Statistics for Astronomers: Lecture 05, 2019.02.21

Prof. Sundar Srinivasan

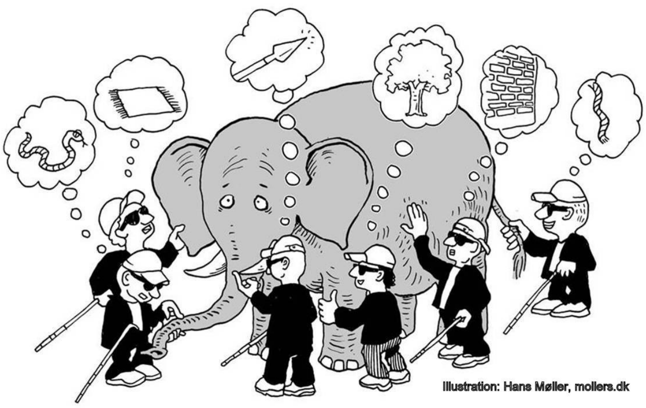
IRyA/UNAM

Recall: Populations and samples

If an experiment results in a random variable X whose probability distribution is $P_{X3pt}(x)$ (discrete) or $p(X)$ (continuous), we say that X is drawn from the PMF/PDF: $X \sim P(X)$ or $X \sim p(X)$.

Population: the underlying probability distribution.
Sample: the results of a finite number of experiments/draws from the population (a subset).

Due to the fact that we perform the experiment a finite number of times, the sample may not be able to faithfully reproduce the population – sample statistics (statistical quantities derived from the sample) are only guesses at the corresponding population statistics (values derived from the population).



"More data is required."

Convention: Greek symbols for population statistics (*e.g.*, μ, σ) and Latin symbols for sample statistics (*e.g.*, $\bar{x}$, s).

Recall: Variance

The mean ($E[X]$) is a measure of central tendency of the distribution. We can also quantify the spread of the distribution around this mean, in terms of the deviations $X_i - E[X]$ for each outcome X_i.

However, from the definition of the mean, the sum of deviations is always zero, so we look at either the absolute deviation or the squared deviation.

The variance is the expectation value of the squared deviation (the "mean squared deviation"): $Var(X) = E[(X - E(X))^2] = E[X^2] - (E[X])^2$.
The standard deviation is the square root of the variance ("root mean square deviation"): $\sigma = \sqrt{Var(X)}$.

Some properties of the variance:

1. By definition, non-negative.
2. For any constant α:
   1. $Var(\alpha) = 0$, because $E[\alpha] = \alpha$.
   2. $Var(X + \alpha) = Var(X)$ – *i.e.*, invariant w.r.t. a location parameter.
   3. $Var(\alpha X) = \alpha^2 Var(X)$.
3. For constants α, β and random variables X, Y,
   $Var(\alpha X + \beta Y) = ??$
   Evaluate this expression using the definition of variance in terms of expectation values.

Recall: Covariance and correlation coefficient

Definition (Covariance)

The covariance is a measure of joint variability of two random variables:
$Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$.
From the definitions, $Cov(X, X) = Var(X)$.

Therefore, $Var(\alpha X + \beta Y) = \alpha^2 Var(X) + \beta^2 Var(Y) + 2\alpha\beta Cov(X, Y)$.
If the two variables are uncorrelated, then the third term vanishes. $Cov(X, Y)$ is not scale-invariant, so:

Definition ((Pearson's) Correlation coefficient)

$$\rho_{XY} = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

A correlation coefficient of +1 (-1) signifies perfect (anti)correlation.

Recall: The mean of N uncorrelated random variables

Bienaymé formula
The variance of the sum of N uncorrelated variables is therefore the sum of their variances:
$$Var\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} Var(X_i).$$

Consider N iid random variables. Using the Bienaymé formula to compute the variance in their mean,
$$Var(\bar{X}) = \frac{1}{N^2} Var\left(\sum_{i=1}^{N} X_i\right) = \frac{1}{N^2} \sum_{i=1}^{N} Var(X_i) = \frac{1}{N^2} N\ Var(X) = \frac{\sigma^2}{N}$$

As the number of measurements N increases, the variance on the mean of these N measurements decreases so that that sample mean starts to approach the population mean.

Recall: Binomial Distribution

The probability distribution of the number of successes in a sequence of n independent experiments, with a single success having probability p. A single success in this case is a Bernoulli trial (the Bernoulli Distribution is a special case of the Binomial Distribution with $n = 1$).
The probability of k successes (and $n - k$ failures) in n trials, and therefore the probability distribution, is given by
$$P(X = k) = \binom{n}{k} p^k (1 - p)^{(n-k)} \text{ (Binomial distribution)}$$

Examples of the Binomial Distribution:
The number of heads obtained in n tosses of a fair coin = Binomial$(n, p = \frac{1}{2})$.
The number of "point" masses in a volume fraction V_1/V of space with N points in volume V = Binomial$(N, p = \frac{V_1}{V})$ (Meszaros, A. 1997 A&A 328, 1).

Mean: $E[X] = np$
Variance: $Var[X] = np(1 - p)$
Both are n times the values for the Bernoulli distribution!

Group assignment: Binomial distribution

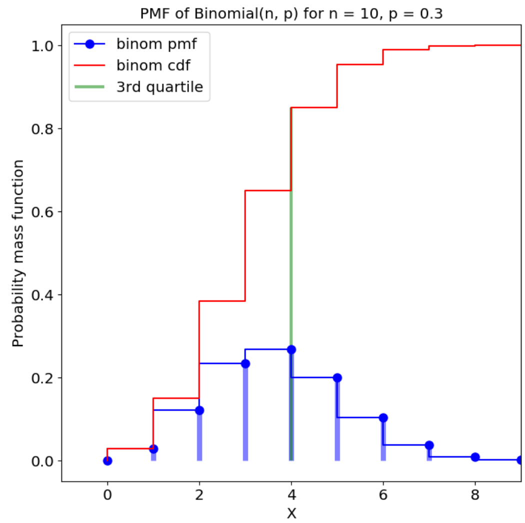
Use the python module `scipy.stats.binom` to answer the following:
1) Plot the PMF and CDF for $n = 10, p = 0.3$.
2) From the plot, what is the approximate location of the third quartile? Is this consistent with the output from the `ppf` method of `scipy.stats.binom`?

Third quartile $= x_3$, such that $CDF(X = x_3) = 0.75$.

But discrete distribution $\Rightarrow$ we need $CDF(X = x_3) \geq 0.75$.

From plot, $CDF(X = 3) \approx 0.65$ and $CDF(X = 4) \approx 0.85$, so $3 < x_3 < 4$.

Check: `scipy.stats.binom.ppf(0.75, n, p)` returns same value!



Code for plot available here.

Poisson distribution = Binomial distribution with small p and large N

A binomial distribution of rare events that are independent of each other and occur at a constant average rate (the average number of events in a fixed number of trials – or per interval – is constant) is called a Poisson distribution. The number of trials in this case is large compared to the number of successes.That is, $p \ll 1$ and $n \gg k$, such that np is finite.

Deriving the PMF: Rewrite the Binomial distribution using $\lambda \equiv np$ (recall: $E[X] = np$ for a Binomial distribution):

$$Binomial(n, k, p) = \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1 - \frac{\lambda}{n}\right)^{n-k} = \lambda^k \times \binom{n}{k} n^{-k} \times \left(1 - \frac{\lambda}{n}\right)^{-k} \times \left(1 - \frac{\lambda}{n}\right)^{n}.$$

Apply the limit $n \to \infty$ to each of the orange terms above:

$$\lim_{n\to\infty} \binom{n}{k} n^{-k} = \frac{1}{k!} \lim_{n\to\infty} \frac{n!}{n^k(n-k)!} = \frac{1}{k!} \lim_{n\to\infty} \frac{n(n-1)(n-2)...(n-(k-1))}{n^k} = \frac{1}{k!},$$

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{-k} = 1, \text{ and } \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{n} = e^{-\lambda}.$$

The Poisson distribution is therefore $Poisson(k; \lambda) = \lambda^k \dfrac{e^{-\lambda}}{k!}$

Moments of the Poisson distribution (contd.)

$$E[X] = \sum_{k=0}^{\infty} k\lambda^k \frac{e^{-\lambda}}{k!} = \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{k\lambda^{k-1}}{k!} = \lambda e^{-\lambda} \frac{\partial}{\partial\lambda}\left(\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\right) = \lambda e^{-\lambda} \frac{\partial}{\partial\lambda} e^{\lambda}$$
$$= \lambda \; (= np, \text{ same as the Binomial distribution!}).$$

Similarly, $Var(X) = \lambda$ (Binomial: $Var(X) = np(1-p)$, with $p \ll 1$).
Therefore, a measurement of N Poisson events has associated with it a standard deviation (an "uncertainty") of $\sqrt{N}$.
This is used to determine the uncertainty in photons received from a source, the surrounding background, and also to compute the uncertainty due to dark current.

Thus, the parameter λ is interpreted as the mean number of events in an interval.
The Poisson distribution describes the probability of a given number of events in a fixed interval, given that the events (a) are independent of each other and (b) occur at a constant rate (*i.e.*, the average rate is the same independent of the location of the interval).

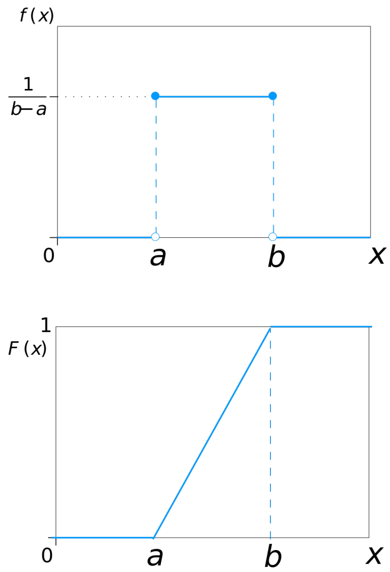
Examples: The probability that (a) a mag 7.0 earthquake hits Mexico City within the next ten years, (b) two supernovae will go off in the Milky Way within the next 100 years, (c) 3 photons from a target will hit a telescope detector within the next second, or (d) A sample containing ^{137}Cs nuclei will produce 15 decays in the next minute. The first application of the Poisson distribution was to the Prussian army's "death by horse kick" data.

Continuous probability distributions

Uniform distribution

Probability per unit interval = constant $\Rightarrow p(x) = \frac{1}{b-a}\mathbb{I}_{\{a \leq x \leq b\}}(x)$

If X is a uniform random variable, then $X \sim U[a,b] = (b-a)Y + a, \text{where } Y \sim U[0,1]$.
(range has to be finite if total probability is normalised!)



PDF (top) and CDF (bottom).
Credit:user:IkamusumeFan/CC BY-SA 3.0

$$E[X] = \frac{a+b}{2}$$

$$Var(X) = \frac{1}{12}(a+b)^2$$

CDF: $F(x) = \frac{x-a}{b-a}$

Median: $x_m \text{ such that } F(x_m) = 0.5$.

From symmetry of PDF, $x_m = E[X] = \frac{a+b}{2}$.

Exponential distribution

A process in which events are independent of each other and occur at a constant average rate is called a Poisson point process. The number of such events in a given interval is a Poisson random variable (it follows a Poisson distribution) and a sequence of such variables is a Poisson process.

A variable that measures the interval between two events in a Poisson point process follows the Exponential distribution.

The Exponential distribution has pdf $p(X = x) \equiv Exp[\lambda] = \lambda e^{-\lambda x}$.
$E[X] = \frac{1}{\lambda}$ and $Var(X) = \frac{1}{\lambda^2}$
(The mean interval between successive Poisson events is equal to the inverse of the mean rate at which the events occur.)

Normal distribution

The pdf of the normal (or Gaussian) distribution is given by

$N(\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$.

The standard version of the distribution, centered at $x = 0$ with standard deviation $= 1$, is obtained by setting $\xi = \frac{x-\mu}{\sigma}$: $\varphi(\xi) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{\xi^2}{2}\right] \Longrightarrow \mathrm{N}(\mu, \sigma) = \frac{1}{\sigma}\, \varphi\left(\frac{x-\mu}{\sigma}\right)$.

$Z \sim \varphi(z) \iff \mathrm{X} \equiv \sigma Z + \mu \sim N(\mu, \sigma)$ (standard normal deviate and normal deviate).

The CDF of the Standard normal distribution, $\Phi(x)$, is given by

$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} \exp\left[-\frac{t^2}{2}\right] dt = \frac{1}{2}\left[1 + erf\left(\frac{x}{\sqrt{2}}\right)\right]$,

where $erf(x)$ is the error function: $erf(x) = \frac{2}{\sqrt{\pi}} \int\limits_{0}^{x} \exp\left[-t^2\right] dt$.

Central Limit Theorem

Consider a sequence $\{X_i\}, i = 1, 2, \cdots, n$ of iid* random variables drawn from some distribution with mean μ and variance σ^2. The sample mean is $S_n = \sum\limits_{i=1}^{n} X_i/n$. How well does this sample mean estimate the population mean μ?

If we estimate the sample mean M times (*i.e.*, we generate M samples of m numbers each and compute the sample mean each time), how are the sample means distributed?

Let us first define a scaled version of the sample mean: $s_n = (S_n - \mu)/\sigma(S_n)$. Since the X_i are iid variables, $\sigma(S_n) = \sigma/\sqrt{n}$ (Bienaymé Formula). Therefore, $s_n = \sqrt{n}(S_n - \mu)/\sigma$.

The Central Limit Theorem (CLT) says that, for large n, this quantity approaches the standard normal distribution. That is,

$\lim\limits_{n\to\infty} \sqrt{n}\frac{(S_n - \mu)}{\sigma} = N(0, 1) = \varphi(s_n)$.

In terms of the CDF, $\lim\limits_{n\to\infty} P(\sqrt{n}\frac{(S_n - \mu)}{\sigma} \leq z) = \Phi\left(\frac{z}{\sigma}\right)$.