Statistics for Astronomers: Lecture 06, 2019.02.25

Prof. Sundar Srinivasan

IRyA/UNAM

Recall: Poisson distribution

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.

The Poisson distribution is therefore $Poisson(k;\lambda) = \lambda^k \dfrac{e^{-\lambda}}{k!}$
Moments: $E[X] = \lambda$ and $Var(X) = \lambda$.
The parameter λ is interpreted as the mean number of events in an interval.

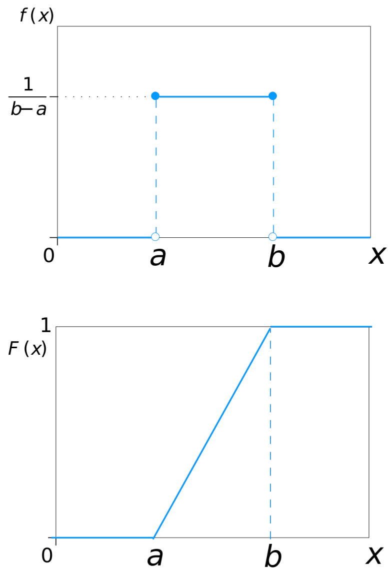
The Poisson distribution describes the probability of a given number of events in a fixed interval, given that the events (a) are independent of each other and (b) occur at a constant rate (*i.e.*, the average rate is the same independent of the location of the interval).

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.

Recall: Uniform distribution

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

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



PDF (top) and CDF (bottom).

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

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

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

Median: x_m such that $F(x_m) = 0.5$.

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

Recall: 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.)

Recall: Normal distribution

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

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

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

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

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

$\Phi(x) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{x} \exp\Big[-\dfrac{t^2}{2}\Big] dt = \dfrac{1}{2}\Big[1 + erf\Big(\dfrac{x}{\sqrt{2}}\Big)\Big]$,

where $erf(x)$ is the error function: $erf(x) = \dfrac{2}{\sqrt{\pi}} \displaystyle\int_{0}^{x} \exp\Big[-t^2\Big] dt$.

Recall: 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 $S_n = \sum_{i=1}^{n} X_i/n$. 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,

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

In terms of the CDF, $\lim\limits_{n\to\infty} P\big(\sqrt{n}\dfrac{(S_n-\mu)}{\sigma} \le z\big) = \Phi\Big(\dfrac{z}{\sigma}\Big)$

Midterm exam

Tentative date: Monday, 25th March (about a month from now).

Functions of random variables

Distributions of functions of random variables - I

Often, we draw random variables from known distributions, and manipulate them in some form. In general, the distribution of the result may not be the same.

Although in many cases it's not easy, there are a few ways to figure out the final distribution.

If $X \sim p_X(x)$ and $Y = f(X)$, what is $p_Y(y)$?

If $X, Y \sim p_{XY}(x, y)$ and $Z = f(X, Y)$, what is $p_Z(z)$?

More complicated cases...

Distributions of functions of random variables - II

Method 1 (derivative of inverse function): $Y = f(X) \implies X = f^{-1}(Y)$.

For any intervals dx and dy, we have that $p_X(x)dx = p_Y(y)dy$. $\implies p_Y(y) = \dfrac{p_X(f^{-1}(y))}{\left|\dfrac{dy}{dx}\right|}$

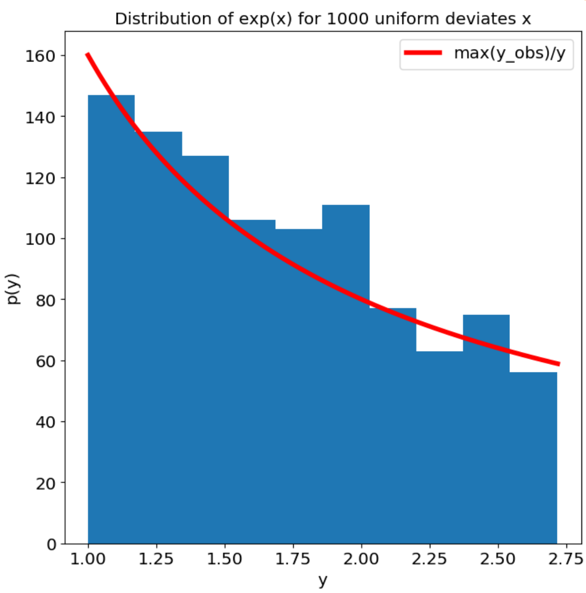
Example: exponential of a uniform deviate

$X \sim U[0, 1]$ and $Y = e^{-X}$.

$p_X(x) = 1$.

$\left|\dfrac{dy}{dx}\right| = y$.

$\implies p_Y(y) = \dfrac{1}{y}, \quad 1 \leq y \leq e$.



Code for plot available here.

Distributions of functions of random variables - III

Method 2 (convolution): (valid for linear combinations) If $f(X, Y)$ is linear in both X and Y, and if $Z = f(X, Y)$, we can write the pdf of Z as a convolution of the pdfs of X and Y. First, write one of the original variables as a function of the other two: $Y = g(X, Z)$. Use this to rewrite the convolution:

$$p_Z(z) = \int_{-\infty}^{\infty} dx\ p_X(x)\ p_Y(y(x, z))$$

Example: sum of two uniform deviates

$X, Y \sim U[0, 1]$ and $Z = X + Y \Longrightarrow Y = Z - X$.

x and y go from 0 to 1, so z goes from 0 to 2. Split this interval into [0, 1] and [1, 2].

For $z \in (0, 1)$, we need $z - x \geq 0$, so $x < z$: $\int_0^z dx\ p_X(x)\ p_Y(z - x) = \int_0^z dx = z$.

For $z \in (1, 2)$, we need $z - x \leq 1$, so $x > z - 1$: $\int_{z-1}^{1} dx\ p_X(x)\ p_Y(z - x) = \int_{z-1}^{1} dx = 2 - z$.

$$\Longrightarrow p_Z(z) = \begin{cases} z & : 0 < z \leq 1 \\ 2 - z & : 1 < z \leq 2 \end{cases}$$

Can be extended to a linear combination of N random variables!

Group assignment

Recall: If $X \sim U[0, 1]$, then $E(X) = 0.5$, $Var(X) = 1/12$.

If $N = 2$ random numbers are generated from $U[0, 1]$ and their mean is computed, what are $E[\bar{X}]$ and $Var(\bar{X})$?

If $X_1, X_2 \sim U[0, 1]$,

$$E[\bar{X}] = E\left[\frac{X_1 + X_2}{2}\right] = \frac{1}{2}\Big(E[X_1] + E[X_2]\Big) = \frac{1}{2}\left(\frac{1}{2} + \frac{1}{2}\right) = \frac{1}{2}$$

$$Var(\bar{X}) = Var\left[\frac{X_1 + X_2}{2}\right] = \frac{1}{2^2}\Big(Var(X_1) + Var(X_2)\Big) = \frac{1}{4}\left(\frac{1}{12} + \frac{1}{12}\right) = \frac{1}{12N} \text{ for } N = 2$$

(We could've also used the Central Limit Theorem to get the same results)

Generate $N = 2$ uniform random numbers from $U[0, 1]$ and find their mean. Do this $M = 1000$ times and plot the sampling distribution.

For comparison, overlay the Normal distribution that has the same expectation value and variance.

Repeat for $N = 3$ and $N = 10$.

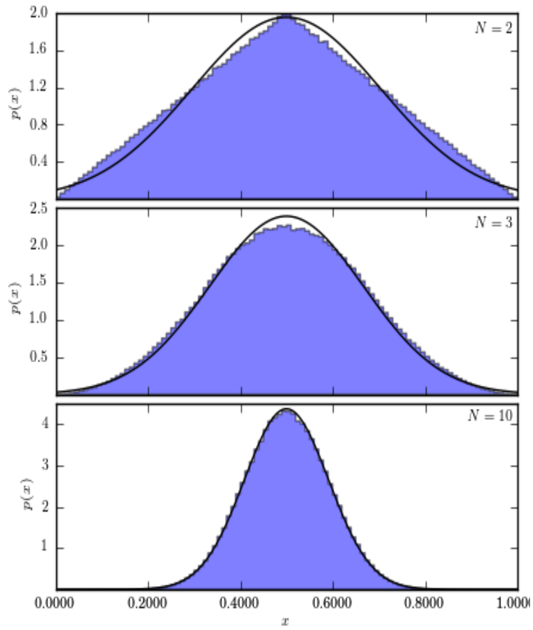


Fig 3.10 from Ivezić+'s (AstroML) book: realisations for $N = 2$ (top), 3 (centre), and 10 (bottom).

Group assignment: one possible solution

To avoid `FOR` loops, vectorise the code. For example, you could use `numpy.random.uniform` to generate a 2×1000 vector:

```
x = numpy.random.uniform(0., 1., size = (2, 1000))
y = numpy.mean(x, axis = 0)
```

The histogram for y is the required result.

Introduction to statistical inference

What should we expect of our statistics?

(Wall & Jenkins Sec. 3.2)

Recall: Populations are summarised by parameters and samples are summarised by statistics.
A statistic is a summary of a sample drawn from the underlying population, and it is an estimate of a parameter.

Some characteristics of a "good" statistic:

1. Efficiency: reproduce parameter with as few samples as possible.
2. Robustness: reproduce parameter accurately by being insensitive to outliers in the sample.
3. Lack of bias: expectation value of the statistic = true parameter value.
4. Consistency: reproduces true parameter value for very large sample size.

Describing probability distributions

(Ivezić et al., "Statistics, Data Mining, and Machine Learning in Astronomy")

A distribution can be described using parameters that describe:

1. location (*e.g.*, $E[X]$, median),
2. scale/width/spread (*e.g.*, σ, $Var(X)$, MADM, interquartile range (IQR)),
3. shape (e.g., skewness, kurtosis), and
4. the CDF: p% quantile (p is called percentile). q_p such that

$$\frac{p}{100} = \int_{-\infty}^{q_p} p(x)dx$$

The median is an example of such a quantile (q_{50}).

Statistics involving quantiles (median, MADM, IQR, percentiles) are robust to outliers, but may be less efficient.

Versions of the above can also be computed from samples. Two important examples: sample mean ($\bar{x}$) and sample standard deviation (s).

Sample mean and sample variance

If unknown, the population mean can be estimated from the sample: $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

We can compare a single observation x_i to two quantities: the population mean and the location estimate using the sample.
Error: deviation from the population mean, $x_i - \mu$ (in general, called bias).
Turns out $E[X - \mu] = 0$ (CLT), so $\bar{x}$ is an unbiased estimator of μ.
Residual: deviation from a sample-based estimate of location m: $x_i - m$.

It can be shown that the sum of squares of residuals, $\sum_{i=1}^{N} \left(x_i - m\right)^2$, is minimum for $m = \bar{x}$.
The variance, therefore, is defined as the expectation value of the squares of residuals from the sample mean.

While the sample mean has N degrees of freedom (it is based on x_i measurements), the $\#dof$ for the variance depends on whether the population mean was given ($\#dof = N$) or had to be estimated from the data ($\#dof = N - 1$).

Bessel's correction for sample variance

The sample variance should be defined as $E[(\text{residuals from sample mean})^2]$. In terms of the x_i, this becomes

$$Var(X) = \frac{1}{N}\sum_{i=1}^{N} \left(x_i - E[X]\right)^2 = \frac{1}{N}\sum_{i=1}^{N} \left(x_i - \mu\right)^2$$

This form of the variance has N degrees of freedom.

However, as on the previous slide, if μ was unknown and had to be estimated from the sample, this is one constraint on the N x_i: $\frac{1}{N}\sum_{i=1}^{N} x_i = \text{constant}$.

The sample variance now only has $N-1$ degrees of freedom, so the correct expression should be

$$Var(X) = \frac{1}{N-1}\sum_{i=1}^{N} \left(x_i - E[X]\right)^2 = \frac{1}{N-1}\sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2$$ (Bessel's correction).

The Wikipedia article on Bessel's correction shows a few ways to prove that this is the correct form of the sample variance.
Note that this form/correction is only required if μ is unknown.

Estimators

Values of parameters can be guessed from finite samples by computing statistics called estimates. The "rule" that specifies how to compute these estimates are called estimators. There are estimators for point as well as interval estimates (later).

Notation Parameter: θ. Estimator for θ: $\hat{\theta}$.
If X is a random variable, $\hat{\theta}(X)$ is a function of the variable and $\hat{\theta}(x)$ is the value of $\hat{\theta}(X)$ for $X = x$.

As with a single sample point, we can compare the estimate to the parameter being estimated, or to the expectation value of the estimate:
(Parameter) Error: $e(x) = \hat{\theta}(x) - \theta$. We can estimate $E[e]$ and $E[e^2]$:

$$E[e] = E[\hat{\theta}(x) - \theta] \equiv \text{Bias}.$$

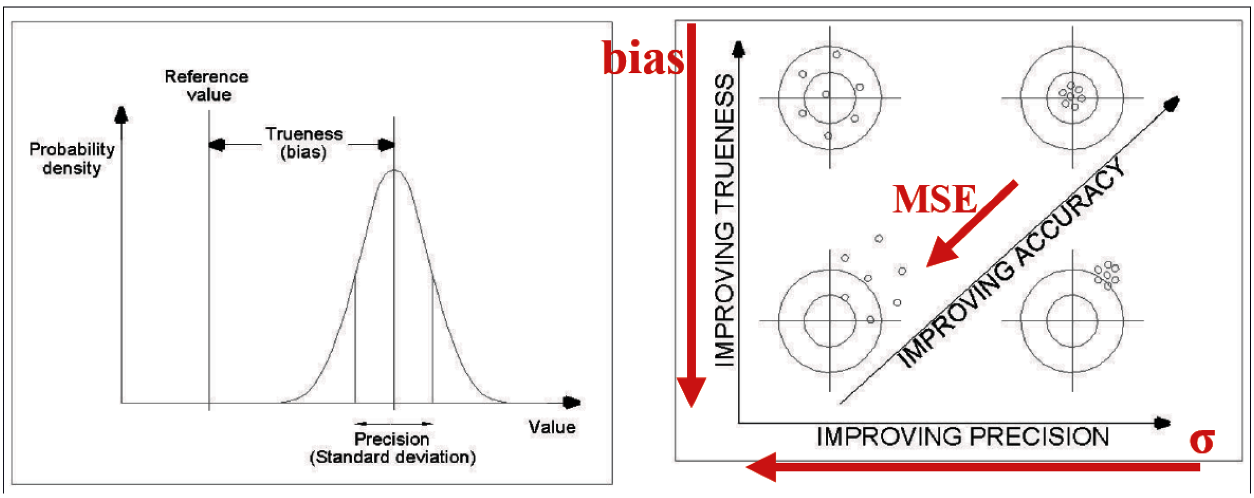
$$E[e^2] = E[(\hat{\theta}(x) - \theta)^2] \equiv \text{Mean square error (MSE)}.$$

(Sampling) Deviation: $d(x) = \hat{\theta}(x) - E[\hat{\theta}(x)]$. $E[d] = 0$, but we can estimate $E[d^2]$:

$$E[d^2] = E[(\hat{\theta}(x) - E[\hat{\theta}(x)])^2] \equiv \text{Variance}.$$

We can show that $\mathrm{MSE}(\hat{\theta}) = V(\hat{\theta}) + B(\hat{\theta})^2$.

Estimators (contd.)



Source: Ivezić+ AstroML book.