Topic 1: Probability Distributions in Ecology and Evolution

Learning Objectives:

- Define a trial, and random variables, and describe discrete/continuous probability distributions using PMF/PDF and CDFs
- Use probability rules to draw biological conclusions
- Describe one example of where each of the following distributions arises in ecology or evolution:
  - Discrete probability distributions
    - Poisson distribution
    - Bernoulli distribution
    - Binomial distribution
    - Geometric distribution
  - Continuous probability distributions
    - Normal distribution
    - Beta distribution
    - Exponential distribution
    - Gamma/Erlang distribution
- Define the moments and central moments of a probability distribution and derive the relationships between them. Use these definitions to draw biological conclusions using the distributions listed above.
- Sample randomly from any discrete and/or continuous probability distribution given its CDF.

Lecture 1.1 Probability

Definitions

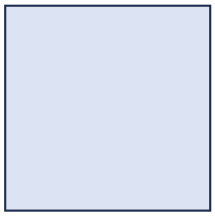
A **trial** is a natural occurrence or event (e.g., birth/death) or an experimental outcome (e.g., time to detection) that can have more than 1 (possibly infinite) outcome. The set of possible outcomes of an experiment is called the **sample space** and is often denoted as S.

Example 1.1: Body mass
Body mass of a black bear. Sample space: $[0kg, 500kg]$. The largest black bear on record was recorded in New Brunswick weighing approximately $500kg$.

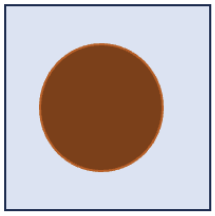
Discussion: What are some other examples of trials in ecology, evolution, and epidemiology? And what are their sample spaces?

We can graphically represent trials and sample spaces with **Venn Diagram**.

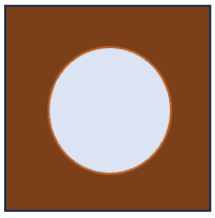
Sample Space, S



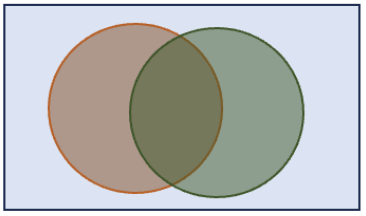
Event/Subset A



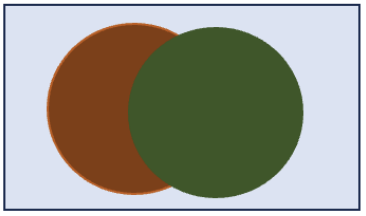
Complement, A^C



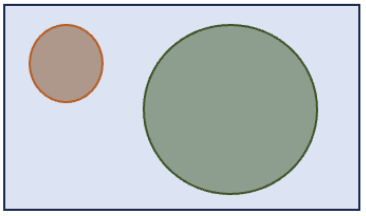
Intersection, $A \cap B$



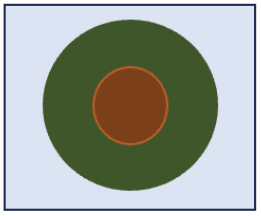
Union, $A \cup B$



Mutually Exclusive



Subset, $A \subset B$



Probability Properties

Probabilities of Mutually Exclusive Outcomes: Suppose a trial can result in one of a set, X, of possible outcomes, x_i. The outcomes are said to be "mutually exclusive" and must satisfy

$$\sum_{i \in X} \Pr(X = x_i) = 1$$

Probabilities of Complements: Consider a subset $A \subset X$ of the sample space. Then:

$$\Pr(X = A) + \Pr(X = A^C) = 1$$

Independent Outcomes: Consider two mutually exclusive subsets $A \subset X$ and $B \subset X$ of the sample space. Then:

$$\Pr(A \cap B) = \Pr(A)\Pr(B)$$
$$\Pr(A \cup B) = \Pr(A) + \Pr(B)$$

Inclusion-Exclusion Rule: Consider two non-mutually exclusive subsets $A \subset X$ and $B \subset X$ of the sample space. Then:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(AB)$$

Example 1.2: Breast cancer

The probability of developing breast cancer for women between the ages of 50-60 is $\Pr(\text{cancer}) = 0.03$

1. **What is the probability that 4 independent patients ALL develop breast cancer?**

$$0.03 \times 0.03 \times 0.03 \times 0.03 = 0.03^4 = 8.1 \times 10^{-7}$$

2. **What is the probability that ANY of them develop breast cancer?**
   patient A OR patent B OR patent C Or Patient D

$$0.03 + 0.03 + 0.03 + 0.03 = 0.12 = 12\%$$

3. **What is the probability that ONLY one of them develops breast cancer?**

(patient A AND not patent B AND not patent C AND not Patient D)
OR
(not patient A AND patent B AND not patent C AND not Patient D)
OR
...

$$\underbrace{\binom{4}{1}}_{\#\text{ of options}} * 0.03 * (1 - 0.03) * (1 - 0.03) * (1 - 0.03)$$
$$= 4 * 0.03 * (1 - 0.03)^3 = 0.109$$

Conditional Probability: Consider two events A and B. The probability that A occurs **given** that event B occurred we have:

$$\Pr(A|B) = \frac{\Pr(AB)}{\Pr(B)}$$

If A and B are **independent** then:

$$\Pr(A|B) = \frac{\Pr(AB)}{\Pr(B)} = \frac{\Pr(A)\Pr(B)}{\Pr(B)} = \Pr(A)$$

Total Probability:

$$\Pr(X = x) = \sum_{y \in Y} \Pr(y)\Pr(x|y)$$

Bayes' Theorem: Bayes' theorem gives a relationship between the conditional probabilities $\Pr(A|B)$ and $\Pr(B|A)$

$$\underbrace{\Pr(A|B)}_{\text{posterior}} = \frac{\overbrace{\Pr(B|A)\Pr(A)}^{\text{likelihood}\quad\text{prior}}}{\underbrace{\Pr(B)}_{\text{marginal prob.}}}$$

where A refers to the "model" and B refers to the "data"

Example: 1.3 Brest cancer cont.

The brca1 gene has two different variants, called alleles, individuals with the "mutant" allele have an increased probability of developing breast cancer (the non-mutant allele is known as the "wild-type"). The probability of having the brca1 mutant allele is 0.02. Suppose that carrying this allele doubles your chance of having breast cancer. From above, we have that the probability of developing breast cancer is $\Pr(\text{cancer}) = 0.03$ (including both individuals with and without the brca1 allele).

1. **What is the probability of developing cancer given that you carry the brca1 allele?**

What do we know?

$$\Pr(\text{mutant}) = \Pr(M) = 0.02$$
$$\Pr(\text{cancer}) = \Pr(C) = 0.03$$
$$\Pr(C|M) = 2\Pr(C|M^C)$$

where M^C is the "wild-type".

What do we want to know?

$\Pr(C|M)$

Let's use the total probability:

$$\Pr(C) = \Pr(M)\underbrace{\Pr(C|M)}_{x} + \Pr(M^C)\underbrace{\Pr(C|M^C)}_{\frac{1}{2}x}$$

Solving for x:

$$\Pr(C|M) = \frac{2\Pr(C)}{2\Pr(M) + \Pr(M^C)} = \frac{2 \times 0.03}{2 \times 0.02 + (1 - 0.02)} = 0.059$$

So 2% of the population with the mutant allele has a 5.9% chance of developing cancer and 98% of the population with the wild type allele has a 2.9% chance of developing cancer such that the population-wide probability of developing cancer is 3%.

2. **A patient has breast cancer, what is the probability that that patient carries the brca1 mutant allele?**

What probability are we looking for?

$$\Pr(M|C)$$

This is other conditional probability than what we have, so let's use Bayes' theorem:

$$\Pr(M|C) = \frac{\Pr(C|M)\Pr(M)}{\Pr(C)} = \frac{0.059 \times 0.02}{0.03} = 0.039$$

Random Variables

A **random variable**, X, is a real-valued function defined on the sample space, S

$$X : S \to \mathbb{R} = (-\infty, \infty)$$

Random variables assign a numerical value to the outcome of a trial providing an ordering to the outcomes. If the range of the random variable is finite or countably infinite the r.v. is **discrete**. If the range is an interval (finite or infinite) the r.v. is **continuous**.

Example: 1.4 fly survival

In an evolution experiment survival of a mutant fly is measured on day 5.

1. **What are the possible outcomes of this trial?**

"survive" and "die"

2. **What would be a reasonable r.v. for this case?**

$$X = \begin{cases} 0 & \text{die} \\ 1 & \text{survive} \end{cases}$$

3. **Is this a discrete or continuous r.v.?**

discrete

Example 1.5: DBH

Tree growth is often measured by the trunk "diameter at breast height" (DBH)

1. **What are the possible outcomes of this trial?**

min diameter $= 0$ and max diameter $\approx 11m$ (from a giant sequoia)

2. **What would be a reasonable r.v. for this case?**

$$X = [0, 11]$$

3. **Is this a discrete or continuous r.v.?**

continuous

Let $s \in S$ be an outcome of the experiment. We will use the shorthand $X = x$ to denote the event $\{s : X(s) = x, s \in S\}$. In other words, we will refer to an event by its value as a random variable. Given that the r.v. also gives an ordering to the events we will also use the shorthand: $X \leq x$ to denote the set of events $s : X(s) \leq x, s \in S$

Example 1.6: Three-spine Stickleback Armour

Three-spine Sticklebacks are small fish that live in lakes throughout British Columbia and the Pacific coast. A key trait of these fish is the amount of skeletal armour they have. Skeletal armour is controlled by the *Eda* gene. If an individual has two copies of the 'C' allele at this locus and hence is heavily armoured. If they have one 'C' and one 'l' allele or two 'l' alleles they are lightly armoured.

1. **Define a r.v. describing the genetics of this system**
   Let's define the r.v. by counting the number of 'C' alleles

$$X = \begin{cases} 0 & \text{ll} \\ 1 & \text{Cl} \\ 2 & \text{CC} \end{cases}$$

2. **Express the probability that an individual is heavily armoured?**

$$\Pr(X = 2)$$

3. **Express the probability that an individual is lightly armoured?**

$$\Pr(X \leq 1)$$

Lecture 1.2 Discrete Distributions

Definitions

The **cumulative distribution function** (cdf) of a random variable X is a function $F : \mathbb{R} \to [0, 1]$ defined by:

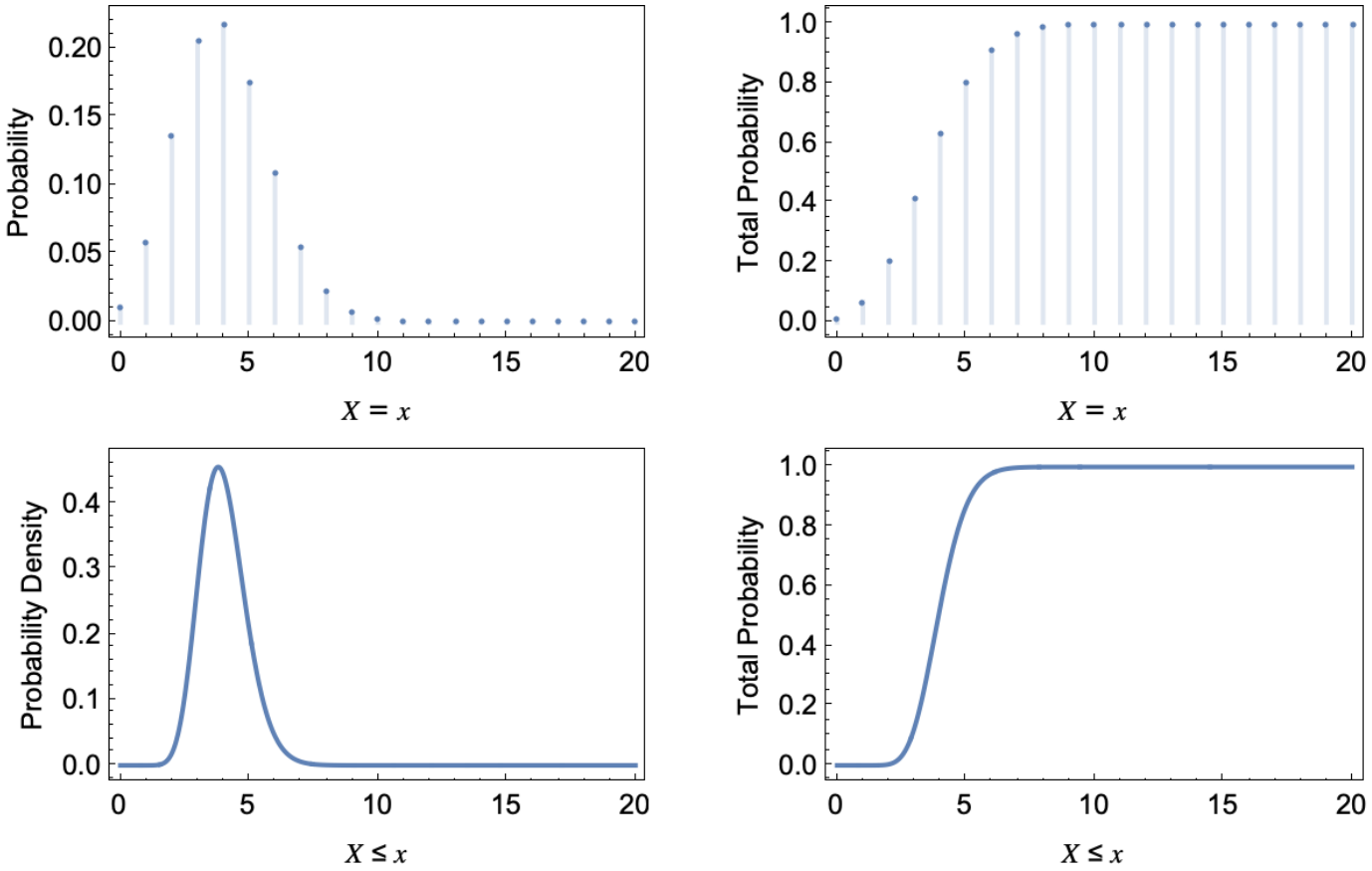
$$F(x) = Pr\left((-\infty, x]\right)$$

Suppose X is a discrete r.v., then the function $f(x) = \Pr(X = x)$ is called the **probability mass function** (pmf). Note that the range of $f(x)$ is $[0, 1]$

Suppose X is a continuous r.v. with cdf F. Then the function $f : \mathbb{R} \to [0, \infty)$ such that:

$$F(x) = \int_{-\infty}^{x} f(y)dy$$

then the function $f(x)$ is known as the **probability density function** (pdf) of X



Joint Probability Distributions

Some trials result in a pair of outcomes, (X, Y). In such cases, the trial is described by the **joint probability distribution** of X and Y.

Discussion: What are examples of biological trials with paired outcomes?

Useful Probability Distributions

Each experiment technically has its own unique sample space and corresponding probability distribution. Many of these "empirical distributions" are very similar and it can be useful to approximate them with one of a set of standard distributions with known properties. Below are some such distributions, as such each of these distributions is described by an idealized experiment. Wikipedia is a useful resource for probability distributions, we will discuss distributions used throughout the rest of this course and some of their most useful properties

1. The Poisson Distribution $X \sim \mathcal{P}(\lambda)$

Motivation: If events occur randomly at a constant rate λ, the probability that x events occur in one unit of time is Poisson distributed.

Domain: $X \in \mathbb{N}$

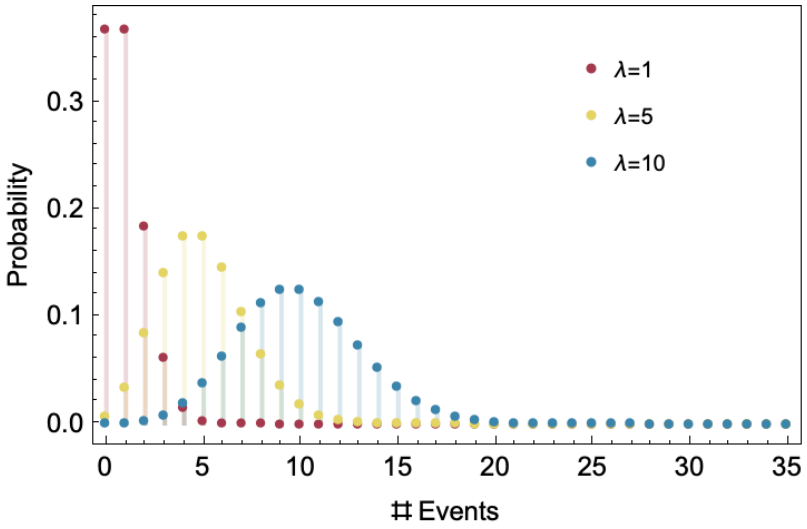
Parameters: $\lambda \in \mathbb{R}^+$ Note: Rates must be greater than 0

PMF: $\Pr(x) = \frac{\lambda^x e^{-\lambda}}{x!}$

The Poisson distribution has the unique property that its mean equals its variance:

Mean: λ

Variance: λ



2. The Bernoulli Distribution $X \sim \mathcal{B}er(p)$

Motivation: Consider a binary trial with sample space of a trial is $X = \{0, 1\}$ where the probability of a **success** ($x = 1$) is p, then the outcome of the trial is Bernoulli Distributed

Domain: $X \in \{0, 1\}$

Parameters: $p \in [0, 1]$

PMF:

Discussion: What do you think the PME of a Bernoulli distribution with $p = 0.25$ looks like?

$$\Pr(x) = \begin{cases} 1 - p & \text{if } x = 0 \\ p & \text{if } x = 1 \end{cases}$$

Mean: p

Variance: $p(1 - p) = pq$

Example: 1.7 Tongue Rolling

Tongue rolling is a dominant trait determined by a single gene with two alleles, (0: no roll, 1: roll).

We often denote the two alleles as 'A' and 'a'. By being dominant we have:

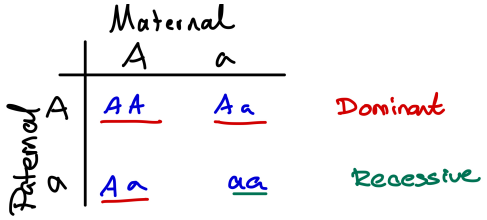
AA: Roll, Aa: Roll, aa=Flat

1. **Suppose your father is heterozygous for tongue rolling, what is the probability that you inherit the dominant allele?**

By the principle of *random segregation* which of your father's alleles you inherit is random. So $p = 0.5$.

2. **Suppose both your parents are heterozygous for this gene. What is the probability that you can roll your tongue?**

To understand this we have to build what is known as Punnett Square.



$p = 0.75$

---

3. The Binomial Distribution $X \sim \mathcal{B}(n, p)$

Motivation: The number of successes, X, among n Bernoulli trials with success probability p is binomially distributed

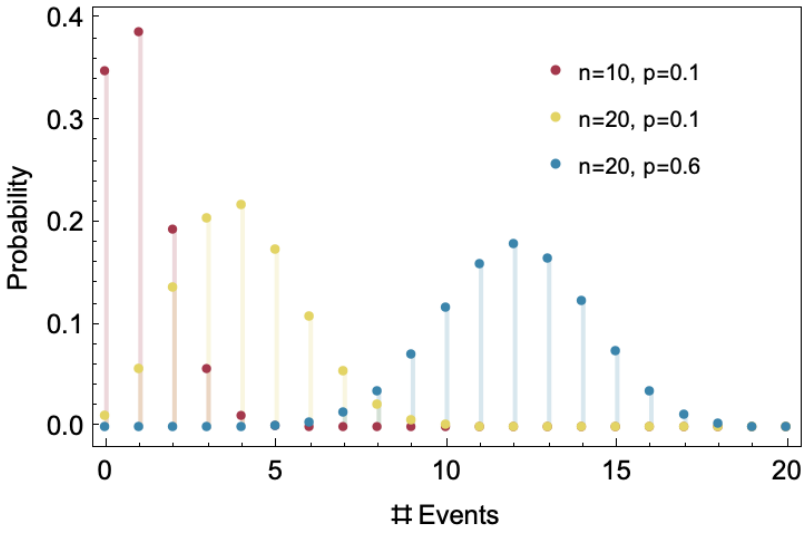
Domain: $X \in \{0, 1, \ldots, n\}$

Parameters: $p \in [0, 1]$

PMF: $\Pr(x) = \binom{n}{x} p^x (1-p)^{n-x}$

Mean: np

Variance $np(1-p) = npq$



Example: 1.8 Mendel's Peas

Peas have 4 chromosomes (this is why they were so good for Mendel to study). Mendel studied 1 trait per chromosome (wrinkly/smooth, pink/white, etc. The principle of independent assortment states that whether you pass on your mom's or your dad's chromosome is independent between chromosomes.

1. **What is the probability that a pea plant passes on all of its mom's chromosomes to its offspring (note: we are ignoring recombination here)?**

The probability of passing on any one of the mom's chromosomes is $p = 0.5$. Hence the number of material chromosomes inherited, X, is Binomially distributed with $n = 4$ and $p = 0.5$

$$\Pr(X = 4) = \binom{4}{4} 0.5^4 (1 - 0.5)^{4-4} = 0.0625$$

2. **What is the probability that YOU pass on x of your mom's chromosomes?**
   Humans have 23 chromosomes

$$\Pr(X = x) = \binom{23}{x} 0.5^x (1 - 0.5)^{23-x}$$

Lecture 1.3 Continuous Distributions

Useful Probability Distributions Cont.

1. Normal Distribution $X \sim \mathcal{N}(\mu, \sigma)$

Motivation: When there are many sources of errors in an experiment with a continuous outcome X, the observed value x is normally distributed around its expected value μ with some variance σ^2 resulting from these cumulative effects.

Domain: $X \in \mathbb{R}$

Parameters: $\mu \in \mathbb{R}\backslash\text{quad } \sigma \in \mathbb{R}^+$

PDF: $\Pr(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

CDF $\Pr(X \leq x) = \frac{1}{2}\text{Erfc}\left(\frac{\mu - x}{\sqrt{2}\sigma}\right)$

Mean: μ

Variance σ^2

The **Standard Normal Distribution** is a special case where $\mu = 0$ and $\sigma = 1$.

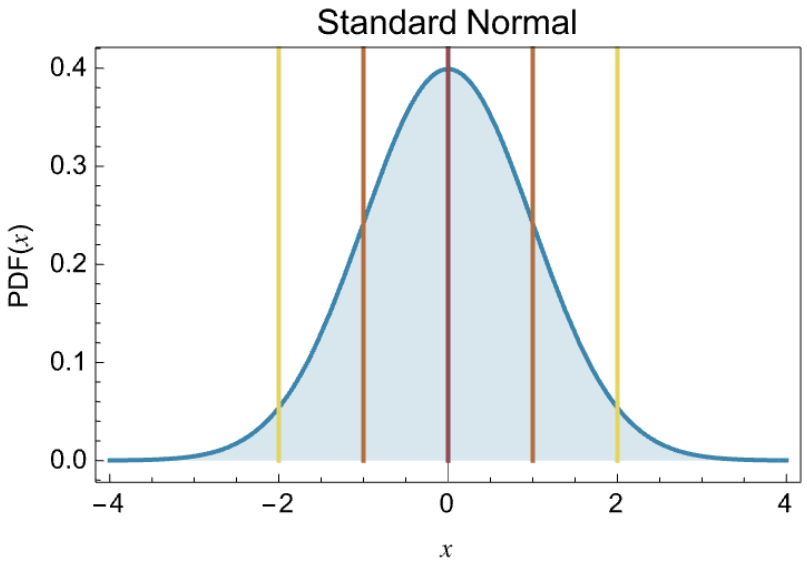


Fig: Red shows mean, orange $\pm 1\text{SD}$ and yellow $\pm 2\text{SD}$. Note that $\approx 95\%$ of the probability area falls in the $\pm 2\text{SD}$ window.

2. Exponential Distribution $X \sim \mathcal{E}xp(\lambda)$

Motivation: If events occur at a constant rate λ, the time until the next event occurs, X, is exponentially distributed.

Domain: $x \in \mathbb{R}^+$

Parmaters: $\lambda \in \mathbb{R}^+$

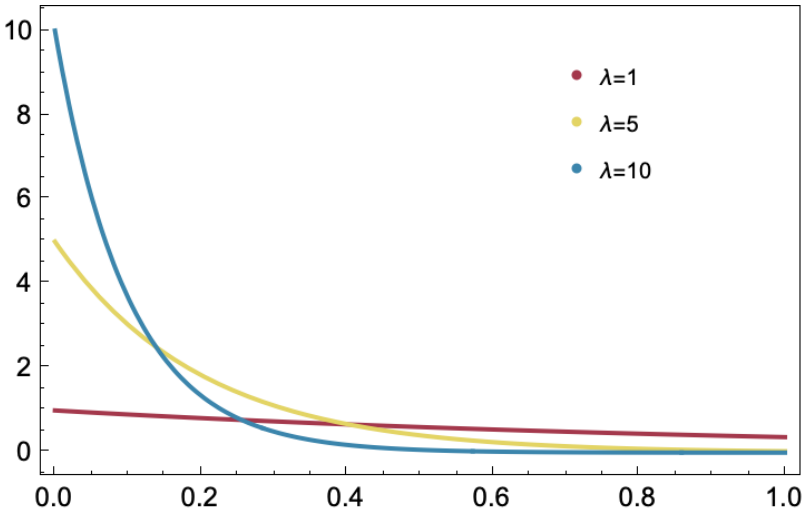
PDF: $Pr(x) = \lambda e^{-\lambda x}$

CDF: $Pr(X \leq x) = 1 - e^{-\lambda x}$

Mean: $\frac{1}{\lambda}$

Variance: $\frac{1}{\lambda^2}$

Skew 2 (the exponential distribution is **right skewed** meaning that it has a long right tail)



The **heterogeneous exponential distribution** extends this to the case with a time-varying rate, $\lambda(t)$. Its pdf is given by:

$$Pr(x) = \lambda(x)e^{-\int_0^x \lambda(t)dt}$$

Example 1.9: Senescence

Not all organisms senesce, for example, brewer's yeast. Yeast cells die at an approximately constant rate throughout their lives. If a yeast cell dies at a rate $\lambda = \frac{1}{3}\frac{1}{\text{days}}$.

1. **What is the expected lifespan of a yeast cell?**

Life expectancy=mean waiting time to death=$\frac{1}{\lambda} = 3 \text{ days}$

2. **How long would a yeast cell have to live to be considered "statistically" old?**

How long does it take for 95% of yeast cells to die? In other words, we want:

$$\Pr(X \leq x) = 0.95$$

using the CDF we have:

$$1 - e^{-\lambda x} = 0.95 \Rightarrow x = 8.99 \text{ days}$$

---

3. Erlang Distribution

Motivation: If events occur at a constant rate λ the waiting time, X, until the k^{th} event is Erlang distributed.

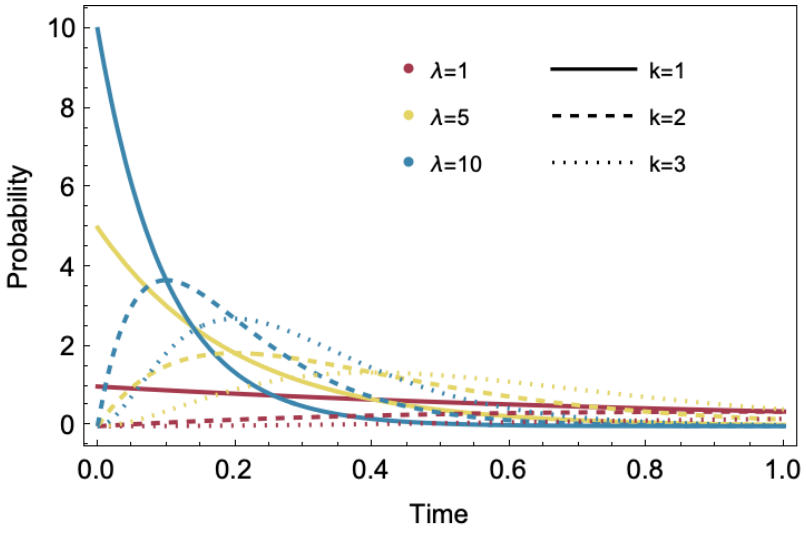
Domain: $X \in \mathbb{R}^+$

Parmaters: $\lambda \in \mathbb{R}^+ \backslash \text{quad } k \in \mathbb{Z}^+$

PDF: $Pr(x) = \frac{\lambda^k x^{k-1} e^{\lambda(-x)}}{k!}$

Mean: $\frac{k}{\lambda}$

Variance: $\frac{k}{\lambda^2}$



The **Gamma Distribution** extends this to allow for non-integer k

---

Example 1.10: Metamorphosis

Many organisms have distinct age classes (e.g., Caterpillar, butterfly). Suppose that the average time to metamorphosis for a tadpole is 14 weeks.

1. **Modelling the time to metamorphosis with an exponential distribution, how long does it take for 10%, 50%, 75%, and 95% of tadpoles to mature?**

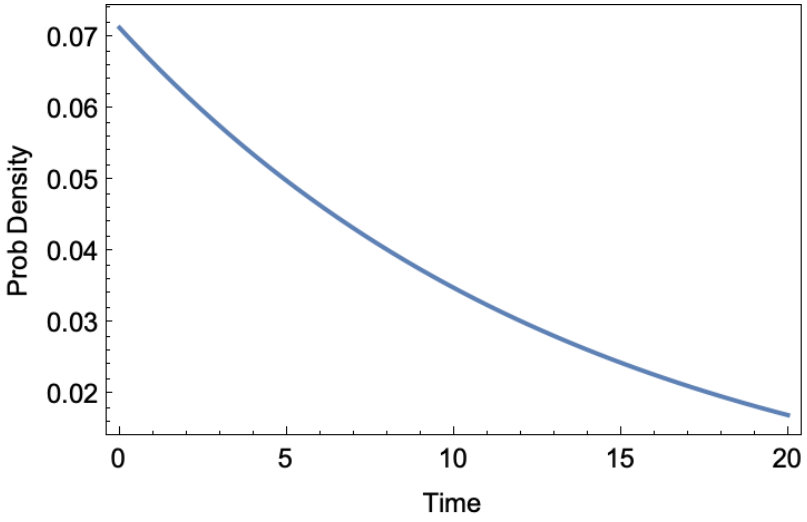
What is the value of λ?

$$\lambda = \frac{1}{14} = 0.071$$

$$CDF(x) = 1 - e^{-\lambda x} = Y \quad Y = 0.1, 0.5, 0.75, 0.95$$

Solving we have $x = \frac{\ln(1-Y)}{-\lambda}$

2. **Draw the distribution of times to maturity. What is the most likely maturation time (e.g., what is the mode of the distribution)?**



Discussion: Does this make sense?

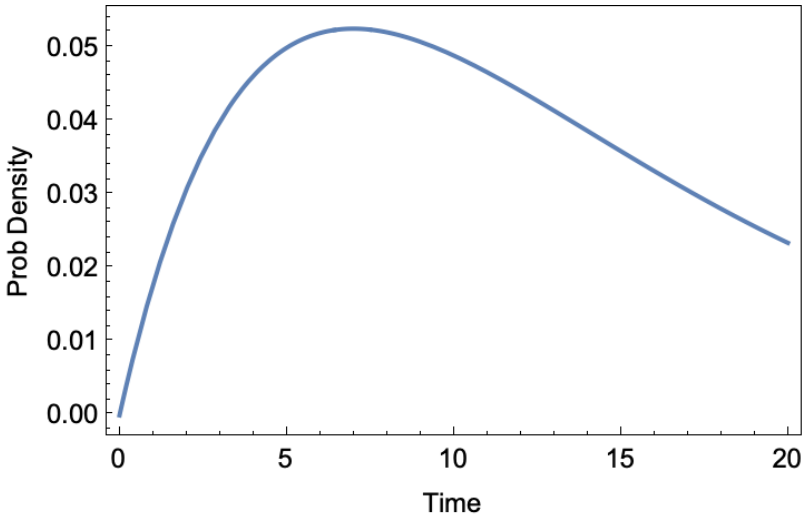
3. **Now model the time to maturation by including a hidden event (e.g., an Erlang distribution with $k = 2$). How long does it take 10%, 50%, 75%, and 95% of tadpoles to mature?**

What is the value of λ?

$$\lambda = \frac{2}{14} = 0.14$$

you have to go twice as fast to make it through 2 events in the same amount of time.

4. **Draw this distribution. What is the mode?**



Mode ≈ 7: most common time to metamorphosis.

Discussion: Does this make sense?

---

4. Uniform Distribution $X \sim \mathcal{U}(a, b)$

Motivation: If the outcome of an event is random and bounded between a minimum a and maximum b value, this outcome is uniformly distributed

Domain: $X \in [a, b]$

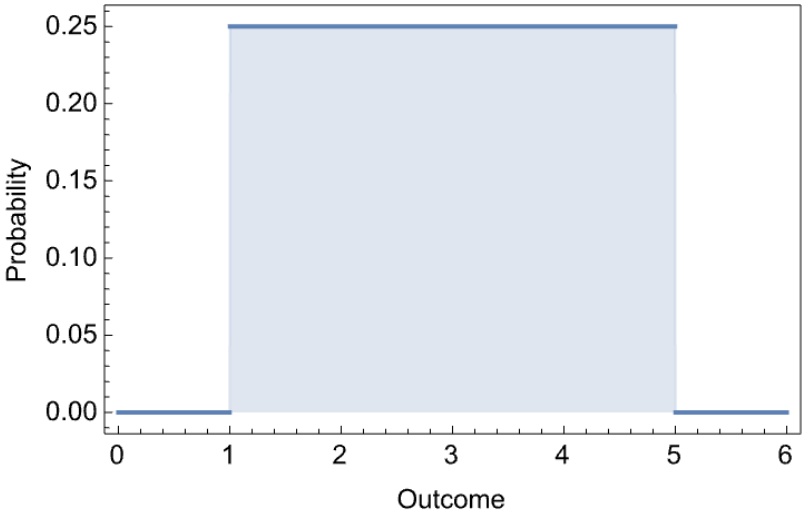
Parameters: $a \in (-\infty, b]$

PDF: $Pr(x) = \frac{1}{b-a}$

CDF: $Pr(X \leq x) = \frac{x-a}{b-a}$

Mean: $\frac{a+b}{2}$

Variance: $\frac{(b-a)^2}{12}$

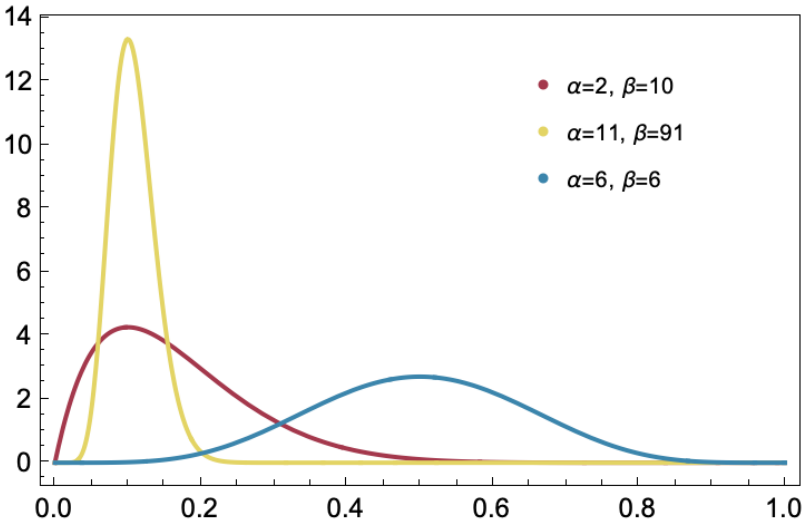


5. Beta Distribution $X \sim \mathcal{B}eta(\alpha, \beta)$

Motivation: Consider n Bernoulli trials in which there were k successes. The probability that the true probability of success was x given the data is Beta distributed with parameters $\alpha = k + 1$ and $\beta = n - \alpha + 2 = n - k + 1$.

Parameters: $\alpha \in \mathbb{N}$ & $\beta \in \mathbb{N}$

Mean: $\frac{\alpha}{\alpha+\beta}$



Discussion: What does each of these distributions mean in terms of the underlying binomial data?

Red: 1 success, 10 trials
Yellow: 10 successes, 100 trials
Blue: 5 successes, 10 trials

Example 1.11: Survival Probability

A researcher is running an experiment on pollution stress on Arctic Char. In two experimental sites (one polluted the other not) each with 30 fish, they found that 5 fish in the polluted site perished over the summer while only 2 fish in the clean site perished.

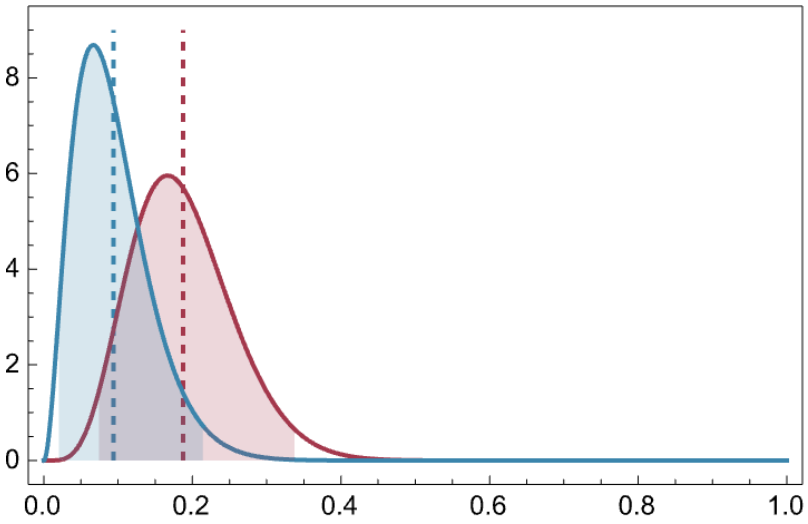
1. **Draw the distribution of mortality probabilities in both sites**

see below

2. **What is the expected mortality probability in each site? Add these points to the plot.**
   $\mu_{po} = \frac{5+1}{30+2} = 0.1875$
   $\mu_{cl} = \frac{2+1}{30+2} = 0.09375$

see below

2. **Draw the 95% Confidence Interval (CI) for the mortality probability in the polluted site.**



Lecture 1.4 Moments

Definitions

The PDF/PMF gives the full description of a distribution. But these functions can be cumbersome or may not have a known form. Alternatively, a distribution can be defined by its **moments**. While to have a full description of the distribution we may need a lot (possibly ∞) moments, like a Taylor Series we can often capture the major features of a distribution with only the first few moments. To define a moment we first have to define an **expected value**.

Suppose X is a continuous random variable with pdf $f(x)$. Then the expectation of X, denoted as $E[X]$, is defined as:

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

For a discrete r.v. X with probability function $\Pr(x)$ we have:

$$E[X] = \sum_{x \in S} x \Pr(x)$$

This definition can be extended to include the expectation of a function of a random variable, $E[g(X)]$

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx$$

Expectation Rules:

1. Constants

$$E[aX] = aE[X]$$

2. Addition/Subtraction

$$E[X + Y] = E[X] + E[Y]$$

3. Powers

$$E[X^2] \neq E[X]^2$$

Example 1.12: Expectation Rules

What is $E[x - 2x^2]$?

$$E[x - 2x^2] = E[x] - 2E[x^2]$$

The First Moment: the Mean

The **mean** of a random variable, often denoted is its expectation.

$$\mu_X = E[X]$$

Example 1.13: Mean of the Bernoulli Distribution

1. Consider the Bernoulli distribution, $X \sim \mathcal{B}er(p)$. Show that $\mu_X = p$.

$$E[X] = \sum_{x \in \{0,1\} x \Pr(x)} = 0 * (1-p) + 1 * p = p$$

2. What is $E[X - 2X^2]$?

$$E[x - 2x^2] = E[x] - 2E[x^2] = p - 2\left(0^2 * (1-p) + 1^2 * p\right) = -p$$

Raw vs. Centered Moments

The n^{th} **raw moment** is the expectation of the n^{th} power of the r.v.

Moment	Expression
First	$E[x] = \mu$ (Mean)
Second	$E[x^2]$
Third	$E[x^3]$

We can also consider the second (and higher) raw moments of a jointly distributed outcome: $E[x, y]$

The n^{th} **centered moment** is $E[(x-\mu)^n]$

Moment	Expression
First	$E[(x-\mu)] = 0$
Second	$E[(x-\mu)^2]$ (Variance)
Third	$E[(x-\mu)^2]$ (Skew)

We can also consider the second (and higher) centered moments of a jointly distributed outcome: $E[(x-\mu_x)(y-\mu_y)] = Cov(x,y)$ is called the **covariance**.

We can perform a change of variables to convert from centered to raw moments. For the second moment, we have:

$$\begin{aligned} E[(x-\mu)^2] =& E[x^2 - 2x\mu + \mu^2] = E[x^2] - 2E[x]\mu + \mu^2 \\ =& E[x^2] - \mu^2 \end{aligned}$$

Hence the second centered moment can be written in terms of the first and second raw moments.

For the third moment, we have:

$$\begin{aligned} E[(x-\mu)^3] =& E[x^3 - 3x^2\mu + 3x\mu^2 - \mu^3] = E[x^3] - 3\mu E[x^2] + 3\mu^2 E[x] - \mu^3 \\ =& E[x^3] - \mu E[x^2] + 2\mu^3 \end{aligned}$$

Hence the third centered moment can be written in terms of the first second and third raw moments.

Example 1.14: Expected waiting time

Suppose that the time it takes a seed to germinate is exponentially distributed with rate λ. Show that the mean (expected) time until germination is $\frac{1}{\lambda}$.

The PDF of the exponential distribution is:

$$\Pr(x) = \lambda e^{-\lambda x}$$

The expectation then is:

$$E[X] = \int_0^\infty x\lambda e^{-\lambda x} dx$$

To do this integral we use integration by parts: $\int uv = uv - \int v du$

$u = x$	$v = -e^{-\lambda x}$
$du = dx$	$dv = \lambda e^{-\lambda x} dx$

$$\begin{aligned} E[X] &= -xe^{-\lambda x}\Big|_0^\infty - \int_0^\infty -e^{-\lambda x} dx \\ &= -xe^{-\lambda x}\Big|_0^\infty - \frac{1}{\lambda} e^{-\lambda x}\Big|_0^\infty \\ &= (0-0) - \frac{1}{\lambda}(0-1) = \frac{1}{\lambda} \end{aligned}$$

Example 1.15: Evolutionary rescue

Suppose you are running an evolution experiment where you expose $n = 10$ E. coli lines to stressful conditions and monitor their ability to adapt. Suppose the probability of adapting before going extinct (known as evolutionary rescue) is $p = 0.3$.

1. **Use the definition of an expectation to show that the expected number of lines to be rescued is $\mu = np$.**

$$E[x] = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x}$$

Shift the sum index by 1 (the first term is simply 0) and then write out the binomial coefficient

$$\begin{aligned} E[x] &= 0 + \sum_{x=1}^{n} x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\ &= \sum_{x=1}^{n} \cancel{x} \frac{n(n-1)!}{\cancel{x}(x-1)!(n-x)!} p \times p^{x-1} (1-p)^{n-x} \\ &= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x} \end{aligned}$$

Let $y = x - 1$ and $m = n - 1$ then note that $m - y = n - x$

$$E[x] = np\underbrace{\sum_{y=0}^{m} \frac{m!}{y!(m-y)!} p^y (1-p)^{m-y}}_{\text{sum over a binomial dist=1}} = np$$

2. **Given that the Bernoulli variance is $Var(X) = np(1-p)$, use the relationship between centered and raw moments to derive the second raw moment of the Binomial Distribution**

$$Var(x) = E[x^2] - E[x]^2$$
$$np - np^2 = E[x^2] - (np)^2$$

Solving we have:

$$E[x^2] = np(np - p + 1)$$

Law of Total Variance

For random variable X, Y, and Z

$$\text{Var}(X) = E_Z[Var(X|Z)] + Var_Z[E[X|Z]]$$

In other words, the variance in an outcome is the average variance + the variance in the averages.

Example: Phenotypic Variance

A classic equation/assumption from quantitive genetics is that a phenotype (P) is equal to the genetic contributions (G) plus the environmental contributions (E)

$$P = G + E$$

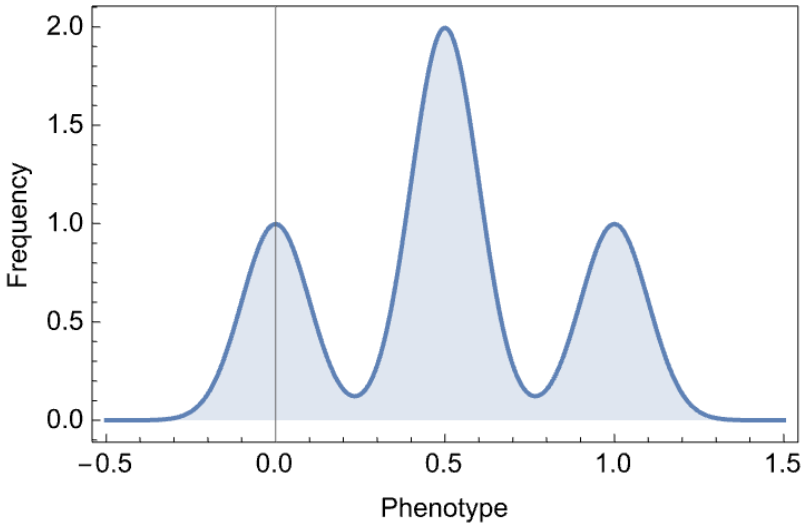
Suppose that a trait is determined by a single bi-allelic locus with genotypes "AA", "Aa", and "aa" such that the mean phenotypes of the three genotypes are:

genotype	mean pheno.	frequency
AA	1	0.25
Aa	0.5	0.5
aa	0	0

But environmental noise adds variation about each of these means according to a normal distribution with mean 0 and standard deviation 0.1.

1. What does the distribution of phenotypes look like in the whole population?

$$P \sim 0.25\mathcal{N}(0, 0.1) + 0.5\mathcal{N}(0.5, 0.1) + 0.25\mathcal{N}(1, 0.1)$$



2. What is the mean phenotype?

The mean is completely independent of the environmental noise so we have

$$E[P] = 0.25 \times 0 + 0.5 \times 0.5 + 0.25 \times 1 = 0.5$$

3. What is the phenotypic variance?

Let's start by calculating $Var_{Geno}[E[P|\text{Geno}]]$

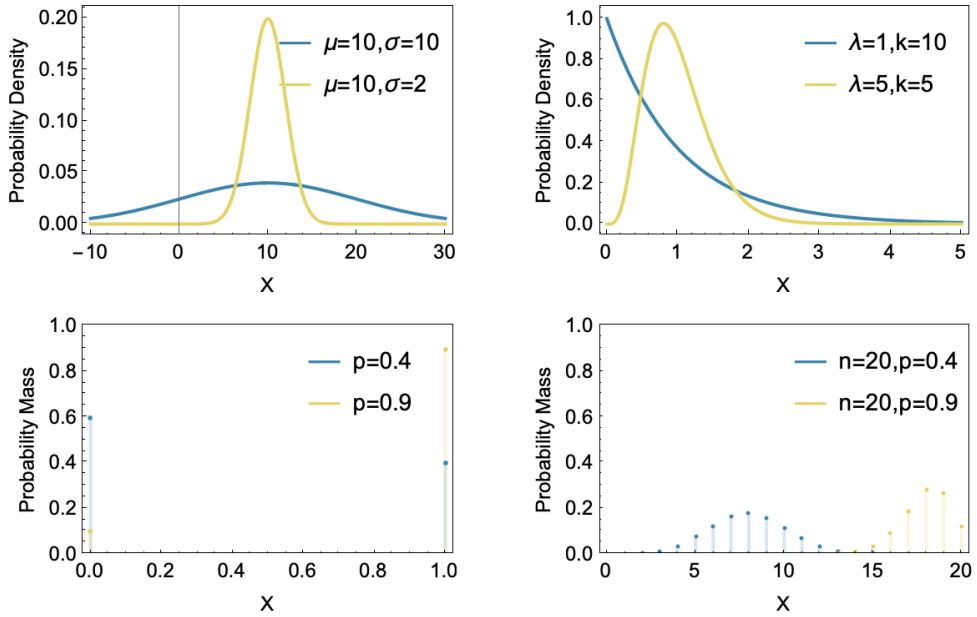
$$\begin{aligned} &Var_{Geno}[E[P|\text{Geno}]] \\ &\quad = 0.25(0 - 0.5)^2 + 0.5(0.5 - 0.5)^2 + 0.25(1 - 0.5)^2 \\ &\quad = 0.125 \end{aligned}$$

Plus the variance given the means $E_{geno}[Var(P|\text{geno})] = 0.1^2$

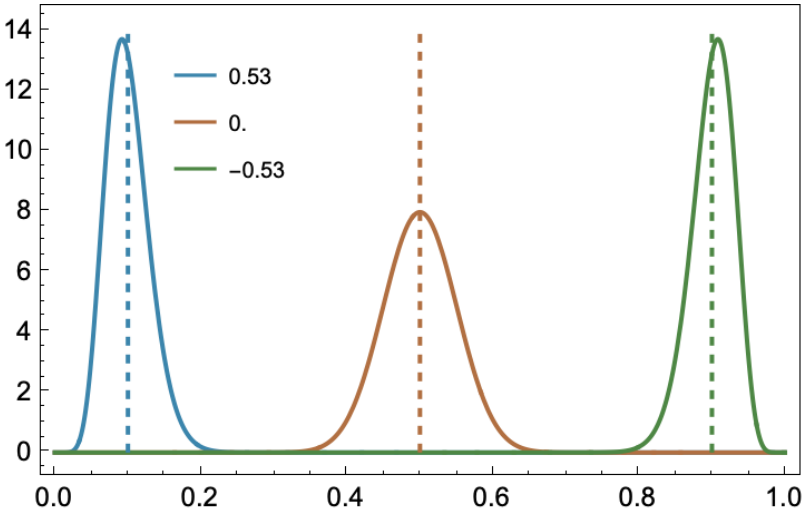
Such that the total variance is 0.135

Visualizing Moments

Let's compare some discrete (Bernoulli and Binomial) and continuous distributions (Normal and Erlang) with a small variance (yellow) against those same distributions with the same mean but larger variance (blue).



What does 'Skew' mean? Skewness is a measure of the asymmetry of a probability distribution. A *negative skew* indicates that the **left tail** of the distribution is longer or fatter than the right tail, and the bulk of the values is concentrated on the right side. In other words, the distribution is **skewed to the left** (e.g. the green distribution below).



Right skewed distributions (blue curve) have a mean (dashed line) greater than the mode. For **left skewed** distributions (green curve) the mean is less than the mode.

Lecture 1.5 Sampling From Probability Distributions

Programming in Python

Jupyter Notebooks are a powerful and interactive computing environment that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. The name "Jupyter" is a combination of three core programming languages it supports Julia, Python, and R.

Key Features of Jupyter:

- **Interactive Computing**: you can write and execute code in small, manageable sections called **cells**.
- **Multiple Language Support** While Jupyter originated from the combination of Julia, Python, and R, it has grown to support a wide range of programming languages.
- **Rich Text Support:** In addition to code cells, Jupyter Notebooks support markdown cells (including latex), enabling the inclusion of formatted text, equations, images, and hyperlinks.
- **Visualization:** Jupyter integrates seamlessly with popular Python libraries.
- **Data Exploration:** You can easily load and manipulate data using libraries like Pandas and NumPy. The ability to mix code with narrative text makes it an excellent tool for data exploration and analysis.
- **Collaboration and Sharing:** Notebooks can be exported to various formats, such as HTML or PDF, or shared online through platforms like GitHub.

Jupyter at SFU

You can access Python and jupyter using the SYZYGY server

Python Arrays

In Python, the term "Python arrays" can be ambiguous because there are two main types of data structures that are commonly referred to as arrays: Python lists and NumPy arrays.

- **Python Lists:**
  - Python lists are a built-in data type in Python and are quite flexible. They can contain elements of different data types and can dynamically grow or shrink in size.
  - Lists are versatile but may not be as efficient for numerical operations as NumPy arrays.

- Example: `python_list = [1, 2, 3, 4, 5]}]`

- **NumPy Lists:**
  - NumPy is a powerful numerical computing library for Python. One of its key features is the NumPy array, a multi-dimensional array object.
  - NumPy arrays are homogeneous and typically contain elements of the same data type, which allows for more efficient numerical operations.
  - NumPy provides a wide range of mathematical functions that operate on entire arrays without the need for explicit loops.

Sampling from Focal Probability Distribution

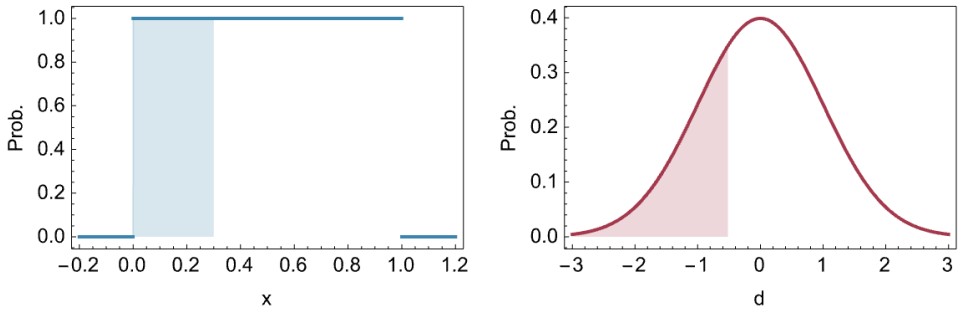
Python: Lecture1_5.ipynb

Step 1: Sample a random number x^* from a uniform Distribution $X \sim U(0,1)$

Step 2: create a mapping from this $U(0,1)$ to our focal distribution $\mathscr{D}$ by equating their CDFs such that:

$$CDF_U(x^*) = CDF_{\mathscr{D}}(d^*)$$

So to draw a random variable, x from distribution $\mathcal{D}$ we have:



The inverse of the CDF is used so often that it has a name the **Percentile Point Function** (ppf).

$$\text{ppf}(x) = y : \text{cdf}(y) = x$$

Example 1.16: Ebola

There have been 12 significant Ebola outbreaks in the last 20 years.

1. **Modelling the waiting time between outbreaks with an exponential distribution what is λ?**

$$\lambda = \frac{12}{20}\frac{1}{\text{years}} = 0.6\frac{1}{\text{years}}$$

2. **What is the expected time until the next outbreak?**

$$\frac{1}{\lambda} = 1.6\bar{6}\text{years} \quad 1\text{year } 8\text{months}$$

3. **Given the random # $u^* = 0.612$ from $U(0,1)$, what is the corresponding random waiting time until the next outbreak?**

$$\text{ppf}_{exp}(0.623) = 1.578 \approx 1\text{year } 7\text{months}$$

4. **The most recent outbreak started in September 2022 in Uganda. Simulate the sequence of the next 5 outbreaks and draw them below.**

Uniform	Exponential	Cumulative Time
$u_1 = 0.17$	$x_1 = 0.327$	$t_1 = 0.327$ (4 mon)

Uniform	Exponential	Cumulative Time
$u_2 = 0.129$	$x_2 = 0.229$	$t_2 = 0.556$ (6.6 mon)
$u_3 = 0.213$	$x_3 = 0.399$	$t_3 = 0.955$ (11.5 mon)
$u_4 = 0.562$	$x_4 = 0.399$	$t_4 = 2.83$ (2 yr 4 mon)
$u_5 = 0.803$	$x_5 = 2.17$	$t_5 = 5.04$ (5 yr 0.5 mon)