



Statistics Unit 1

Topic: Variance, inference & sampling

- Date: @February 22, 2024
- Lane: *Probability, Statistics*

What it covers

- Why variation is *the* foundational concept in statistics
- Expectation and Moments. Why central moments
- Algebra of variance. Why this measure of dispersion?
- Inference from samples
- Variance and the normal distribution $N(\mu, \sigma^2)$

Requires

Random variables, expectation distributions.

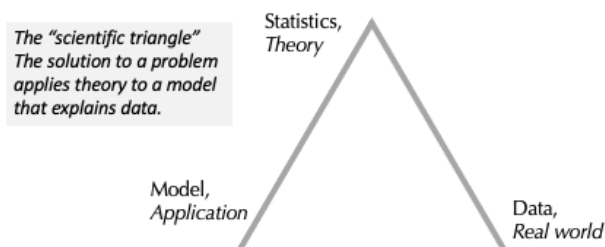
Required by

Testing, Estimation

Density estimation

Statistics - The quantification of errors

The key premise of statistics is that a *measure of accuracy of a dataset can be inferred from the data itself*. This is a strong assumption.



Statistical Inference posits an unobservable, "latent", certain quantity that is the source of stochastic (e.g. uncertain) *observable* quantities. *Inference* is the process of computing the value of the likely unobservable quantity from the set of observed values. For example, assuming a binomial distribution, $k \sim \text{Binomial}(n, p)$ for k successes out of n attempts, can one infer the p of the distribution?

[inference picture of observables v/s non-observables]

The unobservables are expressed as expected values, *estimation* is the method to infer them from sample *statistics*. It is possible because of the way statistics converge to the expected value as the size of the sample increases.

Moments of a distribution

Moments generalize the expected value of an r.v. to the powers of the expected value:

$$E[X^n] = \int x^n f(x) dx$$

Any distribution is fully determined by its series of moments.

Central moments are *standardized* moments, by making them invariant to the mean - the first moment: $\mu = E[X]$. So the central moments are $\int (x - \mu)^n f(x) dx$. *Centralizing* it around the mean obtains a value independent of the mean, and μ is the minimum for any centralizing value, $(X - c)$.



The second central moment is called the *variance*. $\sigma^2(X) = E[(X - \mu)^2]$.

A bit of algebra simplifies this to $\sigma^2(X) = E[X^2] - E[X]^2$. Variance is always non-negative.

Variance

Statisticians use many terms to express variation — “deviation”, “dispersion”, “spread”... *Variance* has a specific meaning given by the *central second moment*. Because $c^2\sigma^2(X) = \sigma^2(cX)$, it's units are units of the random variable. Its square root $\sigma(X)$ is called the *standard deviation* that is convenient because it is in the same units as X , so can be plotted on the same axis.

The additive property of variance makes it the preferable measure when working with random variables:

$$\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y) \text{ when } X \perp Y.$$

Mean and variance properties hold even in the presence of higher moments, but there may be “other things” going on with the higher moments not captured by mean and variance.

So far everything applies to the “ideal”, unobserved quantities. To be useful one needs to link these to actual observed data.

Sampling

How does one make any claims about the value of unobserved moments, and so about the actual but unobserved distribution of their values?

The expectations of mean \bar{x} and variance $\overline{\sigma^2}$ are estimated from sample averages, taking advantage that expectation of X is linear, for the mean:

$$\bar{x} = 1/n \sum_{i=1}^n x_i \rightarrow E[X]$$

In physical terms, this is the “center of mass” of the distribution.

And so for the variance of $\sigma^2(\bar{x})$ assuming the X are independent:

$$\sigma^2(\bar{x}) = \sigma^2((1/n) \sum_{i=1}^n x_i) = 1/n^2 \sum_{i=1}^n \sigma^2(x_i) \rightarrow \sigma^2(X)/n$$

where we plug in $\sigma^2(x) = 1/n \sum_{i=1}^n x_i^2 - \mu^2$. In physical terms this is the “moment of inertia” of the distribution.

As described, the quantities on the left of the arrow are observed statistics, and on the right of the arrow are their desired, but unobserved expectations.

Inference via sampling

Again we are “working backwards” - from the sample set values, given by these sums, to infer the properties of the distribution of individual items. Individuals belong to the “population” from which a finite sample has been taken.

The variance of the mean \bar{x} decreases as $1/n$, which means that the standard deviation decreases as $1/\sqrt{n}$. This gives us a good rule of thumb for the increase in the accuracy of an average as the sample increases.

[expression of variance of samples - “analysis” of variance]

How fast do these converge?

The *Law of Large Numbers* gives a guarantee on the convergence, for any well-enough behaved distribution that have finite mean and variance. There are two ways to look at convergence:

- At a given value of n how far off is the estimate of the mean? Besides an estimate of the variance one might want additional assurances. In short, one can compare the density with a $k e^{-X}$ function, for some value of k . (Note - There are formal results such as “Hoeffding's Inequality” that bound $P(X \geq t) \leq k e^{-t}$ for some k .)

- As n increases how does the error of the mean decrease? This is the above mentioned $1/\sqrt{n}$ property. The standard deviation of the sample mean is called the *standard error*.

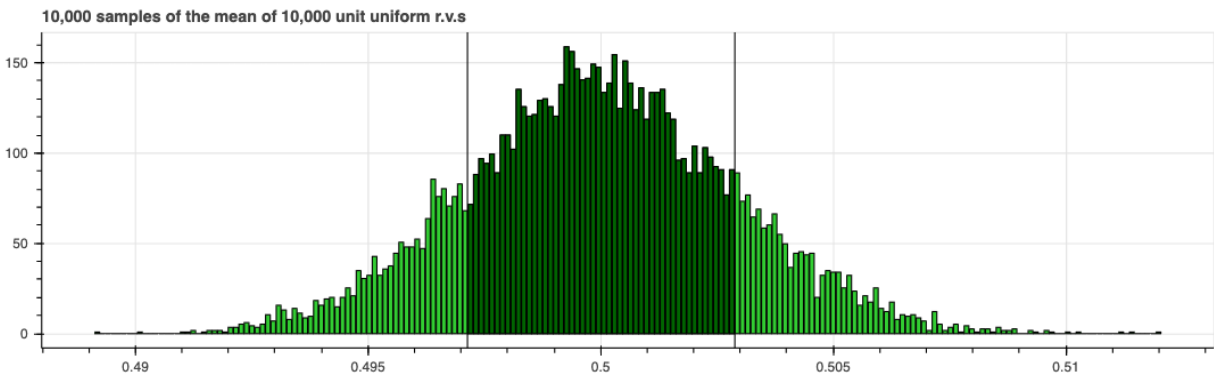
Distributions for which variance doesn't converge have extreme heavy tails — those for which the e^{-x} doesn't apply.

Variance and the normal distribution

The normal is the unique distribution that is entirely determined by it's mean and variance. Conversely if we know (or assume, as an approximation) that all higher moments of a distribution are near zero and can be ignored, then we can treat the distribution as normal. This is useful in the limit of large samples, since by the Central Limit Theorem, *All samples with finite variance converge to the normal as the sample size increases*. Another way to think of this is that the higher moments approach zero faster than the variance.

Estimating an interval from a sample variance

This is a simulation of 10000 runs of the sum of 10000 random numbers between 0 and 1. (A total of 1E8 draws from a random variable.) From the data, it's estimated mean and variance are..., compared to the known values (since we know how the sample is generated)



Calculating probabilities from densities

$$P(a < X < b) = \int_a^b f(x)dx = F(x)|_a^b$$

The CDF is the integral of the PDF. Note the left hand side is a probability, the right hand side is a function.

For discrete r.v.s we speak of a “probability mass function” (PMF) rather than a PDF. It’s possible to have a combined PMF - PDF. Then the CDF will have discontinuities.

In this simulation, for one standard deviation below the mean and one standard deviation above:

a = 0.4971 , b= 0.5028 , $P(a < X < b) = 0.6803$

For the normal distribution, $\Phi(1) - \Phi(-1) = 0.6826$.

Assignment

Read: Evans, Chapter 3.3, Chapter 4.1, 4.5, Chapter 5

References


Some background - A series on making statistics both understandable and precise

Manipulative numbers: How we use statistics to get what we want | David Spiegelhalter
<http://www.huxleysummit.org>

Think numbers are neutral? They're far from it! David Spiegelhalter, Professor of Public
<https://www.youtube.com/watch?v=oUs1uvsz00k>

DAY1/14 Probability & Statistics with Prof David Spiegelhalter

This video forms part of a mathematics course on Probability & Statistics by Prof David Spiegelhalter held at AIMS South Africa in 2012.

 https://www.youtube.com/watch?v=r_OizY4uOmA&list=PLTBqohhFNBE9jRhvdtnuxj9FiOtDONqoy

