Statistics for Astronomers: Lecture 04, 2019.02.18

Prof. Sundar Srinivasan

IRyA/UNAM

Recall: Bayesian inference versus frequentist inference

Bayesian inference relies on assuming prior knowledge of the hypotheses/parameters of interest. In the Bayesian interpretation, probability is a *degree of belief*. In this sense, Bayesian probabilities are subjective.

Frequentist inference is, by contrast, considered objective as it does not incorporate/assume priors for the parameters that are the subject of inference.

Recall: Random variables

Random: Uncertain, no "pattern" can be detected.

Randomness: A measure of uncertainty of the outcome of an experiment. Some sources of "true" randomness – initial conditions of the experiment (*e.g.*, throwing dice, chaos) and environmental effects (*e.g.*, Brownian Motion, dark current).

Random variable: A function that assigns a numerical value to each distinct outcome. Allows us to compute probabilities.

Random process: A sequence of random variables whose outcomes don't follow a pattern. Their evolution can, however, be described probabilistically.
Each observation results in a random variable associated with the process.
The collection of random variables in such a process have two attributes:
an index drawn from an index set, and a numerical value drawn from a state space.
Each outcome is mapped to a unique element of the index set, and each unique outcome is mapped to a unique value in the state space.

Random variable: A function that maps the sample space to the state space; $X : \Omega \to S$.

Probability distribution: A function that maps a random variable to a real number; $p : X \to \mathbb{R}$.

Recall: Discrete and continuous probability distributions

A discrete distribution is also called a probability mass function (PMF).
The probability of the random variable spanning a range of values is the sum of the PMFs for each value. In the two-dice example from the previous slide, $P(4 \leq X \leq 7) = \sum_{x=4}^{7} P(X = x)$.
The PMF for a single value of the random variable, $X = x$, is equal to the probability that $X = x$.

Continuous distributions are called probability density functions (PDFs).
The probability of the random variable spanning a range of values is the integral of the PDF over the range. For example, if $p(X)$ is the PDF,

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(x) dx.$$

While the PDF for a single value of the random variable, $X = x$, may be nonzero, the probability that $X = x$ is zero ($x_1 = x_2$ in the integral above).

Notation: I will use P for total probability or PMF, and p for the PDF.
Note: I won't abbreviate "probability distribution function", so that "PDF" is unambiguous.

Recall: Cumulative distribution function (CDF)

Definition (Cumulative distribution function)

A function $F_X(x)$ of a random variable X such that $F_X(x)$ is the probability that $X \leq x$.

Definition (Quantile function)

The inverse of the CDF, a function $Q(p)$ that returns the value of x such that $F_X(X \leq x) = p$.
e.g., $Q(p = 0.5)$ is the median (equal "mass" on either side of $x = Q(0.5)$).
$Q(p = 0.25)$ and $Q(p = 0.75)$ are the first and third quartiles.

Recall: Expectation value

Definition (Expectation value)

The expectation value of any function $g(X)$ of a random variable X, represented by $E[g(X)]$, is the weighted average of $g(X)$, with the weights being the associated probabilities:

$$E[g(X)] = \sum_{i=1}^{N} g(x_i)\ P(X = x_i) \text{ (discrete case)}$$

$$E[g(X)] = \int_{t=-\infty}^{t=\infty} g(x)\ p(x)\ dx \text{ (continuous case)}$$

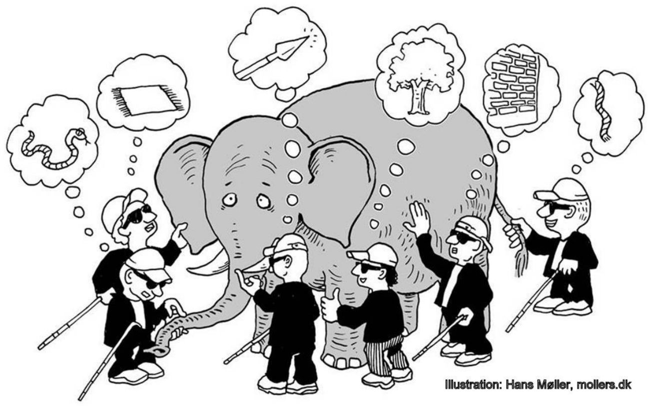
The expectation value is a linear operator, so that, for any two functions $g(X)$ and $h(X)$ and scalars α and β,
$E[\alpha g(X) + \beta h(X)] = \alpha E[g(X)] + \beta E[h(X)]$.

Populations and samples

If an experiment results in a random variable X whose probability distribution is $P_X(x)$ (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).



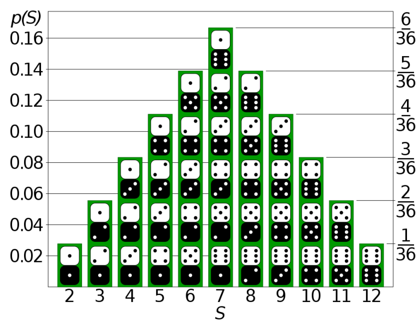
"More data is required."

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

Populations and samples (contd.)

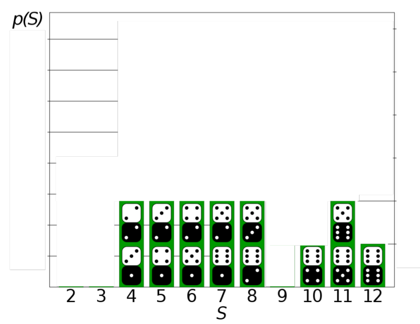
Consider the two-dice example from before:

Population
$\mu = 7.00$
$\sigma = 2.41$
$(= \sqrt{Var(X)})$



(Tim Stellmach/Public Domain)

Sample
$\bar{x} = 7.43$
$s = 2.58$
*discussed later!



Possible sample from 14 throws of the dice.

The sample distribution almost seems uniform. This skews both the mean and the variance to larger values in comparison to the population statistics.

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.

Covariance and correlation coefficient

From previous slide, $Var(\alpha X + \beta Y) = \alpha^2 Var(X) + \beta^2 Var(Y) + 2\alpha\beta E[(X - E[X])(Y - E[Y])]$.
What is the interpretation of the third (mixed) term?

Consider:
If $X \perp Y$, then $(Y - E[Y])$ should not depend on $(X - E[X])$ for any given (X, Y) pair $\Rightarrow$ the term quantifies a dependence between X and Y.
If X and Y are not independent, then the product $(X - E[X])(Y - E[Y])$ is positive if both deviations are in the same direction, and negative if the variables deviate from their means in opposite directions.

The third term is the average of these products of X and Y deviations.

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.

Covariance and correlation coefficient (contd.)

The sign of $Cov(X, Y)$ probes a linear relationship between the two variables X and Y.

Example (Bernoulli random variable)

$Cov(X, X) = p(1-p) > 0$, and $Cov(X, \frac{1}{X+1}) = -\frac{p(1-p)}{2} < 0$.

However, the magnitude isn't as useful, as the covariance is not scale-invariant: $Cov(\alpha X, \beta Y) = \alpha\beta Cov(X, Y)$.

We can define a scale-invariant of $Cov(X, Y)$ instead:

Definition ((Pearson's) Correlation coefficient)

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

In terms of ρ, we can reexamine the above Bernoulli example:
$\rho_{XX} = 1$. A correlation coefficient of $+1$ signifies perfect correlation and a value of -1 means perfect anticorrelation.

Uncorrelated random variables

If random variables X and Y are such that $Cov(X, Y) = 0$, then they are said to be uncorrelated.

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,

$$Var(\bar{X}) = \frac{1}{N^2} Var\left(\sum_{i=1}^{N} X_i\right) = \frac{1}{N^2}\sum_{i=1}^{N} Var(X_i) = \frac{1}{N^2} N Var(X) = \frac{\sigma^2}{N}$$

Example: two-dice problem again. X and Y are random variables corresponding to the values on the first and second dice after each throw. We record $X + Y$ each time. A single measurement of this sum has variance $= 5.83$ (homework!). The average of ten measurements has variance $\frac{5.83}{10} = 0.583$. 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.

Some standard probability distributions

Recall: Bernoulli Distribution

A Bernoulli random variable takes one of only two values: 1 and 0, with probabilities p and $1 - p$ respectively. It is the result of an experiment that asks a single yes-no question.
The variable therefore has state space $S = \{1, 0\}$, with an associated probability distribution given by

Definition (Bernoulli Distribution)

$P(X = 1) = p$ and $P(X = 0) = 1 - p$ (Bernoulli Distribution). We can abbreviate this:
$P(X = x) = p^x (1 - p)^{1-x}\ \mathbb{I}_{x\in\{0,1\}}(x)$, with $\mathbb{I}(x)$ the Indicator (or Heaviside) function.

Example of a Bernoulli random variable: outcome of tossing a single (not necessarily fair) coin.

Mean: $E[X] = 1 \times P(X = 1) + 0 \times P(X = 0) = 1 \times p + 0 \times (1 - p) = p$
Variance: First, $E[X^2] = 1^2 \times P(X = 1) + 0^2 \times P(X = 0) = 1^2 \times p + 0^2 \times (1 - p) = p$
$\Rightarrow Var[X] = E[X^2] - (E[X])^2 = p - p^2 = p(1 - p)$

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 state space $S = \{0, 1, 2, 3, ..., k-1, k\}$. 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: $E[X] = np$ (demonstrated on following slide)

Variance: $Var[X] = np(1-p)$

Both are n times the values for the Bernoulli distribution!

Expectation value of a Binomial Distribution

Recall:

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \quad (1) \qquad k\binom{n}{k} = n\binom{n-1}{k-1} \quad (2)$$

Expectation value for the binomial distribution:

$$\begin{aligned}
E(X) &= \sum_{k=0}^{n} k\binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} k\binom{n}{k} p^k (1-p)^{n-k} \text{ ($k = 0$ term vanishes)} \\
&= \sum_{k=1}^{n} n\binom{n-1}{k-1} p^k (1-p)^{n-k} \text{ (using Equation (2))} \\
&= \sum_{s=0}^{n-1} n\binom{n-1}{s} p^{s+1} (1-p)^{n-s-1} \text{ (setting $s = k-1$)} \\
&= np\sum_{s=0}^{n-1} \binom{n-1}{s} p^{s} (1-p)^{n-1-s} = np \text{ (using Equation (1))}
\end{aligned}$$

Similarly, we can compute $Var[X]$ using

$$k(k-1)\binom{n}{k} = n(n-1)\binom{n-2}{k-2}$$