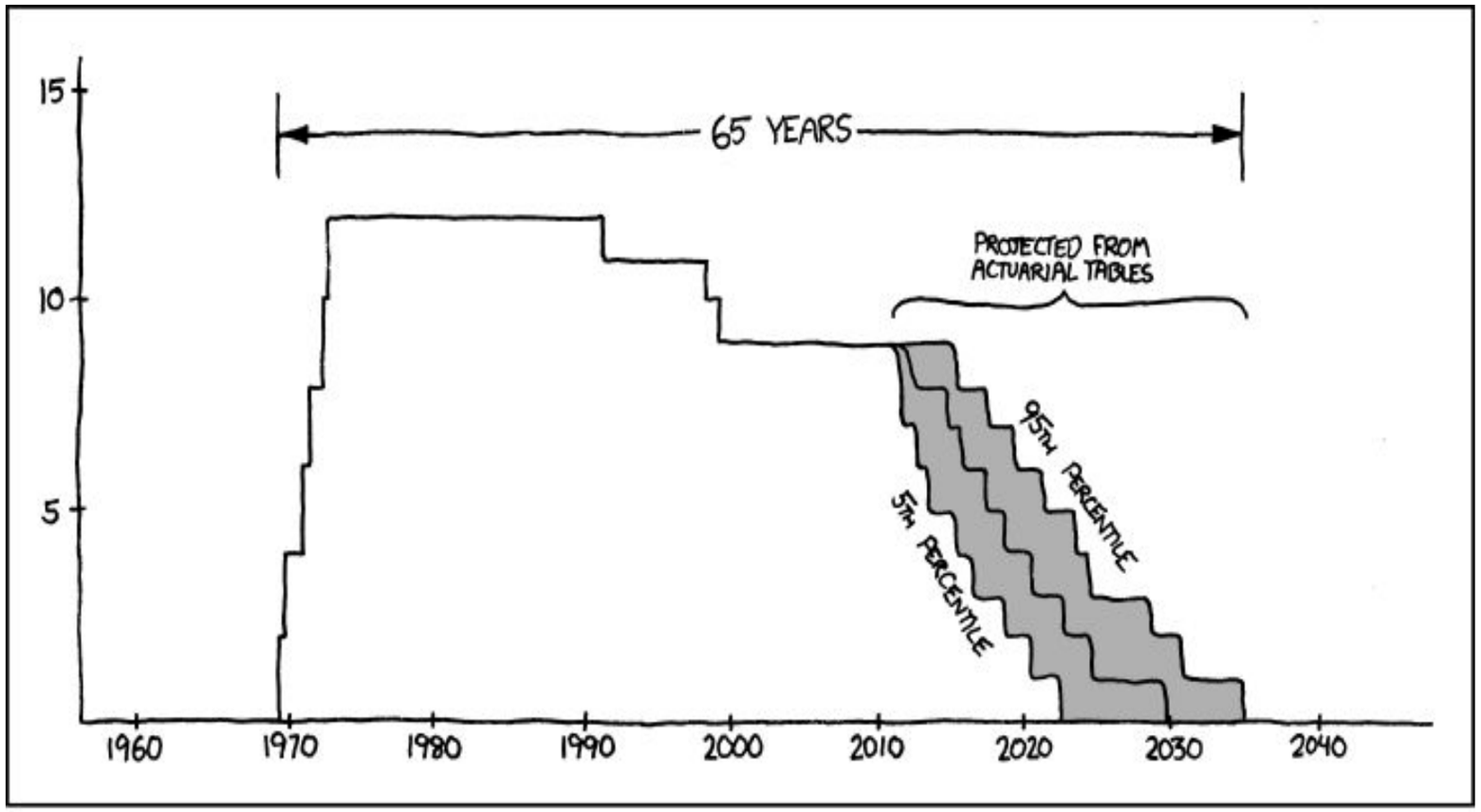
Can You Survive This Workshop? Brief Introduction to Survival Analysis

GRACE HOPPER CELEBRATION

ANITA B.ORG

Download slides & Jupyter notebook for this workshop here: github.com/zuzannna/can-you-survive-this



Marianne Hoogeveen
Bowery Farming

Zuzanna Klyszejko
Wayfair

How to get the most out of this workshop?

During the workshop:
Look out for those labels on the slides which indicate **interactive questions** or **group exercises**

go to hsay.co/s/LUDYK to submit your answers

After the workshop:
Go to our repo and work through additional exercises & simulations:

github.com/zuzannna/can-you-survive-this

find a handout on your table

What are typical data science questions?

go to hsay.co/s/LUDYK to submit your answers

1. Which of my customers are likely to convert soon?

2. What is the probability my patient will have a relapse of symptoms?

3. What proportion of cars produced in our factory is likely to break down after a year? (So we can make sure our one year guarantee is profitable)

4. Is this a cat or a hot dog?

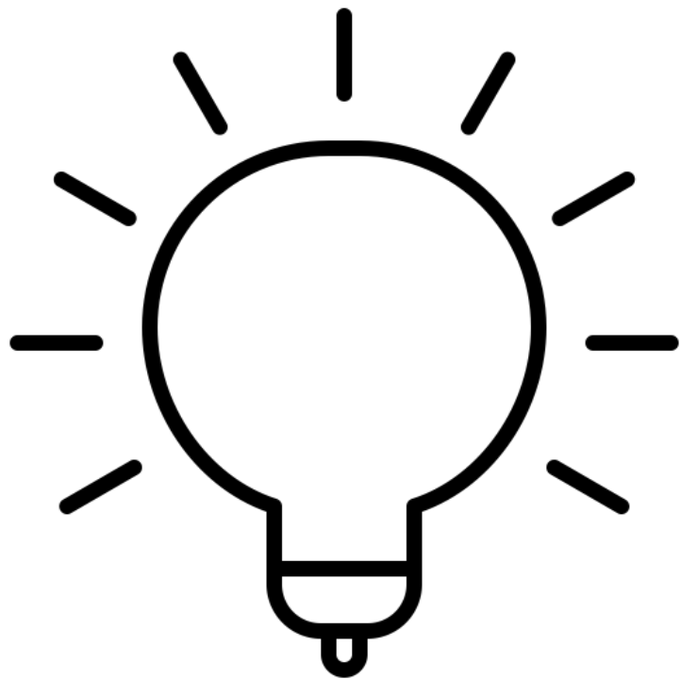
5. All of the above.

Sometimes it is more useful to ask *how long*?

Which of my customers are likely to convert soon?

How long does it take for a customer to convert?

What is the probability my patient will have a relapse of symptoms?

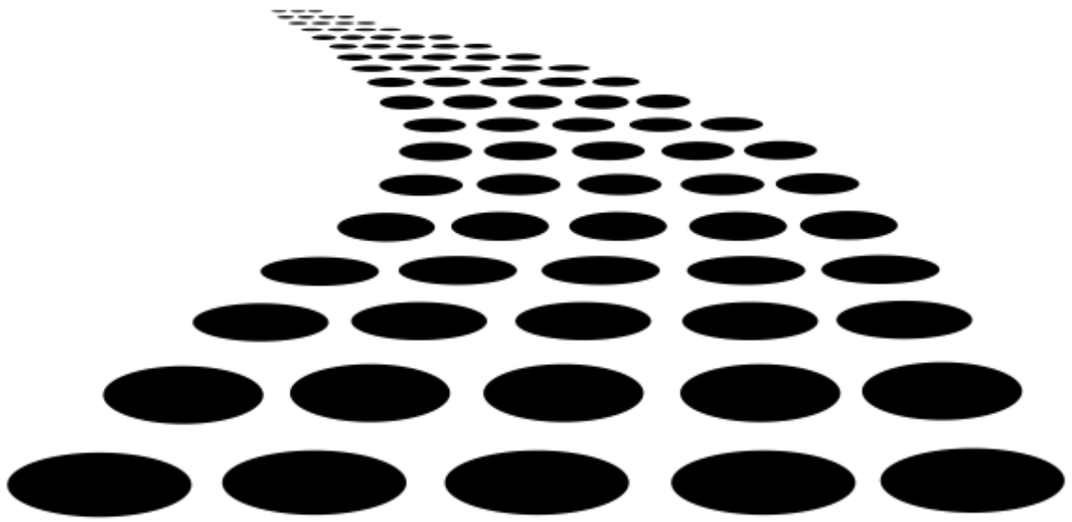


What proportion of cars produced in our factory is likely to break down after a year?

How long should we guarantee our cars won't break?

How long will my patient live for?

What is survival analysis?



Survival analysis deals with estimating duration until an event happens. Events such as...

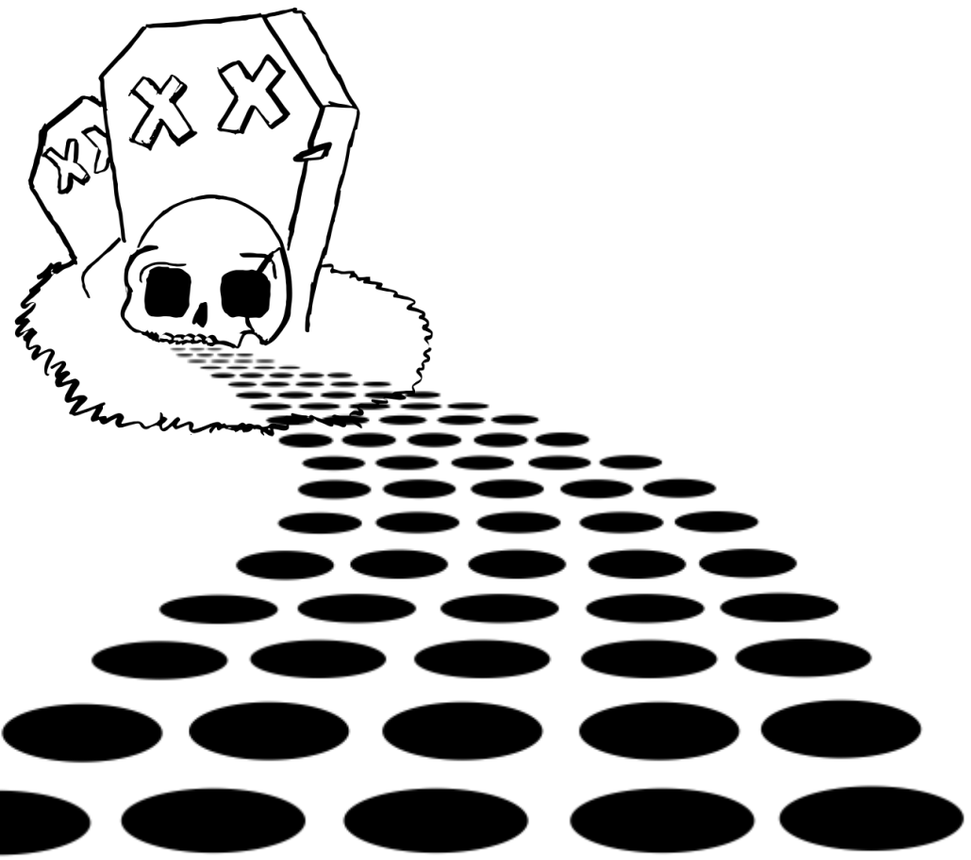
Survival analysis deals with estimating duration until an event happens. Events such as...

time until your car breaks down.

Survival analysis deals with estimating duration until an event happens. Events such as...

time until your customer is fed up with your marketing emails and unsubscribes.

Survival analysis deals with estimating duration until an event happens. Events such as...

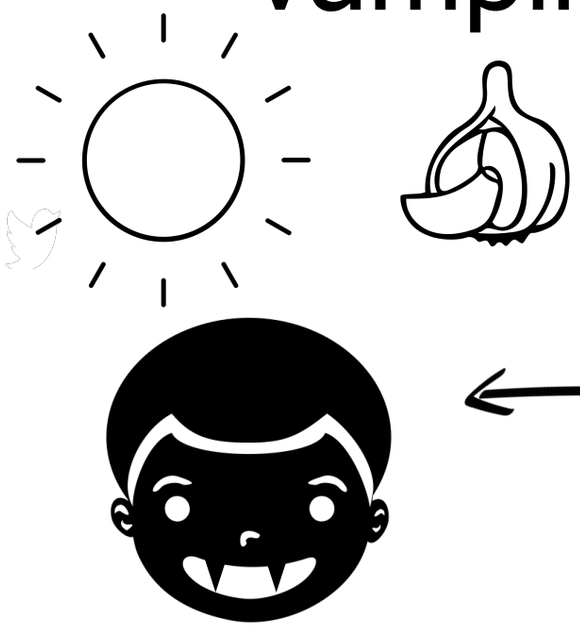


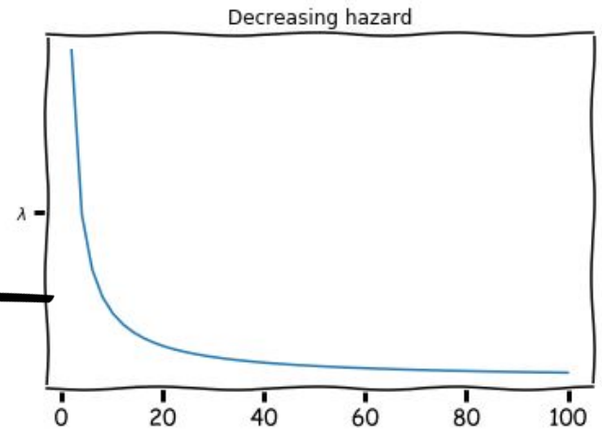
... time until we die.
Speaking of dying and undead...

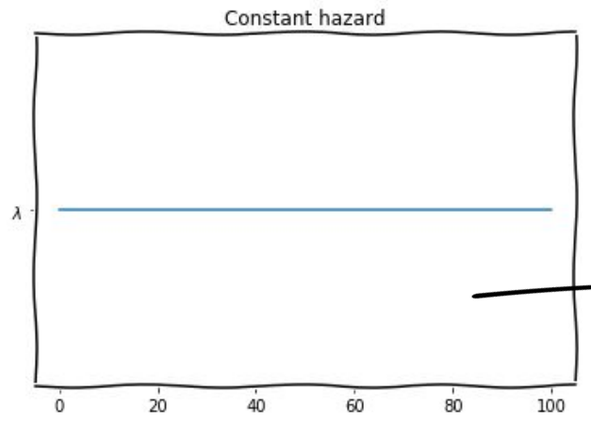
Meet our vampire who will help us understand survival analysis.

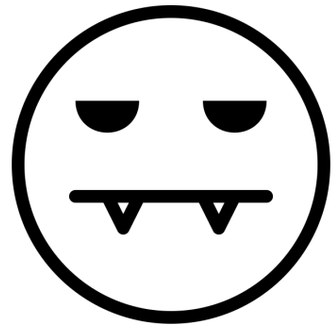
Vampire life is not an easy one. There are a lot of perils they need to avoid to survive for an eternity: garlic, sun rays, angry mobs and depression in old age.

We can quantify the risk of dangerous events occurring with different probability across vampire's life using e.g. Weibull function



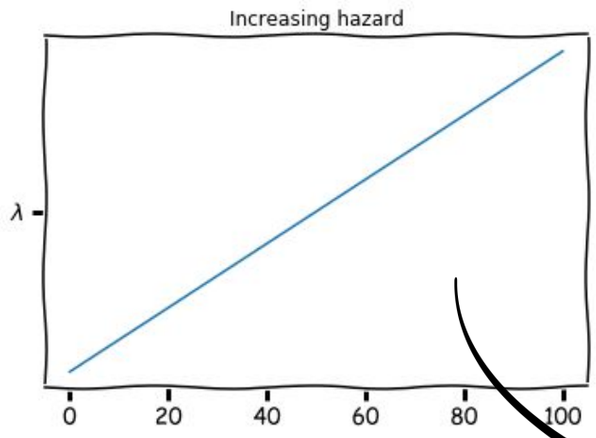






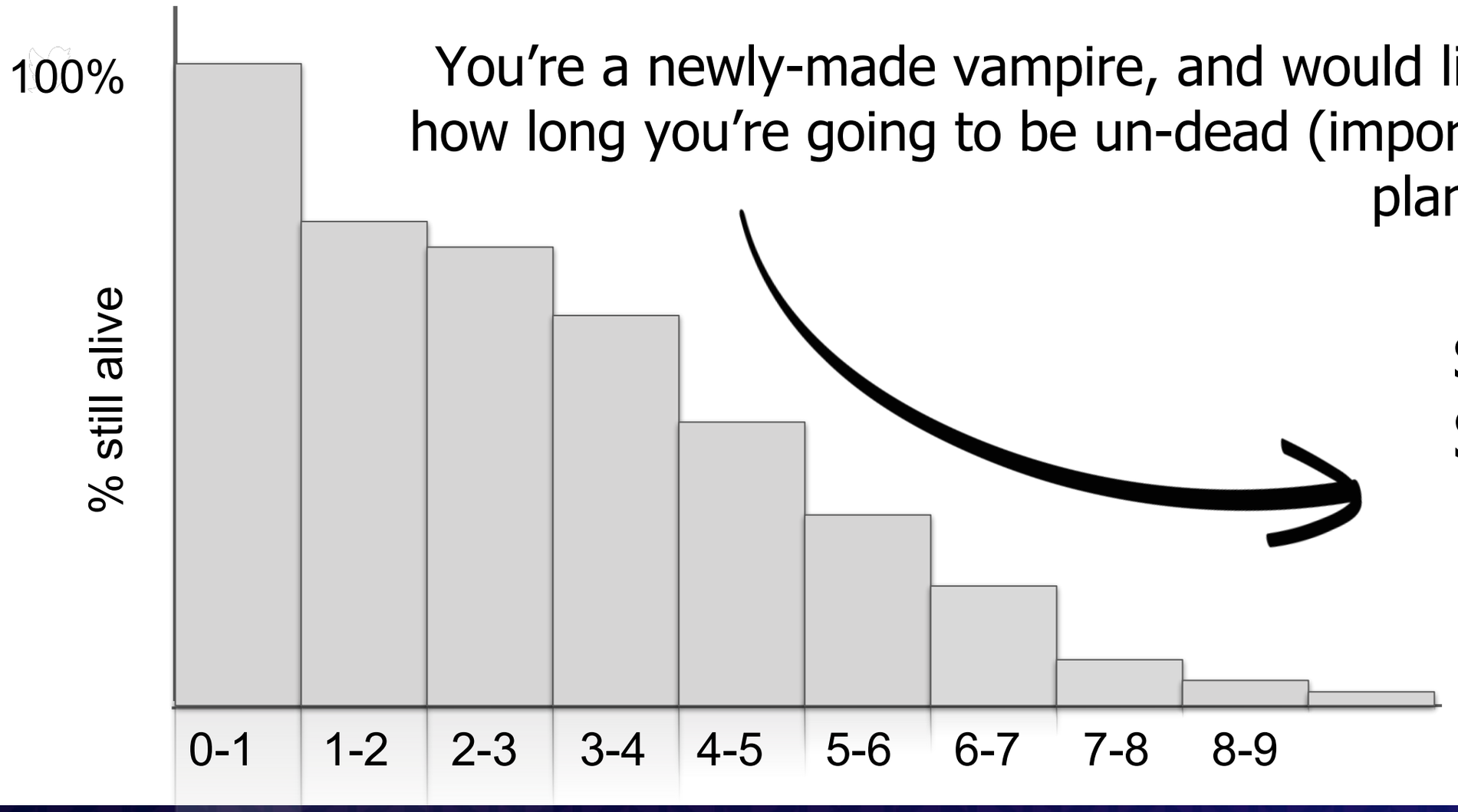
… OR nothing ever changes.

...OR young vampire is inexperienced and might try garlic sauce and sun bathing before they learn.



… OR in the old age vampire gets very depressed and suicidal.

As a new vampire you can leverage survival function to predict your expected length of life



You're a newly-made vampire, and would like to have an idea how long you're going to be un-dead (important for retirement planning, for instance)

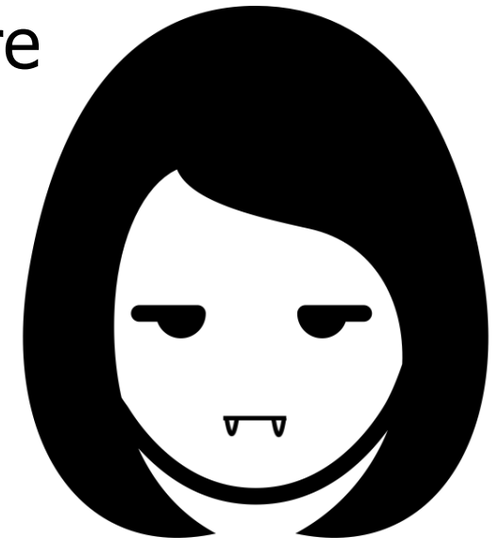
Survival function:

$S(t) = P(T > t)$

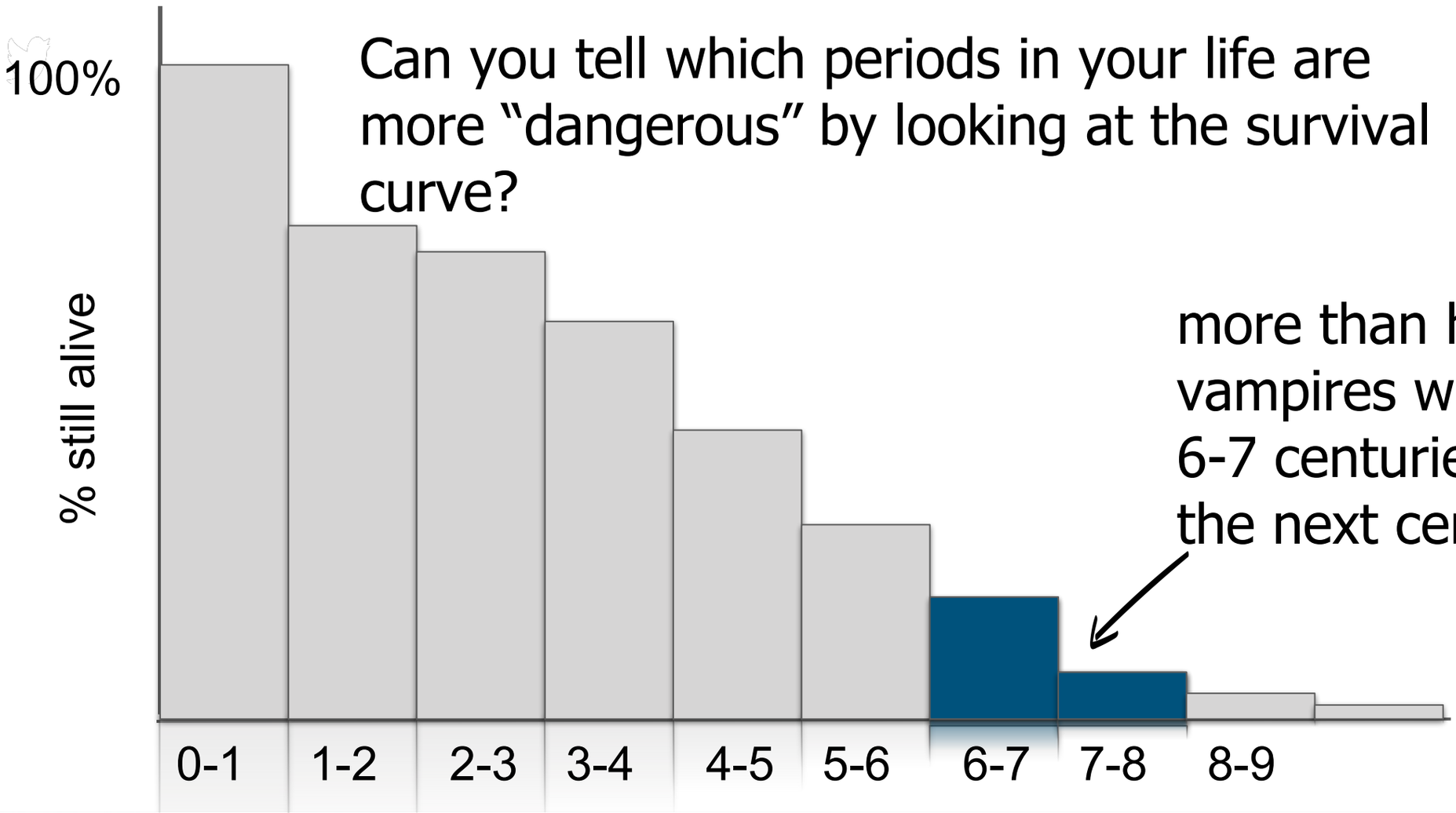
age at time of death (centuries)

There must be better ways to deliver actionable insights than just the survival function

You have now been a vampire for a while. Your first years were challenging, and you wonder whether the worst is behind you.

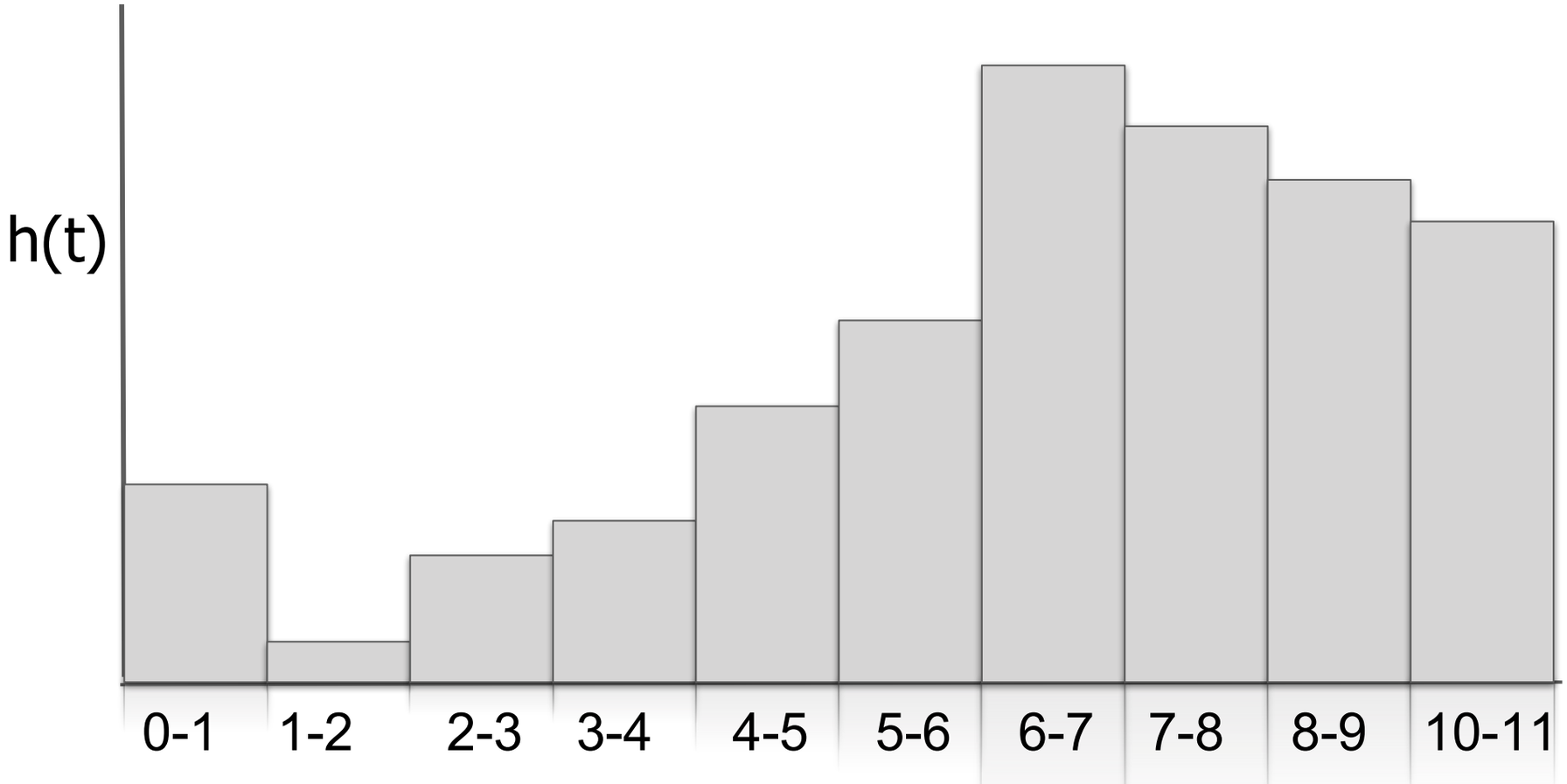


Can you tell which periods in your life are more "dangerous" by looking at the survival curve?



Hazard Rate clearly shows periods of time *t* when the danger drastically increases in comparison to *t - 1*

Hazard rate gives you the probability density you will die at a given age, **given that you have made it to that age**

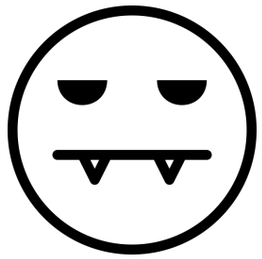


$$h = \frac{\Delta S}{S}$$

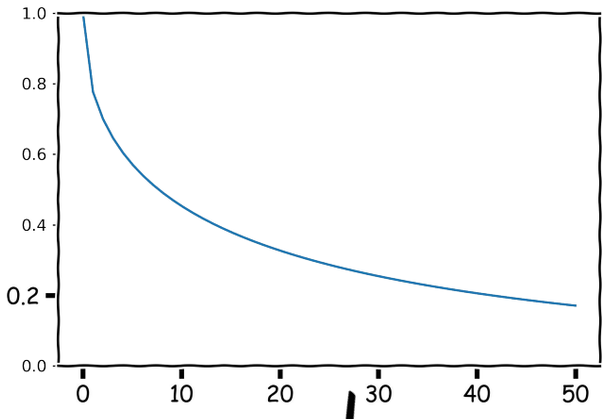
("normalized" pdf)

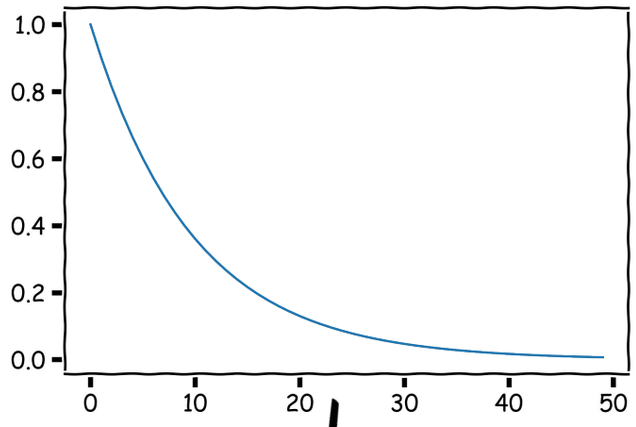
age at time of death (centuries)

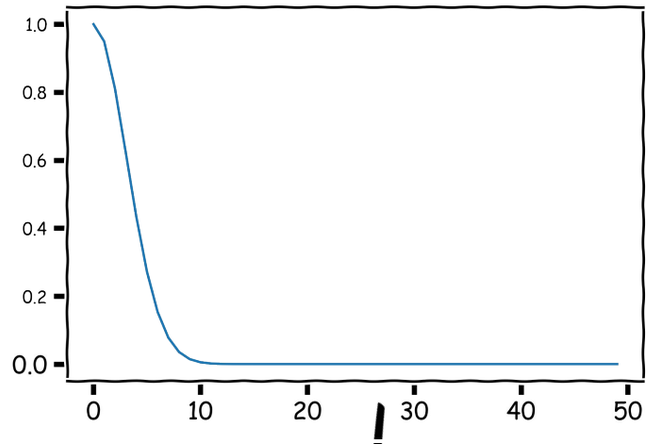
Example: power-law hazard (Weibull)



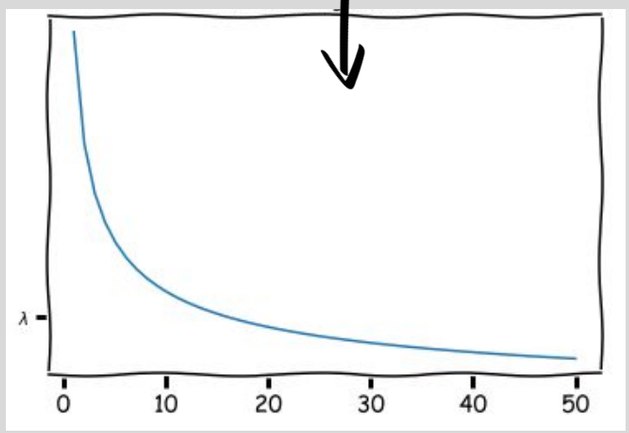
Survival function

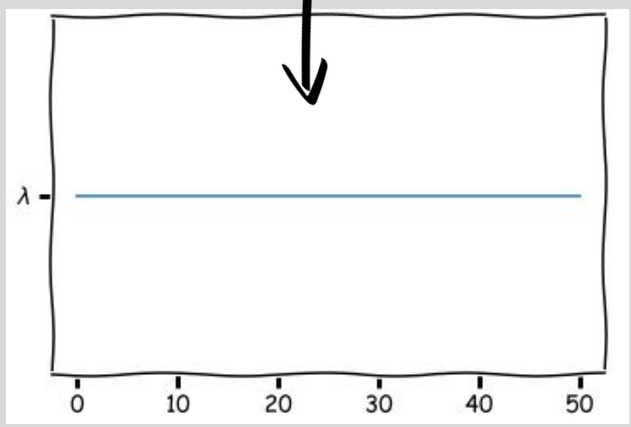


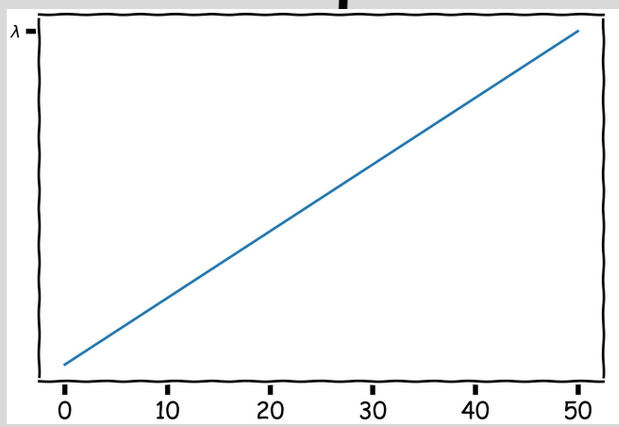




Hazard rate

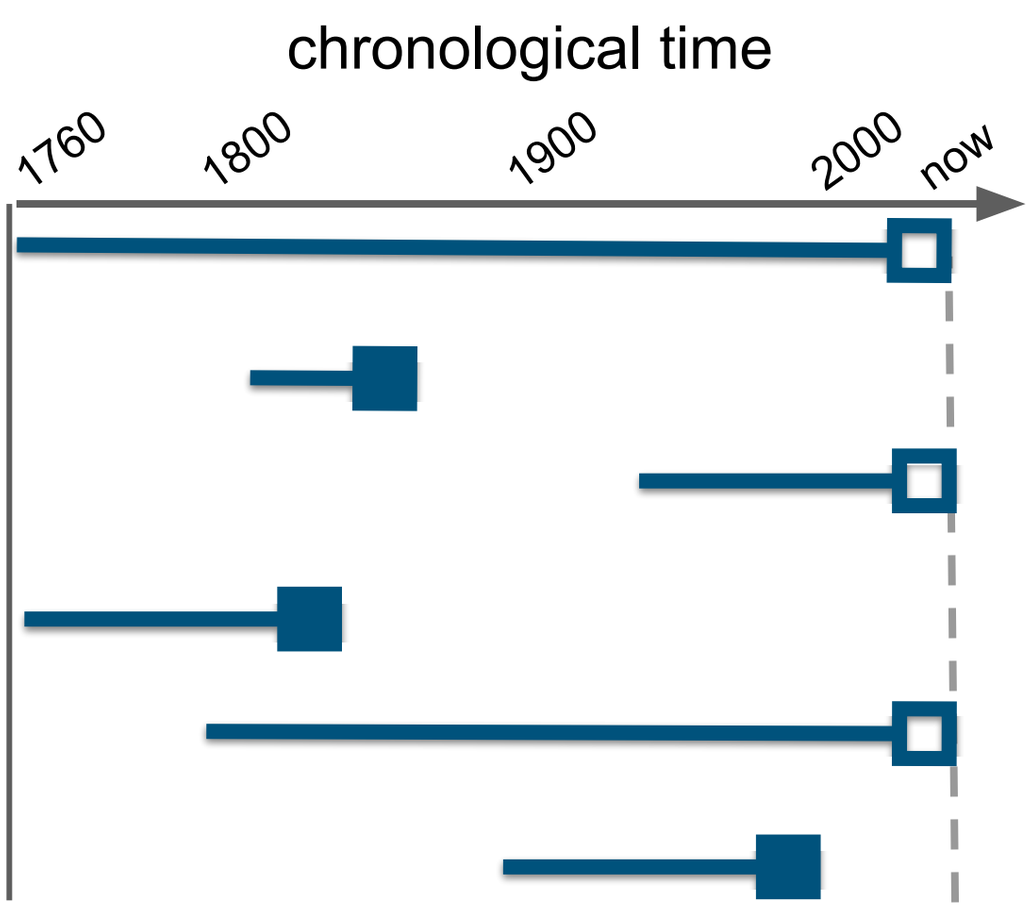


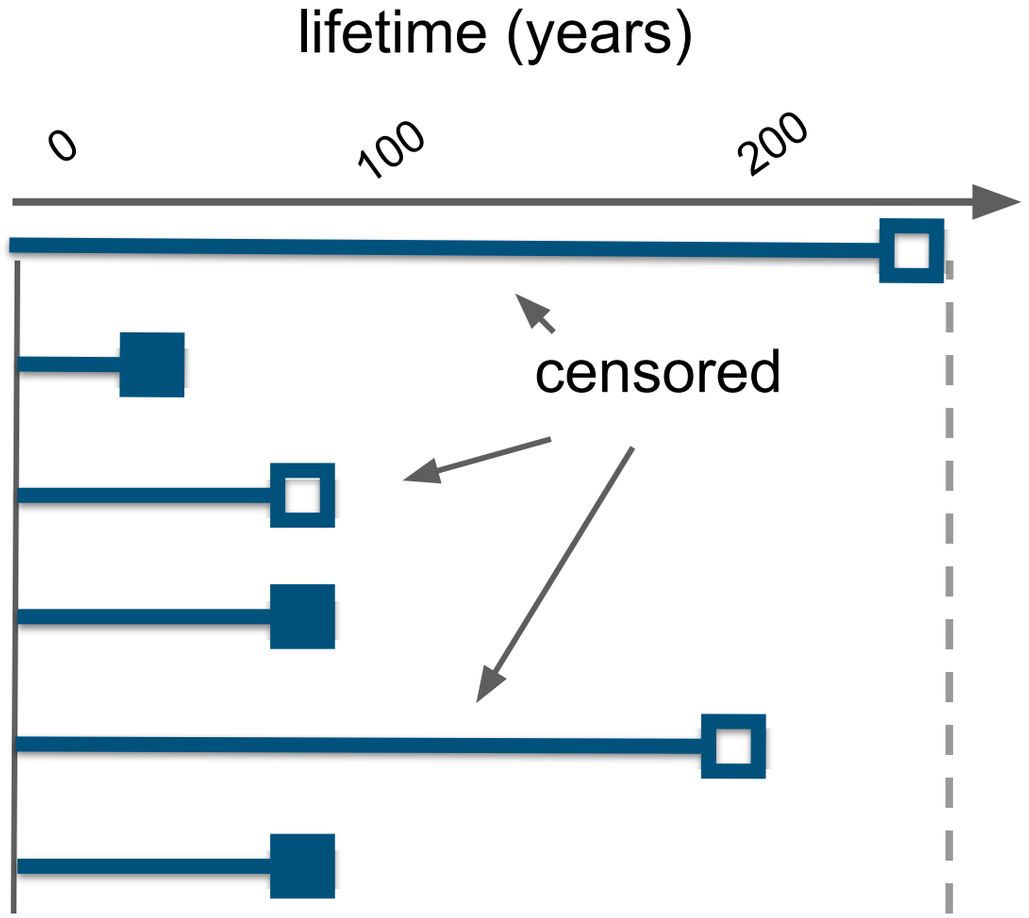




Duration data can have partially missing information which biases conclusions

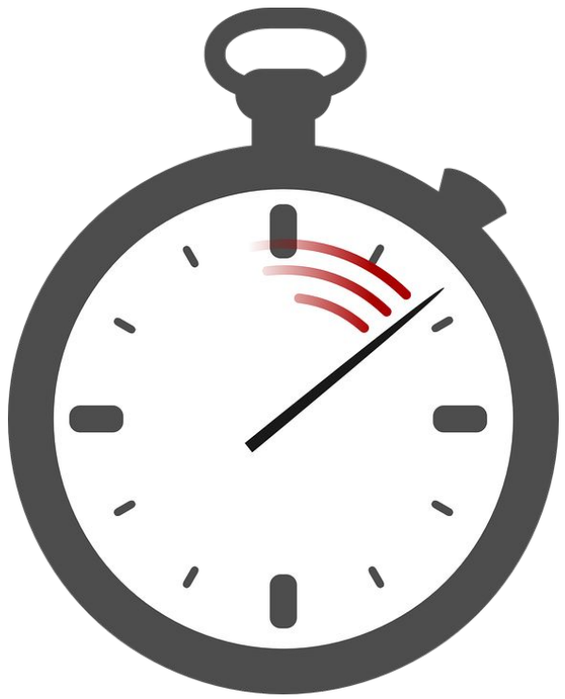
The end date can be missing from the data. We call that "censoring"





Breakout Session: Censoring

find a handout on your table



5 mins

Censoring breakout session: discussing the results

go to hsay.co/s/LUDYK to submit your answers

1. You are in charge of measuring exact lap times in a car race. For the first lap, you have sensors that precisely measure when a car crosses the end line, but for the starting times, all you know is when the starting sign was given.

What is this an example of?

Censoring breakout session: discussing the results

go to hsay.co/s/LUDYK to submit your answers

2. You are trying to measure the width of your sofa with a piece of string that has a known length. It comes just slightly short. You simply write down the length of the string with a note that the sofa is actually slightly wider.

What is this an example of?

Censoring breakout session: discussing the results

go to hsay.co/s/LUDYK to submit your answers

3. When measuring pollutants in our water a researcher may not care about (or instruments may not be able to detect) the level of pollutants if it falls below a certain threshold (e.g., .005 parts per million). In this case, any pollutant level below .005 ppm is reported as “<.005 ppm.

What is this an example of?

Censoring breakout session: discussing the results

go to hsay.co/s/LUDYK to submit your answers

4. Everyday you commute to work using a city-wide rental bike system. [...] You usually set a timer which you turn on before checking out the bike and stop the moment you dock it but today you forgot to stop it and when you got to your office it was still running at 55 minutes.

What is this an example of?

Censoring breakout session: discussing the results

go to hsay.co/s/LUDYK to submit your answers

5. You are working on a hydroponic farm and it is recommended that you use a special plant food on Boston lettuce leaves for a short amount of time, starting on the 14th day after planting it. [...] Today is 9/14 and your colleague who was doing the planting says it happened either on 09/01 or 09/02.

What is this an example of?

Censored data can bias our conclusions about lifetimes

Study: ~1000 vampires and their age at time of death (or current age if alive) over period of 500 years - **orange curve**

Asked Alice (a clairvoyant vampire from the *Twilight* saga) to predict final ages of all currently undead vampires - **blue curve**

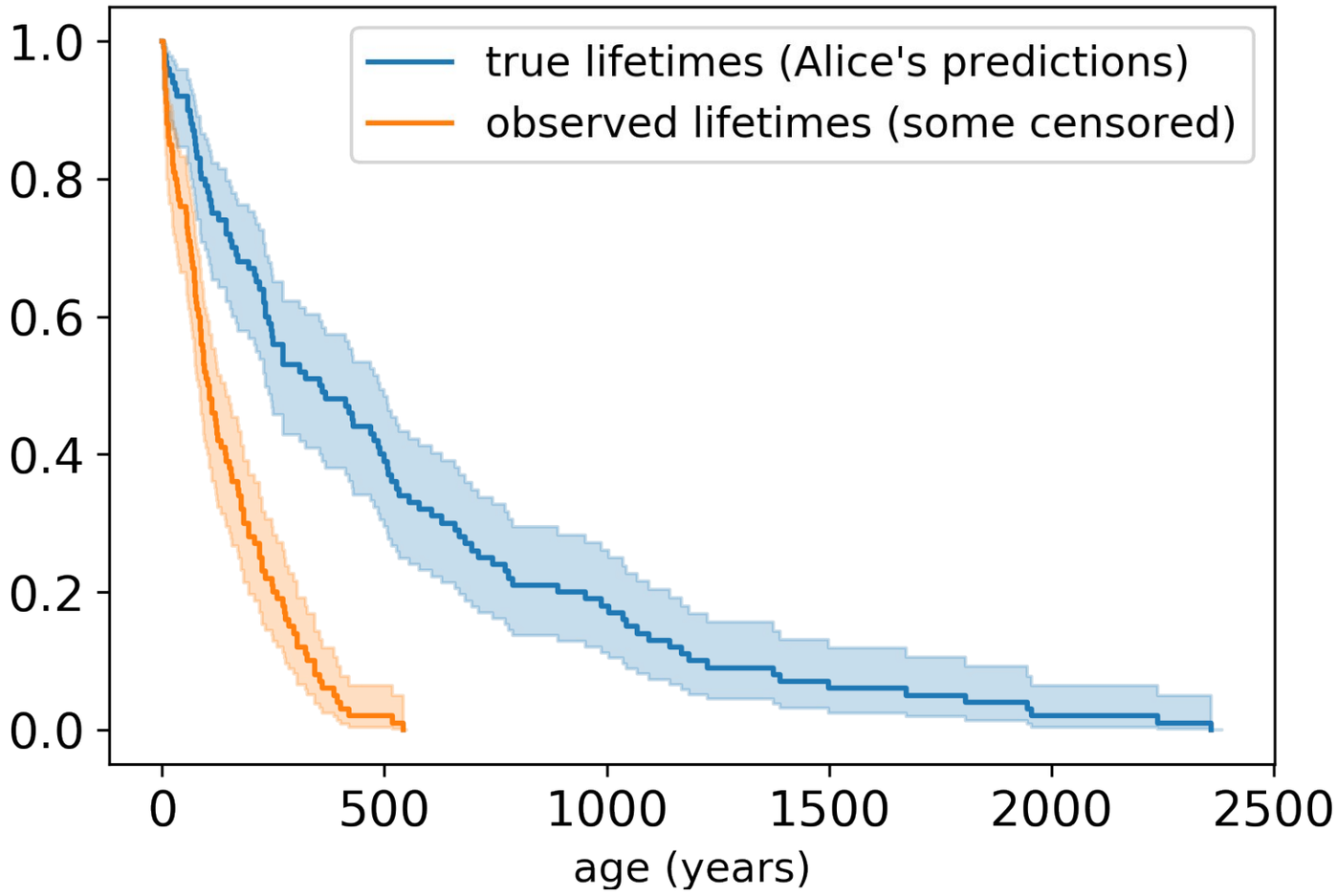
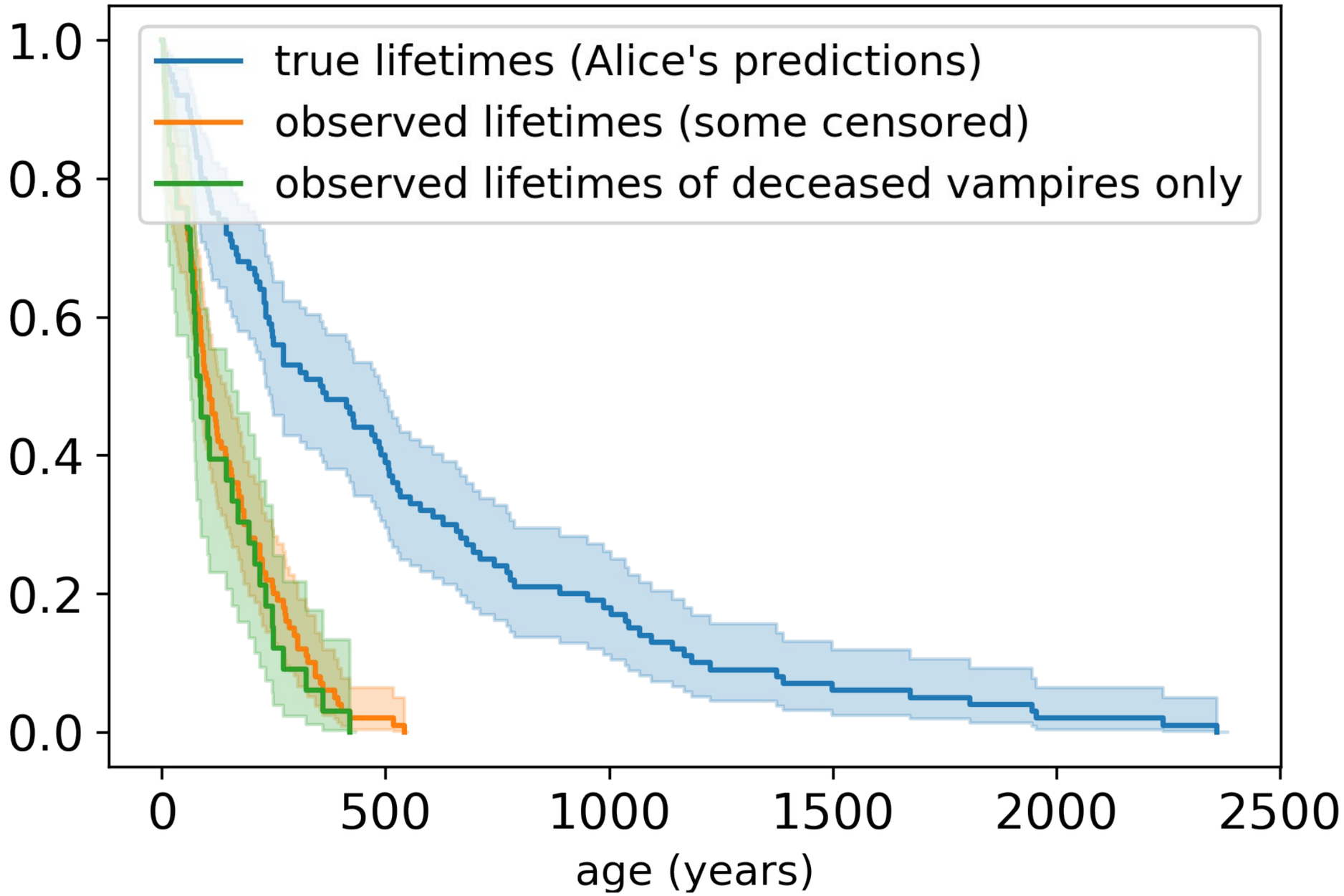
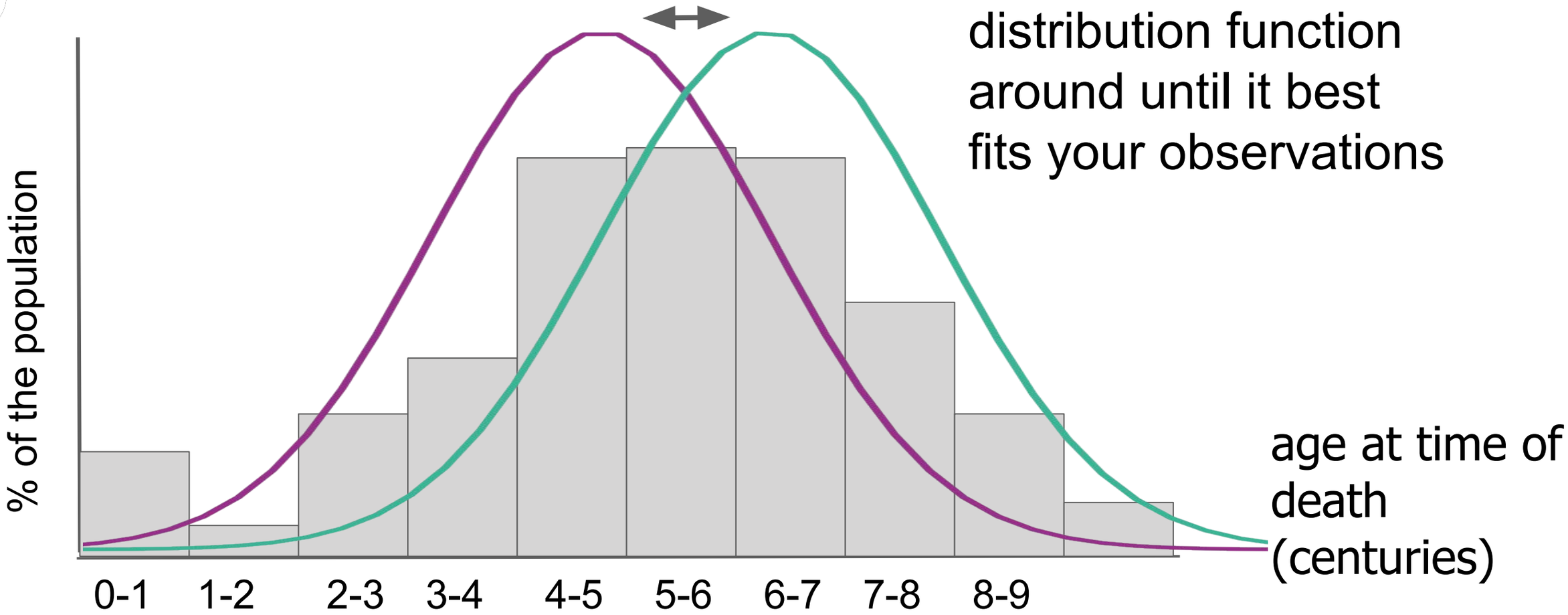


image (modified from):
https://www.flickr.com/photos/makelessnoise/3477085794

If we simply remove all censored observations we still end up with a bias

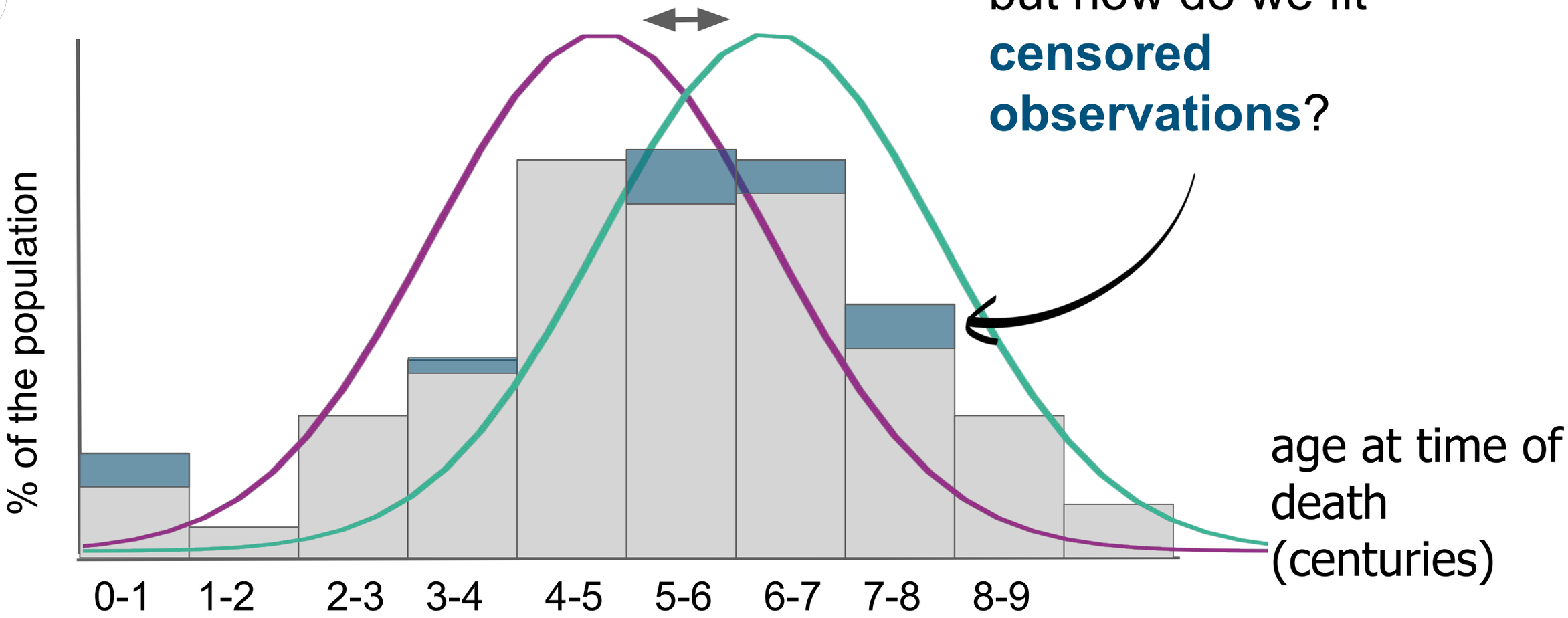


Without censoring you can use likelihood estimation to fit a probability density function to your data



move the probability distribution function around until it best fits your observations

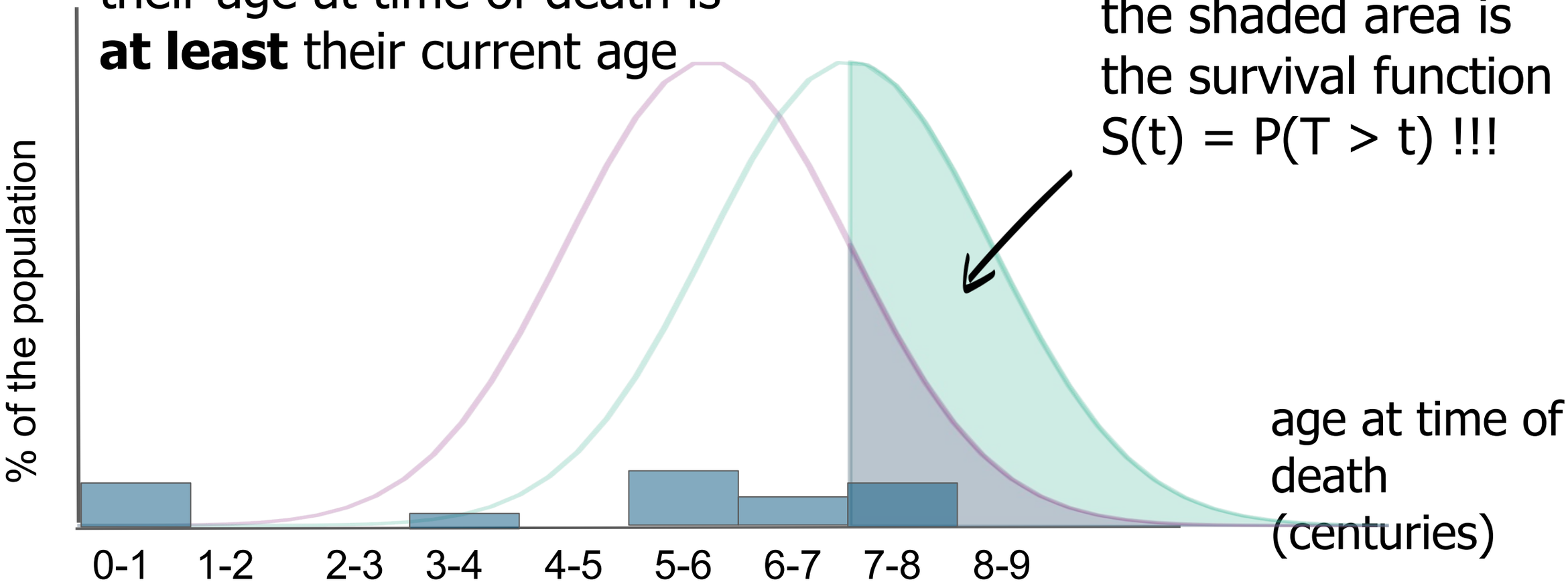
Without censoring you can use likelihood estimation to fit a probability density function to your data



but how do we fit **censored observations**?

But with censored observations you must use a cumulative distribution

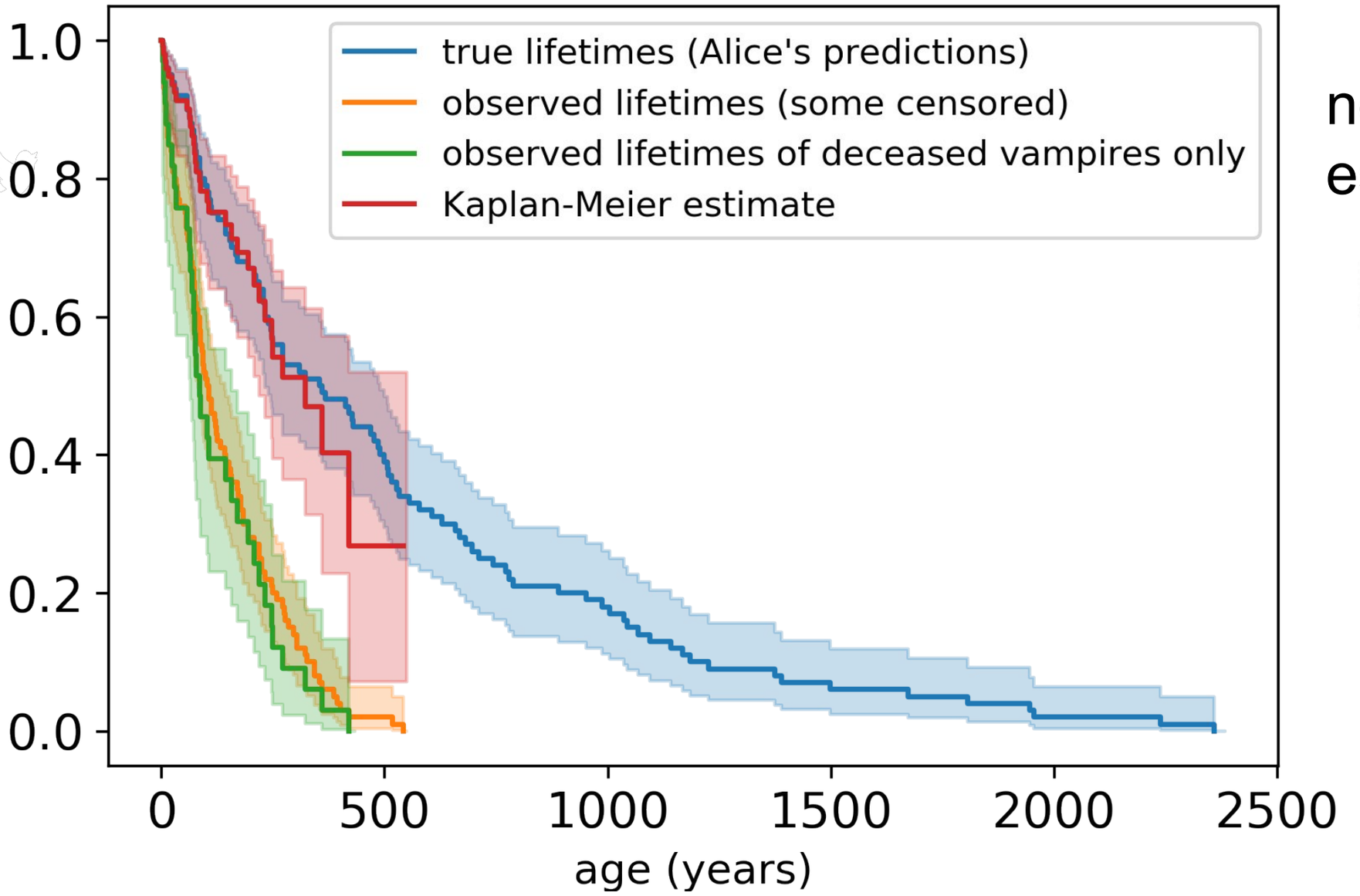
for vampires who are still un-dead, we only know that their age at time of death is **at least** their current age



the shaded area is the survival function S(t) = P(T > t) !!!

age at time of death (centuries)

Kaplan-Meier estimator does a good job of accounting for censoring



non-parametric estimator

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

censored observations are not included in d_i but are present in n_i

Breakout session: Blinking experiment

find a handout on your table

STEP 1: Introduce yourself to your neighbor (or two!)

STEP 2: Designate the experimenter and arm them with a time measuring device.

STEP 3: The experimenter measures how long each person within the group can keep their eyes open without blinking. **Maximum time: 30 seconds.**

STEP 4: One of the participants switches with the experimenter and measures how long they can last.

STEP 5: Enter your group's results here: https://goo.gl/P4oX9q.

Experiment: discussion

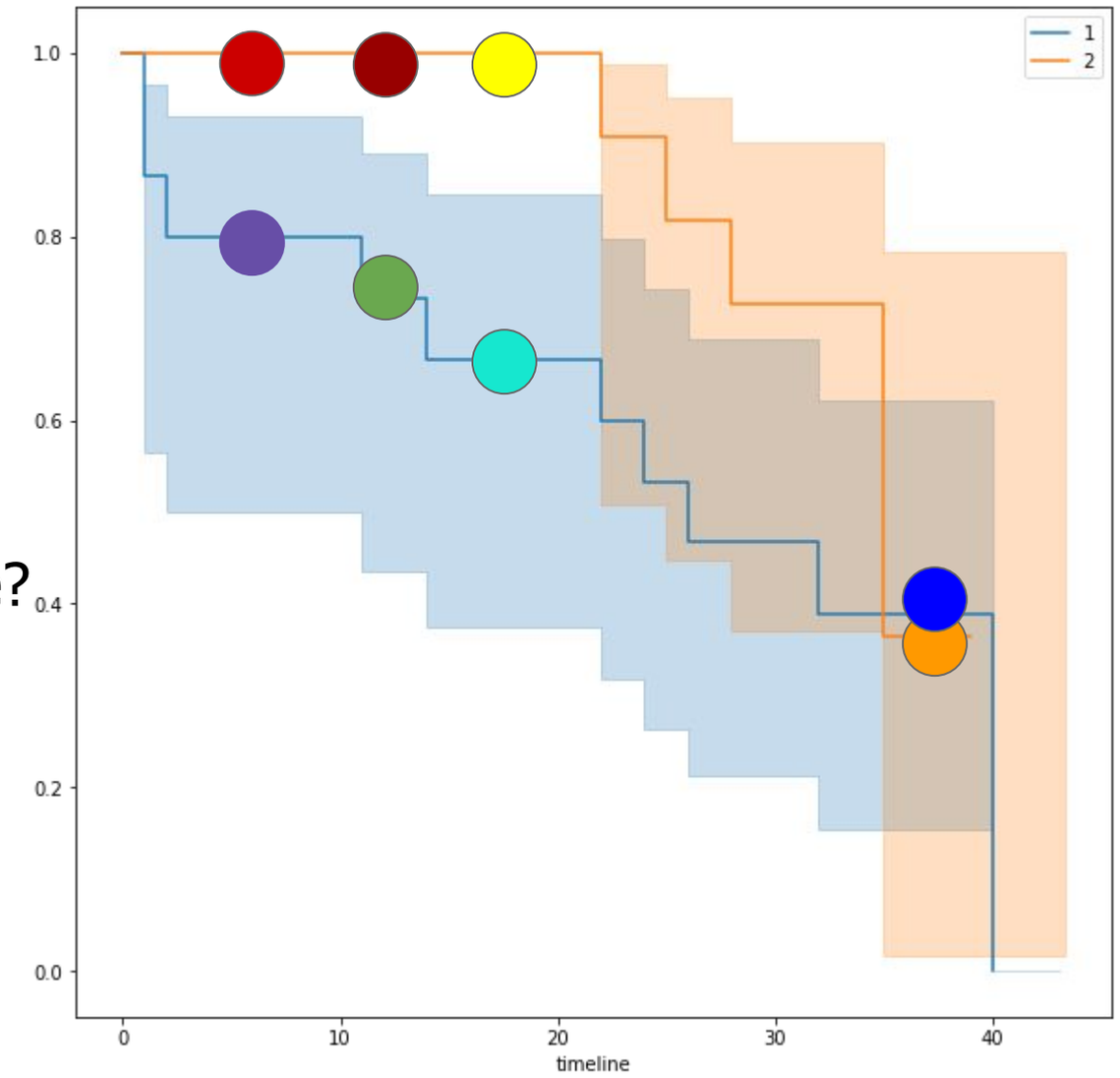
What are the sources of potential bias in our measurements?

What would you add as covariates?

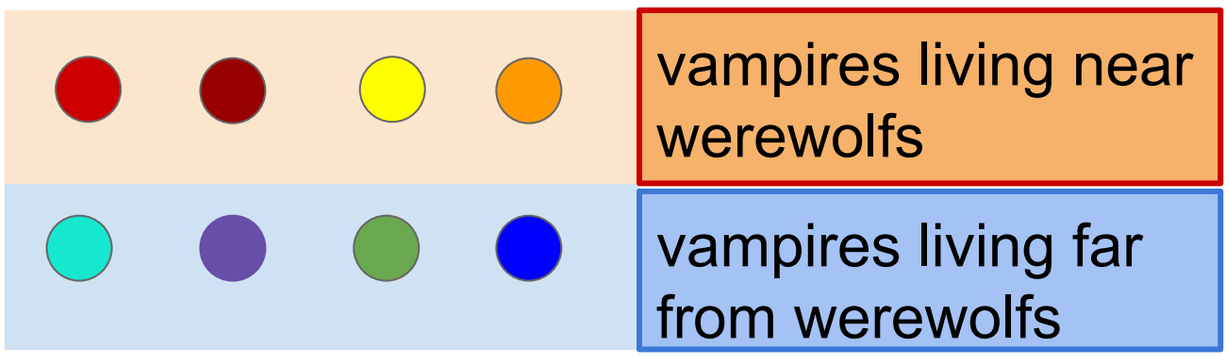
Can you think of other experiments you could run?

How can we measure differences between the groups? **Log rank test**!

Example question: Are there any differences in survival among vampires who live within 25 mile radius from large werewolf groups and those who don't?



Which of the pair of samples has a higher value?
1 if warm color, 0 if cool color, 0.5 if a tie.



0 1 1 0.5

Cox proportional hazards model lets us predict individual hazard rates based on a set of features

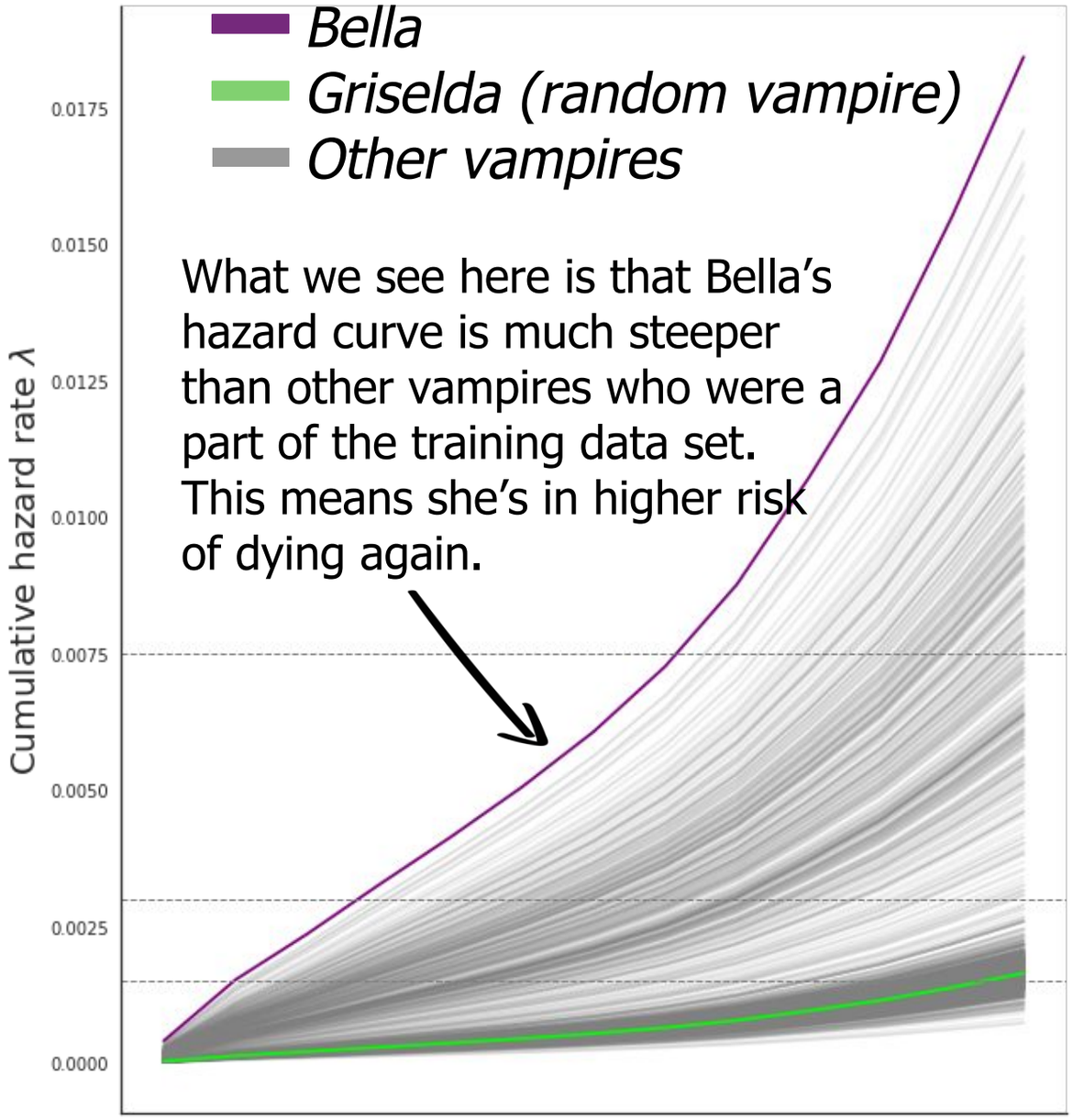
Features for Cox regression model

$$\lambda(n|X_i) = \lambda_0(n)exp(X_i \cdot \beta)$$

Get **coefficient values** β

Example: using Cox model we could calculate hazard rate for newly undead Bella Swan from the Twilight saga based on her *age*, *proximity to werewolf groups, size of her vampire coven* and *potassium content in her blood at the time of the change*. We need to calculate beta weights based on data from a population of vampires and apply them to Bella's data.

Using beta weights calculated on the whole population we can predict *relative* hazard of a single person (Bella Swan) during her lifetime

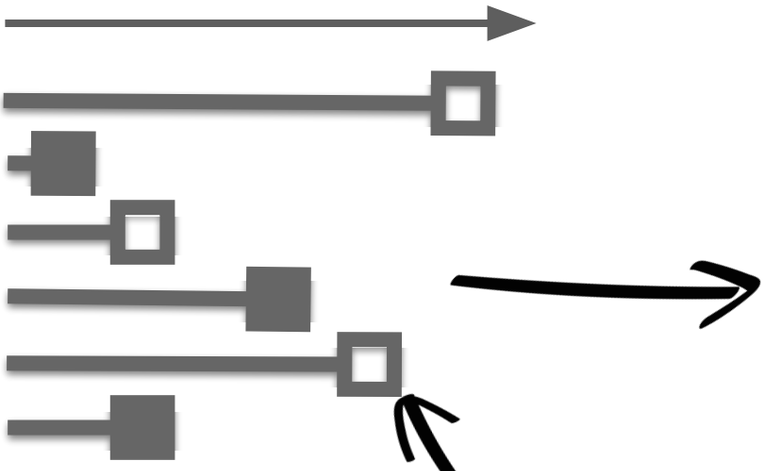


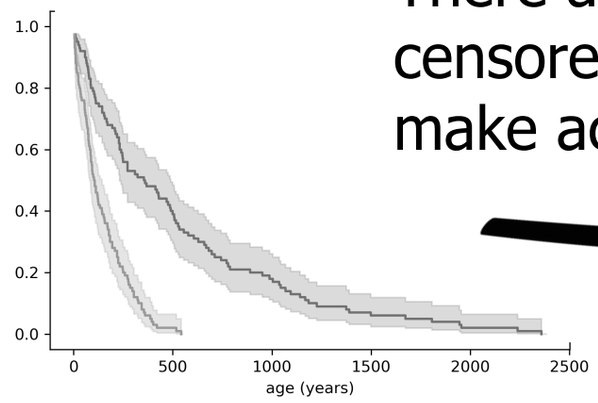
Features for Cox regression model

$$\lambda(n|X_i) = \lambda_0(n)exp(X_i \cdot \beta)$$

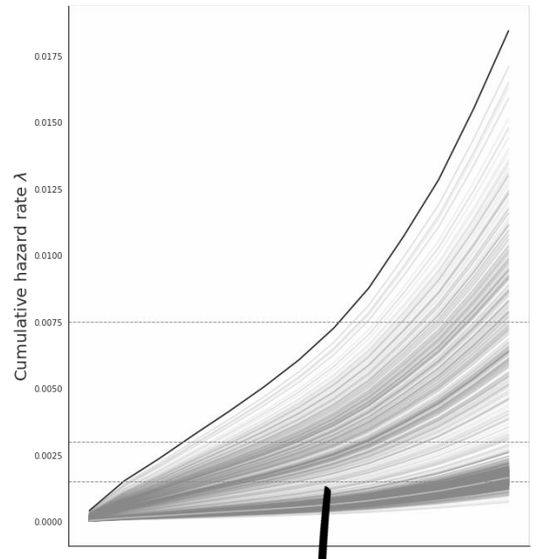
Get **coefficient values** β

Compute **individual hazard curves λ_i** with $X_i \cdot \beta$ for Bella and others





There are ways to deal with censored observations to make accurate predictions.



When you deal with time duration data and don't account for censored observations your results will be biased

Wait, what was this workshop about?

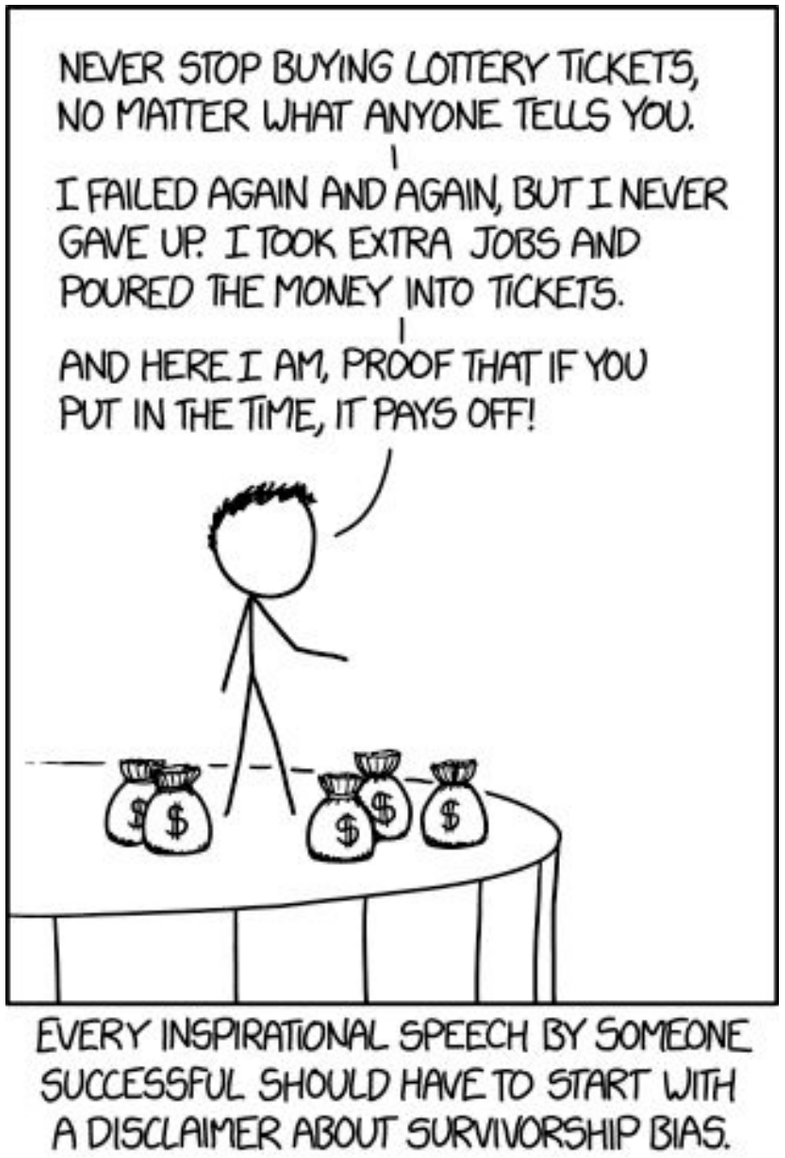
Maybe you will come with the next method?

Keep us posted!

To continue your adventure with survival analysis:
Go to our repo and work through additional exercises & simulations.
All our teaching materials are there too in case you want to teach this workshop to your friends!

github.com/zuzannna/can-you-survive-this

Thank You For Surviving This Workshop!



Marianne Hoogeveen
Bowery Farming

Zuzanna Klyszejko
Wayfair