x represents the possible values the random variable (X) can take.

$f(x)$ is the probability density function of the random variable X. It describes how likely different values of X are.

$xf(x)$ ~~means~~ represents weighting each value of x by its probability - this means we are adjusting each value of x by how common it is.

Week 4 Notes Part 1 Error of the Mean

Code ▾

An indefinate integral finds the general antiderivative of a function. ie. sums up a continuous function over a range of values.

$$\int f(x)dx$$

dx ← represents an infinitesimally small change in x.

$f(x)$ ↑ function that we're integrating.

$$\int \dots dx$$

Author: Brendan Gongol

Last update: 03 January, 2023

1 Overview

In this note we will discuss the (expected) error involved in the estimation of true underlying mean by using the sample mean. We will start with reviewing the variance and its properties as we will need this machinery in order to derive important results, both in this section, and in the following weeks; then we will characterize the spread (e.g. expected error) of the sample mean considered as a random variable; finally we will run a simple numerical simulation and observe how the derived formula works in practice.

2 Properties of Variance

Variance is simply the average of the squared differences b/w each value + the mean. Squaring makes sure that all deviations are positive values + larger deviations are weighted more heavily.

Before we proceed with the main topic of this Note, let us revisit some properties of the variance. Please give this section much attention as we will need the variance and its properties on many occasions in the weeks to come, in particular when discussing such important topics as explained-vs-unexplained variance, F-test, and ANOVA. The variance of a random variable X distributed with $f(x)$ is defined as

(1) Variance measures how much the values of X spread out from the mean. It is the average of the squared distances b/w each value of X and the mean.

$$Var(X) = \underline{E[(x-\mu)^2]} = \int (x-\mu)^2 f(x)dx$$

where $E[]$ denotes the "expected value", and μ is the expected value of x (population mean) defined as

(2)

$$\mu = E[x] = \int x f(x)dx$$

In other words, for a variable with an arbitrary distribution, the variance is simply the expected value of the squared deviation from the mean. Note that what we are considering here are true underlying values of the mean and variance as defined by the distribution function $f(x)$ of the random variable, not their sample estimates. The standard deviation of the random variable X is defined as

Expected value = population mean. ie, if observe X multiple times you would expect to see its average value.

SD. $$\sigma_x = \sqrt{Var(X)}$$

Taking the √ ~~[illegible]~~ brings the measure back to the same unit of the original data - making it more interpetatable than the variance.

Side note: It does not hurt to reiterate one more time that, as it follows from the above definitions, mean and variance (and hence the standard deviations) of a random variable X are fixed values, calculated from an characterizing the fixed (but probably unknown in many practical applications) underlying distribution $f(x)$. In contrast, the estimators such as sample mean and sample standard deviation are calculated for a given, randomly drawn sample, and hence they are random variables themselves with their own distribution functions. When all we are given is a sample (realization) of a random variable X, we use those estimators in order to try inferring (at least approximately) the true underlying values, so the properties of the distribution functions of the estimators are of much interest (i.e. how close, on average, we expect our estimation to be to the true underlying value? do you remember the homework problem dedicated to this very question?). True μ + Var : fixed values from the entire distribution
sample mean ($\bar{x}$) + sample SD : estimates based on a sample, which are themselves random variables with their own distributions.

Since the expected value of a sum is a sum of expected values, we can also rewrite the variance as

(3) Var(x) can be rewritten ~~[illegible]~~ using the expected values properties

expands to

$$Var(X) = E[(x-\mu)^2] = E[x^2 - 2\mu x + \mu^2]$$

E x each element

$$Var(x) = \underbrace{E[x^2] - E[2\mu x] + E[\mu^2]}$$

and since $E[x] = \mu$ we get $E[x^2] - \mu^2$

$$Var(x) = E[x^2] - 2\mu E[x] + \mu^2 = \boxed{E[x^2] - \mu^2}$$

this is shite! See handwritten notes.

where we also used the fact that for any constant value a,

(4)

$$\overbrace{\boxed{E[ax]} = \int axf(x)dx = a\int xf(x)dx = \boxed{aE[x]}}^{=}$$

and

(5)

$$\overbrace{\boxed{E[a]} = \int af(x)dx = a\int f(x)dx = \boxed{a}}^{=}$$

multiplying a random variable by a constant multiplies its expectant value by that constant, and the expectant value of a constant is just the constant itself.

(since $f(x)$ is normalized probability density, i.e. $\int xf(x)dx = 1$).

If we rescale variable X, i.e. consider random variable aX instead, where a is some constant, then it is clear that the mean of the new variable will be equal to $a\mu$, and using (3) we obtain when you multiply all your data by a constant, the spread (variance) increases by the square of that constant.

(6)

$$\overbrace{\boxed{Var(aX)} = E[a^2x^2] - a^2\mu^2 = a^2(\underbrace{E[x^2] - \mu^2}) = \boxed{a^2Var(X)}}^{=}$$

same as $E[(x-\mu)^2]$ ①

The variance of a sum of two random variables is given by

(7)

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

where the covariance describes the cross-correlation between the random variables:

(8) Covariance measures how much two random variables vary together

positive cov : ↑X = ↑Y
negative cov : ↑X = ↓Y
Zero cov : the two variables are uncorrelated - their movements dont systematically affect the other

Equation 3

constant vs random variable ~~for~~ ~~expectation~~ within the expectation operator $E[\cdot]$

$\boxed{E[x] = \mu}$ x is a random variable so it's expectation $E[x]$ is defined by integrating over the probability density function, which gives the mean (μ).

$\boxed{E[\mu] = \mu \quad , \quad E[\mu^2] = \mu^2}$ μ itself is a constant (its fixed) as it represents the true ~~mean~~ mean ~~of~~ from the entire distribution. Since it is a constant, its expectation is itself.

∴ b/c constants can be pulled out of the expectation operator, for any constant a and random variable X, we have $E[aX] = a \times E[X]$. This is why in the variance example:

$$E[x^2] - \underline{E[2\mu x]} + E[\mu^2] = E[x^2] - \underline{2\mu \cdot E[x]} + E[\mu^2]$$

then, since $E[x] = \mu$ and $E[\mu^2] = \mu^2$ we can ~~cancel out to~~ write:

$$E[x^2] - 2\mu \times \mu + \mu^2$$

$$= E[x^2] - 2\mu^2 + \mu^2$$

$$= E[x^2] - \mu^2$$

∴ Var(x) can be calculated by taking the expected value of the square of x and subtracting the mean.

$$Cov(X,Y) = E[(x - \mu_x)(y - \mu_y)]$$

In the equation above, μX and μY are the population means (expected values of x and y) for X and Y random variables, respectively. If the random variables X and Y are independent, their covariance is zero (as a side note, correlation between X and Y is defined simply as normalized covariance $cor(X,Y) = Cov(X,Y)/\sqrt{Var(X)Var(Y)}$, so that we are simply stating here that independent variables have zero correlation; note that the opposite is not necessarily true!). We will omit the proof, although it is quite simple and you can try proving this fact yourself or looking the proof up if you are interested.

For independent variables we thus have instead of (7):

independent variables we just take the sum of all variances.

(9)

$$Var(X+Y) = Var(X) + Var(Y)$$

which can be generalized to an arbitrary number of independent variables:

(10)

$$Var(X_1 + \ldots + X_n) = Var(X_i) + \ldots + Var(X_n)$$

3 Standard Error of the Sample Mean

We have observed and discussed many times by now that sample statistics such as, e.g. sample mean, sample variance, slope of the regression line etc are random variables. Indeed, their values are computed from and thus depend on the particular sample randomly drawn from the underlying distribution. If we are going to draw another sample, it is impossible to predict next value of sample statistics we are going to observe in that sample with certainty, and the expected results can be described only as a probability distribution over the possible values.

In this section we will take a closer look at a sample mean. The sample mean for any given sample $x_1, \ldots, x_n$ of size n is given by

(11)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$E[x] : \mu$

and we know that the sample mean is the estimator for the underlying population mean (2). As the sample size n increases the sample mean (11) converges to population mean, $\bar{x} \xrightarrow[n\to\infty]{} \mu$ ("law of large numbers").

Since sample mean is a random variable, this convergence occurs in probabilistic sense: by drawing a sample we still can only get a random, unpredictable value of the sample mean $\bar{x}$, but the distribution of $\bar{x}$ becomes tighter with increasing n, so our estimation becomes progressively more accurate as n increases. Let us examine the distribution of the sample mean in more detail.

In order to do that, let us rewrite (11) in terms of random variables, rather than specific realization xi in a particular sample. The random variable representing the sample mean is defined as

(12)

Random variable for sample mean $\rightarrow$ x

$$\bar{X} = \frac{(X_1 + X_2 + \ldots + X_n)}{n}$$

Note that this is just the change in notation. The meaning of (12) is still the same as (11): for each particular realization $x_1, \ldots, x_n$, the value $\bar{x}$ is defined as (11), but we need to think now in terms of random processes, not their particular realizations, so this notation is more appropriate. Note that each datapoint xi is represented in (12) by an independent random variable (random process) X_i. Since we are building our sample by drawing (independently) n times from the same underlying distribution $f(x)$, all random variables X_i in (12) are independent and identically distributed.

X ↗

By using properties (10) and (6) of the variance, we can immediately see that the variance of the random variable $\bar{x}$ is given by $\boxed{Var(aX) = a^2 Var(X)}$, here $a = \frac{1}{n}$ so $(\frac{1}{n})^2 = \frac{1}{n^2}$

(13) Variance of sample mean

$$Var(\bar{X}) = Var(X_i/n + X_2/n + \ldots + X_n/n) = \frac{1}{n^2}(Var(X_1) + \ldots + Var(X_n))$$

But our random variables X_i are identically distributed (they represent independent drawings from the same underlying distribution), hence their variances are the same (and are determined solely by that underlying distribution). Since we have the sum of n identical terms in (13), we can arrive to the final conclusion as

(14) The variance of the sample mean $\bar{X}$ for the random variable X is $\frac{Var(X)}{n}$. This means the variance of the sample mean decreases as the number of independent draws (n) increases.

$$Var(\bar{X}) = \frac{1}{n^2} \cdot nVar(X) = Var(X)/n$$

∴ to improve accuracy of the estimate of the population mean, need to increase sample size which ↓ variance of the sample mean making the estimate more reliable.

where we dropped the index i from the random variable X, for which we are calculating the sample mean. Since the standard deviation of a distribution (true standard deviation) is the square root of the variance, we can also take the square root of both sides of (14) and rewrite it as

(15) SD of sample mean $\bar{X}$ is the $\sqrt{\ }$ of its Variance $\sigma_{\bar{X}} = \sqrt{Var(\bar{X})} = \sqrt{\frac{Var(X)}{n}} = \frac{\sigma_X}{\sqrt{n}}$

$$\sigma_z = \frac{\sigma_x}{\sqrt{n}}$$

This extremely important relationship we just derived tells us the following: suppose we have a random variable X with standard deviation σ_X; if we repeatedly draw samples of size n of that random variable and compute sample means for each of those samples, the distribution of the resulting sample means is going to have standard deviation $\sigma_x/\sqrt{n}$. In other words, this is a "typical" error we expect to make when we draw a sample of size n and use its sample mean as the approximation to the true underlying mean.

The formula (15) is a formal, rigorous explanation to a widely appreciated fact: if we want to accurately measure some quantity in the presence of random variation/noise, we have to perform multiple measurements and the stronger the noise σ_x, the more measurements we need in order to achieve any given accuracy. Indeed, by performing multiple measurements and averaging them we calculate the sample mean, and the expected error (deviation from the true underlying value we are trying to assess through multiple measurements) has the scale of the standard deviation of the mean σ_z (15). If we need to achieve some fixed accuracy ε, then we need to perform at least n measurements, where n is determined from $\sigma_x/\sqrt{n} \sim \varepsilon$, i.e. $n \sim \sigma_x^2/\varepsilon^2$ and σ_x is the standard deviation (level of noise) of the quantity we are trying to measure.

The standard deviation of the sample mean is also known as standard error of the mean (SEM). It represents the typical error that we expect when using the sample mean to estimate the population mean.

SD + SEM.

① SD of a distribution

The standard deviation (σ_x) of a random variable (X) is the sqrt of its variance ($Var(X)$).

$$\sigma_x = \sqrt{Var(X)}$$

② Var of sample mean

The variance of the sample mean ($\bar{x}$) is:

$$Var(\bar{X}) = \frac{Var(X)}{n}$$

③ SD of sample mean

By taking the sqrt of both sides we get:

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$

④ Symbols

X - represents the random variable

x - represents ~~the~~ a specific instance/realization of the random variable X

$\bar{X}$ - represents the sample mean, which is a random variable itself.

σ_X - standard deviation of the random variable X *Also known as the population SD. Describes the spread of the entire populations values.

$\sigma_{\bar{X}}$ - SD of the sample mean $\bar{X}$, also known as the SEM.

z - In this context σ_z is used to denote the SEM.

σ_x - SD calculated from the sample values drawn from the population. It serves as an estimate of the population SD (σ_X).

⑤ Standard error of the mean (SEM):

$$\sigma_z = \frac{\sigma_x}{\sqrt{n}}$$

The SEM (σ_z) is calculated using the SD of the sample values (σ_x) because we often dont know the true population SD (σ_X). The σ_x is an estimate based on the data that we have.

⑥ Accuracy + sample size

– accuracy referrs to how close our sample mean is to the true population mean, a smaller ε = higher accuracy.

To achieve a certain accuracy (ε), we need to perform enough measurements such that SEM $\sim \varepsilon$:

$$\frac{\sigma_x}{\sqrt{n}} \sim \varepsilon \xrightarrow{\text{solving for } n} n \sim \frac{\sigma_x^2}{\varepsilon^2}$$

as ε^2 is the denomenator - if we want a low specific accuracy value, we need a very high n value.

* SEM is inversly proportional to the $\sqrt{n}$. ($SEM/\sigma_z \sim \frac{1}{\sqrt{n}}$)

so if we want a 10 fold improvement in accuracy $= \frac{1}{10} \times \varepsilon$

$SEM \sim \frac{1}{\sqrt{n}}$, $SEM^2 \sim \frac{1}{n}$, $(\frac{1}{10})^2 \sim \frac{1}{n}$, $\frac{1}{100} \sim \frac{1}{n}$, $n \sim 100$ fold increase

then we need a 100-fold increase in sample size.

Another important consequence of (15) is that the SEM is inversely proportional to the square root of the sample size. Thus if we want, for instance, to achieve 10-fold improvement in the accuracy of the estimation of the underlying mean, we have to increase the sample size 100-fold.

$SEM = \frac{1}{\sqrt{n}}$

$SEM^2 = \frac{1}{n} \rightarrow (\frac{1}{10})^2 = \frac{1}{n}$

10 fold = reduction by 10 = 1/10

$\frac{1}{100} = \frac{1}{n}$

$n = 100$

4 Numerical Simulation

In order to further develop our understanding and intuition (and practice some more R), let us work through a numerical simulation. We want to simulate sample mean as random variable, thus for a given sample size n we want to perform (as usual) multiple sample drawings, compute the means of each sample and see how these sample means are distributed. In order to characterize this distribution we will compute its mean (i.e. mean of the sample means) and standard deviation (standard deviation of the sample means). We will run this experiment for a few different sample sizes and compare the standard deviation of the sample means (SEM), as function of the sample size n, to the value predicted by the formula (15). Note that this procedure is very similar to the homework problem you were solving earlier, but this time we are studying the distribution of the sample means rather than sample standard deviations.

goal.

Hide

example

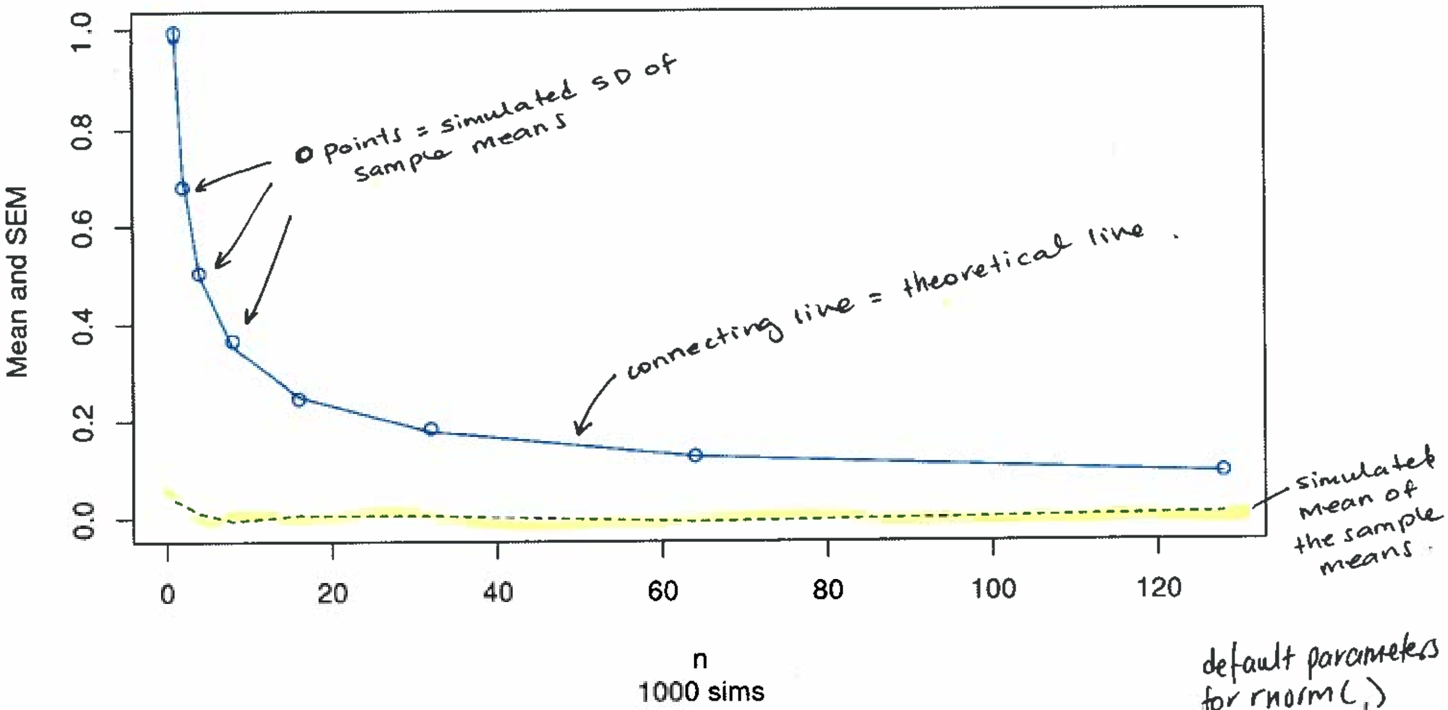
```
smpl.sizes <- c(1,2,4,8,16,32,64,128) # sample sizes
r.mu <- numeric() will contain the average sample means for each sample size
r.sem <- numeric() will contain the standard error of the sample means for each sample size
n.sim <- 1000  number of simulations
for ( i.smpl in smpl.sizes ) {
 r.mat <- matrix(rnorm(i.smpl*n.sim),  creates a matrix of 1000 rows x 4 columns of
 nrow=n.sim,ncol=i.smpl)                   random normal values
 mu.tmp <- apply(r.mat,1,mean) - calculates the mean for each row. Returns a vector of 1000 means
 r.mu <- c(r.mu,mean(mu.tmp)) - calculates the mean of mu.tmp + appends it to r.mu
 r.sem <- c(r.sem,sd(mu.tmp)) - calculates SD of mu.tmp (which represents the SEM for this
}                                        sample size) and appends it to the r.sem vector.
plot(c(min(smpl.sizes),max(smpl.sizes)),
 c(min(c(r.mu,r.sem)),            • initializes empty plot w x axis ranging from
 max(c(r.mu,r.sem))),               min-max of sample sizes + y axis ranging
 type="n",ylab="Mean and SEM",      from the min-max values of r.mu + r.sem
 xlab="n",sub=paste(n.sim,"sims")) • type = "n" means no points/lines drawn yet
                                    • Sub adds a subtitle to the plot.
points(smpl.sizes,r.mu,type="l",lty=2) - line plot of smpl.sizes against r.mu
points(smpl.sizes,r.sem,col="blue") - adds points for standard errors against
points(smpl.sizes,1/sqrt(smpl.sizes),                                    sample sizes
 type="l",col="blue")        ↳ adds a plot of the theoretical standard error of the
                             mean (1/sqrt(smpl.sizes))
```

rnorm (i.smpl (4) x n.sim (1000)) → generates 4000 random values drawn from a normal distribution, with mean=0 SD=1 (default parameters).

apply (x, margin, function) → x = array/matrix, margin = subscripts that function will be applied to (1 = row, 2 = columns), function = the function that will be applied.

lty = line type. 1 = solid (default), 2 = dashed, 3 = dotted etc.

comparing the emperical standard error (r.sem) with the theoretical standard error (1/sqrt(smpl.sizes))



default parameters for rnorm()

In the code shown above, we first define a set of few sample sizes we want to try. In the loop that follows, for each sample size `i.smpl` we select `n.sim` random samples of size `i.smpl` from the normal distribution with mean 0 and standard deviation $\sigma_X = 1$ (note how we select data for all `n.sim` samples at once and organize them into a matrix with each row representing one of `n.sim` samples we have drawn; it's just another trick that you may find appealing - or not; we could instead run an internal loop `1:n.sim` in order to generate each of the `n.sim` resamplings individually). Inside the loop, we first compute the sample means for each sample drawn (apply `mean()` across rows, `MARGIN=1`); then we compute and save the mean and standard deviation of those sample means.

In the final plot (shown below), we draw the simulated mean of the sample means (i.e. where the distribution of the sample mean is centered) and the simulated standard deviation of the sample means (how much spread that distribution has), as a function of sample size. Last `points()` command adds the theoretical line: since we were drawing samples from the normal distribution with `σ=1` , we expect, according to (15), that the standard error of the sample mean will follow the curve `1/sqrt(smpl.sizes)` . As we can see from the plot, the simulated SEM follows the theoretical prediction perfectly and indeed decreases with increasing sample size.

*

So if we pulled random ~~variables from~~ values from a normalized distribution with a $\sigma = 2$ we'd expect $\frac{2}{\sqrt{smpl \cdot sizes}}$.