MIE237

Neil Montgomery

2016-01-08

more review

Covariance (more on this later)

Covariance is hard because it's a property of the *joint distribution* of two random variables X and Y:

$$\mathrm{Cov}(X, Y) = \mathrm{E}((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))$$

But it's nice to have an informal sense of its meaning. The main thing is to have a concept of *positive* versus *negative* covariance.

Positive (negative) covariance means larger values of X tend to coincide with larger (small) values of Y.

Positive covariance: Suppose we plan to pick a student at random from the class and measure their height X and weight Y.

Negative covariance: Suppose we plan to pick two students at random from the class and measure the height of the first X_1 and the height of the second X_2.

Important terminology to keep straight

Word	Informal meaning	Mathematical model
Population	"All possible numerical outcomes of the random process under consideration"	Random variable X with some distribution
Sample	"A subset of the population"	$X_1, X_2, \ldots, X_n$ independent with the same distribution
Dataset	$x_1, x_2, \ldots, x_n$, a *realization of a sample*	Nothing to model

Example

"Amount of iron in engine oil from a fleet of haul trucks, in parts per million."

Word	Informal meaning	Mathematical model
Population	All possible values of iron in oil in ppm.	$X \sim N(\mu, \sigma^2)$
Sample	A plan to sample oil readings from the trucks	$X_1, X_2, \ldots, X_n$ independent $N(\mu, \sigma^2)$
Dataset	24, 31, 2, 14, 4, … 9	

"Statistics"—the general problem and an approach to its solution

The distrubution of the random variable used to model the population might be only partly specified.

- $N(15, 4)$
- $N(\mu, 4)$
- $N(\mu, \sigma^2)$
- "Some distribution with a mean and a variance."

Problem: get the missing info.

Solution: plan to gather a sample $X_1, \ldots, X_n$

Use probability to make inferences about the unknowns.

"Parameter"

These unknown constants are to be called *parameters*.

Examples: $N(\mu, \sigma^2)$, $Exp(\lambda)$, Binomial(n, p).

(I've found engineers sometimes use "variable" for this concept and use "parameter" to mean "a thing related to what I'm interested in". We will usually use "variable" to mean "a column from a dataset".)

"Sample" and "Statistic"

A sample is a sequence of random variables (often independent and identically distributed) $X_1, \ldots, X_n$ with n as the *sample size*. I suggest *conceptualing* a "sample" as a *plan to gather data*.

A *statistic* is any function of the sample. So it follows that a statistic is a *random variable*. So statistics have their own distributions, means, variances etc.

Examples:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \text{Sample average}$$

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}{n-1} \qquad \text{Sample variance}$$

Common point of confusion

A *sample* is a plan to gather data. An actual dataset is the realization of this plan.

But a *dataset* is not random. It is a collection of constants—what was actually observed.

Say a dataset has a variable named x with observed values $x_1, \ldots, x_n$.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{(observed) sample average}$$

$$s^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1} \qquad \text{(observed) sample variance}$$

Key facts about $\overline{X}$

Mean and variance of $\overline{X}$

For *any* sample $X_1, \ldots, X_n$ that are independent and have the same distribution with mean μ and variance σ^2:

$$\mathrm{E}\left(\overline{X}\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(X_i\right) = \frac{1}{n}n\mu = \mu$$

$$\mathrm{Var}\left(\overline{X}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(X_i\right) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

$$\mathrm{SD}\left(\overline{X}\right) = \frac{\sigma}{\sqrt{n}}$$

IF the sample is from a normal distribution

Suppose further that $X_1, \ldots, X_n$ is a sample from a $N(\mu, \sigma^2)$ distribution. The following (equivalent) statements are then exactly true:

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$$

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

But even if the sample is *not* from a normal distribution…

"Fundamental Theorem of Statistics"

For population modeled with a distribution with mean μ and variance σ^2. Consider a sample $X_1, \ldots, X_n$.

$$P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq u\right)$$

$$\lim_{n\to\infty} P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq u\right)$$

$$\lim_{n\to\infty} P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq u\right) = P(Z \leq u)$$

where $Z \sim N(0,1)$. (Real name: central limit theorem "CLT")

The real reason why the CLT is so important

The convergence is really fast, so that:

$$P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq u\right) \approx P(Z \leq u)$$

for n "large enough", which depends mostly on the *skewness* of the underlying distribution.

n large enough	2	10	30	60
underlying shape	Normal	Symmetric	Some kewness	More skewness

Summary of exact and approximate results

Statistic	Exact distribution if population is normal	"for n large enough"
$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$	$N(0,1)$	$N(0,1)$
$\frac{(n-1)S^2}{\sigma^2}$	χ^2_{n-1}	???
$\frac{\overline{X}-\mu}{S/\sqrt{n}}$	t_{n-1}	t_{n-1}
$\frac{S_1^2/\sigma_1^2}{S^2/\sigma_2^2}$	$F_{n-1,m-1}$	???