Probability distributions

Alexander Khanov

PHYS6260: Experimental Methods is HEP
Oklahoma State University

September 15, 2023

Mean, variance, covariance

- mean $=$ expected value of x

$$\text{discrete}: \langle x \rangle = \sum_i P_i x_i \quad \text{continuous}: \langle x \rangle = \int f(x) x \, dx$$

- variance $=$ expected value of $(x - \langle x \rangle)^2$

$$\sigma^2 = \left\langle (x - \langle x \rangle)^2 \right\rangle = \left\langle x^2 \right\rangle - \langle x \rangle^2$$

- standard deviation $\sigma = \sqrt{}$ of variance
- covariance of two variables x and y:

$$\text{cov}(x, y) = \langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle = \langle xy \rangle - \langle x \rangle \langle y \rangle$$

- correlation coefficient of two variables:

$$\rho(x, y) = \text{cov}(x, y)/(\sigma_x \sigma_y)$$

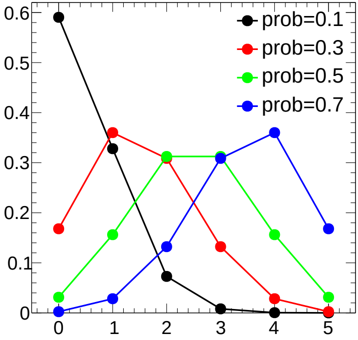
Binomial distribution

- Consider an experiment which has two possible outputs: 1 (with probability p) and 0 (with probability $q = 1 - p$). In a series of n experiments, what is the probability to get k 1's and $(n - k)$ 0's?
  - the number of ways to chose k experiments out of n (disregarding order) is $\dfrac{n!}{(n-k)!k!}$
  - k experiments have probability p and $n - k$ experiments have probability q, so the probability to get k 1's and $(n - k)$ 0's is

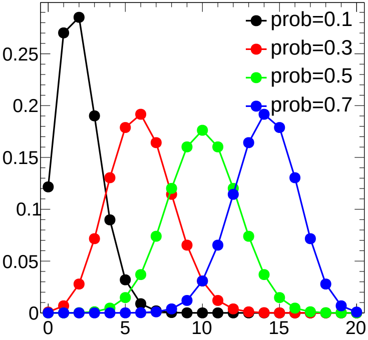
$$P(k) = \frac{n!}{(n-k)!k!} p^k q^{n-k}$$

  - check the normalization: $\sum_{k=0}^{n} P(k) = (p + q)^n = 1$
- Binomial distribution has mean value np and variance $np(1 - p)$

Examples of binomial distribution



$n = 5$



$n = 20$

Multinomial distribution

- If an experiment has d possible outcomes and there are n trials then the probability of getting $n_1, \ldots, n_d$ outcomes of type $1, \ldots, d$ is

$$P(n_1, \ldots n_d) = \frac{n!}{n_1! \ldots n_d!} p_1^{n_1} \ldots p_d^{n_d}, \quad \sum_{i=1}^{d} p_i = 1, \quad \sum_{i=1}^{d} n_i = n$$

  - this distribution describes bin contents of a histogram with d bins and the total number of entries n
  - each individual bin content follows binomial distribution, but contents of different bins are correlated

What happens when n becomes large?

- If p is not close to either 0 nor 1, then according to Stirling's formula,

$$\frac{n!}{(n-k)!k!}p^k q^{n-k} \approx \frac{n^n}{e^n}\frac{e^{n-k}}{(n-k)^{n-k}}\frac{e^k}{k^k}p^k q^{n-k} = \left(\frac{np}{k}\right)^k \left(\frac{nq}{n-k}\right)^{n-k}$$

  - using logarithm expansion $\ln(1+x) = x - \frac{x^2}{2} + \ldots$ one can show that

$$\ln\left[\left(\frac{np}{k}\right)^k \left(\frac{nq}{n-k}\right)^{n-k}\right] \approx -\frac{(k-np)^2}{2npq}$$

- If p is small so that $\lambda = np$ is small compared to n, then

$$\frac{n!}{(n-k)!k!}p^k q^{n-k} = \frac{n(n-1)\ldots(n-k+1)}{k!}\left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} =$$

$$\frac{n}{n}\frac{n-1}{n}\ldots\frac{n-k+1}{n}\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k} \approx \frac{\lambda^k}{k!}e^{-\lambda}$$

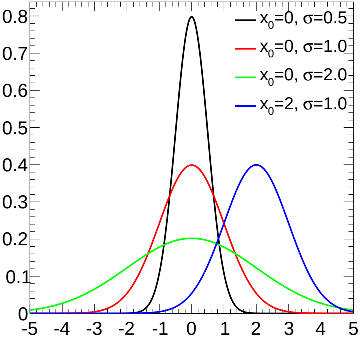
Gaussian distribution

- If the number of experiments n is large and p is not close to $0/1$ then the binomial distribution becomes Gaussian (or normal)

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-x_0)^2}{2\sigma^2}}$$

A sum of a large number of independent variables is approximately normally distributed, no matter what is the underlying distribution of the variables (central limit theorem).

- Gaussian distribution has mean value x_0 and variance σ^2
- Binomial distribution can be approximated by a Gaussian distribution with $x_0 = np$ and $\sigma^2 = npq$



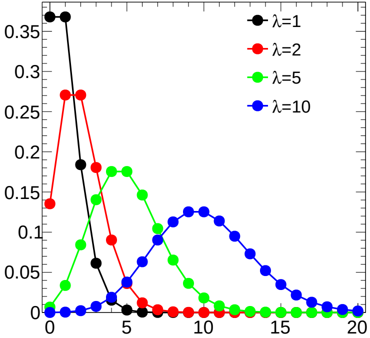
Poisson distribution

- If the number of experiments n is large and p is small but their product $\lambda = np$ is moderate (1–10) then the binomial distribution becomes Poisson

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$P(k)$ is the probability of a given number of events to occur in a fixed interval of time if these events occur with a known average rate λ independently from each other

- Poisson distribution has both mean value and variance equal to λ
- Poisson distribution is the limit case of binomial distribution when $n \to \infty$ and np remains fixed
- If λ is large then Poisson distribution becomes very similar to Gaussian

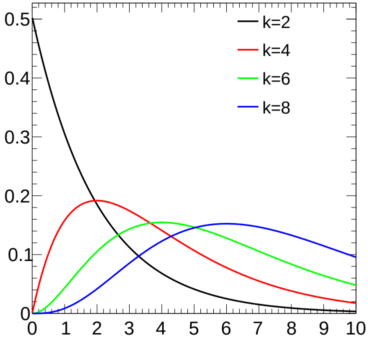


χ^2 distribution

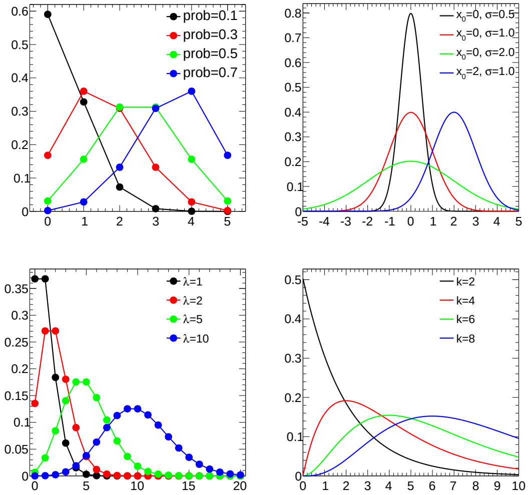
- This is the distribution of a sum of the squares of k independent standard normal random variables ($x_0 = 0$, $\sigma = 1$)

$$f(x) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$$

- If k variables x_i are distributed normally then $\sum_i \frac{(x_i - x_{0i})^2}{\sigma_i^2}$ is distributed as χ^2
- χ^2 distribution has mean value k and variance $2k$
- per central limit theorem χ^2 becomes Gaussian as k increases
  - in practice, $\sqrt{2\chi^2}$ is much closer to a Gaussian, with mean of $\sqrt{2k-1}$ and unit variance



Summary of distributions



Joint probability

- Joint probability distribution of a pair of random variables is probability distribution of all possible pairs of outcomes
  - this extrapolates to any number of variables

A pair of coins

	heads	*tails*
heads	1/4	1/4
tails	1/4	1/4

Multivariate normal distribution ($\mathbf{x} = x_1, \ldots, x_k$):

$$f(\mathbf{x}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{x_0})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{x_0})\right)}{\sqrt{(2\pi)^k \det \boldsymbol{\Sigma}}}$$

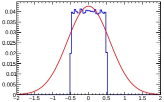
- x and y are independent if and only if $f_{xy}(x,y) = f_x(x)f_y(y)$
- The method of transformations: Let $\mathbf{x} = x_1, \ldots, x_k$ be continuous random variables with joint probability density $f_{\mathbf{x}}(\mathbf{x})$. Let $\mathbf{x} = \mathbf{h}(\mathbf{y})$. Then $f_{\mathbf{y}}(\mathbf{y}) = f_{\mathbf{x}}(\mathbf{h}(\mathbf{y}))|J|$, where $J = \partial\mathbf{h}/\partial\mathbf{y}$.
  - Example: a sum of two independent random variables $z = x + y$

    $$\left\{\begin{array}{lcl} x & = & x \\ y & = & z - x \end{array}\right., \; J = -1, \; f_{xz}(x,z) = f_{xy}(x, z-x) = f_x(x)f_y(z-x)$$

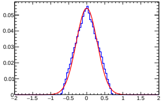
    Integrating out x, for z we get $f_z(z) = \int f_x(x)f_y(z-x)\,dx$

Sum of random variables

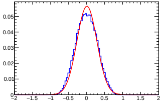
- If x is normally distributed with mean x_0 and variance σ_x^2, y is normally distributed with mean y_0 and variance σ_y^2, and x and y are independent, then $z = x + y$ is normally distributed with mean $z_0 = x_0 + y_0$ and variance $\sigma_z^2 = \sigma_x^2 + \sigma_y^2$
  - If x and y are correlated, z is still normally distributed with mean $z_0 = x_0 + y_0$ and variance $\sigma_z^2 = \sigma_x^2 + \sigma_y^2 + 2\rho\sigma_x\sigma_y$
- Addition of independent random variables also works for Poisson and χ^2 distributions
  - in fact the opposite is also true: if z is Gaussian (Poisson) distributed and x and y are independent then both x and y are also Gaussian (Poisson) distributed
- Central limit theorem: sum of a large number of any random variables is approximately normally distributed



x_1



$\frac{1}{\sqrt{2}} \sum_{i=1}^{2} x_i$



$\frac{1}{\sqrt{3}} \sum_{i=1}^{3} x_i$

Ratio of random variables

- Ratio of two independent standard normal random variables follows Cauchy distribution $f(x) = \dfrac{1}{\pi(1+x^2)}$
- Mean of Cauchy distribution is undefined (as are all momenta)
  - $\displaystyle \lim_{T\to\infty} \int_{-T}^{+aT} \frac{x}{\pi(1+x^2)}\, dx = \lim_{T\to\infty} \frac{1}{2\pi} \ln \frac{1+a^2T^2}{1+T^2} = \frac{\ln a}{\pi}$
- A sum of Cauchy distributed variables is Cauchy distributed, so the central limit theorem fails here