

Uncertainty and sampling

Jonathan Shapiro

Department of Computer Science

Announcements

1. I am putting the slides as Jupyter notebooks on Blackboard. You will need to download them onto a system where you have Jupyter set up.
2. I am working on other formats on Blackboard as well.

Overview

Sampling from a population

1. A very important scenario in data science is:

to infer properties of large population from the properties of a sample from that population.

2. This introduces uncertainties in any conclusions we draw from the sample about the larger population.

Sometimes one has all the data there is

- For example, all of the sensor data from a particular piece of infrastructure may be available.
- However, the goal is probably to infer properties from new data arriving in the future.

So, the present is a sample of the future (but not always a representative sample)



Assumptions and notation

Assumptions

1. There is a "parent" population, which is essentially infinite.
2. The sample is drawn randomly from the parent population.

a) Thus, the sample is representative of the parent population

b) Thus, the sample is Independently and identically distributed ("IID").

3. The order of drawing the sample does not matter

a) The data is "exchangeable".

Are the assumptions valid?

1. The population is never infinite, but is often large compared to the sample size.
2. Getting a truly representative sample can be very difficult.
3. The order the data was collected may matter. For example, suppose it is an opinion survey and some event takes in the middle of the survey which could affect opinions.

The population probability distribution

- The population is viewed as a probability distribution with:

mean: μ ;

standard deviation: σ (or variance σ^2);

neither do we typically know.

- Expectations over the parent probability will be denoted with calligraphic E, viz $\mathcal{E}[\cdot]$.

Properties of the sample

- Data:** is N real numbers sampled from the parent population, denoted x_i , with $i = 1, \dots, N$.
- Sample mean:** (aka empirical mean) denote as \hat{m} with

$$\hat{m} \equiv \frac{1}{N} \sum_i^N x_i.$$

- Sample variance:** (aka empirical variance) denote as \hat{s}^2 and

$$\hat{s}^2 \equiv \frac{1}{N} \sum_i^N (x_i - \hat{m})^2.$$

- Sample standard deviation:** $\hat{s} = \sqrt{\hat{s}^2}$ (obviously).

Empirical measurements are denoted with the little "hats", $\hat{\cdot}$.

Notice that the sample statistics are random variables. With a different sample, they would have different values.

Sample statistics as population estimators

- The sample mean is an unbiased estimator of the population mean,

$$\mathcal{E}(\hat{m}) = \mu.$$

- The sample variance is **not** an unbiased estimator of the population variance (it underestimates),

$$\mathcal{E}(\hat{s}^2) = \frac{N-1}{N} \sigma^2.$$

- The unbiased estimator corrects for this,

$$\hat{s}_{\text{unbiased}}^2 = \frac{N}{N-1} \hat{s}^2 = \frac{1}{N-1} \sum_i^N (x_i - \hat{m})^2.$$

We will almost always use the unbiased estimator. So, to simplify notation,

1. denote the *unbiased estimator* as \hat{s} ;
1. denote the *biased estimator* as \hat{s}_b ; b for biased.

Standard error in the mean (SEM)

- Standard error in the mean, also called *SEM* (in python, e.g.), *standard error*, or *stderr*.
- It is the uncertainty in the estimate of the mean.
- The formula is the standard deviation divided by the square root of the sample size,

$$\text{sem} = \frac{\hat{s}}{\sqrt{N}}.$$



The University of Manchester

Standard error in the mean

Standard error in the mean (SEM)

- This is a very important quantity in data science and statistics.
- Typically, we want to estimate some property of the parent population.
- The best estimate is the mean of that quantity. (Usually! There are exceptions.)
- The standard error in the mean tells you the uncertainty in your estimate of that quantity.

Standard deviation versus standard error in the mean

- **Standard deviation:**

- a) Measures the variation in the sample.
- b) Can be used as the uncertainty in any individual in the population.
- c) As the sample size increases, the sample standard deviation converges to the population (true) standard deviation

$$\text{As } N \rightarrow \infty, \text{ then } \hat{s} \rightarrow \sigma$$

Standard deviation versus standard error in the mean

- **Standard error in the mean:**

- a) Measures the variation in the estimate of the mean.
- b) It is the uncertainty in the estimate of the mean.
- c) As the sample size increases, the sample standard error in the mean converges to 0, because the uncertainty in the mean converges to zero.

$$\text{As } N \rightarrow \infty, \text{ then } \hat{m} \rightarrow \mu, \text{ the population mean}$$

Where does this come from?

Standard error in the mean: is the root-mean-squared deviation between the sample mean and the true mean,

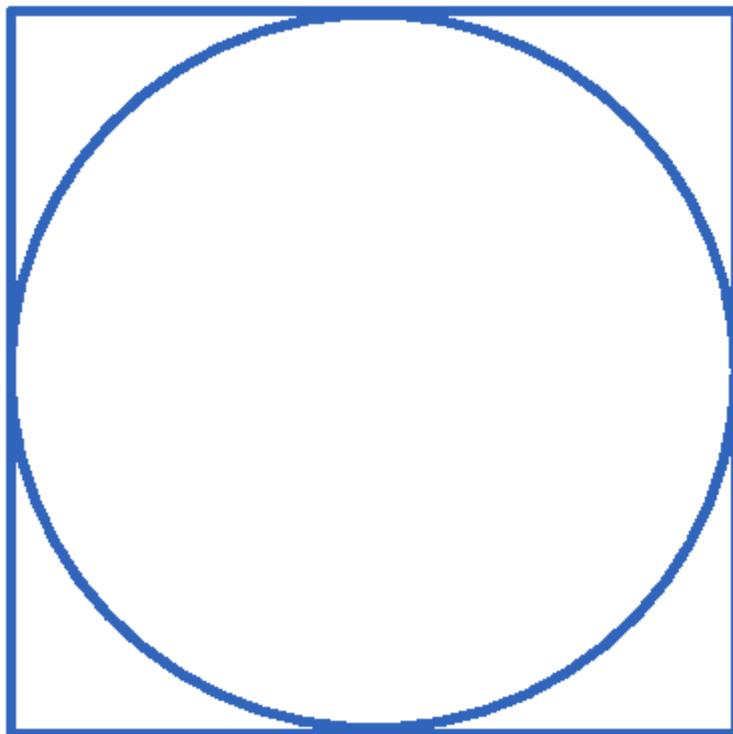
$$\begin{aligned} \text{SEM} &= \mathcal{E}(\hat{m} - \mu)^2, \\ &= \frac{\hat{s}^2}{N}. \end{aligned}$$

(Note: you can do this calculation. It is a little tedious, but involves nothing you don't already know.)

Example: Monte Carlo estimate of $\pi/4$.

We saw last lecture:

Ratio of a circle to an inscribed square:



Value:

$$\frac{\pi}{4} = 0.78539816339744830961566084581988 \dots$$

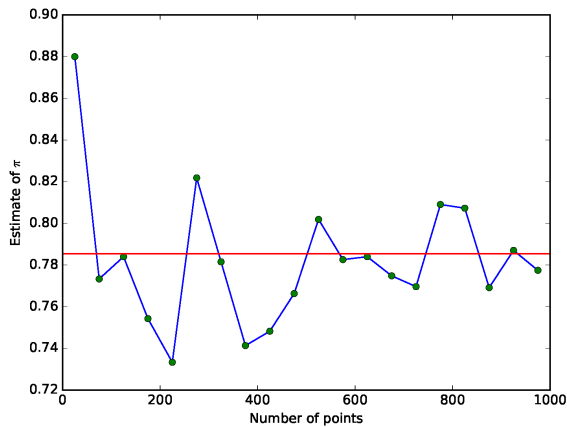
But suppose I did not know the value of π . I will try to estimate it from data.

Estimate the value of $\pi/4$ from data

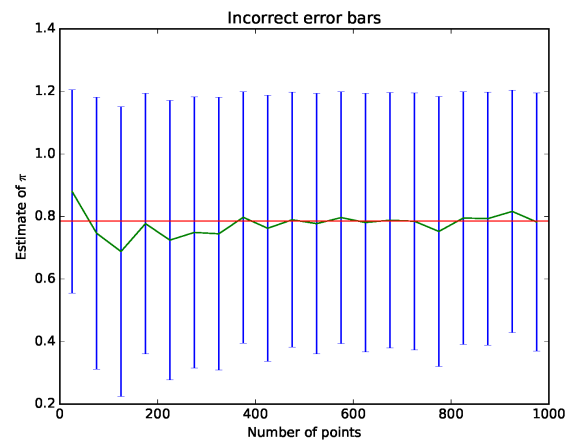
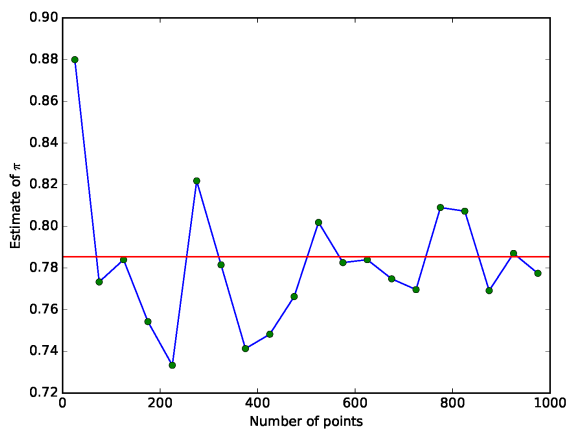
1. Generate N random x, y pairs uniformly in the unit square.
2. Count how many N_o fall inside the circle. ($\sqrt{x^2 + y^2} \leq 1/2$.)
3. Use N_o/N as an estimate of $\pi/4$.

Called a "Monte Carlo" method because it uses randomness to estimate a non-random quantity.

Estimate the value of $\pi/4$ from data



Estimate the value of $\pi/4$ from data



Estimate the value of $\pi/4$ from data

- Using $N = 975$, the estimate is 0.78 ± 0.01 . Actual value $\frac{\pi}{4} = 0.785398 \dots$
- Notice that the sem goes gets smaller and smaller, but slowly (like $1/\sqrt{N}$).
- Whereas, the standard deviation moves towards the true value, which is $\sqrt{p(1-p)}$ with $p = \pi/4$ namely 0.41.
- (It is a Bernoulli process, since the random point is inside the circle with probability $p = \pi/4$.)

Properties of large samples

The law of large numbers

- The law of large numbers states that in the limit of a large number of samples, the empirical mean converges to the population mean.

$$\lim_{N \rightarrow \infty} \hat{m} \rightarrow \mu.$$

- There are different versions in which the strength of the convergence differs in strength.

Combining random variables (Not about large samples, but needed.)

Let x and y be random variables sampled from different parent distributions, such that

$x \sim \rho_x$ with properties μ_x and σ_x . The symbol \sim means "sampled from".

Likewise, $y \sim \rho_y$ with properties μ_y and σ_y .

- Mean of $x + y$ is $\mu_x + \mu_y$;
- Variance of $x + y$ is $\sigma_x^2 + \sigma_y^2$
- Standard deviation of $x + y$ is $\sqrt{\sigma_x^2 + \sigma_y^2}$

Central limit theorem 1

- Consider, N random variables, $x_i, i = 1, \dots, N$,
 - drawn independently from a parent distribution with mean μ and standard deviation σ , but not necessarily a normal distribution,
- Then, in the limit of large N , under weak conditions on the original distribution, the sampling distribution for the mean,

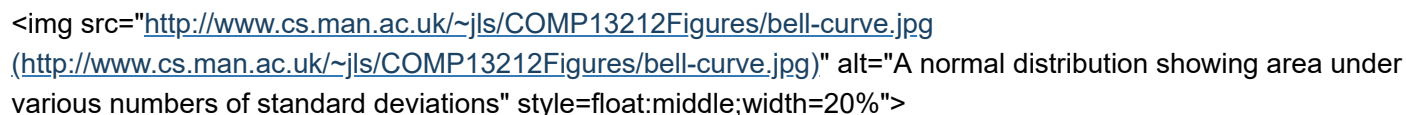
$$\hat{m} = \frac{1}{N} \sum_i^N x_i$$

is a normal distribution with mean μ and standard deviation σ/\sqrt{N} .

The central limit theorem explains several things

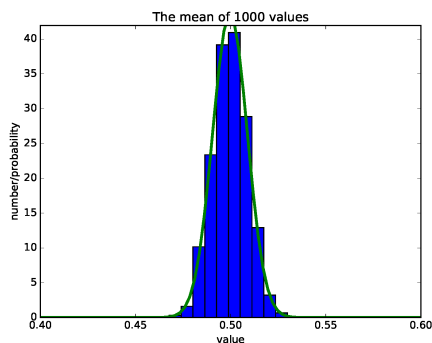
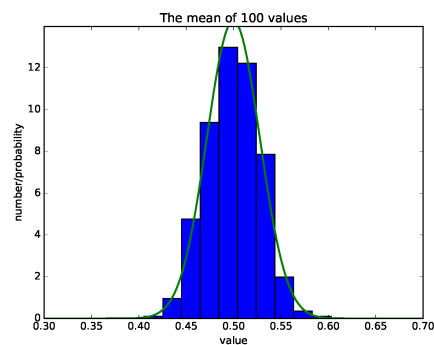
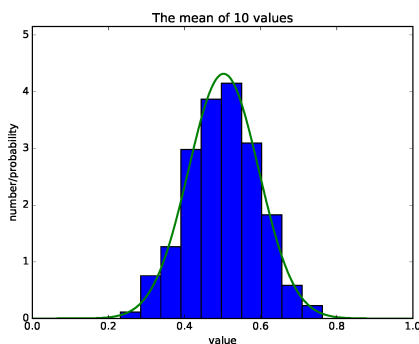
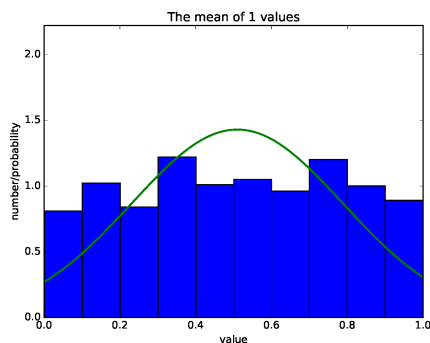
1. It explains the ubiquity of the normal distribution (aka the Gaussian distribution).
2. It also explains the meaning of the standard error in the mean as a measure of uncertainty (see next figure)

The Normal distribution

A normal distribution showing area under various numbers of standard deviations" style="float:middle; width=20%;"/>

Demonstration

I sample N points uniformly from $[0, 1)$ one thousand times.

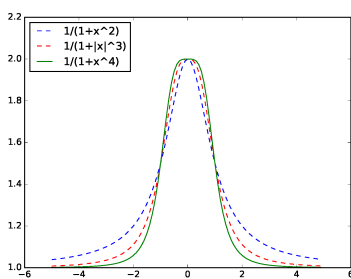


Central limit theorem II

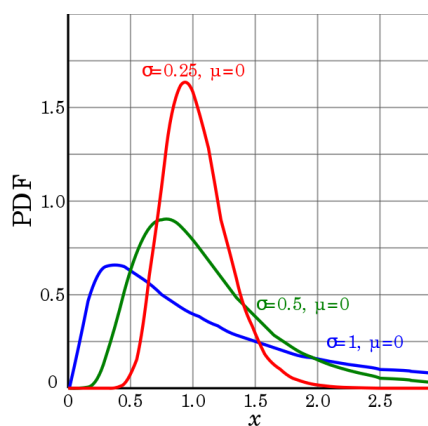
- It is not necessary for all the values to come from the same distribution.
- They can come from different distributions, as long as all the distributions have well-defined means, variances, obey other mild conditions.
- Then, the sampling distribution for \hat{m} is a normal distribution with μ as the sum of the means of the distributions,
- and σ^2 as the sum of the variances of the distributions divided by N . (Variances add, remember.)

When does the central limit not apply?

"Fat-tailed" distributions



"Log-normal" distribution (also fat-tailed).



When does the central limit theorem not apply?

- More generally, with distributions with infinite means or variances,
- or infinite moments arbitrarily higher than the second moment.

WHY is the central limit theorem true?

- ???????
- Entropy (explain)



The University of Manchester

Conclusions

What is important to learn

1. How sample statistics relate to population statistics.
2. How to generate an unbiased estimation of the population standard deviation σ .
3. The standard error in the mean is the measure of uncertainty
4. The importance of the central limit theorem and normal distribution.
5. When the central limit theorem might not apply.

Write in your memory in indelible ink

1. Always use the standard error in the mean for error bars on mean values.
2. Use the standard deviation as the uncertainty in a randomly chosen individual value, and as a measurement of the variability in the population.
3. Variances add (not standard deviations).