Probability, Statistics, and Calculus Review

LOGM 634

9 October 2014

Statistics & Probability

- Statistics
  - Inferring a population's behavior based on the performance of a sample
    - Representative of the population?
    - Sample is unbiased?
- Probability
  - Predicting the future behavior of a single item, based on the population's known performance

Another view

- Probability:
  - The study of the laws of chance
- Statistics:
  - Use probability laws on past information to predict future events

Modeling Uncertainty

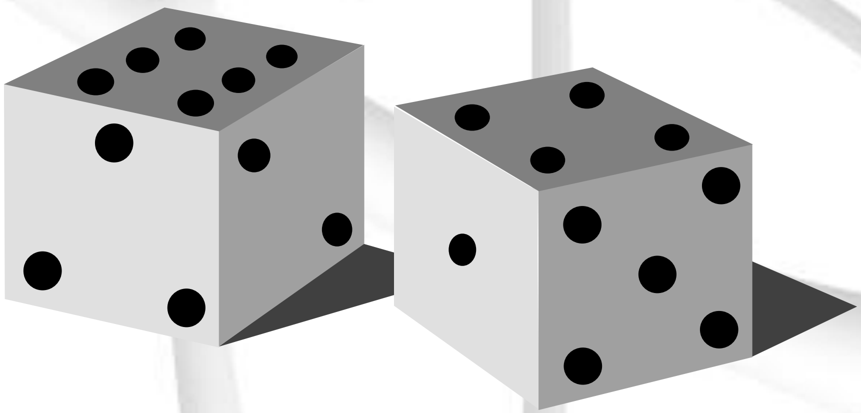
- Basic Method
  - Sample space & events within an experiment
  - Define probabilities of events
  - Compute probabilities of more complex events formed from unions/intersections of the elementary events
- Concept of Random Variables
  - Variable that takes on certain values iaw specific probabilities
  - Specify the probability distribution of a random variable to characterize a random process
- Use both through the course

Key Terms

- Probability
- Experiments
- Sample spaces
- Outcomes
- Events
- Random variables

An Experiment

- Probability analysis is always made in the context of some *experiment*
  - If I exceed my aircraft's weight and balance limits, can I still take off?
  - I bought a ticket -- will I win the lottery tonight?

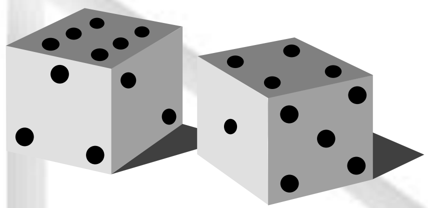


I was just handed two dice -- will I roll a seven?

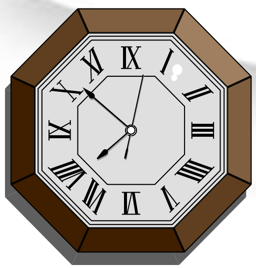
Sample Space

- Collection of all possible outcomes of an experiment
  - Mutually exclusive
  - Collectively exhaustive
- Countable (discrete)

We know something will happen in our next experiment, we're just not sure what!



- Uncountable (continuous)



Events

- Just a collection (subset) of outcomes contained in the sample space
  - one, a few, or all of ‘em
    - Exactly 1 = simple
    - > 1 = compound

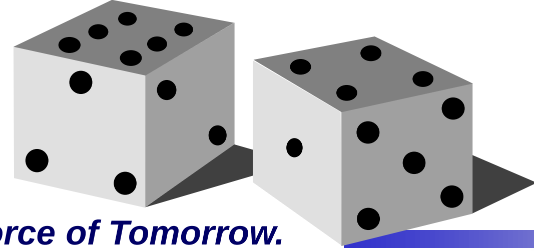
A Random Variable

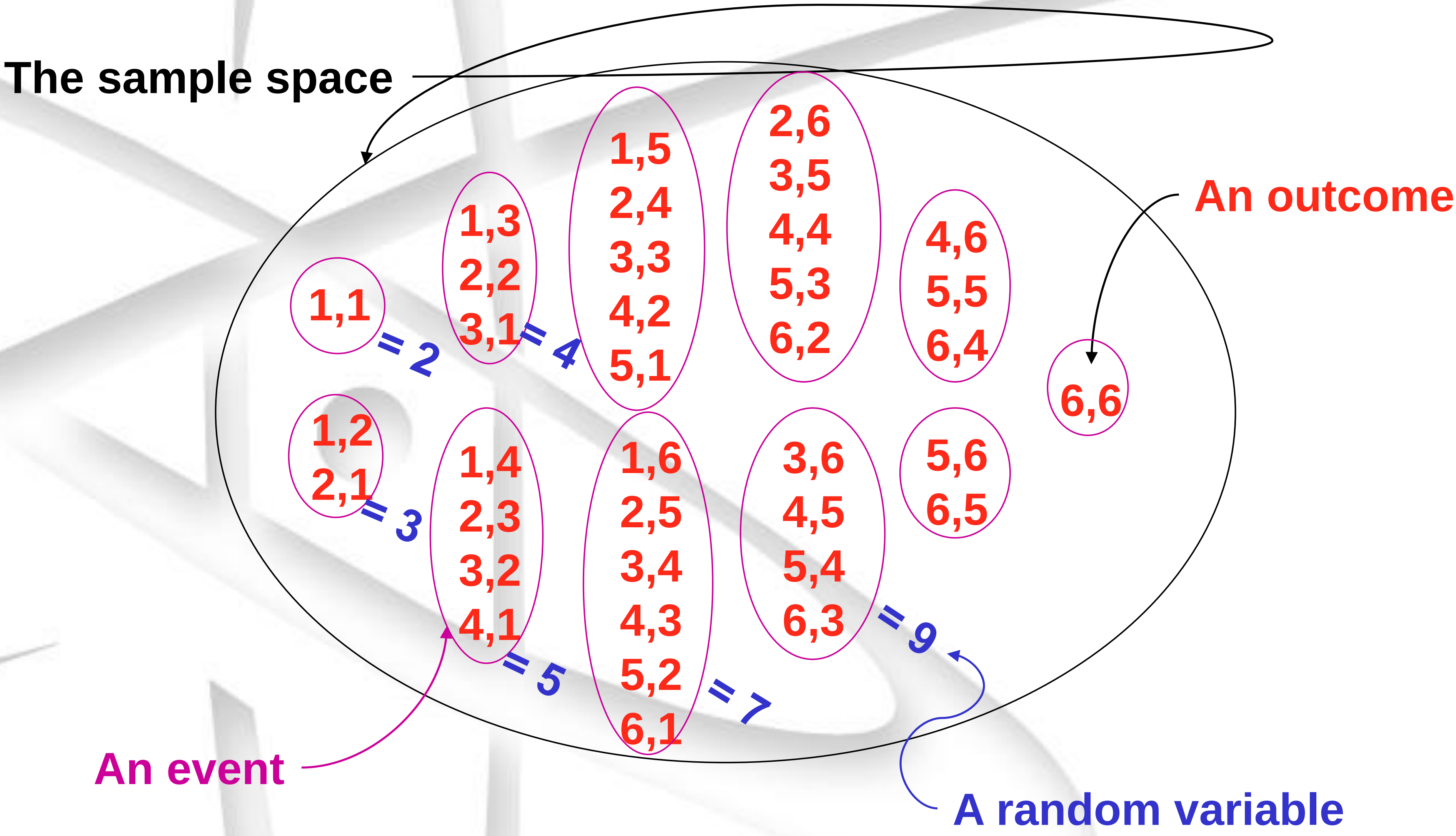
- A variable that takes on numerical values iaw some probability distribution
- May be continuous or discrete
- Nothing random about it – it's not really a variable!
- It's a FUNCTION
  - It maps an event (a single outcome or a set of 'em) to the *real number line,*

(runner breaks tape)

(runners total time)

Roll the dice…





What about the Probability?

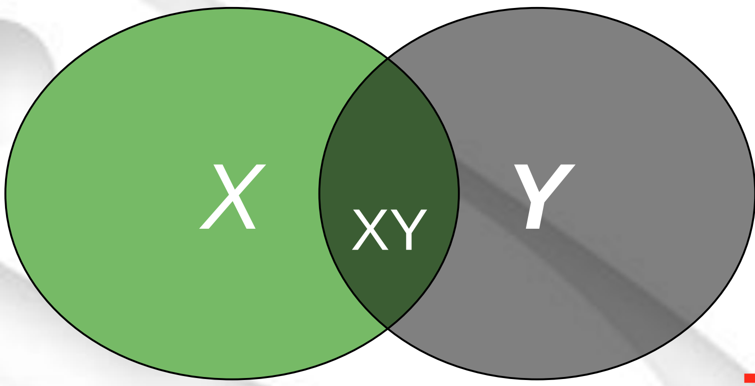
- Probability is the ***likelihood*** of an outcome
  - Call it “X”

- For each event, X, in the sample space, we assume the probability of X occurring has the following properties:
  - for any specified event X, $0 \le P(X) \le 1$
  - If P(X) = 1, then X is a *certainty*
  - If $P(X)$ = 0, then X is an *impossibility*
    - *Doesn’t mean x <u>won’t</u> happen – it’s just highly unlikely*

What about...

- The union:
  - The probability that either *x* or *y* or both occur

$$P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)$$



The union

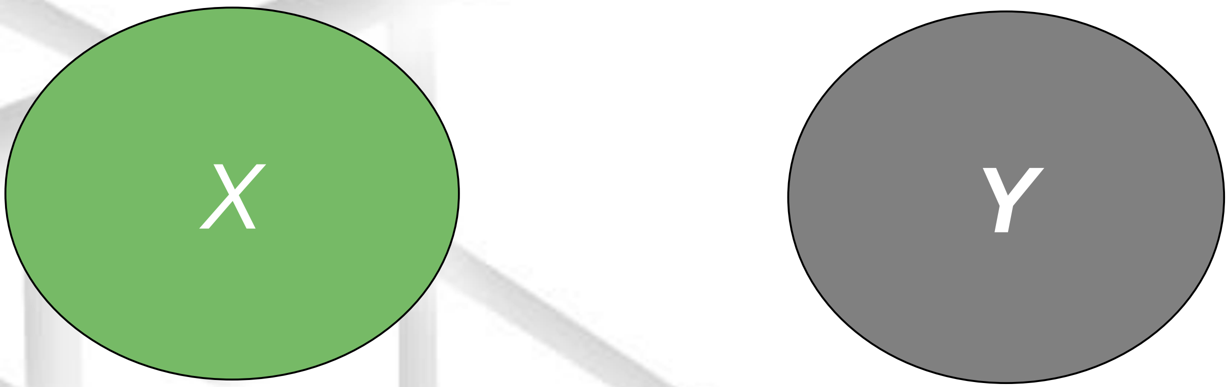
$$P(X \cup Y) = ?$$

The intersection

$$P(X \cap Y) = ?$$

What about…..

- Are *X* and Y *Mutually Exclusive*?



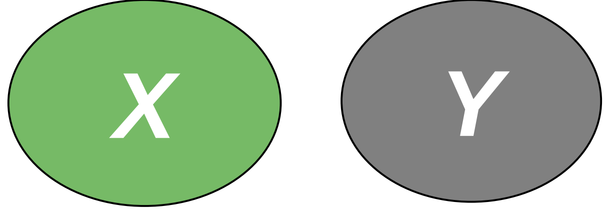
The union

$$P\left(X \bigcup Y\right) = ?$$

The intersection

$$P\left(X \bigcap Y\right) = ?$$

What about…..



If X and Y are mutually exclusive then X and Y cannot occur together and P(X ∩ Y) = 0

So, when X and Y are Mutually Exclusive, then

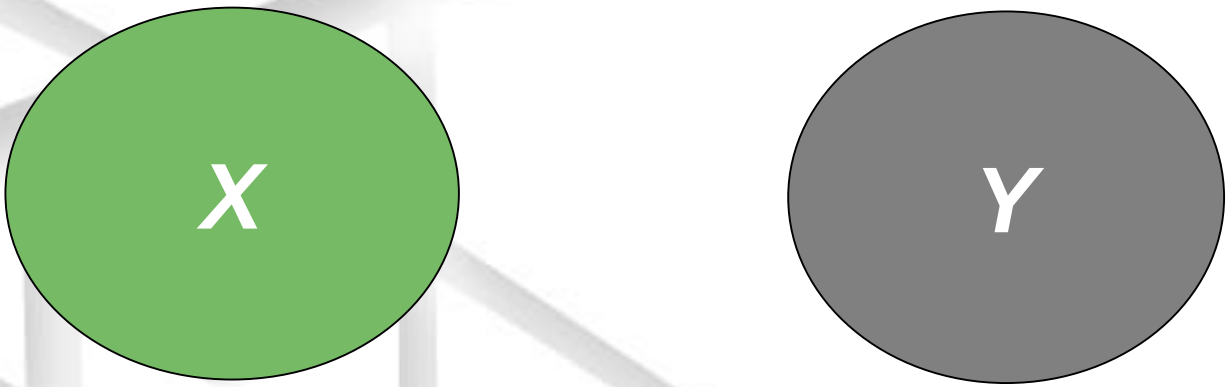
$$P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)$$

$$P(X \cup Y) = P(X) + P(Y) - 0$$

$$P(X \cup Y) = P(X) + P(Y)$$

What about…..

- Are *X* and Y *Independent*?



The union

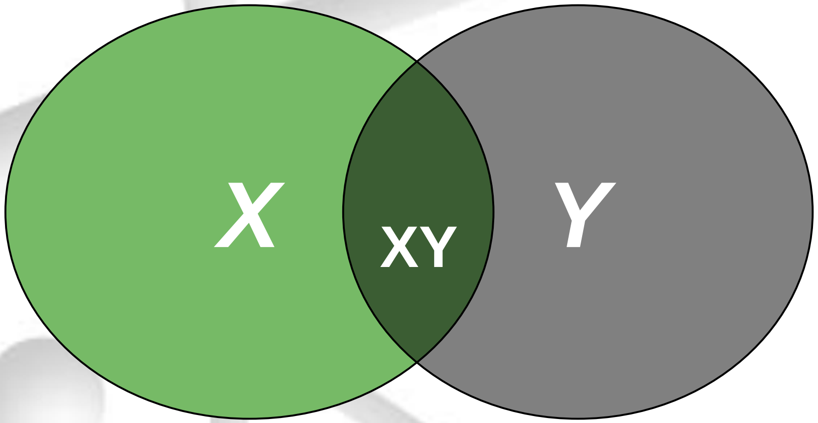
$$P\left(X \bigcup Y\right) = ?$$

The intersection

$$P\left(X \bigcap Y\right) = ?$$

Conditional Probabilities

- Theorem for Conditional Probability:

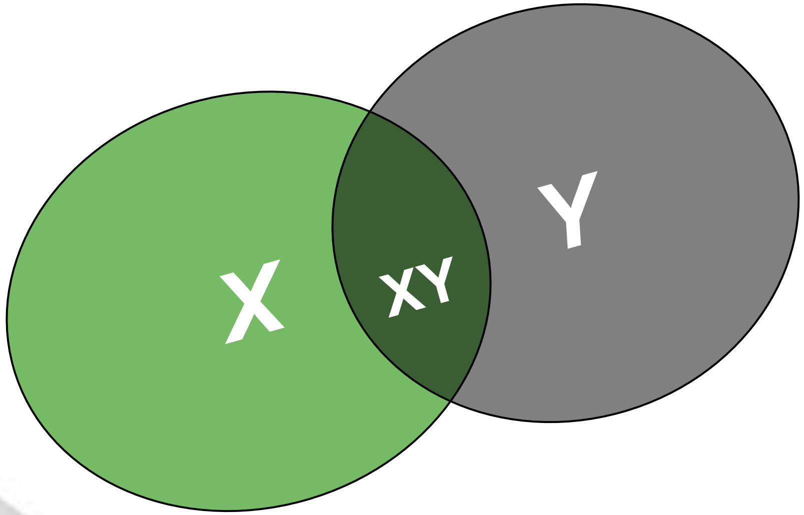


Defn: If $P(Y) > 0$, then

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} \rightarrow P(X \cap Y) = P(X|Y)P(Y)$$

Conditional & Independent

- If X and Y are independent, then neither influences the outcome of the other



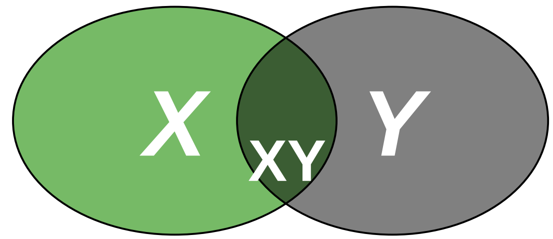
then, $P(X|Y) = P(X)$

and $P(X \cap Y) = P(X|Y)P(Y) = P(X)P(Y)$

What about…..

- So, back to the Union of X and Y,

If X and Y are Independent, then



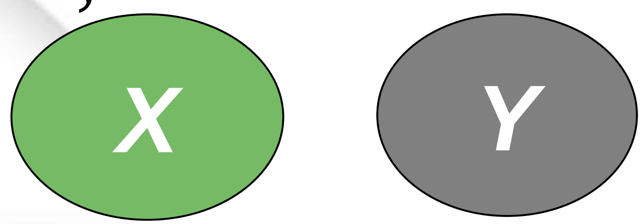
$$P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)$$

$$P(X \cup Y) = P(X) + P(Y) - P(X|Y)P(Y)$$

$$P(X \cup Y) = P(X) + P(Y) - P(X)P(Y)$$

When X and Y are Mutually Exclusive,

$$P(X \cup Y) = P(X) + P(Y)$$



Discrete vs Continuous

- Discrete: when can *count* stuff
  - We can be exact

$$P(X) = P(X = x)$$

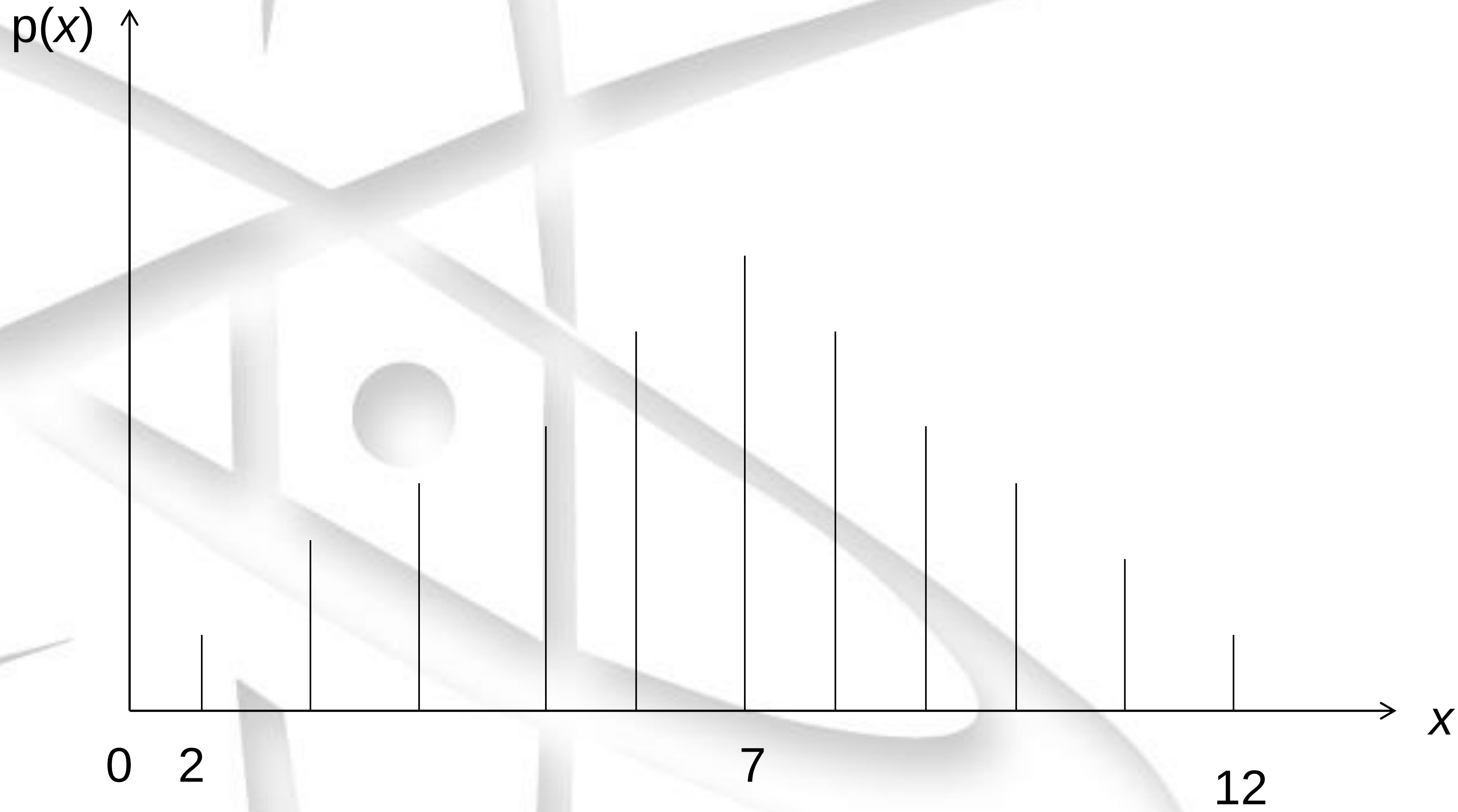
- Continuous: used for measuring (other) stuff
  - We're using some tool (clock, meter, etc.)
    - can *never* be exact
  - We thus need to specify a tolerance or confidence

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

'Formal' Probability

- Discrete: use a probability mass function (pmf)
  - We're interested in the *relative frequency* of each event
  - Example: our dice toss
    - can roll a '7' six of 36 ways -- hence p(7) = 1/6
    - can roll a '2' only 1 way -- hence p(2) = 1/36
    - what about rolling a '13'?
  - What's the big assumption?
    - The dice are *fair* -- each face equally likely to appear on each die
  - Since the sample space is discrete, we can assign a probability to every outcome

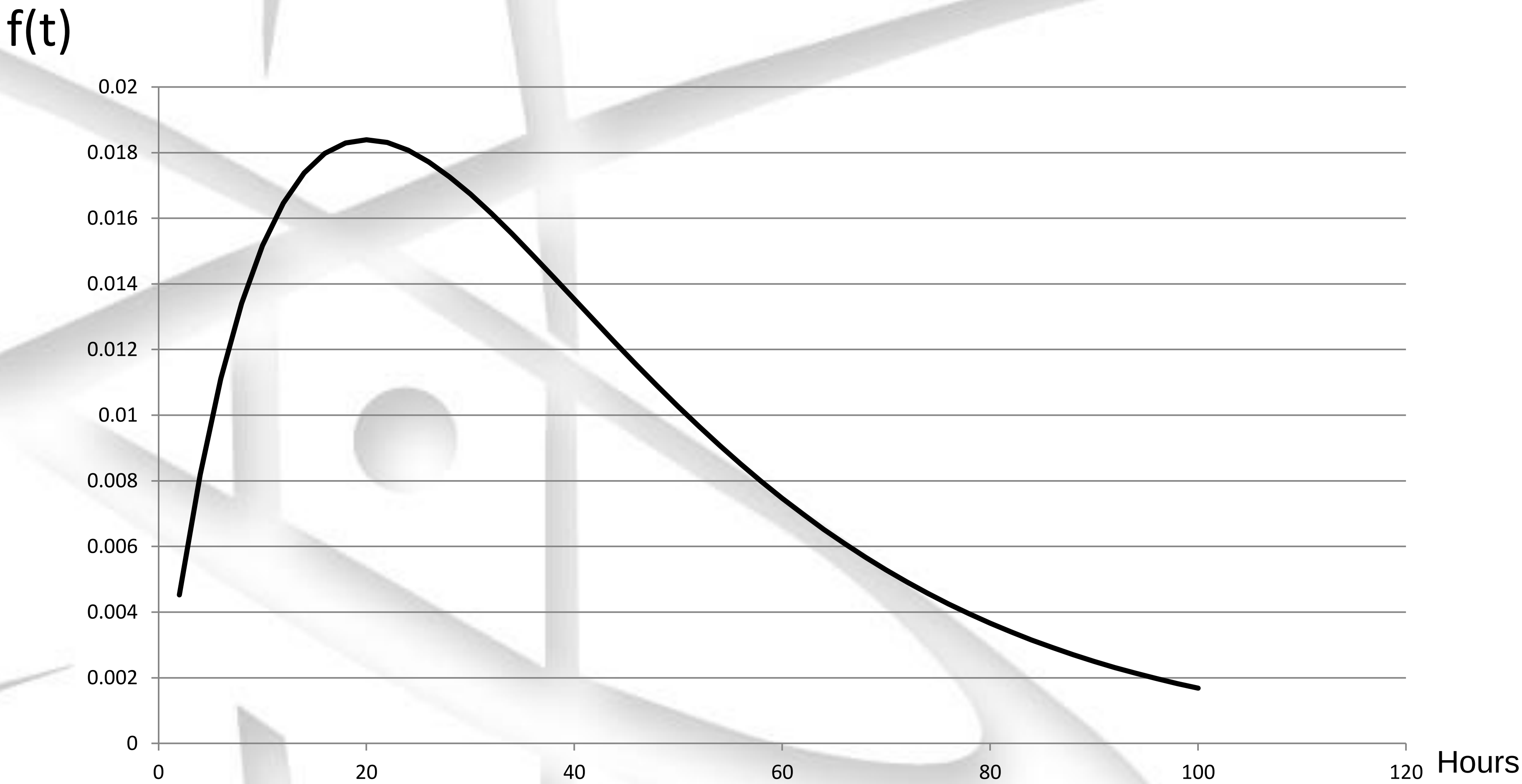
A pmf
(Our dice toss)



Continuous Data

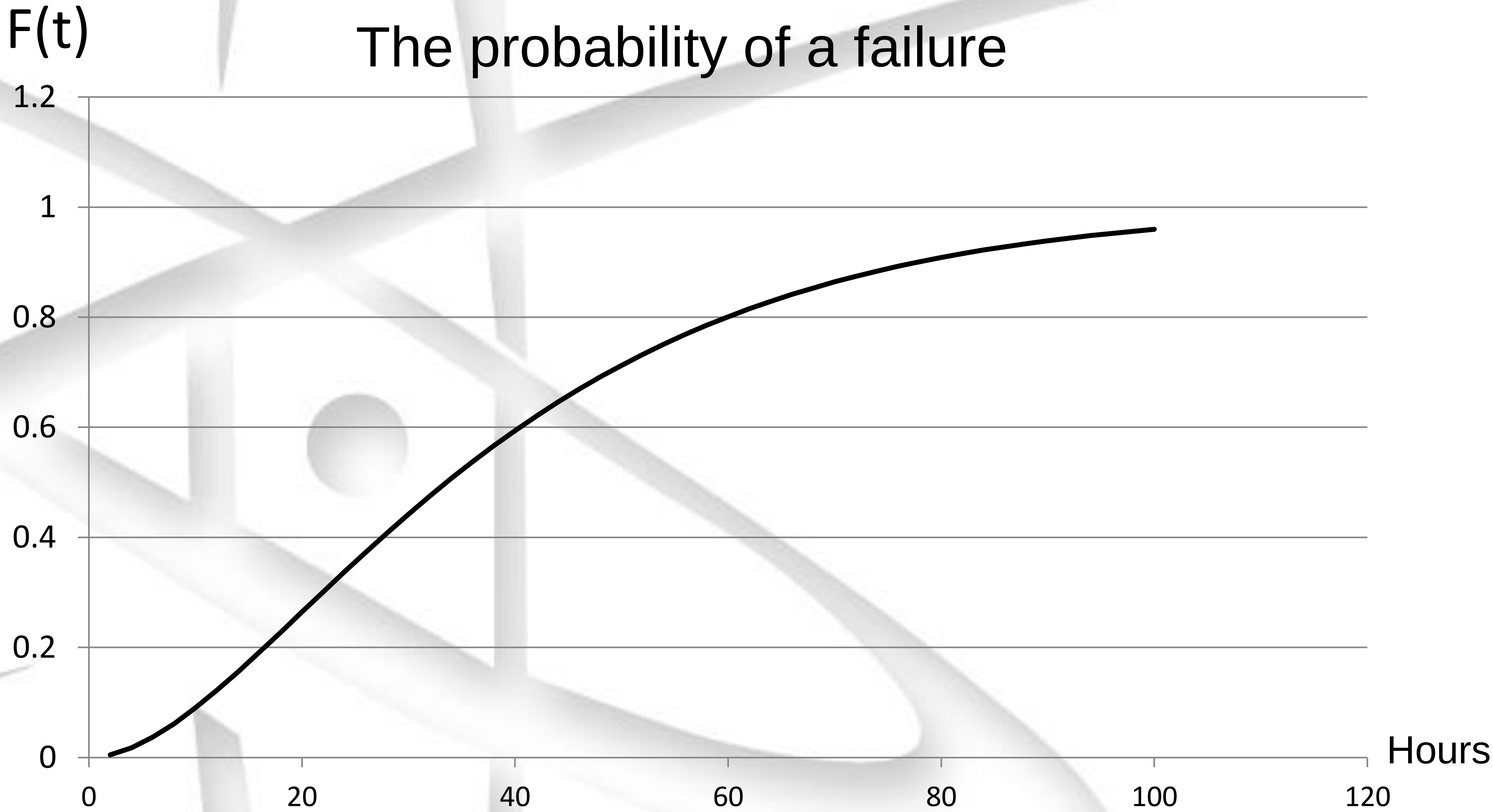
- We use a probability density function (pdf)
- Can not assign a probability to an exact value -- can only assign probabilities to ranges of values
  - Calculus alert! (the pdf is a *derivative*)
- The shape of the pdf curve shows us what values of a *random variable* are most likely to occur
  - I.e., values of x occur most frequently where $f(x)$ is largest

A pdf



The amount of time until a component fails

Graph of a CDF



The cdf

- The *cumulative distribution function* is based on either the pmf or pdf
  - depends on the data type
  - Gives us the probability of a random variable X taking a value equal to, or less than, some specified value *x*

$$F(x) = P\left(X \le x\right) \quad \text{for } -\infty < x < \infty$$

cdf Example

Let X be the number I'll roll on my next toss of two dice.

The probability I'll roll a 7 or less is

$$P\left(X \le 7\right) = \sum_{k=0}^{7} p(x) = 21/36$$

It's easy, in the discrete case -- we just sum up all the probabilities less than or equal to the value of interest.

Example

Let *X* be the amount of time to component failure

The probability the component will fail in 225 hours is written as

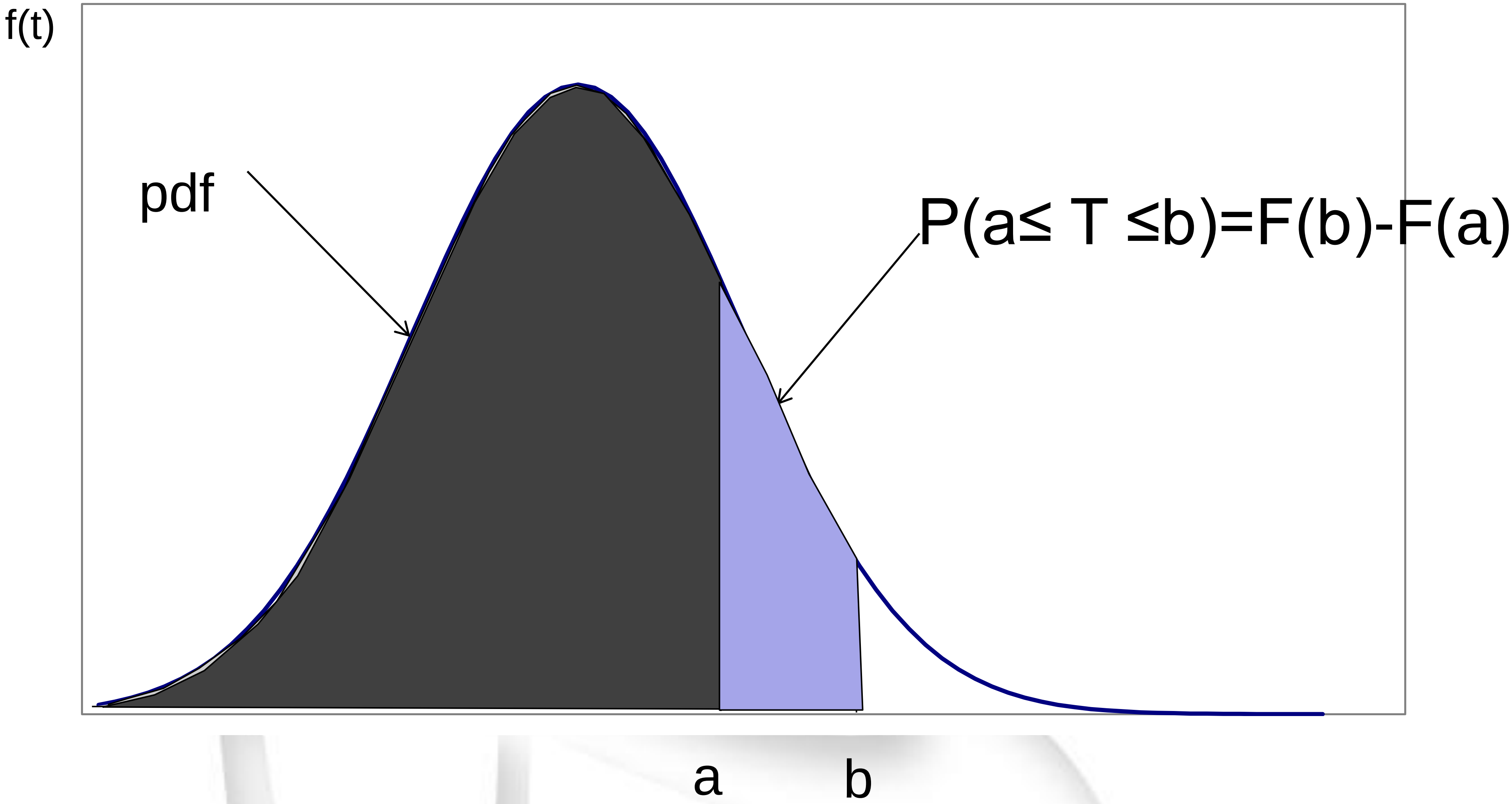
$$P\left(T \leq 225\ hrs\right)$$

Since time is a continuous-valued variable, I would really need to *integrate* some pdf over the interval (0, 225) to compute this probability!

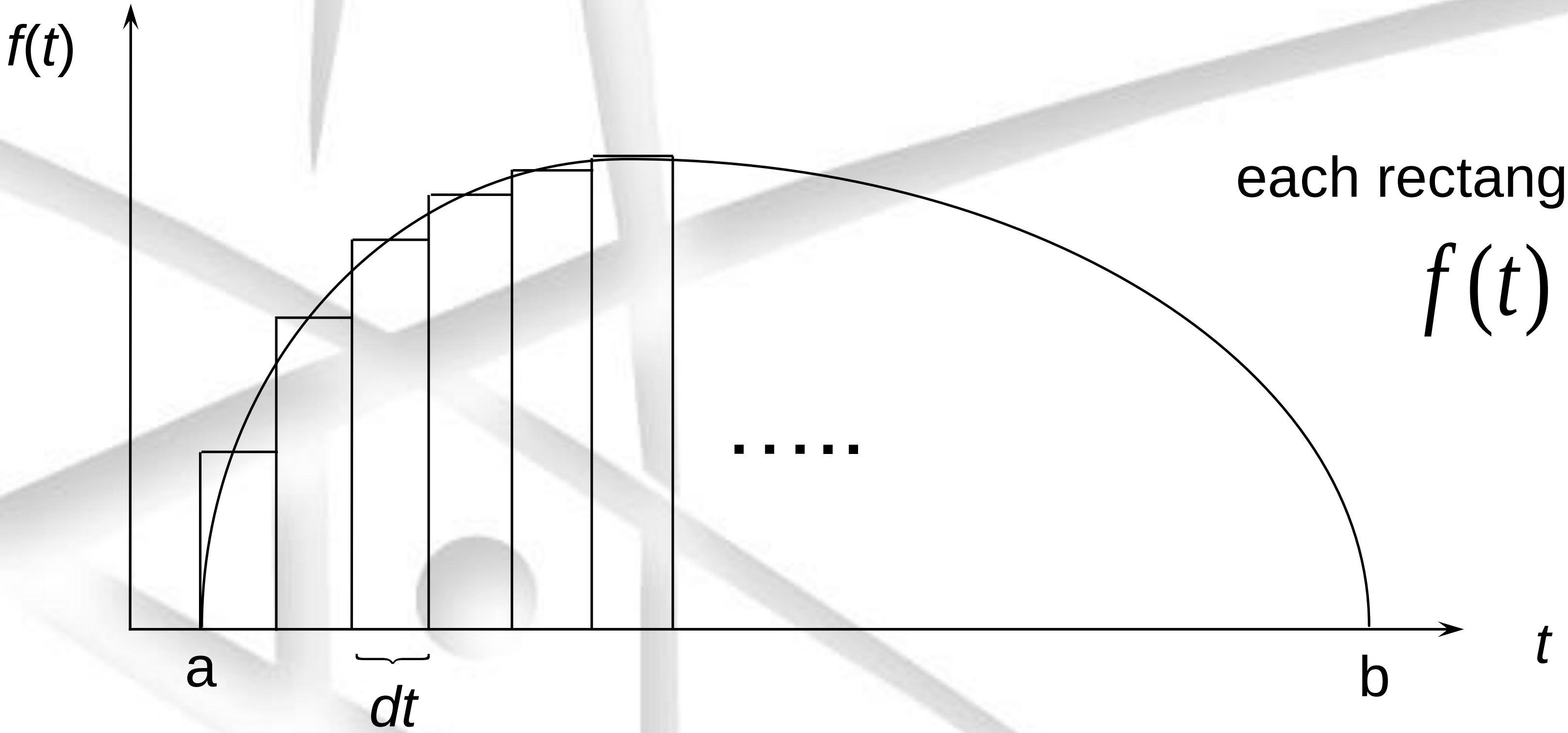
Integrals??!

- An integral is a sum of infinitely small things
- Its usually thought of as the area under a curve
- In our case:
  - The *pdf* (a function) defines the shape of the curve
  - A *cdf* value, $F(t)$, represents the *cumulative* area under the pdf curve within the interval $[-\infty, t]$
  - The *total* area under the pdf curve always equals 1
  - To get the area within some interval [a,b], an integral adds all possible pdf values, each evaluated for a small interval *dt* (for values of *t* within [a,b])

pdf and cdf



Integration



each rectangle has area

$$f(t)\times(dt)$$

An integral as the limit of a sum of rectangles → *dt* wide by *f*(*t*) tall.
As *dt* becomes smaller, the rectangles better approximate the true area.

$$\int_a^b f(t)\,dt \approx \text{sum }\left(\sum\ \right)\text{ of the rectangles over the interval }[\text{a,b}]$$

The Basic Integral

We're integrating with respect to a variable called "*x*".
We evaluate the integral for values of *x* within [a,b]

$$\int_a^b \underbrace{\qquad\qquad}\, dx$$

Indicates the variable involved (integrate wrt)

Integrand: The stuff we insert here—usually a function

- It is the pdf → $f(x)$
- Or, the expectation formula → $x f(x)$
- If we insert nothing, then we assume the integrand is a *constant*, and equal to 1.0

Integral Examples

- The (continuous) cdf is an integral value

$$P\left(T \le a\right) = F(a) = \int_0^a f(t)\,dt$$

- The probability of failure over a specific time interval [a,b]:

$$\begin{aligned} P\left(a \le T \le b\right) &= F(b) - F(a) \\ &= \int_0^b f(t)\,dt - \int_0^a f(t)\,dt \\ &= \int_a^b f(t)\,dt \end{aligned}$$

Some Useful Integrals

$$\int k\,dx \rightarrow kx$$

$$\int kf(x)\,dx \rightarrow k\int f(x)\,dx$$

$$\int \left[f(x) + g(x) \right] dx \rightarrow \int f(x)\,dx + \int g(x)\,dx$$

$$\int x^n\,dx = \frac{1}{n+1} x^{n+1} \rightarrow \int x^3\,dx = \frac{1}{4} x^4$$

$$\int x^{-1}\,dx = \int \frac{1}{x}\,dx \rightarrow \ln x$$

Some More Useful Integrals

$$\int e^{ax}dx = \frac{1}{a}e^{ax} \quad \rightarrow \quad \int e^{2x}dx = \frac{1}{2}e^{2x}$$

$$\int be^{ax}dx = \frac{b}{a}e^{ax} \quad \rightarrow \quad \int 3e^{\lambda t}dt = \frac{3}{\lambda}e^{\lambda t}$$

$$\ln e = 1$$

$$\ln e^{k} = k$$

- The mean is an *expectation*
  - Each possible value, *t*, weighted by its "rectangle area"
  - We use the cdf integrand, but now multiply the integrand by *t*
  - We integrate over all possible values of *t* (from *t* = 0 to ∞)

$$E(T) = \int_0^\infty t\, f(t)\, dt$$

$$MTTF = E(T) = \int_0^\infty R(t)\, dt$$

(A form you'll see soon)

Compare with Discrete Variables

- To help see what the integral is doing, recall the average value (expectation) of a discrete RV:

$$E(X) = \sum_{all\, x} x\, p(x)$$

- In words: the average value of a discrete RV is simply the sum of all possible values of *x*, each weighted by its probability of occurrence *p*(*x*)
- The integral is doing the same thing, except we must deal with the “billions and billions” (infinitely many) *t* values

Variance

- The 'spread' or dispersion about the mean

$$Var(X) = E\left[(X - \mu)^2\right]$$

$$Var(X) = \sum_{all\ x} (x - \mu)^2\, p(x)$$ (discrete)

$$Var(X) = \int_{0}^{\infty} (x - \mu)^2\, f(x)\,dx$$ (continuous)

Variance—An easier way

- The 'spread' or dispersion about the mean

$$Var(X) = E\left[\left(X-\mu\right)^2\right] = E\left[X^2\right] - \mu^2 = E\left[X^2\right] - E\left[X\right]^2$$

- Note: To find the second moment

$$E\left[X^2\right] = \sum_{all\ x} x^2\ p(x) \qquad \text{(discrete)}$$

$$= \int_0^\infty x^2 f\left(x\right) dx \qquad \text{(continuous)}$$

Uniform Distribution, U(a, b)

- Useful when occurrences randomly vary and little else is known about the shape of the distribution

$$pdf : f(t) = \begin{cases} \dfrac{1}{b-a} & \text{if a} \leq t \leq \text{b} \\ 0 & \text{otherwise} \end{cases}$$

Normal Distribution, N(μ, σ²)

- Used for errors of various types

pdf: $f(t) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-(t-\mu)^2 / (2\sigma^2)}$ for $-\infty < t < \infty$

cdf: no closed form $\rightarrow$ use look-up tables

Lognormal Distribution, LN(μ, σ²)

- Time to perform some task

$$\text{pdf:}\quad f(t)=\begin{cases}\left(\dfrac{1}{t\sigma\sqrt{2\pi}}\right)\mathrm{e}^{-(\ln t-\mu)^2/\left(2\sigma^2\right)} & \text{if t} > 0\\ 0 & \text{otherwise}\end{cases}$$

cdf: no closed form $\rightarrow$ use look-up tables

Weibull Distribution, Weibull(β, θ)

- Time to complete some task, time to failure

$$\text{pdf:}\quad f(t) = \begin{cases} \dfrac{\beta}{\theta}\left(\dfrac{t}{\theta}\right)^{\beta\text{ -1}} e^{-(t/\theta)^{\beta}} & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{cdf:}\quad F(t) = \begin{cases} 1 - e^{-(t/\theta)^{\beta}} & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$$

Exponential Distribution Expon(β)

- Constant failure rates, 1/β

$$\text{pdf:}\quad f(t) = \begin{cases} \dfrac{1}{\beta} e^{-t/\beta} & \text{if } \ t \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{cdf:}\quad F(t) = \begin{cases} 1 - e^{-t/\beta} & \text{if } \ t \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

Binomial Distribution, bin(n, p)

- Discrete distribution representing # of successes in *n* independent Bernoulli trials where the probability of success for each trial is p
- The binomial (a set of Bernoulli trials)
  - Probability I'll get exactly 3 '7's on my next 5 tosses of two dice
  - Probability of exactly 1 failure among 5 identical components

$$\text{pmf: } p(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \text{for } x=0,1,\ldots,n$$

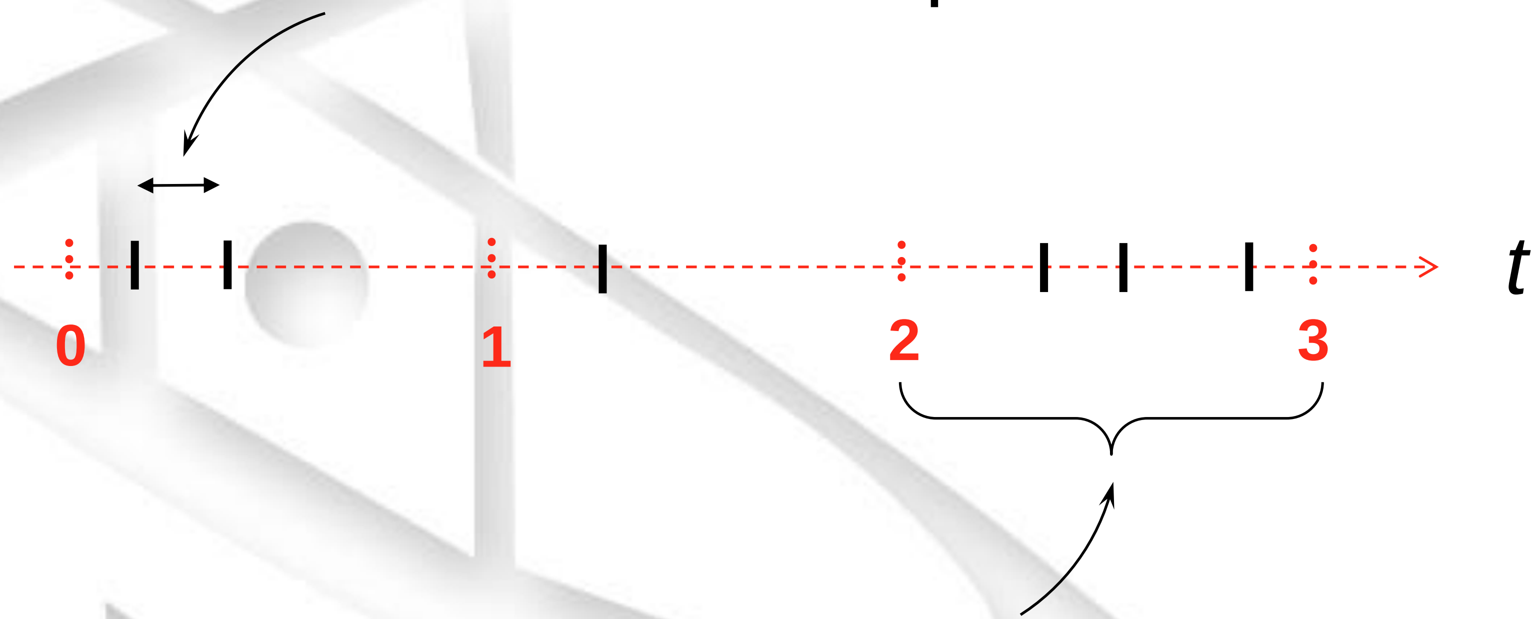
$$\text{where, } \binom{n}{x} = \frac{n!}{(n-x)!\,x!} \qquad (\text{binomial coefficient})$$

A Poisson Process

- A unique relationship between the Poisson pmf and the exponential pdf…
- It's WHEN the time between successive occurrences is exponentially distributed, AND
  - Failure rate = λ
- The number of occurrences in a given time, t, is Poisson distributed
  - Expected # failures over time t = λt

Poisson Process – the idea…

Time between successive occurrences is exponentially distributed with parameter λ



Number of occurrences in a specified time interval t is Poisson distributed with parameter , λt

Poisson Process Assumptions

- Things happen *serially* (no stinkin' batches)
- *Stationarity* (parameters are constant w.r.t. time)
- *Independent* increments:
  - What happens in one time period doesn't affect what happens in a different, non-overlapping period

Poisson Distribution

- The Poisson is a discrete distribution representing the number of random events that occur in an interval of time, when events occur at a constant rate
  - Probability I'll get exactly 5 emails in the next hour, given they average about λ per hour

$$\text{pmf:} \quad p_n(t) = \frac{e^{-\lambda t}\left(\lambda t\right)^n}{n!}$$

Other Useful Things!

$$x^m x^n = x^{(m+n)}$$

$$\frac{x^m}{x^n} = x^{(m-n)}$$

$$\left(x^m\right)^n = x^{mn}$$

$$\left(xy\right)^m = x^m y^m$$

$$\mathrm{Log}_a(x) = y \quad \Longleftrightarrow \quad a^y = x$$

$$\mathrm{Log}_a(a^x) = x$$

$$\mathrm{Log}_a(r^x) = x\,\mathrm{Log}_a(r)$$

$$\mathrm{Log}_a(xy) = \mathrm{Log}_a(x) + \mathrm{Log}_a(y)$$

$$\mathrm{Log}_a(x/y) = \mathrm{Log}_a(x) - \mathrm{Log}_a(y)$$

$$\mathrm{Log}_e(x) = \ln x$$