Important probability distributions in population genetics

Jan van Waaij

December 7, 2022

Credits: all pictures are from Wikipedia.

1 Binomial distribution

The binomial distribution with parameters n and p takes values in $\{0, \ldots, n\}$. It is the distribution of n independent experiments with probability of succes p. So when $X_1, \ldots, X_n$ are zero or one with probability p, then

$$S = \sum_{k=1}^{n} X_k \tag{1}$$

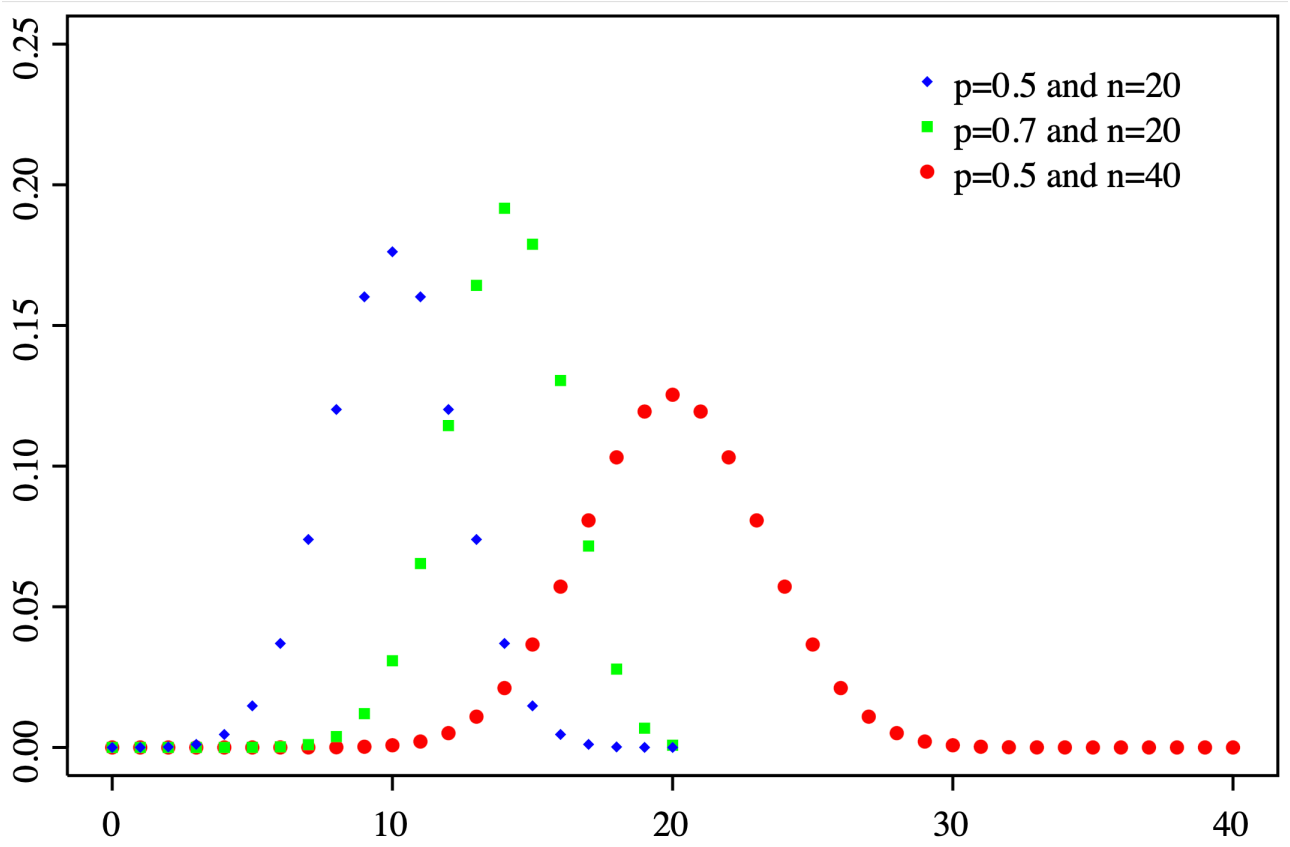
is binomially distributed with parameters n and p.

Probability mass function:

$$\binom{n}{k} p^k (1-p)^{n-k},$$

Difficult to calculate for n large. However, can be well approximated with Stirling's formula, or approximated with the Poisson distribution.

Mean np, variance $np(1-p)$.



1.1 Sums

When $X \sim \text{binom}(n, p)$ and $Y \sim \text{binom}(m, p)$, and X and Y are independent, then we see from eq. (1) that $X + Y \sim \text{binom}(m + n, p)$. In general with $p \neq q$, the sum is not Binomial.

1.2 History

Blaise Pascal 1623–1662, France, for $p = 1/2$.

General p first studied by Jacob Bernoulli 1655-1705, Switzerland.

Figure 1: left: Jakob Bernoulli, right: Blaise Pascal

1.3 An application in population genetics

If the frequency of a certain genotype is p, then the probability of seeing k individuals in a sample of n individuals is binomially distributed with parameters n and p.

1.4 Estimators for p

Maximum likelihood estimator $\hat{p} = \frac{X}{n}$.

The beta distribution is conjugate with the binomial distribution. If

$$\begin{aligned} p &\sim \text{beta}(\alpha, \beta), \\ X \mid p &\sim \text{binom}(n, p), \end{aligned}$$

then

$$p \mid X \sim \text{beta}(\alpha + X, \beta + n - X)$$

In particular, the posterior mean is

$$\frac{X + \alpha}{n + \alpha + \beta},$$

so for instance with $\text{beta}(1, 1) = \text{uniform}([0, 1])$, then the Bayesian estimator is

$$\hat{p}_{Bayes} = \frac{X + 1}{n + 2}$$

Benefit:

$$0 < \hat{p}_{Bayes} < 1,$$

because $p = 0$ or 1 might be unreasonable. Drawback: it is a (slightly) biased estimator.

1.5 Confidence intervals

1.5.1 Normal approximation interval / Wald interval

$\hat{p} \pm \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ Do not use, really bad.

1.5.2 Wilson score interval

Also based on normal approximation, but more cleaver. Is reasonably good.

1.5.3 Clopper-Pearson interval

Let X be binomially distributed with parameters n and p, of which p is unknown. The maximum likelihood estimator and moment estimators are $\hat{p} = X/n$. The exact confidence interval or Clopper-Pearson[1] interval set for $\hat{p}$ is given by $[p_L, p_U]$, where

$$\sum_{k=X}^{n} \binom{n}{k} p_L^k (1-p_L)^{n-k} = \alpha/2$$

and

$$\sum_{k=0}^{X} \binom{n}{k} p_U^k (1-p_U)^{n-k} = \alpha/2$$

Then

$$P_p(p < p_L) = P_p\left(\sum_{k=X}^{n} \binom{n}{k} p^k (1-p)^{n-k} < \alpha/2\right) \leq \alpha/2,$$

and

$$P_p(p > p_U) = P_p\left(\sum_{k=0}^{X} \binom{n}{k} p^k (1-p)^{n-k} < \alpha/2\right) \leq \alpha/2.$$

So $P_p(p \notin [p_L, p_U]) = P_p(p < p_L) + P(p > p_U) \leq \alpha/2 + \alpha/2 = \alpha$.

1.6 Bayesian credible intervals

Note that under the beta prior,

$$p \mid X \sim \text{beta}(\alpha + X, \beta + n - X).$$

Let $q_{\alpha/2}$ and $q_{1-\alpha/2}$ be the $\alpha/2$ and $1-\alpha/2$ quantiles of the beta$(\alpha+X, \beta+n-X)$ distribution. Then $[q_{\alpha/2}, q_{1-\alpha/2}]$ is a $1-\alpha$ credible interval. For $\alpha = \beta = 1/2$, this is the Jeffrey confidence interval.

1.7 Approximations

1.7.1 Poisson approximation

When $n \to \infty$ and $p = p_n$ is so that $np_n \to \lambda \in (0, \infty)$, then binom$(n, p) \rightsquigarrow$ Poisson(λ). This is often used in population genetics when the number of individuals is large and the exact binomial distributions are approximated by Poisson distributions.

So when $X \sim$ binom(n, p), is the probability distribution of having X time succes in n experiments, when you have on average np times succes, and when $Y \sim$ Poisson(λ) is the distribution of having Y times succes in infinity times experiments when you have on average λ times success. So in a sense the Poisson distribution is an binomial distribution with $n = \infty$.

[1] Named after the inventors C.J. Clopper and E.S. Pearson

1.7.2 Normal distribution

By the central limit theorem

$$\frac{X - np}{np(1-p)} \rightsquigarrow N(0,1).$$

But is a bad approximation according to the literature, unless you have exceptionally large N.

2 Poisson distribution

The Poisson distribution (named after the French mathematician Siméon Denis Poisson, 1781 – 1840) is a distribution on all non-negative integers. Has one parameter λ.

The probability mass function is

$$p(k) = \frac{e^{-\lambda}}{k!}\lambda^k, \quad k = 0, 1, 2, \ldots$$

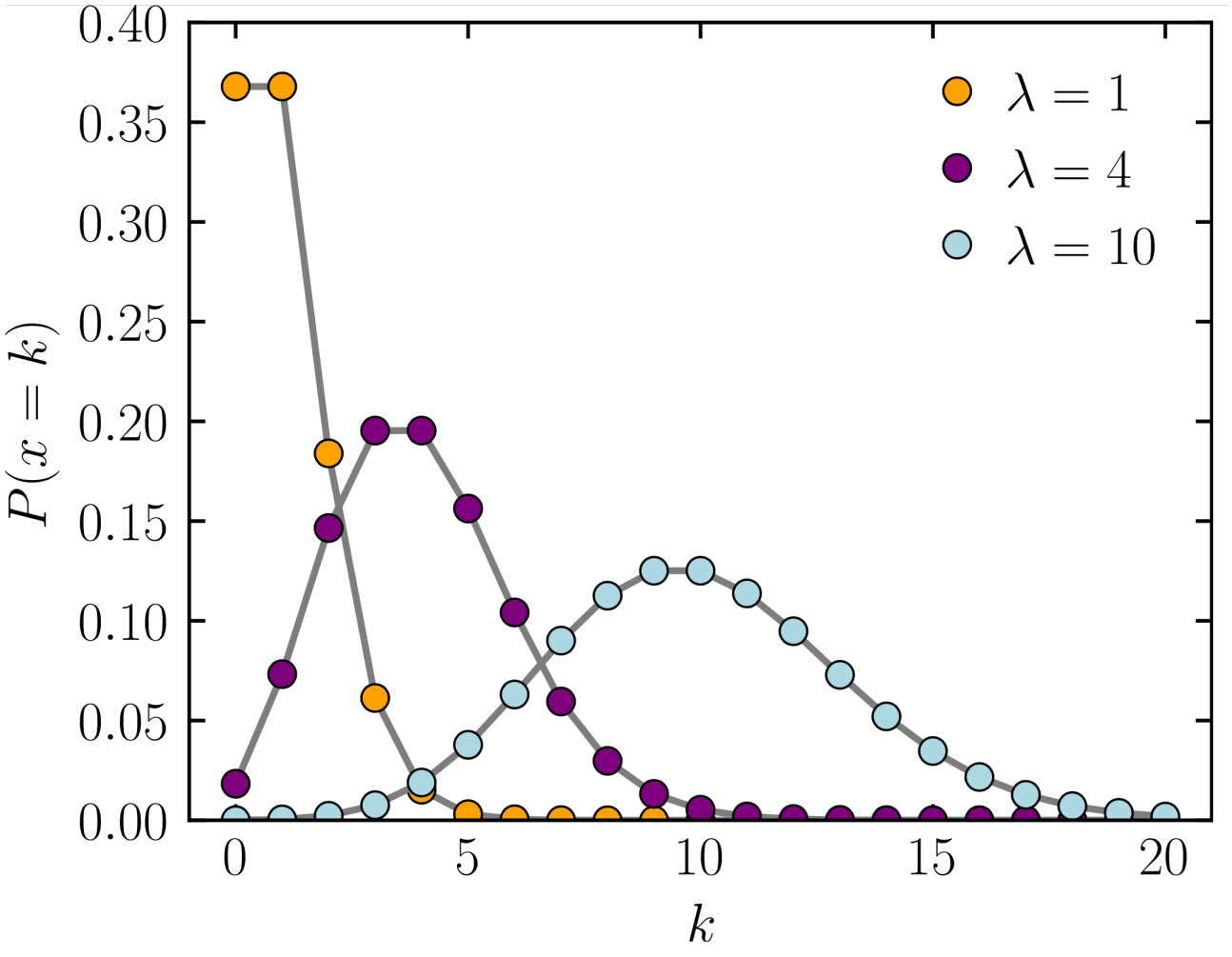
Is based on the Taylor expansion

$$e^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!},$$

so

$$\sum_{k=0}^{\infty} p(k) = 1.$$

Mean: λ, variance λ.



2.1 Relationship with the binomial distribution

The sum of two independent Poisson distributions is again Poisson with parameter the sum of the parameters.

$$P(X_1 \mid X_1 + X_2) \sim \text{binom}\left(X_1 + X_2, \frac{\lambda_1}{\lambda_1 + \lambda_2}\right)$$

2.2 History

It was first introduced by Abraham de Moivre in 1711 (French but fled at young age to England because of Huguenots persecution), publised in a book by Siméon Denis Poisson (1781–1840).

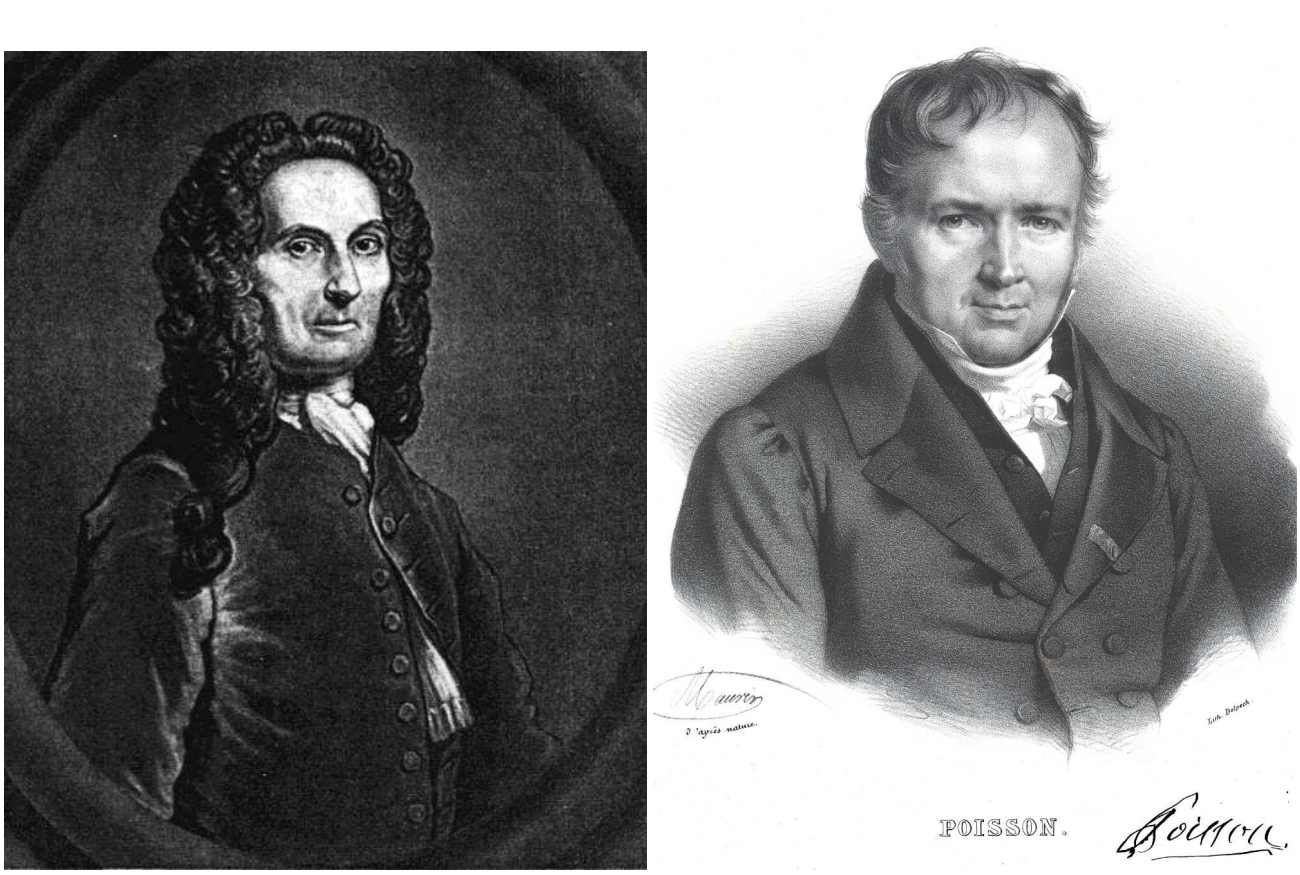


Figure 2: left: Abraham de moivre, right: Siméon Dennis Poisson

2.3 Application in population genetics

In the Wright-Fisher model, you have 2N individuals in each population. Each individual picks his parent randomly. So, each parent can be choosen by $2N$ children with probability $1/(2N)$. We have $2N \cdot \frac{1}{2N} = 1$, so when N is large, the number of children is approximately Poisson distributed with parameter 1.

2.4 Sums

If $X_i \sim \text{Poisson}(\lambda_i)$, are independent, then

$$X_1 + \ldots + X_n \sim \text{Poisson}\left(\lambda_1 + \ldots + \lambda_n\right).$$

2.5 Estimation

If you have $X_1, \ldots, X_n$ random variables that are Poisson distributed with unknown parameter λ, then

$$\hat{\lambda}_{lik} = \frac{1}{n} \sum_{k=1}^{n} X_k$$

is the maximum likelihood estimator.

2.6 Confidence interval

If you have on observation X, then

$$\frac{1}{2}\chi^2(\alpha/2, 2X) \leq \lambda \leq \frac{1}{2}\chi^2(1-\alpha/2, 2X+2),$$

is a confidence interval, where $\chi^2(p, f)$ is the quantile function of the Chi squared distribution with f degrees of freedom.

If you have n observations $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$, then $X_1 + \ldots + X_n \sim \text{Poisson}(n\lambda)$. So

$$\frac{1}{2}\chi^2(\alpha/2, 2(X_1 + \ldots + X_n)) \leq n\lambda \leq \frac{1}{2}\chi^2(1-\alpha/2, 2(X_1 + \ldots + X_n) + 2),$$

so

$$\frac{1}{2n}\chi^2(\alpha/2, 2(X_1 + \ldots + X_n)) \leq \lambda \leq \frac{1}{2n}\chi^2(1-\alpha/2, 2(X_1 + \ldots + X_n) + 2).$$

2.7 Bayesian inference

The Poisson distribution is conjugate with the gamma distribution.

$$\begin{aligned} \lambda &\sim \text{gamma}(\alpha, \beta) \quad \text{shape parameter } \alpha \text{ rate parameter } \beta, \\ X_1, \ldots, X_n \mid \lambda &\sim \text{Poisson}(\lambda), \\ \lambda \mid X_1, \ldots, X_n &\sim \text{gamma}\left(\alpha + \sum_{k=1}^{n} X_k, \beta + n\right). \end{aligned}$$

So Bayesian estimator (posterior mean)

$$\frac{\alpha + \sum_{k=1}^{n} X_k}{\beta + n},$$

credible interval:

$$\left[\gamma_{\alpha/2, \alpha+\sum_{k=1}^{n} X_k, \beta+n}, \gamma_{1-\alpha/2, \alpha+\sum_{k=1}^{n} X_k, \beta+n}\right],$$

where $\gamma_{p,\alpha,\beta}$ is the p-th quantile of the gamma distribution with parameters α and β.

3 Geometric distribution

Parameter p. The number of failures before the first time succes, takes values in $0, 1, \ldots$. It has pdf

$$p(k) = (1-p)^k p.$$

So suppose $X_1, X_2, \ldots$ are experiments with probability of success p, then the smallest k so that $X_{k+1} = 1$ and $X_a = \ldots = X_k = 0$ is geometric distributed with parameter p.

If $Y \sim$ geometric(p), then $Y \geq k$ means that the first k experiments fail. This happens with probability $(1-p)^k$. So $P(Y \geq k) = (1-p)^k$. So

$$P(Y = a + b \mid Y \geq a) = \frac{(1-p)^{a+b}p}{(1-p)^a} = (1-p)^b p = P(Y = b).$$

So the geometric distribution is memoryless.

Mean $\frac{1-p}{p}$ variance $\frac{1-p}{p^2}$.

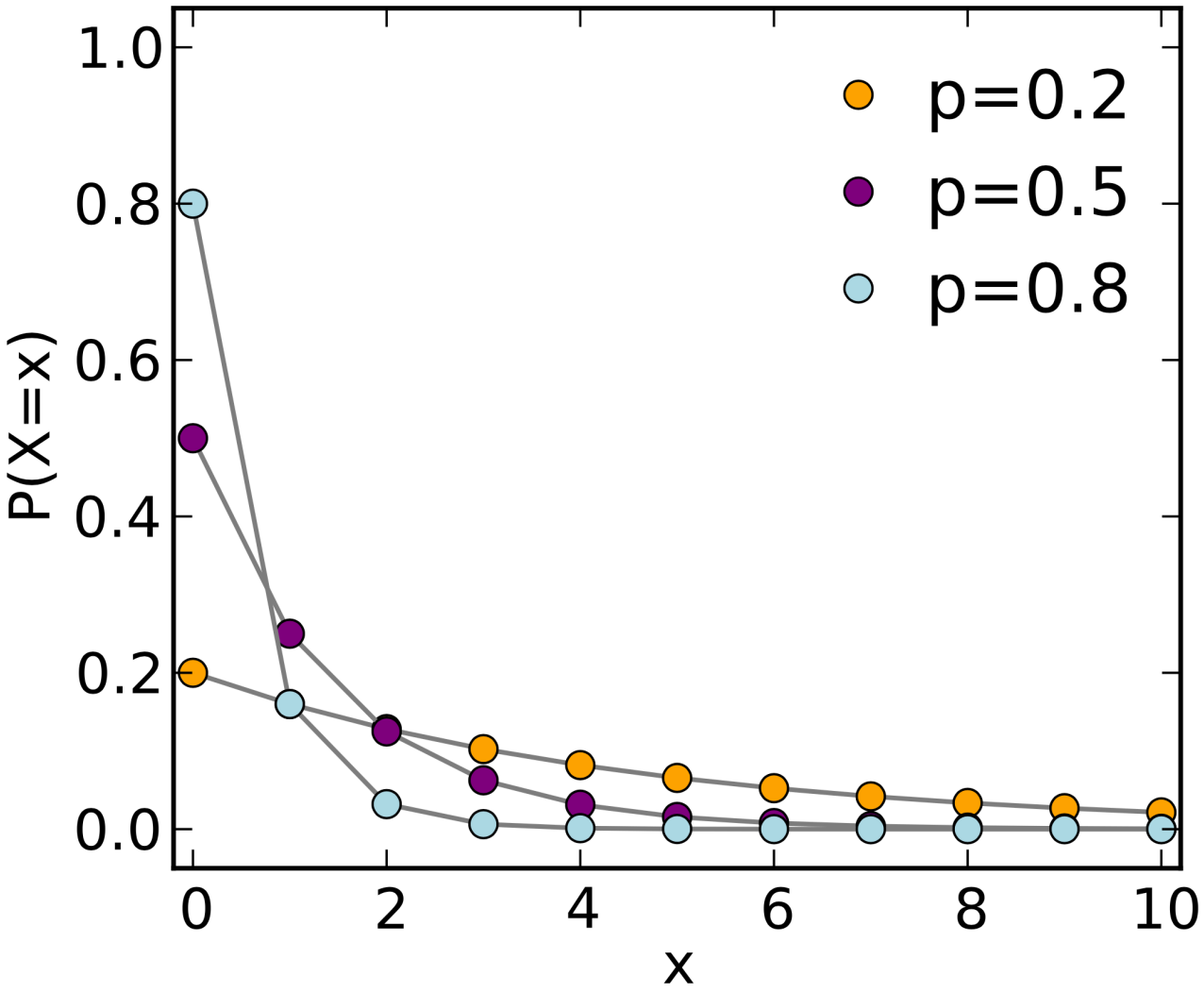


Figure 3: Geometric distribution

3.1 Application in population genetics

In the Wright-Fisher model, with population size $2N$, if you have two individuals, then they have the same parent with probability $1/(2N)$. They have probability of having the same grandparent with probability $1/(2N)$. They have probability of having latest common ancestor in the $k+1$ generation with probability

$$\left(1 - \frac{1}{2N}\right)^k \frac{1}{2N}.$$

Which is the geometric distribution.

3.2 Relation with the exponential distribution

Lemma 1. *Let $Y_n \sim$ geometric(p_n), so that $np_n \to \lambda$. Let $Z_n = \frac{Y_n}{n}$ and let $Z \sim \text{Exp}(\lambda)$, then $Z_n \rightsquigarrow Z$, as $n \to \infty$.*

Proof. Let $x > 0$. Then

$$\begin{aligned} P(Z_n \leq x) =& P(Y_n \leq \lfloor nx \rfloor) \\ =& 1 - P(Y_n > \lfloor nx \rfloor) \\ =& 1 - (1-p_n)^{\lfloor nx \rfloor + 1} \end{aligned}$$

Note that $1 - p_n \to 1$, as $n \to \infty$ and

$$\left(1 - \frac{p_n n}{n}\right)^{nx} \to e^{-\lambda x}, \text{ as } n \to \infty.$$

So $P(Z_n \leq x)$ converges to $P(Z \leq x)$. So Z_n converges weakly to Z. □

3.3 Maximum likelihood estimator

$$\frac{n}{n + \sum_{k=1}^n X_k}.$$

3.4 Bayesian inference

The geometric distribution is conjugate with the beta distribution.

$$\begin{aligned} p &\sim \text{beta}(\alpha, \beta), \\ X_1, \ldots, X_n \mid p &\sim \text{geometric}(p), \\ p \mid X_1, \ldots, X_n &\sim \text{beta}\left(\alpha + n, \beta + \sum_{k=1}^n X_k\right). \end{aligned}$$

So Bayesian estimator for p,

$$\frac{\alpha + n}{\alpha + \beta + n + \sum_{k=1}^n X_k},$$

95% credible interval

$$\left[q_{0.025, \alpha+n, \beta+\sum_{k=1}^n X_k}, q_{0.975, \alpha+n, \beta+\sum_{k=1}^n X_k}\right],$$

where $q_{p,\alpha,\beta}$ is the p quantile for the beta distribution with parameters.

4 Exponential distribution

Continuous distribution on $[0, \infty)$ with density $f(x) = \lambda e^{-\lambda x}$. Then $P(Y \geq x) = e^{-\lambda x}$.

Mean $\frac{1}{\lambda}$, variance $\frac{1}{\lambda^2}$.

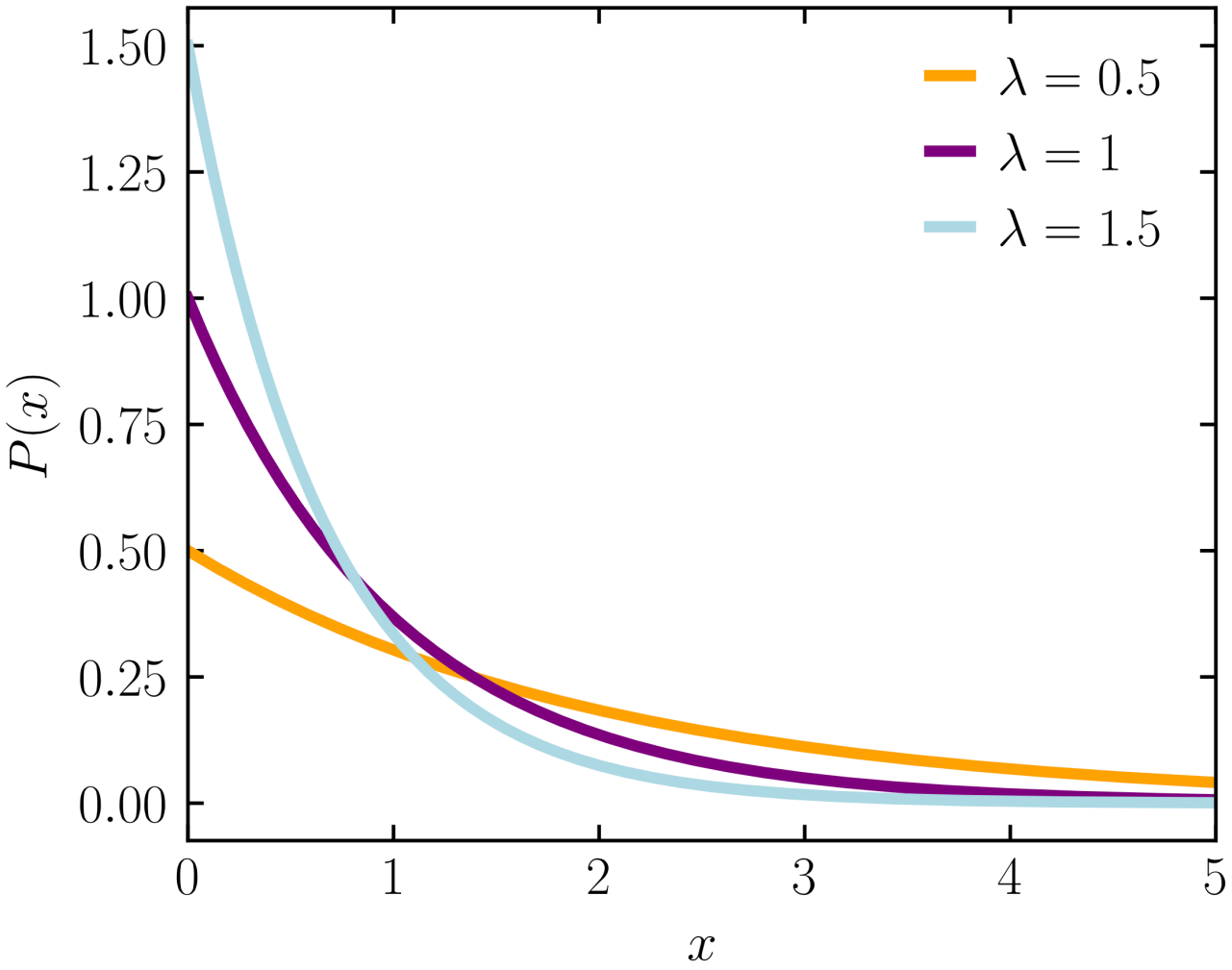


Figure 4: Exponential

4.1 Relation with the geometric distribution

Let $X \sim \text{Exp}(\lambda)$. Then for $k \in \mathbb{N}_0$,

$$\begin{aligned} P(\lfloor X \rfloor = k) &= \int_k^{k+1} \lambda e^{-\lambda x} dx \\ &= -e^{-\lambda x}\Big|_{x=k}^{k+1} \\ &= e^{-\lambda k} - e^{-\lambda(k+1)} \\ &= e^{-\lambda k}(1 - e^{-\lambda}) \\ &= \left(1 - (1 - e^{-\lambda})\right)^k (1 - e^{-\lambda}). \end{aligned}$$

So $\lfloor X \rfloor$ is geometric distributed with parameter $p = 1 - e^{-\lambda}$.

4.2 Memorylessness of the exponential distirbution

Let Y be exponential distributed with parameter λ. Then

$$P(Y \geq a + b \mid Y \geq a) = \frac{e^{-\lambda(a+b)}}{e^{-\lambda a}} = e^{-\lambda b} = P(Y \geq b).$$

So the exponential distribution is memoryless.

4.3 The minimum of two exponentials is exponential

Lemma 2. *Let* $X \sim \text{Exp}(\lambda)$ *and* $Y \sim \text{Exp}(\mu)$ *be independent. Then* $X \wedge Y := \min(X, Y) \sim \text{Exp}(\lambda + \mu)$.

Proof. Let $Z = X \wedge Y$. Let $x \geq 0$. Then

$$\begin{aligned} P(Z \geq x) =& P(X \geq x, Y \geq x) \\ =& P(X \geq x)P(Y \geq x) \\ =& e^{-\lambda x} e^{-\mu x} \\ =& e^{-(\lambda+\mu)x}. \end{aligned}$$

So $Z \sim \text{Exp}(\lambda + \mu)$. □

A Difference between credible intervals and confidence intervals

A.1 Confidence intervals

A confidence interval is an interval $I = I(X)$ (the interval depends on the data) of parameters (more generally a confidence set, which is a set of parameters) so that for all parameters p in your paramater space

$$P_p(p \in I) \geq 1 - \alpha$$

(typically $1 - \alpha = 0.95$).

So you have data X which is generated according to an unknown parameter p, and with X you construct $I = I(X)$ so that $P_p(p \in I(X))$. So you are never sure whether the true parameter is in your data set, only that when you do the experiment 20 times, you expect that in 19 cases the true parameter is in the interval.

Assuming that there is a true parameter p is called the frequentist school of statistics. It was developed in the early 20th century and is more recent than Bayesian statistics. It was developed by Ronald Fisher, Jerzy Neyman, and Egon Pearson.

A.2 Credible intervals

A credible interval / Bayesian confidence set (more generally credible set) is an interval with $1 - \alpha$ posterior probability. So

$$\begin{aligned} p \sim& P \\ X \mid p \sim& P_p. \end{aligned}$$

So P is your believe about the reality, and then you see data and learn from it to update your believe to $P \mid X$.

A credible interval is an interval I so that

$$P(I \mid X) \geq 1 - \alpha.$$

So I is a $1 - \alpha$ adequate description of reality.