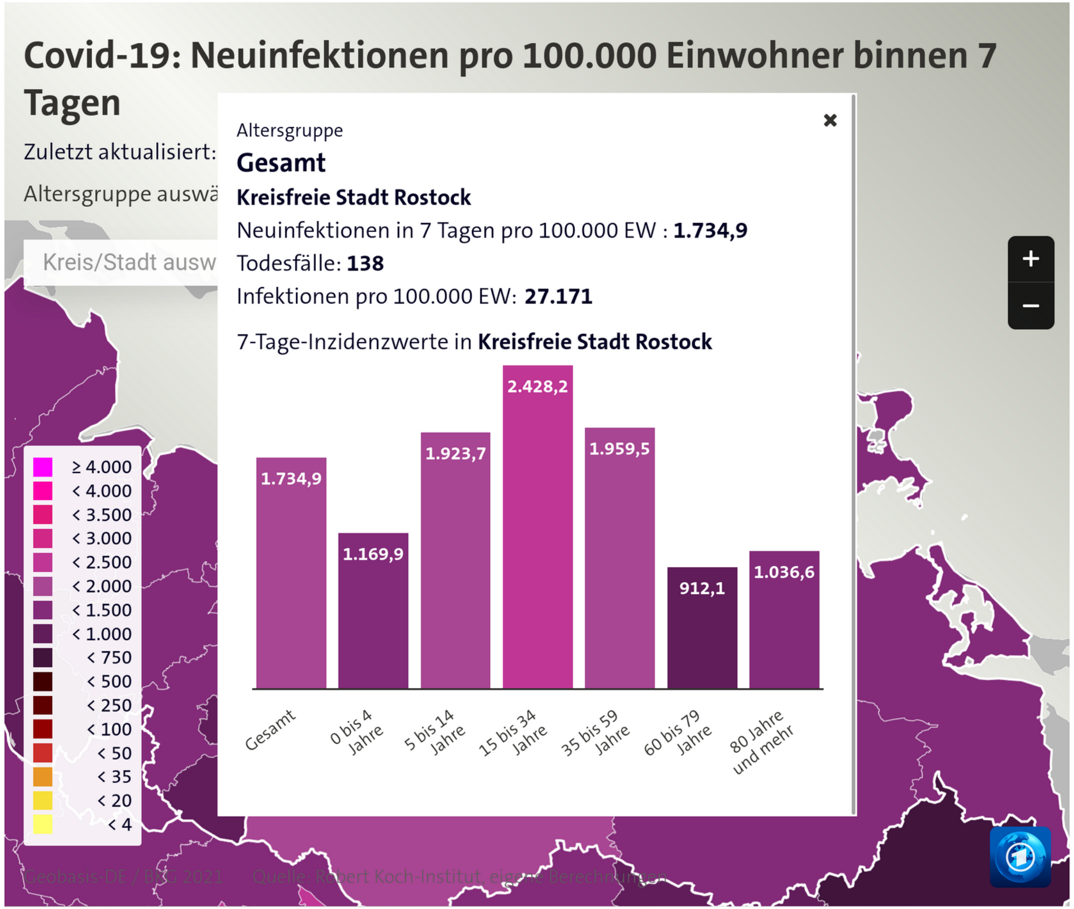
0000-0002-3340-8518

j.schoeley@uni-rostock.de

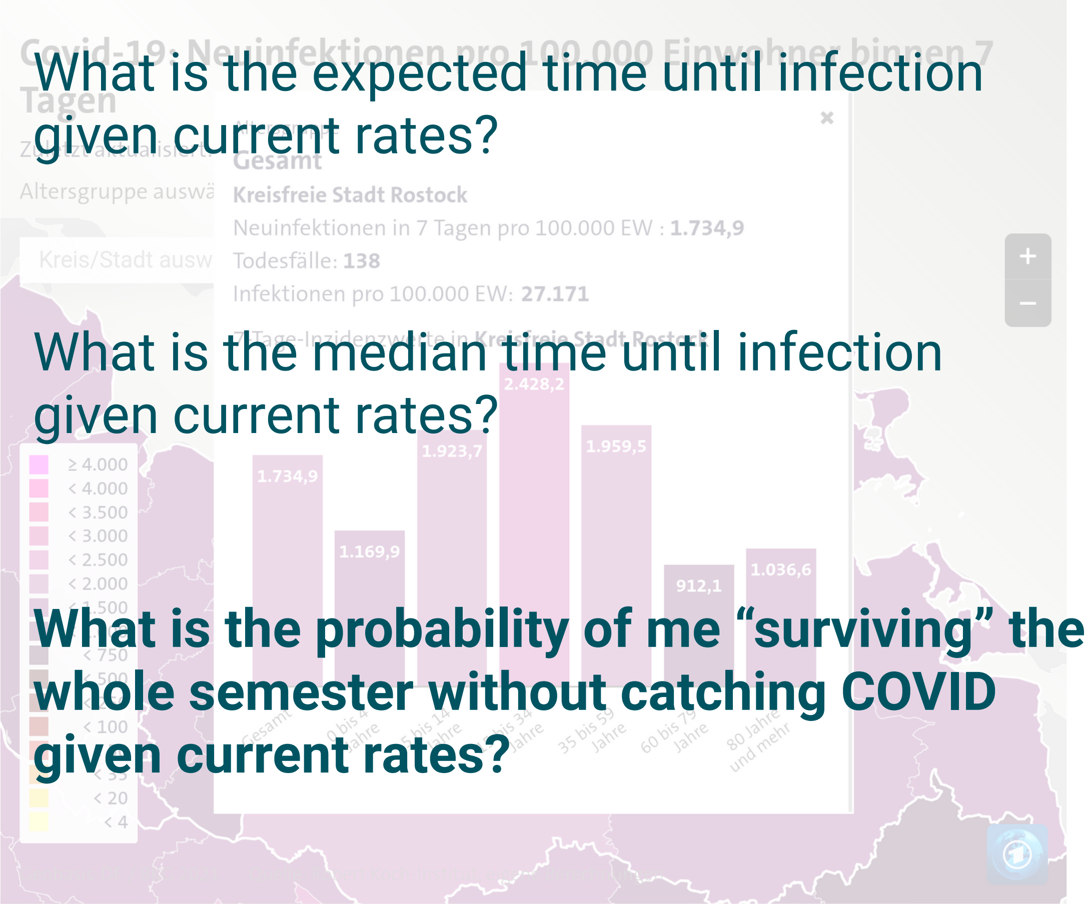
How long until infection?



How long until infection?



How long until infection?

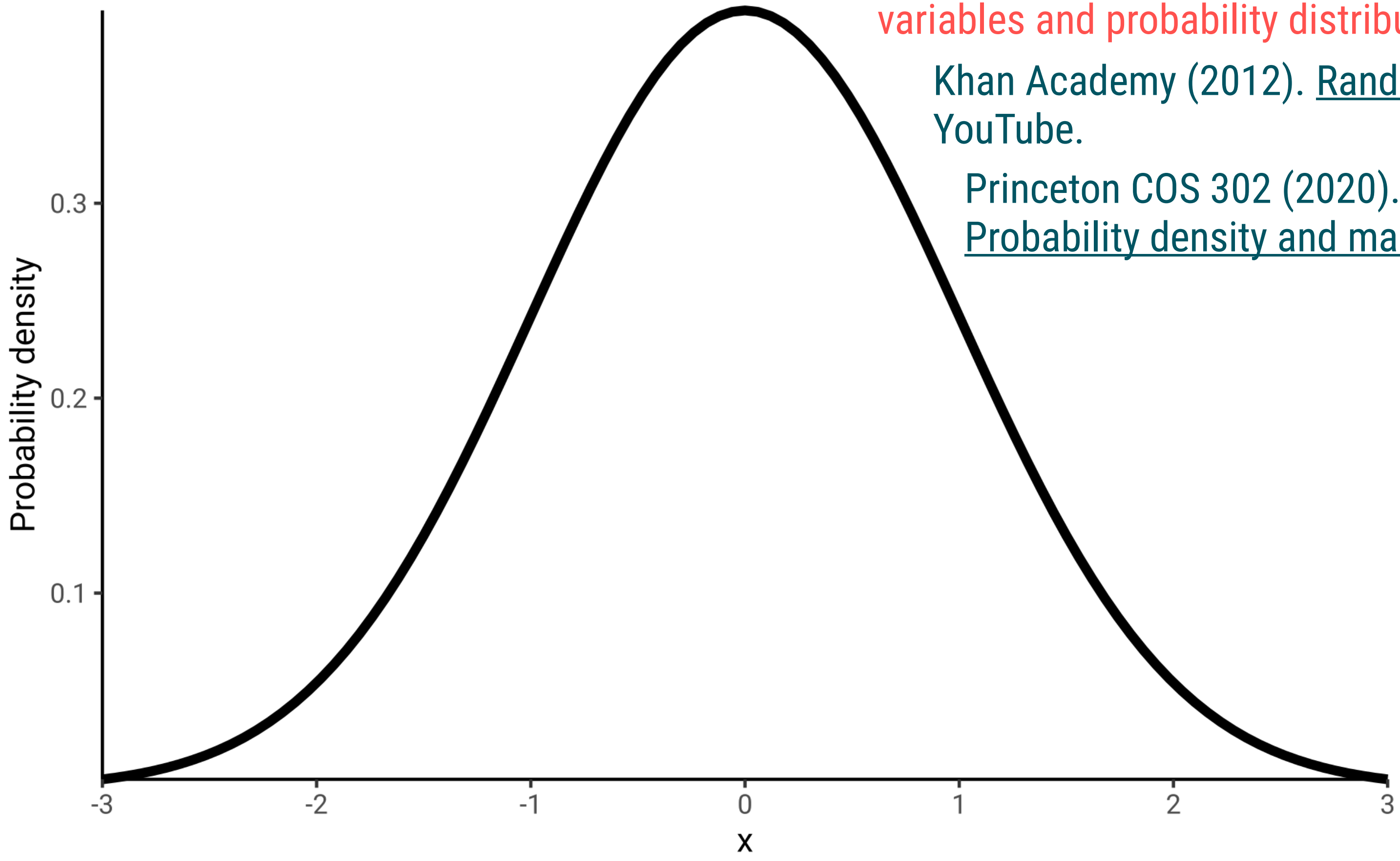


What is the expected time until infection given current rates?

What is the median time until infection given current rates?

What is the probability of me “surviving” the whole semester without catching COVID given current rates?

Recap: Random Variables & Probability Distributions

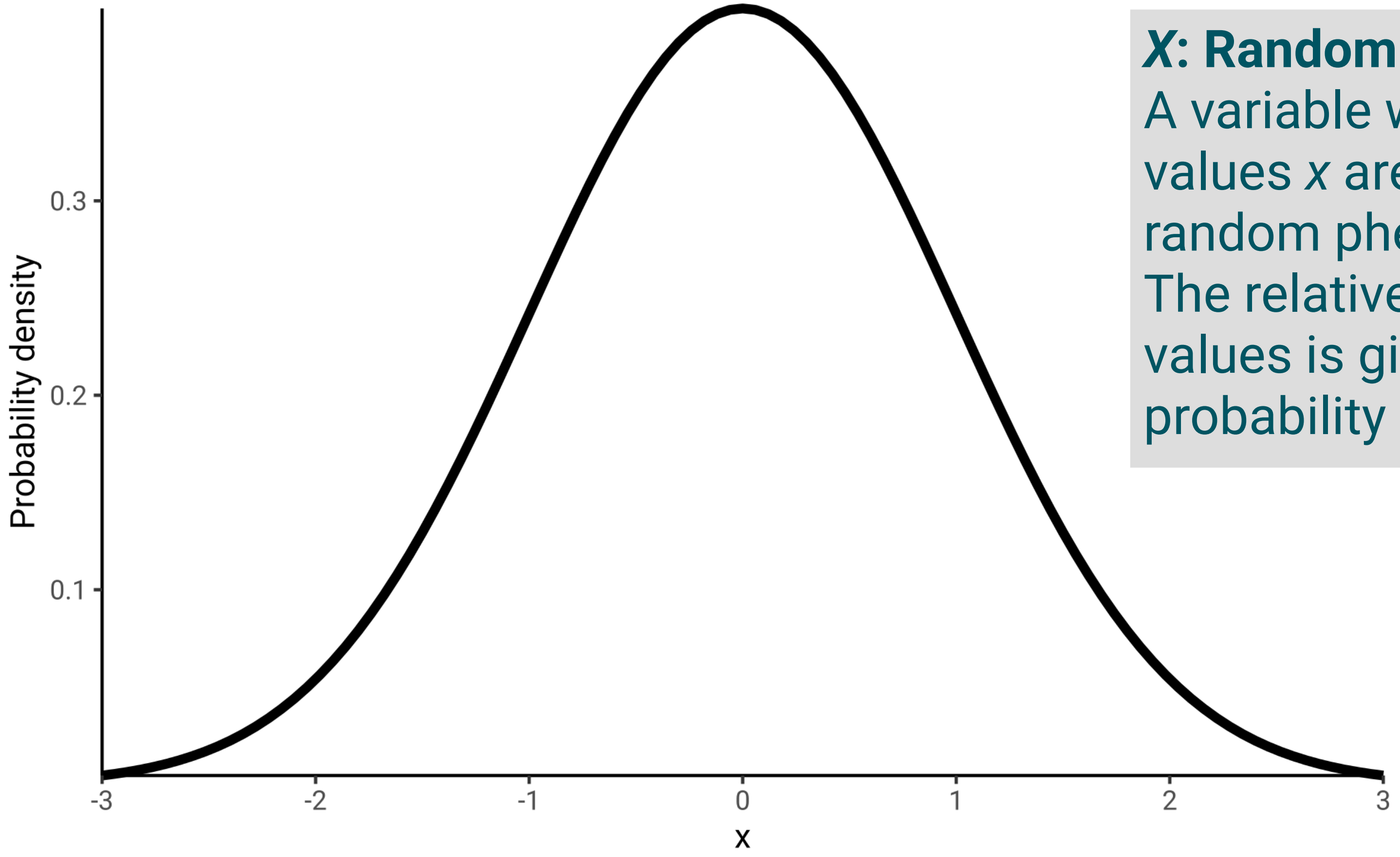


For refreshing your understanding of random variables and probability distributions watch

Khan Academy (2012). Random variables. YouTube.

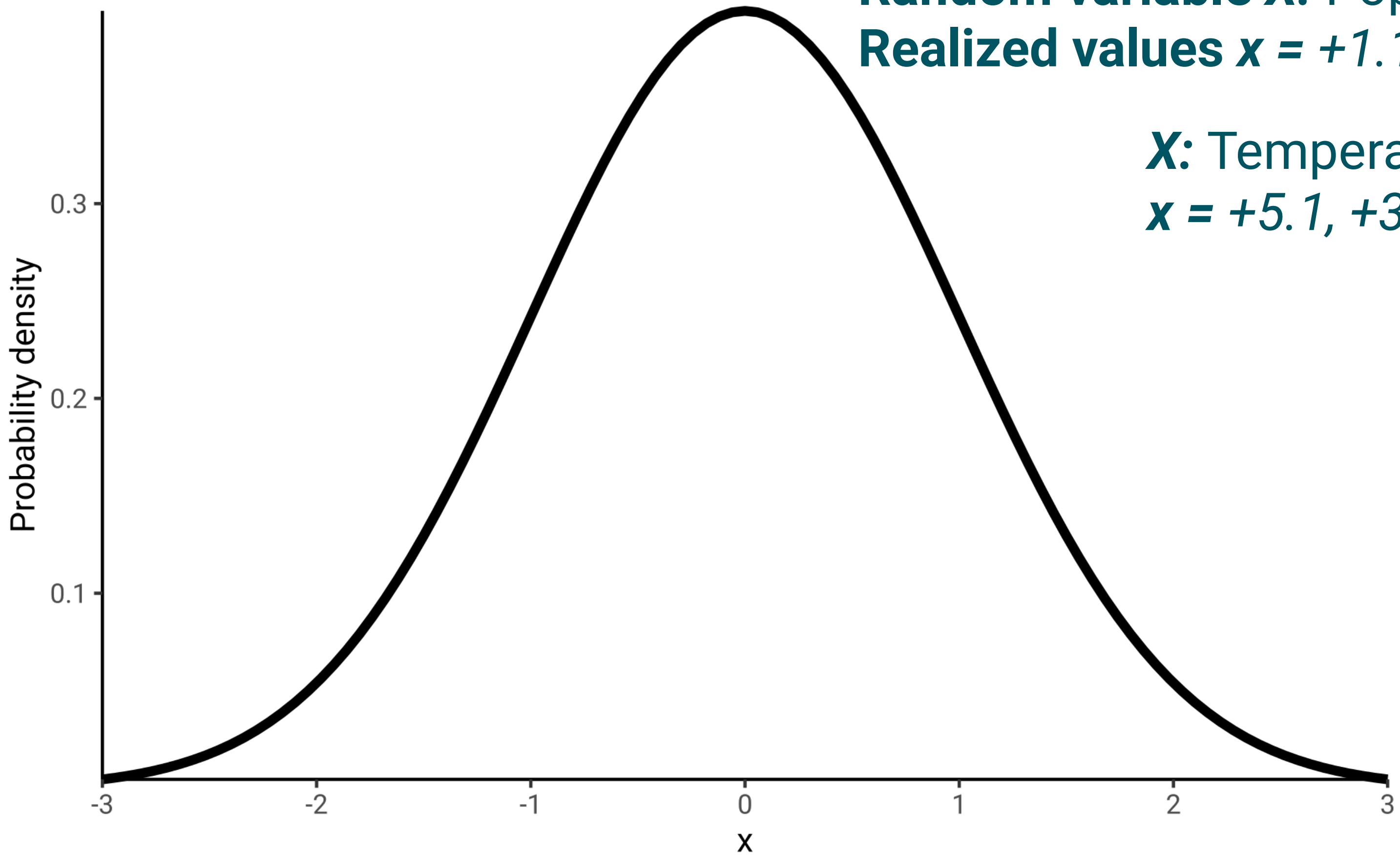
Princeton COS 302 (2020). Probability density and mass functions. YouTube.

Recap: Random Variables & Probability Distributions



***X*: Random Variable**
A variable whose possible values x are outcomes of a random phenomenon.
The relative likelihood of values is given by the probability density $f_X(x)$.

Recap: Random Variables & Probability Distributions

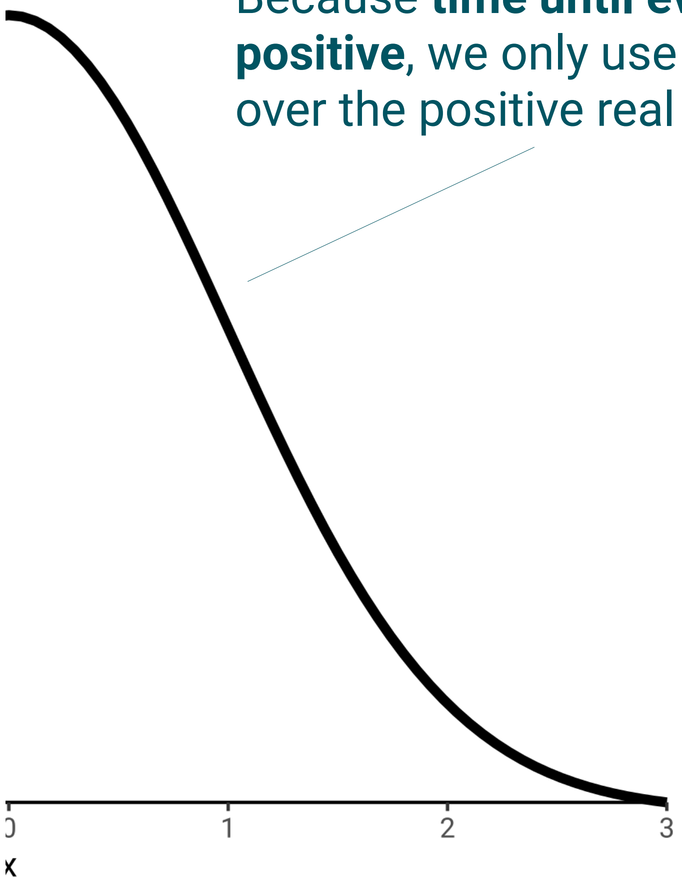


Random variable *X*: Population growth
Realized values *x* = *+1.1%, -0.2%, +0.2%...*

***X*:** Temperature anomaly
***x* =** *+5.1, +3.7, -1.2, -0.2...*

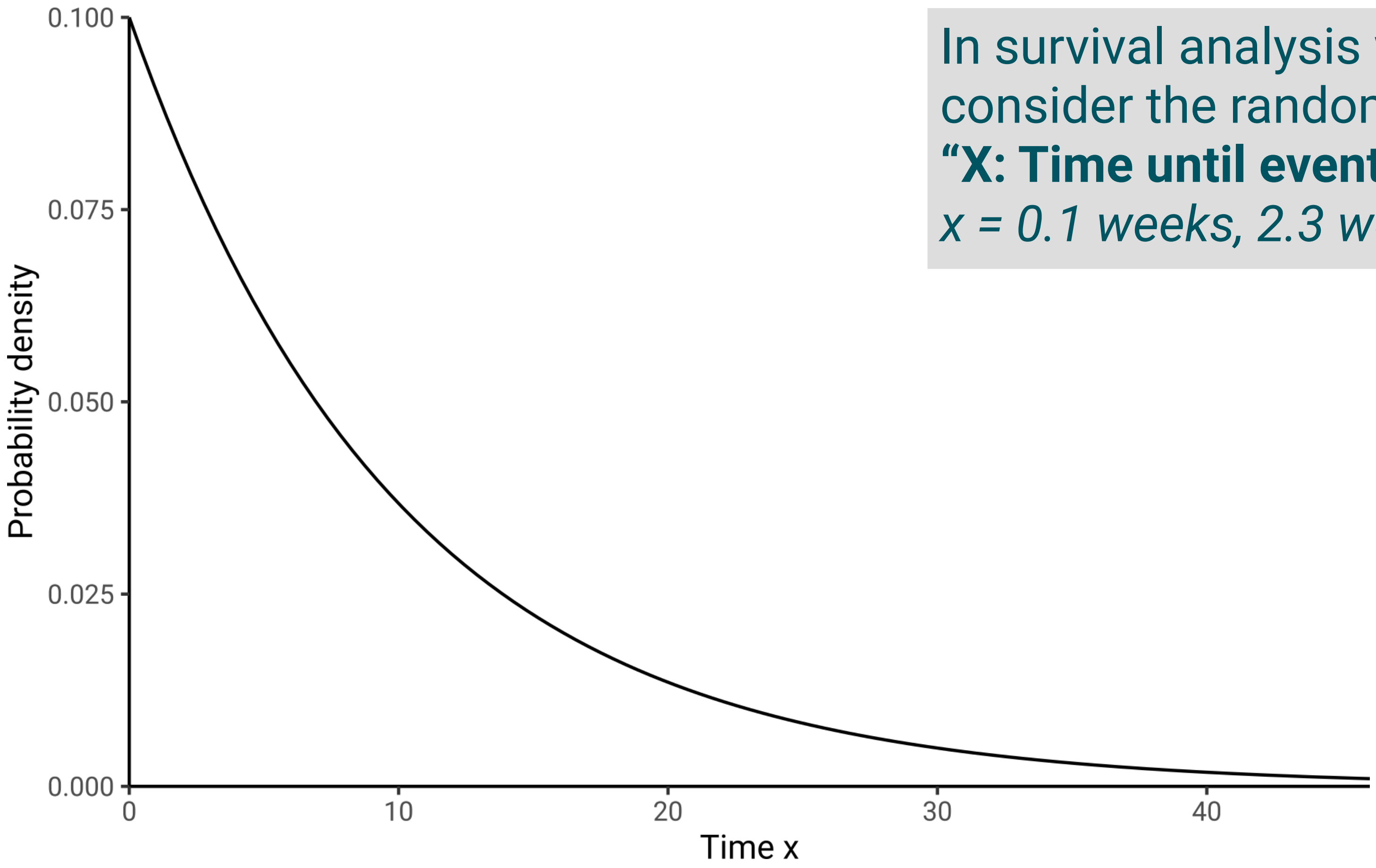
Survival Distributions

In survival analysis we consider the random variable
“X: Time until event”
x = 0.1 weeks, 2.3 weeks…



Because **time until event is always positive**, we only use distributions over the positive real numbers.

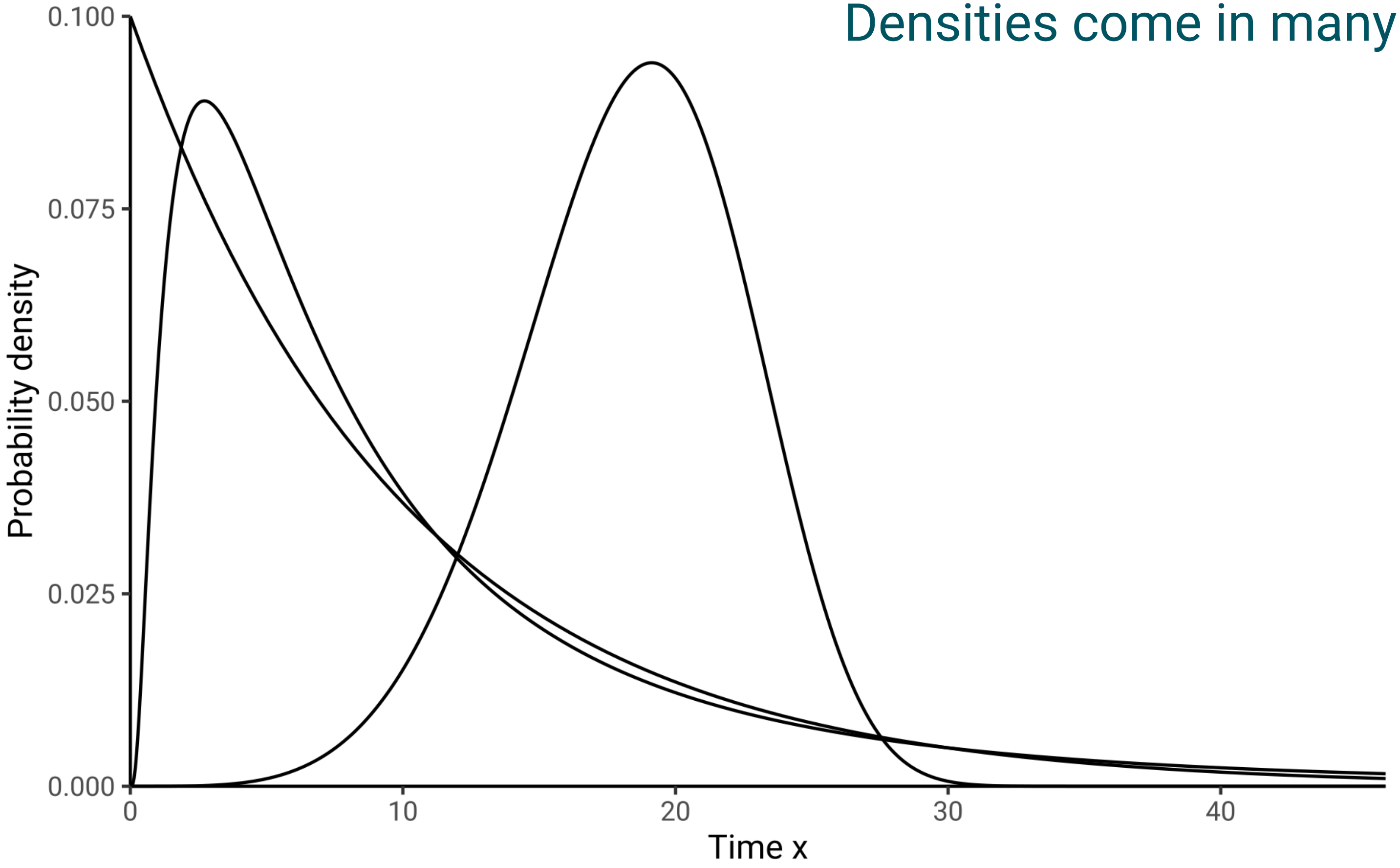
f(x) Density Function



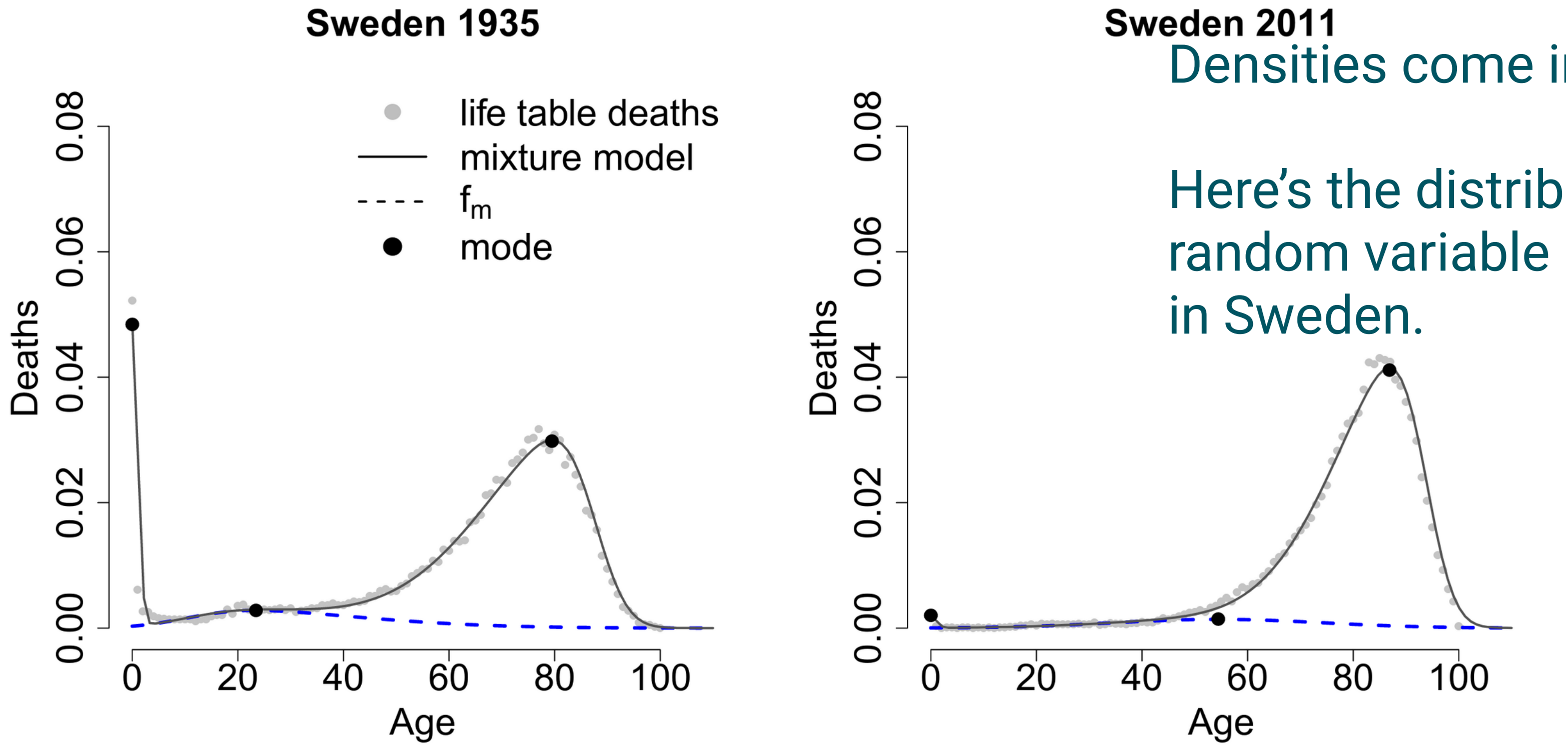
In survival analysis we consider the random variable **"X: Time until event"** *x = 0.1 weeks, 2.3 weeks…*

f(x) Density Function

Densities come in many shapes



f(x) Density Function



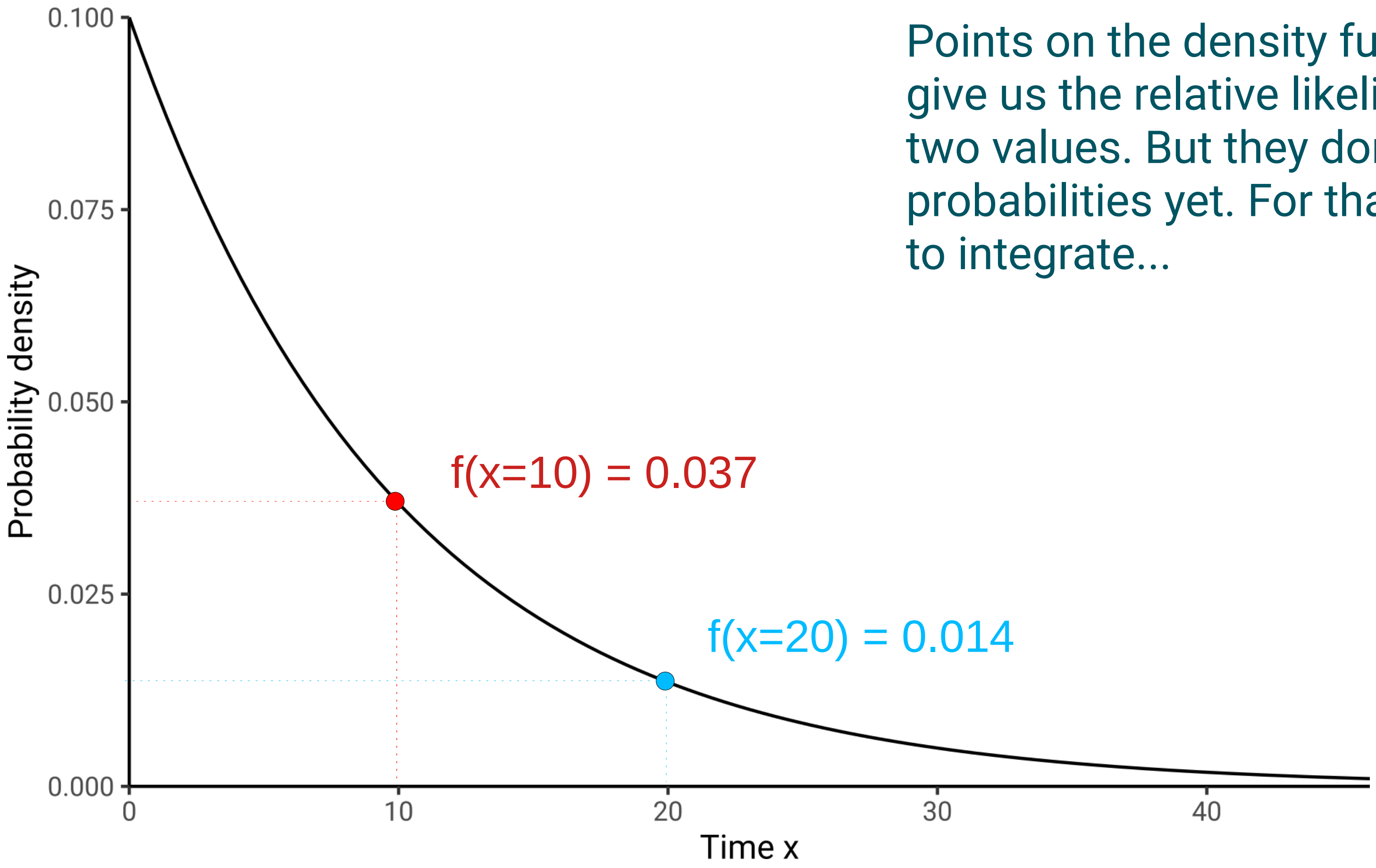
Densities come in many shapes

Here's the distribution of the random variable "years until death" in Sweden.

Fig. 2 Model fit on life table deaths for Sweden in 1935 and 2011. The solid line shows the overall mixture model. The dotted line highlights the fit of the Skew Normal employed to estimate accidental and premature mortality. The big dots point out the three modal ages of the distribution

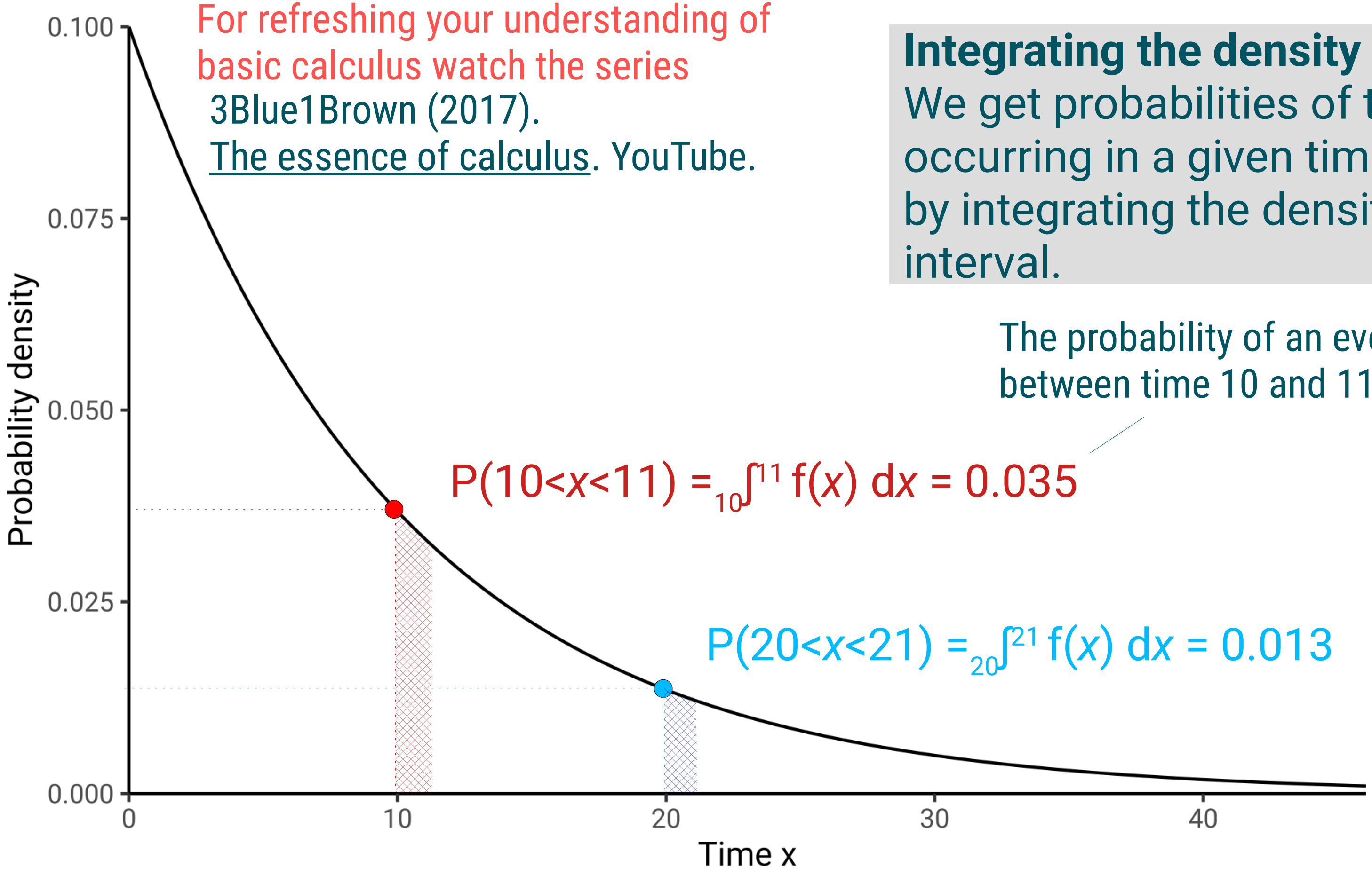
Zanotto etal. (2021). A Mixture-Function Mortality Model.

f(x) Density Function



Points on the density function give us the relative likelihood of two values. But they don't give us probabilities yet. For that we need to integrate...

f(x) Density Function

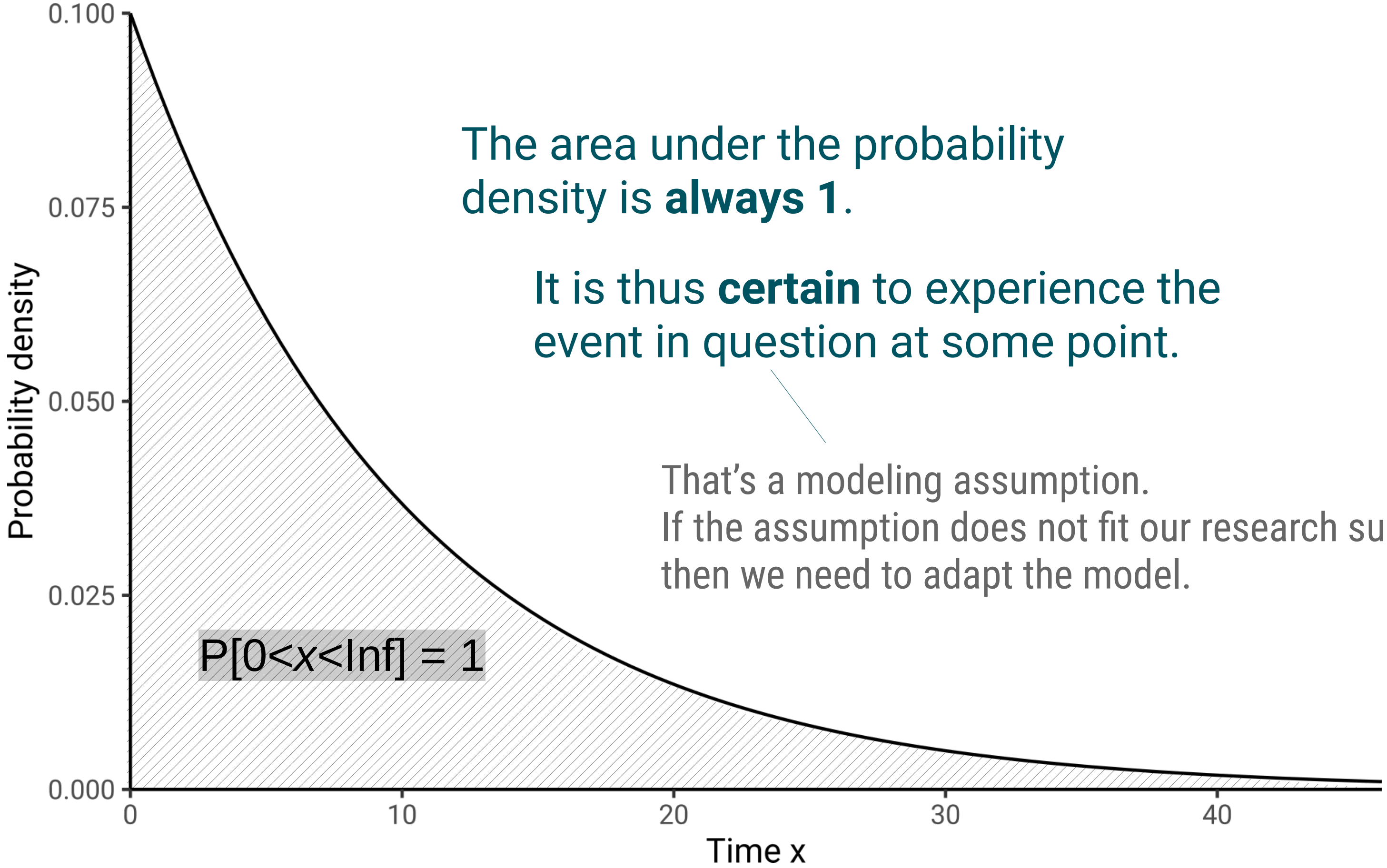


For refreshing your understanding of basic calculus watch the series 3Blue1Brown (2017). The essence of calculus. YouTube.

Integrating the density function
We get probabilities of the event occurring in a given time interval by integrating the density over said interval.

The probability of an event occurring between time 10 and 11 is 0.013.

f(x) Density Function



The area under the probability density is **always 1**.

It is thus **certain** to experience the event in question at some point.

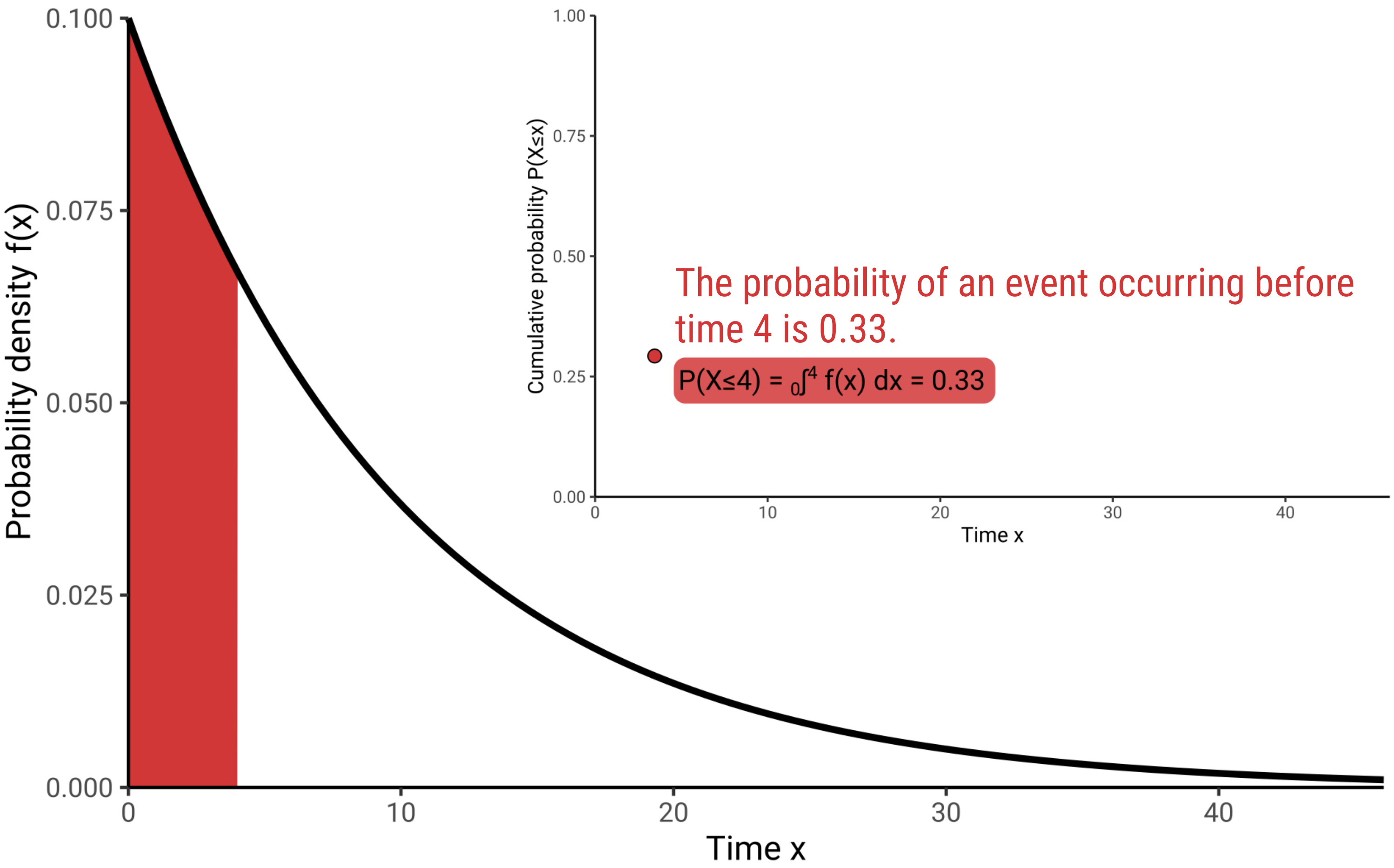
That's a modeling assumption.
If the assumption does not fit our research subject, then we need to adapt the model.

F(x) Distribution Function

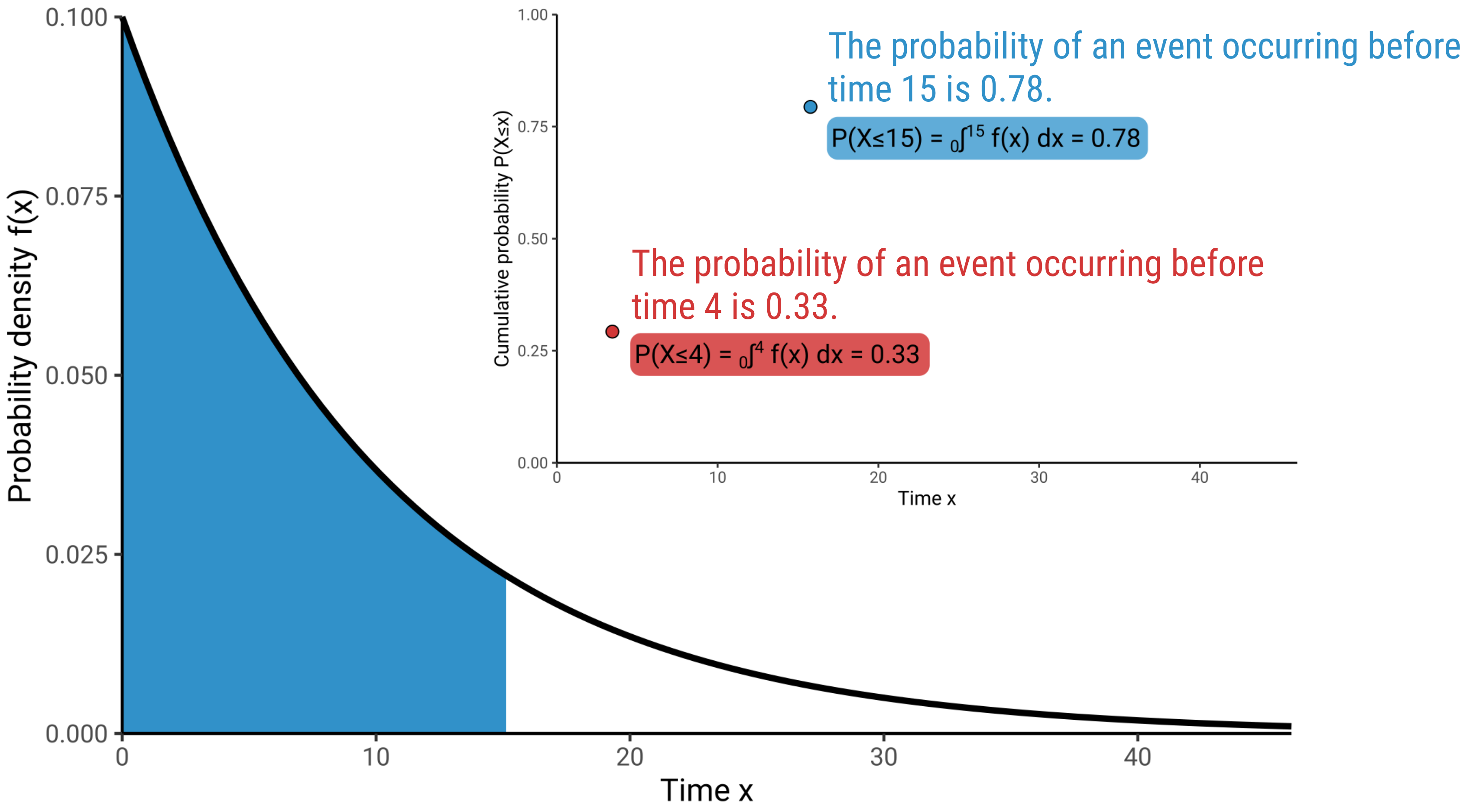
But what about the probability of an event occurring **before** time x?

Distribution Function!

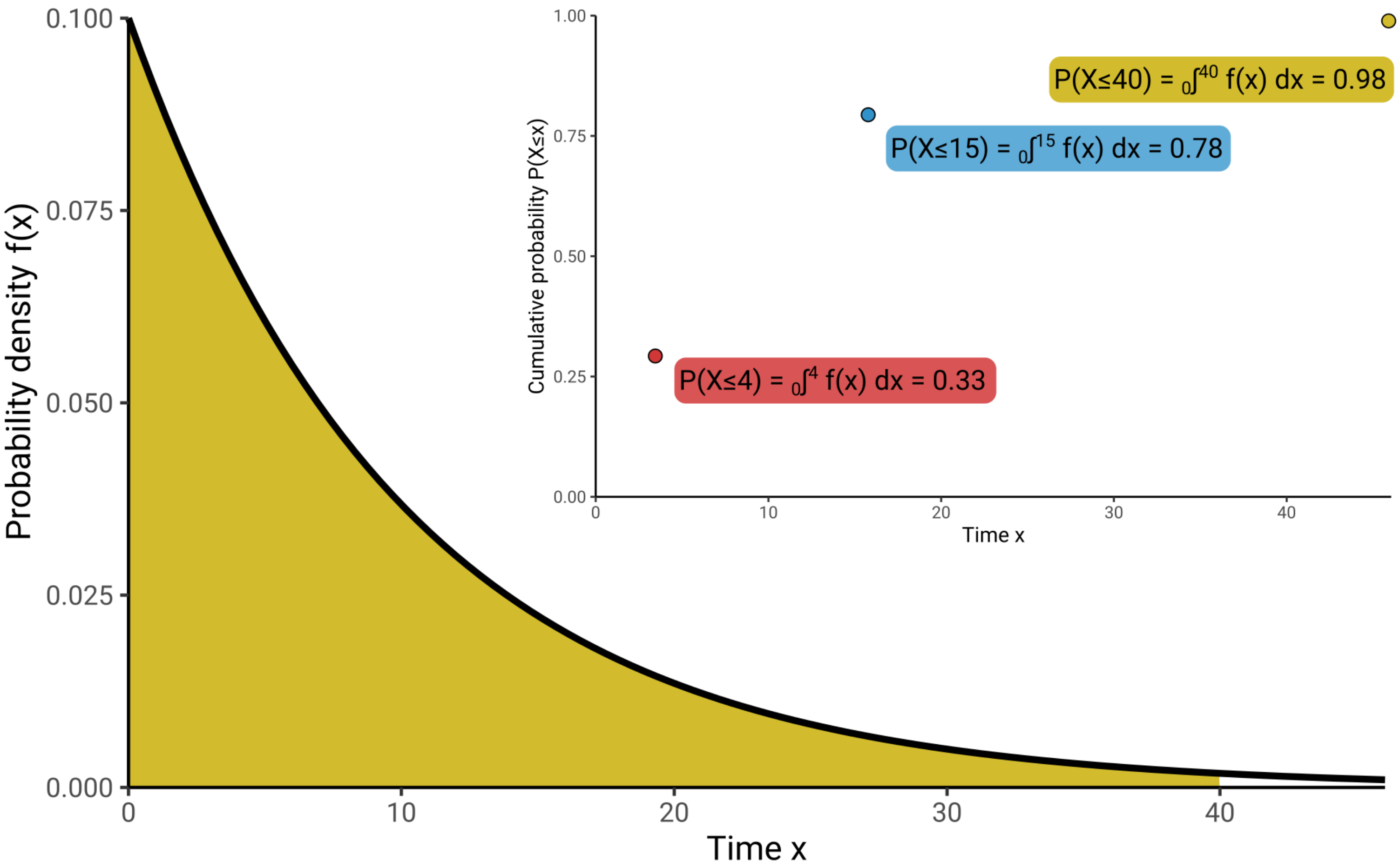
F(x) Distribution Function



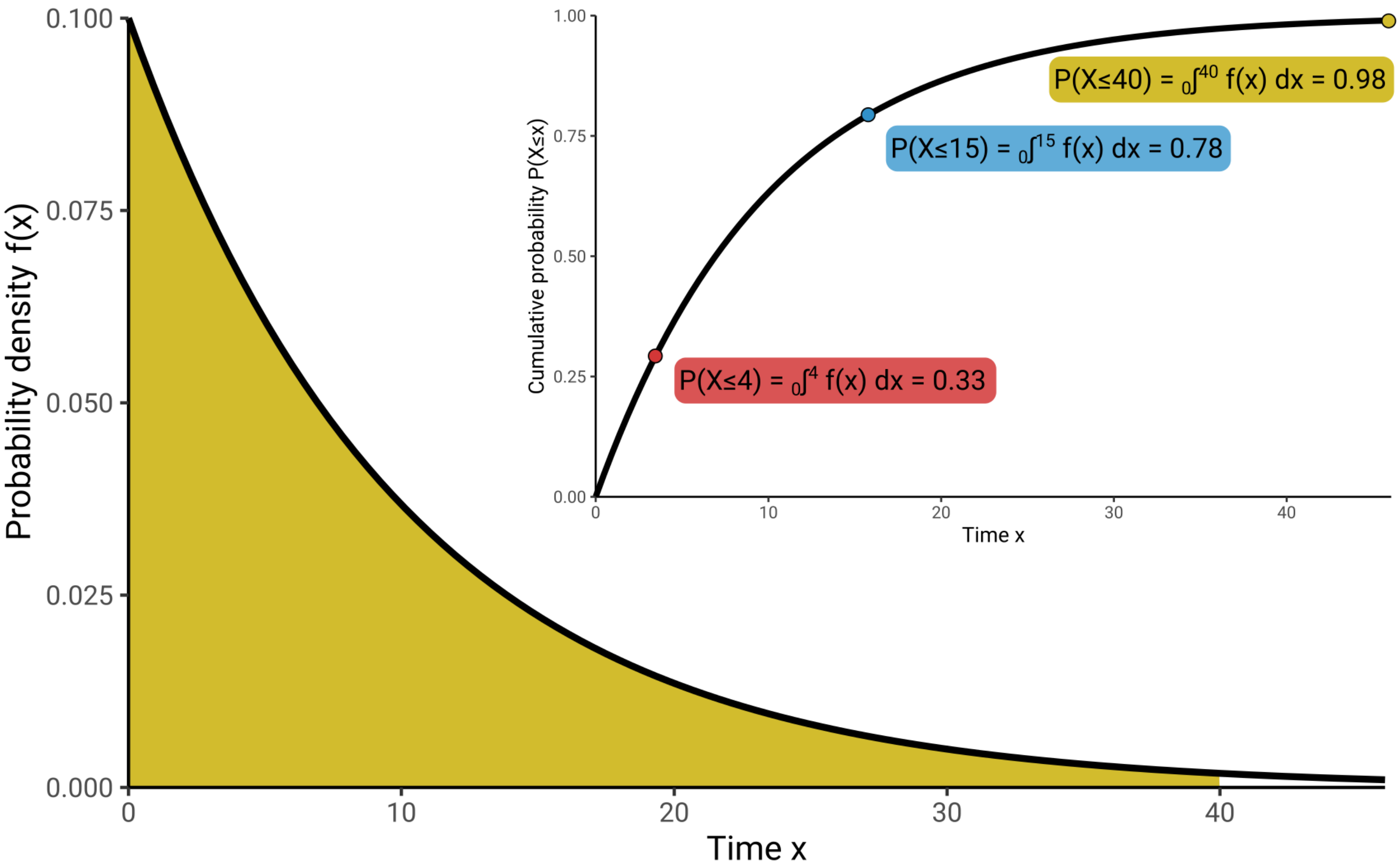
F(x) Distribution Function



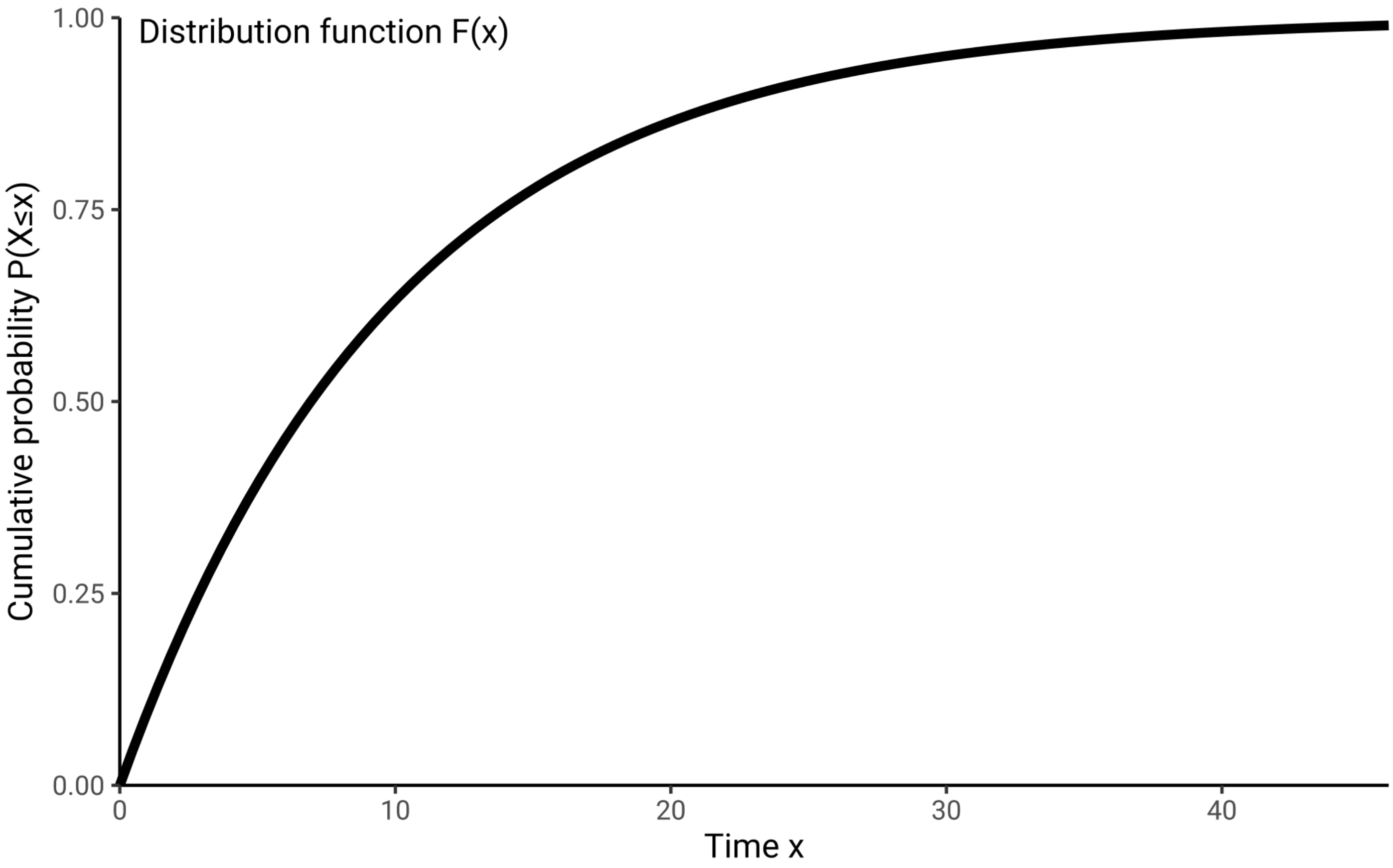
F(x) Distribution Function



F(x) Distribution Function



F(x) Distribution Function



***X*: Time until event**

x = 0.1 weeks, 2.3 weeks...

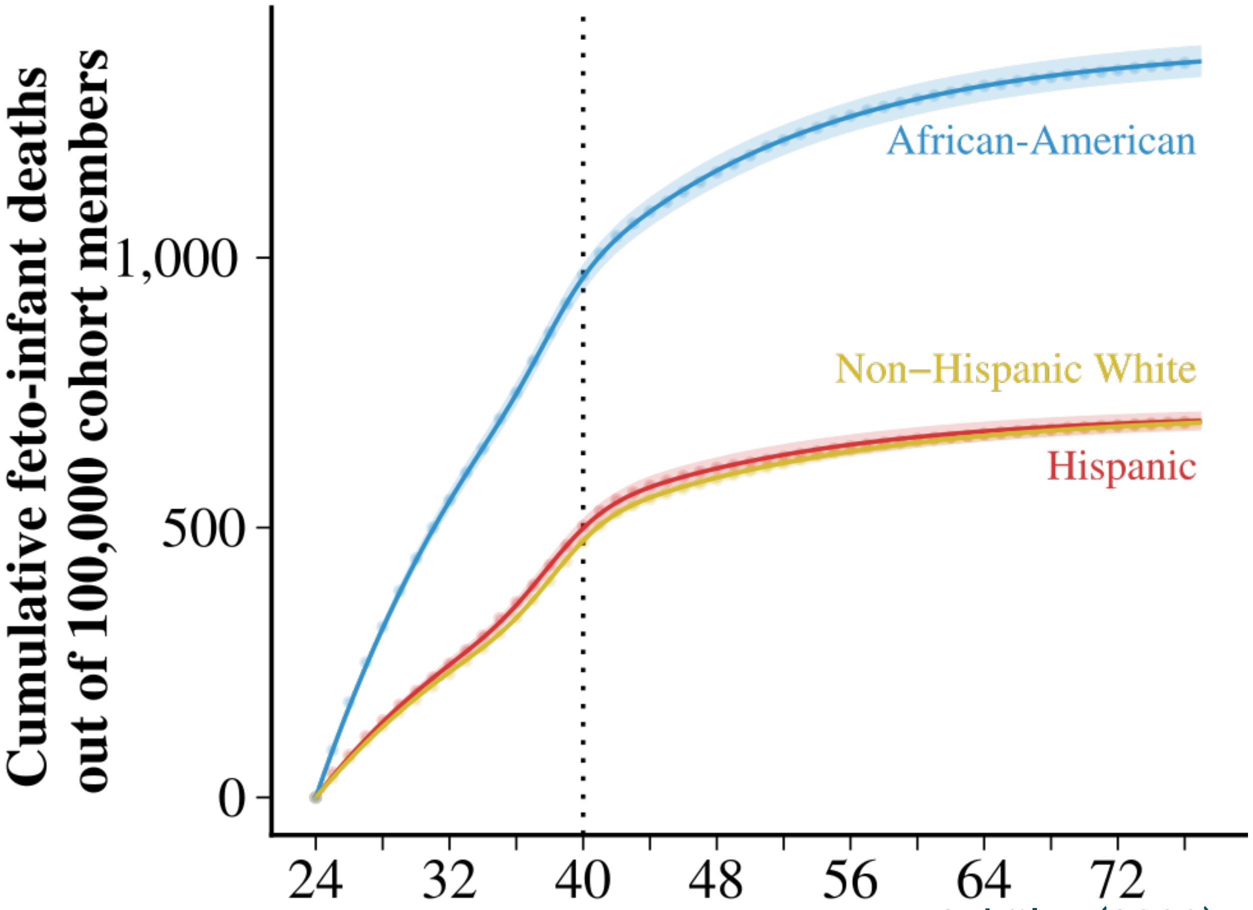
F(*x*): Distribution function

aka Cumulative function

The probability of experiencing the event until time *x*.

$$\mathbf{F}(x) = \mathbf{P}(X \leq x) = \int_0^x \mathbf{f}(x)\,\mathbf{d}x$$

F(x) Distribution Function



Schöley (2020). The dynamics of ontogenescence.

***X*: Time until event**
x = 0.1 weeks, 2.3 weeks...

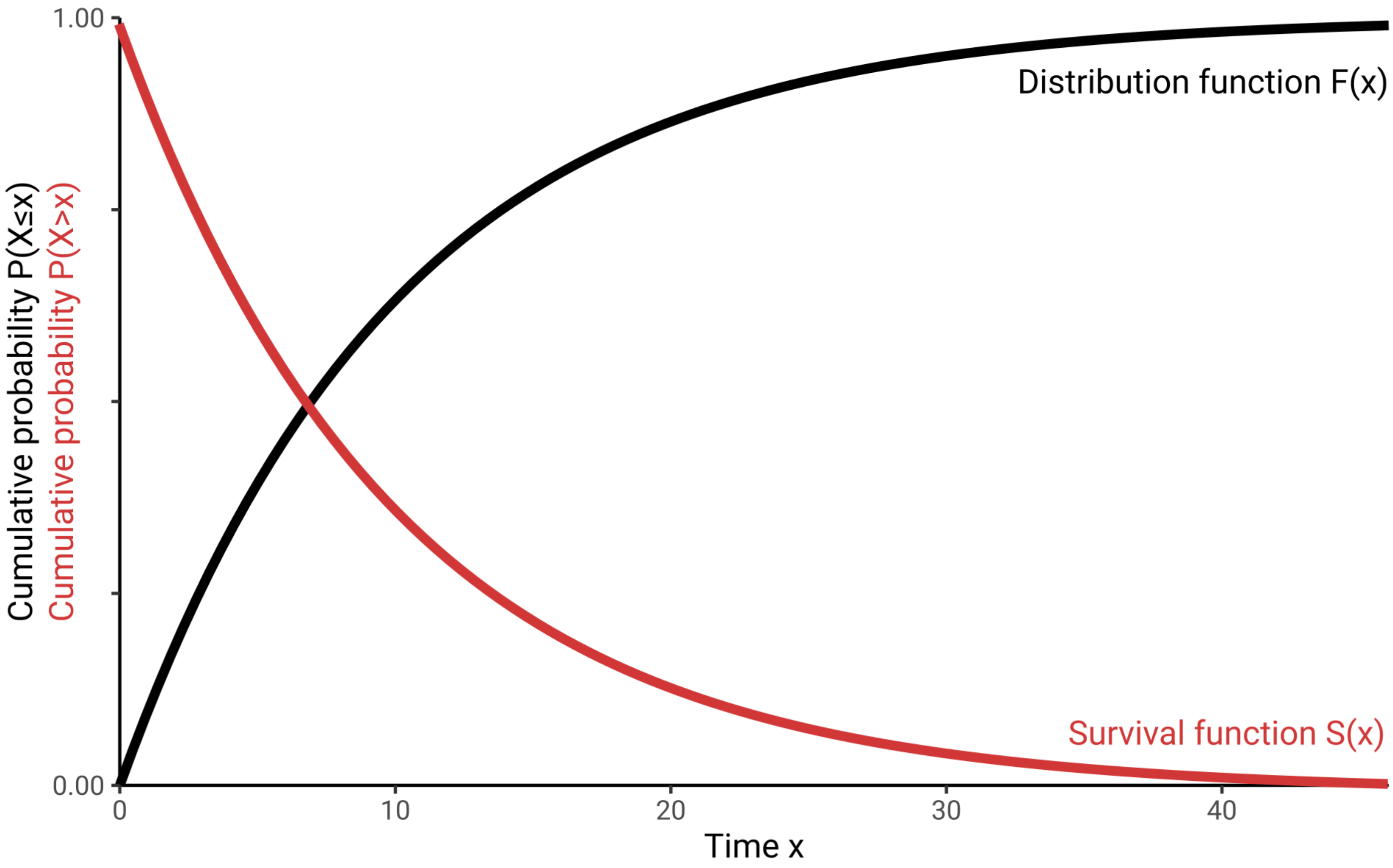
F(*x*): Distribution function
aka Cumulative function
The probability of experiencing the event until time *x*.
$F(x) = P(X \leq x) = \int_0^x f(x)\,dx$

S(x) Survival Function

But what about the probability of an event occurring **after** time x?

Survival Function!

S(x) Survival Function

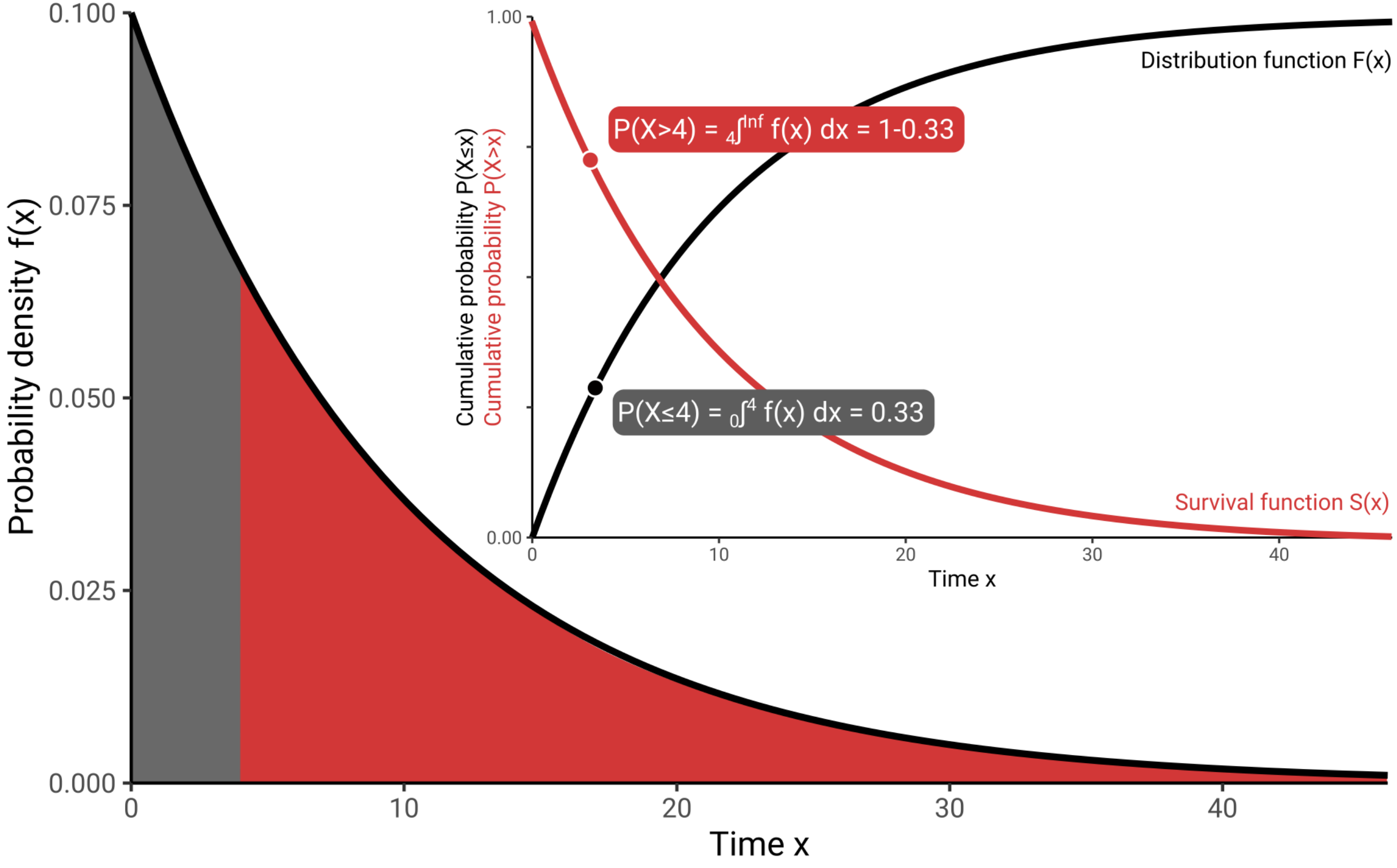


***X*: Time until event**
x = 0.1 weeks, 2.3 weeks...

S(*x*): Survival function
The probability of *not* experiencing the event until time *x*.

$$S(x) = P(X>x) = \int_x^{\text{Inf}} f(x)\,dx$$
$$= 1-F(x) = 1-\int_0^x f(x)\,dx$$

S(x) Survival Function



***X*: Time until event**

x = 0.1 weeks, 2.3 weeks...

S(*x*): Survival function

The probability of *not* experiencing the event until time *x*.

S(*x*) = P(*X*>*x*) = $\int_x^{Inf}$ **f(*x*) dx**

= 1-F(x) = 1-$\int_0^x$ f(*x*) dx

S(x) Survival Function

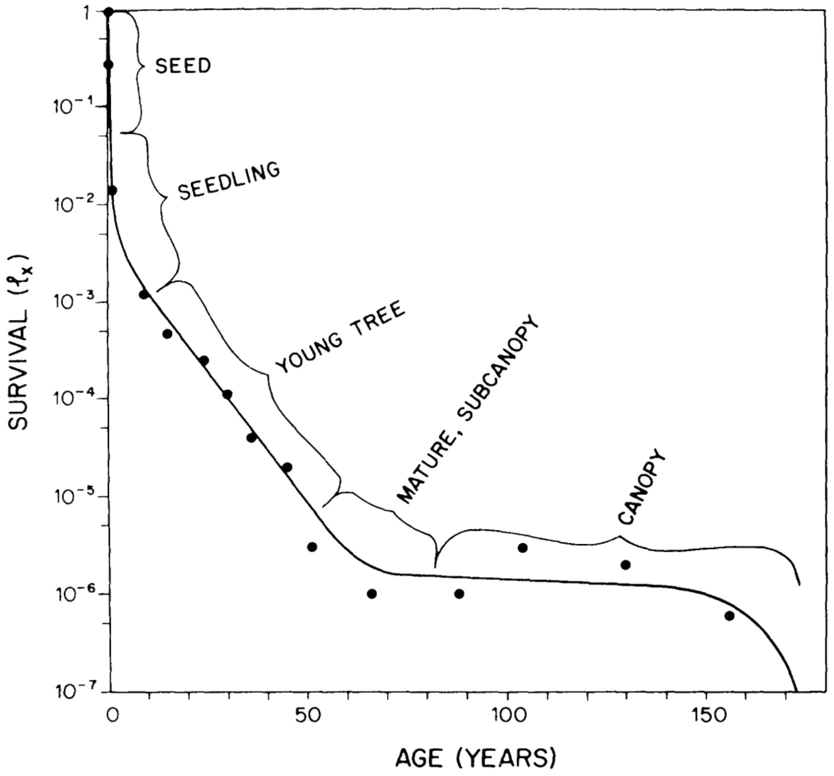


FIGURE 1. Survivorship curve for *Euterpe globosa*, over 7 orders of magnitude. All values of l_x except that at 9 years are completely independent of each other, thus providing an internal check on accuracy.

Valen (1975). Life, Death, and Energy of a Tree.

***X*: Time until event**
x = 0.1 weeks, 2.3 weeks...

S(*x*): Survival function
The probability of *not* experiencing the event until time *x*.

$$\mathbf{S(x) = P(X>x) = \int_x^{Inf} f(x)\,dx}$$

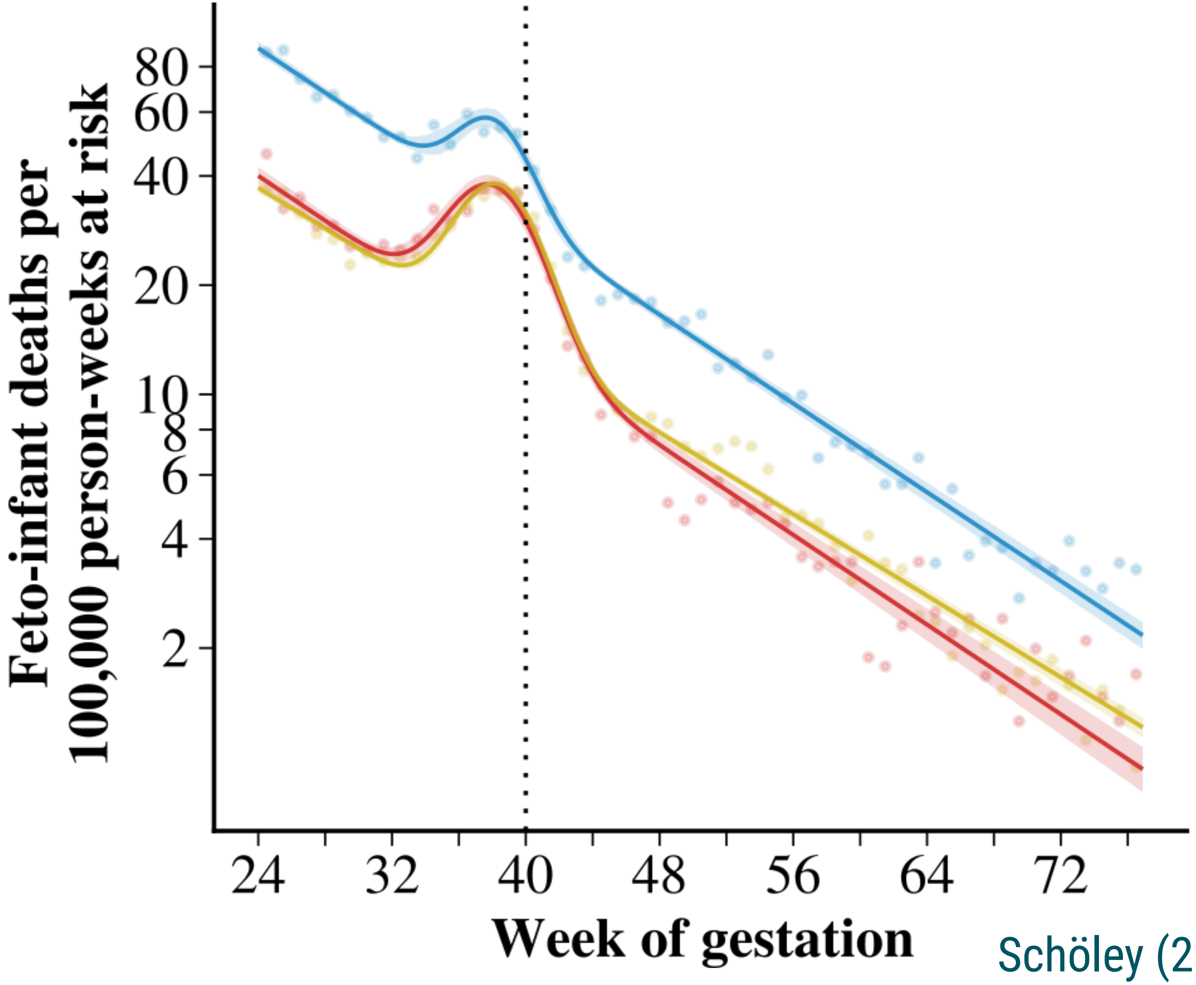
$$= 1-F(x) = 1-\int_0^x f(x)\,dx$$

h(x) Hazard Function

But what about the risk of an event occurring around time x, given that it did not occur before?

Hazard Function!

h(x) Hazard Function



***X*: Time until event**
x = 0.1 weeks, 2.3 weeks...

h(*x*): Hazard function
The instantaneous rate of new events at time *x* among those who did not experience the event yet.

h(*x*) = f(*x*)/S(*x*)
$= \lim_{h \to 0} P(x \leq X < x+h \mid X \geq x)/h$

Schöley (2020). The dynamics of ontogenescence.

h(x) Hazard Function

We have the *rate* of events
2428 infections per 100,000 persons per 7 days
2428 infections per 100,000 persons-weeks
0.02428 infections per 1 person-week

We want the *survival* probability
What is the probability of me "surviving" the whole semester without catching COVID given current rates?

Thus, **we need** to express S(x) in terms of h(x)

***X*: Time until event**
x = 0.1 weeks, 2.3 weeks...

h(*x*): Hazard function
The instantaneous rate of new events at time *x* among those who did not experience the event yet.

h(*x*) = f(*x*)/S(*x*)
$= \lim_{h \to 0} P(x \leq X < x+h | X \geq x)/h$

h(x) Hazard Function

Thus, **we need** to express S(x) in terms of h(x)

$$\begin{aligned} h(x) &= f(x)/S(x) \\ &= F'(x)/S(x) \\ &= [1-S(x)]'/S(x) \\ &= -S'(x)/S(x) \end{aligned}$$

Remember, $F(x) = {}_0\int^x f(x)\,dx$.
Thus, $f(x) = d/dx\, F(x) = F'(x)$.

***X*: Time until event**
x = 0.1 weeks, 2.3 weeks...

h(*x*): Hazard function
The instantaneous rate of new events at time *x* among those who did not experience the event yet.

$h(x) = f(x)/S(x)$
$= \lim_{h\to 0} P(x \leq X < x+h \mid X \geq x)/h$

H(x) Cumulative Hazard Function

Thus, **we need** to express S(x) in terms of h(x)

$$h(x) = f(x)/S(x)$$
$$= F'(x)/S(x)$$
$$= [1-S(x)]'/S(x)$$
$$= -S'(x)/S(x)$$

Remember, $F(x)={}_0\!\int^x f(x)\,dx$.
Thus, $f(x)=d/dx\,F(x) = F'(x)$.

$${}_0\!\int^x h(x) = {}_0\!\int^x -S'(x)/S(x)\,dx$$

Rule for logarithmic derivatives: $\int g'(x)/g(x)\,dx = \log|g(x)| + C$

$$= -\log S(x)$$
$$-{}_0\!\int^x h(x) = \log S(x)$$
$$\mathbf{\exp[-{}_0\!\int^x h(x)] = S(x)}$$

***X*: Time until event**
x = 0.1 weeks, 2.3 weeks...

H(*x*): Cumulative Hazard
The integral of h(x).

$$\mathbf{H(x) = {}_0\!\int^x h(x)\,dx = -\log S(x)}$$

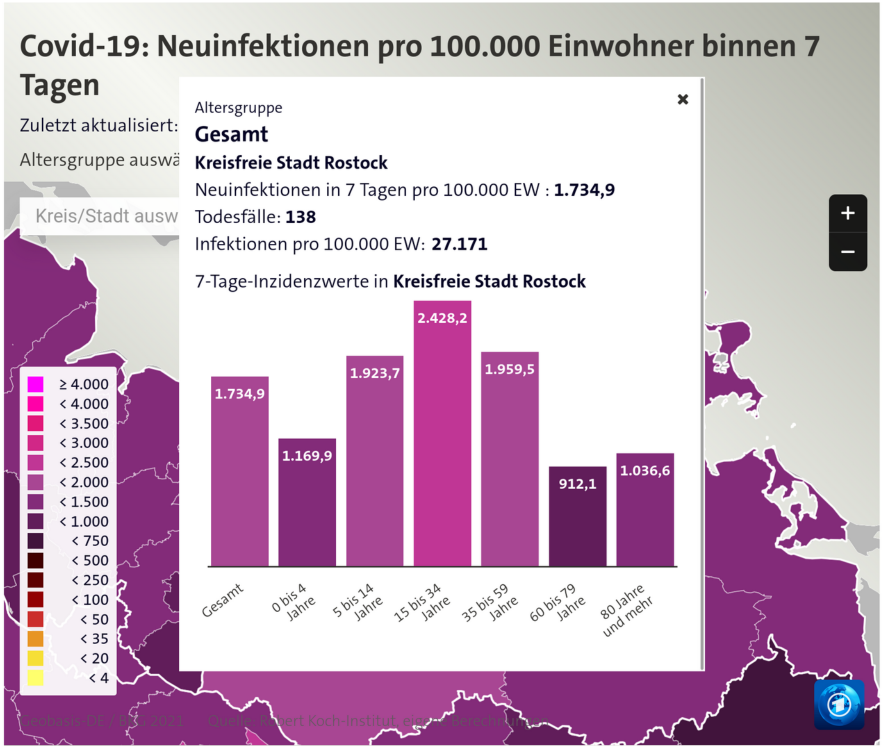
The Exponential Distribution

We have the *rate* of events
2428 infections per 100,000 persons per 7 days
2428 infections per 100,000 persons-weeks
0.02428 infections per 1 person-week

We want the *survival* probability
What is the probability of me "surviving" the whole semester without catching COVID given current rates?

Thus, **we need** to express S(x) in terms of h(x)

$\exp[-{}_0\int^x h(x)dx] = S(x)$



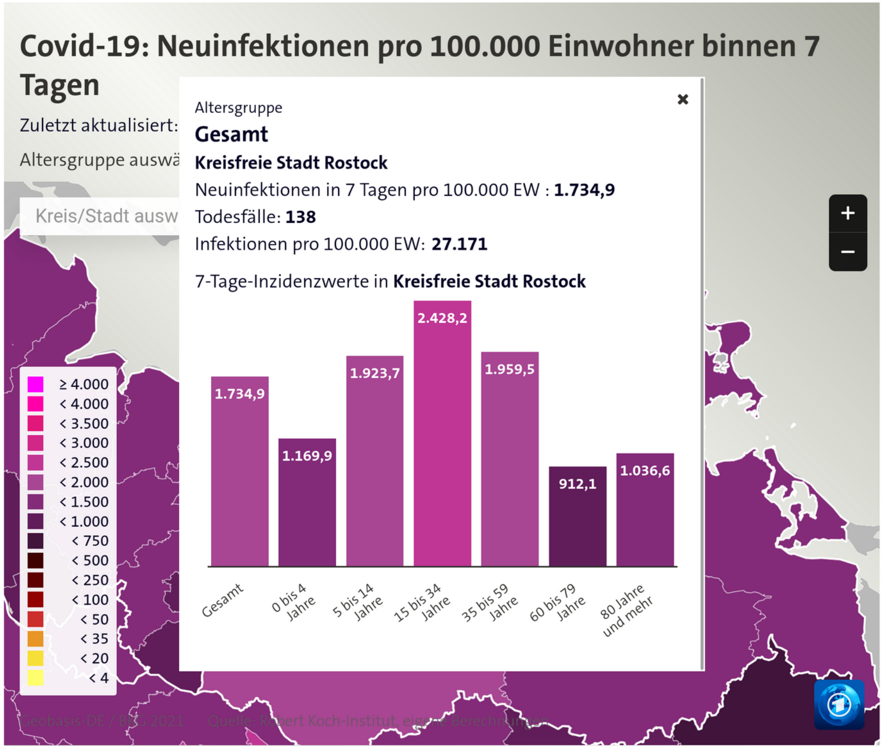
The Exponential Distribution

We have the *rate* of events
2428 infections per 100,000 persons per 7 days
2428 infections per 100,000 persons-weeks
0.02428 infections per 1 person-week

We want the *survival* probability
What is the probability of me “surviving” the whole semester without catching COVID given current rates?

Thus, **we need** to express S(x) in terms of h(x)

$$\exp[-{}_0\int^x h(x)dx] = S(x)$$



But what **shape** should h(*x*) have?

The Exponential Distribution

But what **shape** should h(*x*) have?

We have the *rate* of events

2428 infections per 100,000 persons per 7 days

2428 infections per 100,000 persons-weeks

0.02428 infections per 1 person-week

The Exponential Distribution

But what **shape** should h(x) have?

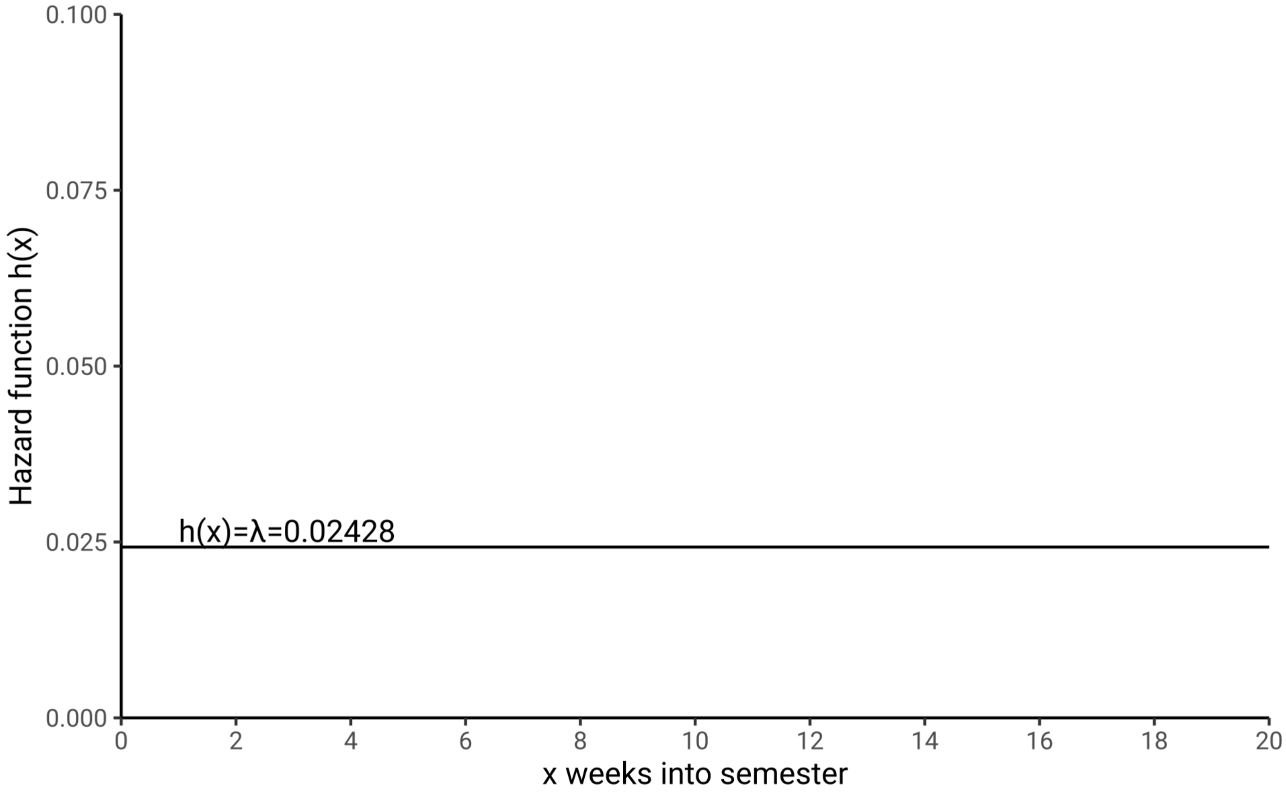
We have the *rate* of events
2428 infections per 100,000 persons per 7 days
2428 infections per 100,000 persons-weeks
0.02428 infections per 1 person-week

Assuming the rate stays constant over time, **h(x) is a constant!**
h(x) = λ = 0.02428

The Exponential Distribution

Assuming the rate stays constant over time, **h(*x*) is a constant!**

$h(x) = \lambda = 0.02428$



The Exponential Distribution

Assuming the rate stays constant over time, **$h(x)$ is a constant!**

$h(x) = \lambda = 0.02428$

Applying the survival identity from earlier...

$\exp[-{}_0\!\int^x h(x)\,dx] = S(x)$

...yields the survival function of the **Exponential Distribution**

$$\boldsymbol{S(x) = \exp(-{}_0\!\int^x \lambda dx) = \exp(-\lambda x)}$$

$= \exp(-0.02428x)$

Exponential distribution
If the hazard does not change over time, the time until event is exponentially distributed.

$\boldsymbol{h(x) = \lambda}$
$\boldsymbol{S(x) = \exp(-\lambda x)}$

The Exponential Distribution

Assuming the rate stays constant over time, **$h(x)$ is a constant!**

$h(x) = \lambda = 0.02428$

Applying the survival identity from earlier...

$\exp[-{}_0\int^x h(x)\, dx] = S(x)$

...yields the survival function of the **Exponential Distribution**

$\mathbf{S(x) = \exp(-{}_0\int^x \lambda dx) = \exp(-\lambda x)}$

$= \exp(-0.02428x)$

What is the probability of me "surviving" the whole semester without catching COVID given current rates?

Exponential distribution

If the hazard does not change over time, the time until event is exponentially distributed.

$\mathbf{h(x) = \lambda}$

$\mathbf{S(x) = \exp(-\lambda x)}$

Survival Identities

In survival analysis we consider the random variable **“*X*: Time until event”**
x = 0.1 weeks, 2.3 weeks…

Survival Identities

In survival analysis we consider the random variable **"*X*: Time until event"**
x = 0.1 weeks, 2.3 weeks…

We express our knowledge about the distribution of *X* in any of these functions. Knowing any single function we can derive all other via the **Survival Identities**.

Survival Identities

In survival analysis we consider the random variable
“*X*: Time until event”
x = 0.1 weeks, 2.3 weeks…

We express our knowledge about the distribution of *X* in any of these functions. Knowing any single function we can derive all other via the **Survival Identities**.

f(*x*): Density function
The relative likelihood of experiencing the event around time *x*.

Survival Identities

In survival analysis we consider the random variable
“*X*: Time until event”
x = 0.1 weeks, 2.3 weeks…

We express our knowledge about the distribution of X in any of these functions. Knowing any single function we can derive all other via the **Survival Identities**.

f(*x*): Density function
The relative likelihood of experiencing the event around time *x*.

$F(x)={}_0\int^x f(x)\, dx$

F(*x*): Distribution function
aka Cumulative function
The probability of experiencing the event until time *x*.
$\mathbf{F(x) = P(X \leq x)}$

Survival Identities

In survival analysis we consider the random variable
"X: Time until event"
x = 0.1 weeks, 2.3 weeks…

We express our knowledge about the distribution of X in any of these functions. Knowing any single function we can derive all other via the **Survival Identities**.

f(x): Density function
The relative likelihood of experiencing the event around time x.

$F(x)={}_0\int^x f(x)\,dx$

$S(x)={}_x\int^{\mathrm{Inf}} f(x)\,dx$

F(x): Distribution function
aka Cumulative function
The probability of experiencing the event until time x.
$F(x) = P(X \leq x)$

$S(x)=1-F(x)$

S(x): Survival function

The probability of *not* experiencing the event until time x.
$S(x) = P(X>x)$

Survival Identities

In survival analysis we consider the random variable **"*X*: Time until event"**
x = 0.1 weeks, 2.3 weeks…

We express our knowledge about the distribution of *X* in any of these functions. Knowing any single function we can derive all other via the **Survival Identities**.

f(*x*): Density function
The relative likelihood of experiencing the event around time *x*.

$F(x)={}_0\int^x f(x)\,dx$

$S(x)={}_x\int^{Inf} f(x)\,dx$

$h(x)=-S'(x)/S(x)$

h(*x*): Hazard function
The instantaneous rate of new events at time *x* among those who did not experience the event yet.

$\mathbf{h(x) = \lim_{h\to 0} P(x \le X < x+h \mid X \ge x)/h}$

F(*x*): Distribution function
aka Cumulative function
The probability of experiencing the event until time *x*.
$\mathbf{F(x) = P(X \le x)}$

$S(x)=1-F(x)$

S(*x*): Survival function

The probability of *not* experiencing the event until time *x*.
$\mathbf{S(x) = P(X > x)}$

Survival Identities

In survival analysis we consider the random variable
“*X*: Time until event”
x = 0.1 weeks, 2.3 weeks…

We express our knowledge about the distribution of *X* in any of these functions. Knowing any single function we can derive all other via the **Survival Identities**.

f(*x*): Density function
The relative likelihood of experiencing the event around time *x*.

h(*x*): Hazard function
The instantaneous rate of new events at time *x* among those who did not experience the event yet.
$h(x) = \lim_{h \to 0} P(x \leq X < x+h \mid X \geq x)/h$

$F(x) = \int_0^x f(x)\,dx$

$S(x) = \int_x^{Inf} f(x)\,dx$

$h(x) = -S'(x)/S(x)$

$H(x) = \int_0^x h(x)\,dx$

$H(x) = -\log S(x)$

F(*x*): Distribution function
aka Cumulative function
The probability of experiencing the event until time *x*.
$F(x) = P(X \leq x)$

$S(x)=1-F(x)$

S(*x*): Survival function

The probability of *not* experiencing the event until time *x*.
$S(x) = P(X > x)$

H(*x*): Cumulative Hazard
The integral of h(x).

Survival Identities

There are many more identities...

Table 2/1: Relations among the six functions describing stochastic lifetime

from \ to	$f(x)$	$F(x)$	$R(x)$	$h(x)$	$H(x)$	$\mu(x)$
$f(x)$	–	$\int_0^x f(z)\,\mathrm{d}z$	$\int_x^\infty f(z)\,\mathrm{d}z$	$\frac{f(x)}{\int_x^\infty f(z)\,\mathrm{d}z}$	$-\ln\left\{\int_x^\infty f(z)\,\mathrm{d}z\right\}$	$\frac{\int_0^\infty z\,f(x+z)\,\mathrm{d}z}{\int_x^\infty f(z)\,\mathrm{d}z}$
$F(x)$	$F'(x)$	–	$1-F(x)$	$\frac{F'(x)}{1-F(x)}$	$-\ln\{1-F(x)\}$	$\frac{\int_x^\infty [1-F(z)]\,\mathrm{d}z}{1-F(x)}$
$R(x)$	$-R'(x)$	$1-R(x)$	–	$\frac{-R'(x)}{R(x)}$	$-\ln[R(x)]$	$\frac{\int_x^\infty R(z)\,\mathrm{d}z}{R(x)}$
$h(x)$	$h(x)\,\exp\left\{-\int_0^x h(z)\,\mathrm{d}z\right\}$	$1-\exp\left\{-\int_0^x h(z)\,\mathrm{d}z\right\}$	$\exp\left\{-\int_0^x h(z)\,\mathrm{d}z\right\}$	–	$\int_0^x h(z)\,\mathrm{d}z$	$\frac{\int_x^\infty \exp\left\{-\int_0^z h(v)\,\mathrm{d}v\right\}\mathrm{d}z}{\exp\left\{-\int_0^x h(z)\,\mathrm{d}z\right\}}$
$H(x)$	$-\frac{\mathrm{d}\{\exp[-H(x)]\}}{\mathrm{d}x}$	$1-\exp\{-H(x)\}$	$\exp\{-H(x)\}$	$H'(x)$	–	$\frac{\int_x^\infty \exp\{-H(z)\}\,\mathrm{d}z}{\exp\{-H(x)\}}$
$\mu(x)$	$\frac{1+\mu'(x)}{\mu^2(x)}\times\mu(0)\times \times\exp\left\{-\int_0^x \frac{1}{\mu(z)}\,\mathrm{d}z\right\}$	$1-\frac{\mu(0)}{\mu(x)}\times \times\exp\left\{-\int_0^x \frac{1}{\mu(z)}\,\mathrm{d}z\right\}$	$\frac{\mu(0)}{\mu(x)}\times \times\exp\left\{-\int_0^x \frac{1}{\mu(z)}\,\mathrm{d}z\right\}$	$\frac{1}{\mu(x)}\{1+\mu'(x)\}$	$\ln\left\{\frac{\mu(x)}{\mu(0)}\right\}+ +\int_0^x \frac{1}{\mu(z)}\,\mathrm{d}z$	–

Rinne (2008). The Weibull Distribution.

Recap

For survival distributions and identities read

Klein & Moeschberger (2003). Survival Analysis. Sections 2.1–2.4.

For refreshing your understanding of basic calculus watch the series

3Blue1Brown (2017). The essence of calculus. YouTube.

For refreshing your understanding of random variables and probability distributions watch

Khan Academy (2012). Random variables. YouTube.

Princeton COS 302 (2020). Probability density and mass functions. YouTube.

Homework

Choose a time-to-event setting that interests you and look up a constant rate related to that setting. What is the time scale for your setting? When does the time-to-event start? When have have half of the population experienced the event given the chosen rate?

Example: Today we looked at the time until I catch COVID. I choose the rate 2,428 infections per 100,000 persons per 7 days from the local COVID incidences and assumed this rate to be constant. The timescale was “weeks into the semester” and it starts at the first week of the semester. I used the survival function of the exponential distribution to calculate the time until the probability of catching COVID reached 50%.

$S(x) = \exp(-\int_0^x \lambda dx) = \exp(-\lambda x)$

Materials for this lecture

github.com/jschoeley/survival_analysis-ur-ss22

Jonas Schöley

@jschoeley

0000-0002-3340-8518

j.schoeley@uni-rostock.de