Lecture Notes: Point Estimation

October 31, 2022

1 Introduction

A statistical model of a process consists of identifying suitable random variables, and specifying the probability distributions associated with the random variables. Let us consider the process of radioactive decay of a substance and let us assume that we are interested in predicting the waiting time between successive decay events. A statistical model of this process involves defining the continuous random variable X which is the time between two successive decay events in seconds, and specifying the probability distribution of X, which for example could be $f(x) = \lambda e^{-\lambda x}$, with $\lambda = 4$. Given this model, we can make probabilistic predictions regarding the outcome of the process. For example (i) we can predict the probability that the waiting time between two decay events is larger than 10 seconds.

Often times we do not know all the details of a statistical model. In some cases, we might not know the probability distributions of the random variables we define, yet we might be interested in knowing quantities like mean or variance of the random variable. Or, we may not know the value of a parameter of a probability distribution. E.g For the pdf $f(x) = \lambda e^{-\lambda x}$, we may not know the value of the parameter λ. In the next few weeks we will learn some techniques that use sampled data to infer the details of a statistical model. These techniques fit under the term statistical inference.

2 Random Sampling

The first step in statistical inference is collecting or sampling data. Assume that you are a quality control (Q.C.) inspector working for DiGiorno's, the frozen pizza manufacturer. You are interested in estimating the mean weight of cheese present in a slice. Every day 100,000 slices are made in the manufacturing unit. These 100,000 slices constitute the population, which consists of every single outcome of the process. One way to estimate the mean weight of cheese in a slice, is to simply measure its value in all 100,00

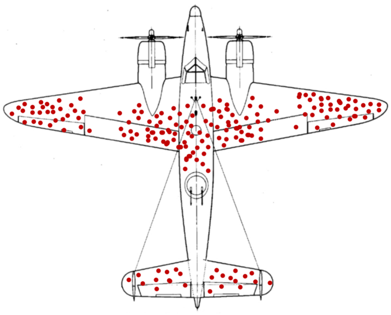


Figure 1: Sketch of a world war II fighter jet. A red dot indicates a site where damage was detected. Source:Wikipedia

slices. This is time consuming and therefore infeasible. In this chapter we will learn how to estimate this quantity (mean weight of cheese in a slice), by just observing a small subset of the population, say a random sample of 100 slices. This estimate is our best guess of the population mean based on the sample data.

A sample can be picked by going to the warehouse and picking 100 boxes from the shelves at random. If the boxes have a serial number between 1 and 100,000, the Q.C. inspector can generate 100 random numbers between 1 and 100,000 using a computer and pick the corresponding boxes. This is a common technique used to ensure that the sample, is representative of the population. Here are a few cases where the sampling procedure in incorrect.

- Jim is interested in learning about the coffee drinking habits of the undergraduate student community. He prepares a questionnaire and sits next to the campus coffee shop. Some of the students who enter and exit the cafe, participate in the survey. Is this a good method to sample the student community ?

  The population that the survey intends to learn about, consists of the university's entire undergraduate students. Some of these students might be regular coffee drinkers, some might drink it occasionally, and some might not drink it at all. Jim has to find a location where the students he meets are a good representation of the student population, for example the entrance to the campus. By sitting next to the campus coffee shop, he most likely will meet only the coffee drinkers in the population. His sampling procedure is therefore biased.

- An engineer at the Faber Castell pencil manufacturing plant, tests the pencils made everyday to ensure that it is of good quality. He tests the first 10 pencils that are made on any given day. Is his method of testing any good ?

  It is possible that the first 10 pencils manufactured at the start of the day, are not representative of the population. The first few pencils might be of better quality than the rest, as the employees are less tired in the morning and chances of error are less. Over the course of the day, it is likely that the quality of output reduces. It is therefore better to pick 10 pencils at random from all the pencils manufactured in a day, to test the quality of the output.

- During the second world war, the United States Air force was losing a lot of its fighter jets during combat. To improve the safety of the jets, a study of the damage suffered by the jets was undertaken. All the jets that returned to base were inspected and the damaged suffered by each was documented. Figure 1 shows all the locations on a jet where damage was detected. The study concluded that the areas with most red dots (damage) be reinforced to improve the safety of the jets. Do you agree with these findings ?

  The findings suffer from what is termed as survivorship bias. Dr. Abraham Wald, a statistician who read the report dismissed its conclusions. He noted that in order to conclude where the jets need reinforcement, it is vital to look at both the jets that returned to base, and ones that were lost in combat. He suggested that the areas with the red dots, need the least reinforcement, as the jets returned to base despite suffering damage. The jets that got hit in the other areas (regions with no red dots) most likely crashed.

Sample Variability

Two quality control inspectors gather their own sample of 30 batteries from the warehouse. The first inspector finds that just 2 batteries are defective. The second inspector find that 10 batteries are defective. Is this to be expected ?
Yes, this is to be expected. Data obtained from a small subset of the population will exhibit variations. As the size of the sample increases, the variability will reduce.

3 Point Estimation

In the rest of this chapter we will learn how to use sampled data to

1. To obtain an estimate for a parameter of a statistical model.
2. To obtain an estimate for the mean value of a random variable.
3. To obtain an estimate for the variance of a random variable.

3.1 Maximum Likelihood Estimation

Let us revisit the process of radioactive decay which we discussed earlier. Assume that the distribution of waiting times between successive decay events is given by $f(x) = \lambda e^{-\lambda x}$, with λ being an unknown. What we have here is a family of probability distributions parameterized by λ. Knowing the exact value of λ will enable us to make probabilistic predictions about a specific radioactive decay process. The maximum likelihood estimation (MLE) technique will enable us to come up with an estimate for λ based on sampled data.

The first step is to build the likelihood function which gives the probability of observing the sampled data. This will be a function of the unknown parameter θ. The maximum likelihood estimate of a parameter θ, written as θ_{MLE}, is that value of θ which maximizes the likelihood function. The two problems discussed below demonstrate how this technique is used.

Problem I: Discrete case

A coin is flipped 100 times. 55 heads we observed and the rest of the outcomes were tails. Find the maximum likelihood estimate for the probability p of obtaining heads on a single toss.

Let us assume that the coin flips are independent and identically distributed Bernoulli trials. The random variable of interest is the number of heads or number of successes (mapping heads to success and tails to failure) X. We know that X follows a Binomial distribution, $Binomial(n = 100, p)$. Note that we do not know the value of p and we are seeking an estimate for it.
The first step is to construct the likelihood function $\mathcal{L}$ which is the probability of observing the sampled data (55 heads). Since X has a binomial distribution

$$\mathcal{L}(X = 55; p) = \binom{100}{55} p^{55}(1-p)^{45} \tag{1}$$

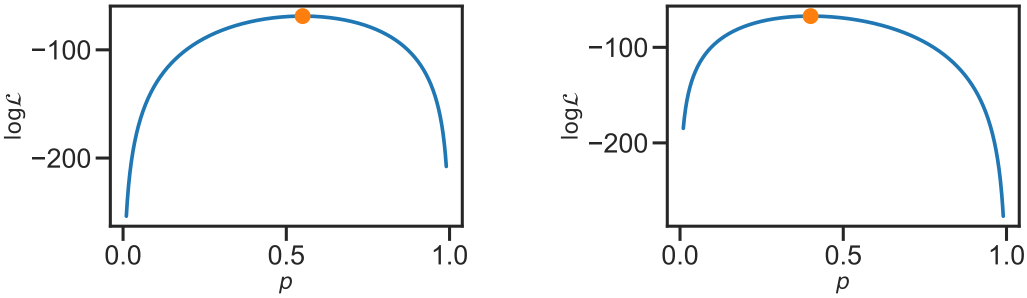
The maximum likelihood estimate of parameter p, which is written as p_{MLE}, is that value of p which maximizes the likelihood function $\mathcal{L}$. Note that $\mathcal{L}$ is not a probability distribution and it is a function of variable p. To find p_{MLE} we set the first derivative of $\mathcal{L}$ to zero (slope hits zero at maximum).

$$\left.\frac{d\mathcal{L}}{dp}\right|_{p=p_{MLE}} = 0 \tag{2}$$

$$\begin{aligned}
\binom{100}{55}\left[55p^{54}(1-p)^{45} - p^{55}45(1-p)^{44}\right] &= 0\\
55p^{54}(1-p)^{45} &= p^{55}45(1-p)^{44}\\
55(1-p) &= 45p\\
55 &= 100p\\
p_{MLE} &= \frac{55}{100} \qquad (3)
\end{aligned}$$

Note that p_{MLE} does not equal p, the true probability of obtaining heads in a single coin toss. It is simply our best estimate of p, based on observing 55 heads in 100 tosses.

Sometimes it might just be more convenient to do pen and paper calculations with $\log\mathcal{L}$, instead of $\mathcal{L}$. Since logarithm is a monotonically increasing function, the maximum of $\log\mathcal{L}$ and the maximum of $\mathcal{L}$ are exactly equal. Let's redo the above problem this time maximizing $\log\mathcal{L}$. The log likelihood can be written as



(a)

(b)

Figure 2: (a) Plot of the log-likelihood function for the case of observing $N = 55$ heads in $M = 100$ coin tosses. (b) Same as (a) but for $N = 40$ heads observed in $M = 100$ coin tosses. The orange dot corresponds to the MLE estimate in both plots.

$$\log \mathcal{L}(X = 55; p) = \log \binom{100}{55} + 55 \log p + 45 \log(1 - p) \tag{4}$$

For generality, let us assume that the coin was tossed M times and N heads were observed. The above equations can then be rewritten as

$$\log \mathcal{L}(X = N; p) = \log \binom{M}{N} + N \log p + (M - N) \log(1 - p) \tag{5}$$

Maximizing the above function and setting it to zero we get

$$\begin{aligned}
\left.\frac{d \log \mathcal{L}}{dp}\right|_{p=p_{MLE}} &= 0 \\
\frac{N}{p} - \frac{M - N}{1 - p} &= 0 \\
\frac{N}{p} &= \frac{M - N}{1 - p} \\
N(1 - p) &= (M - N)p \\
N - Np &= Mp - Np \\
p_{MLE} &= \frac{N}{M}
\end{aligned} \tag{6}$$

The above equation gives a general expression that helps us estimate p_{MLE} for any observed data is therefore called the *maximum likelihood estimator*. Plugging in $N = 55$ and $M = 100$ we get the exact same answer as above.

Figure 2 (a) shows a plot of the likelihood function for the case discussed above. The orange dot is the maxima and is our estimate p_{MLE}. Figure 2 (b) shows another case where 40 heads were observed in 100 coin tosses.

Problem II: Continuous case

Suppose the lifetime of a Samsung battery is modeled by an exponential distribution ($f(x) = \lambda e^{-\lambda x}$). The parameter λ is unknown. A random sample of size 5 was collected. Tests revealed their lifetimes to be 2, 3, 1, 3 and 4 hours respectively. Can you provide an estimate of λ ?

Let X_i be the lifetime of the i th battery, where i goes from 1 to $n = 5$. The random variables X_1, X_2, X_3, X_4, X_5 have taken values 2,3,1,3 and 4 respectively. Let us first construct the likelihood function, which is the probability of observing 2,3,1,3 and 4 as lifetimes of a sample of 5 batteries. The random variables $\{X_i\}$ are independent and identically distributed and therefore their joint pdf is given by

$$\begin{aligned} f(X_1, X_2, X_3, X_4, X_5) &= \lambda e^{-\lambda x_1} \lambda e^{-\lambda x_2} \lambda e^{-\lambda x_3} \lambda e^{-\lambda x_4} \lambda e^{-\lambda x_5} \\ &= \lambda^n e^{-\lambda \sum x_i} \end{aligned} \tag{7}$$

In the above equation $n = 5$ and $\sum x_i = 2 + 3 + 1 + 3 + 4 = 13$. For a random variable Y with pdf $f(y)$, the probability that Y lies in a tiny interval $[y_0, y_0 + dy]$ can be approximated as $\int_{y_0}^{y_0+dy} \simeq f(y_0)dy$. We use this idea to construct the likelihood function $\mathcal{L}$. The probability that X_is lie in a tiny region around the observed data is

$$\mathcal{L} \simeq \lambda^n e^{-\lambda \sum x_i} dx_1 dx_2 dx_3 dx_4 dx_5 \tag{8}$$

The maximum likelihood estimate of λ, is the value of λ, that maximizes the likelihood function $\mathcal{L}$.

$$\left. \frac{d\mathcal{L}}{d\lambda} \right|_{\lambda = \lambda_{MLE}} = 0 \tag{9}$$

$$\begin{aligned} \left[n\lambda^{n-1} e^{-\lambda \sum x_i} - \lambda^n \sum x_i e^{-\lambda \sum x_i} \right] dx_1 dx_2 dx_3 dx_4 dx_5 &= 0 \\ \lambda^n \sum x_i e^{-\lambda \sum x_i} &= n\lambda^{n-1} e^{-\lambda \sum x_i} \\ \lambda_{MLE} &= \frac{n}{\sum x_i} && (10) \\ \lambda_{MLE} &= \frac{5}{13} && (11) \end{aligned}$$

The maximum likelihood estimate of λ is $5/13$. Equation 7 is an expression which can be used to compute the ML estimate for any set of observations. Therefore it is called the maximum likelihood estimator of the process.

3.2 Properties of the MLE Estimator

The MLE technique presented above is easy to implement even for cases when the likelihood $\mathcal{L}$ is a function of more than one variable. For example if the data comes from a population that is normally distributed you will have to estimate two parameters μ and σ. In this case the process of maximizing $\mathcal{L}$ will involve computing two partial derivatives ($\frac{\partial \mathcal{L}}{\partial \mu}$ and $\frac{\partial \mathcal{L}}{\partial \sigma}$); setting both to equal zero and solving the pair of equations. There also exist numerical methods for maximizing $\mathcal{L}$ when it is a function of several variables, but this topic is outside the scope of this course.

Let θ be a parameter that we are trying to estimate and let θ_{MLE} be its maximum likelihood estimator. Let the true value of the parameter be denoted by θ^*. θ_{MLE} satisfies the following properties

- As the sample size of data collected increases θ_{MLE} converges to θ^*
- Let $g(\theta)$ be a function of θ. Then the maximum likelihood estimate of g is given by $g(\theta_{MLE})$.
- θ_{MLE} is a random variable since it relies on sampling data. It has been proven that the probability distribution of θ_{MLE} can be approximated by the normal distribution with mean θ^* and with a standard deviation that is inversely proportional to the square root of the sample size n.

3.3 MLE in Imaging

Let us discuss one final example which will illustrate the utility of MLE. It is based on a similar example presented by Stanley Chan in his book Introduction to Probability and Data Science. It is a toy problem that demonstrates how MLE can be used in imaging biological samples.

Here is how imaging a micron sized biological sample works. The sample is first stained or infused with a dye molecule of a specific kind. These molecules can absorb light of a specific wavelength. Once the electrons in these molecules absorb a photon they get excited and move to a higher energy state. After a while they relax back to the ground state and in the process of relaxation emit photons of a different wavelength. Physicists have established that the number of photons X emitted by a region in the sample per unit time is a Poisson distributed random variable with parameter λ being proportional to concentration of dye molecules in the region. The probability distribution of X is

$$\pi(x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{12}$$

Using sensors capable of detecting and counting photons from a region in the sample you can indirectly estimate the concentration of dye molecules in that region and use this information to construct an image of the sample.

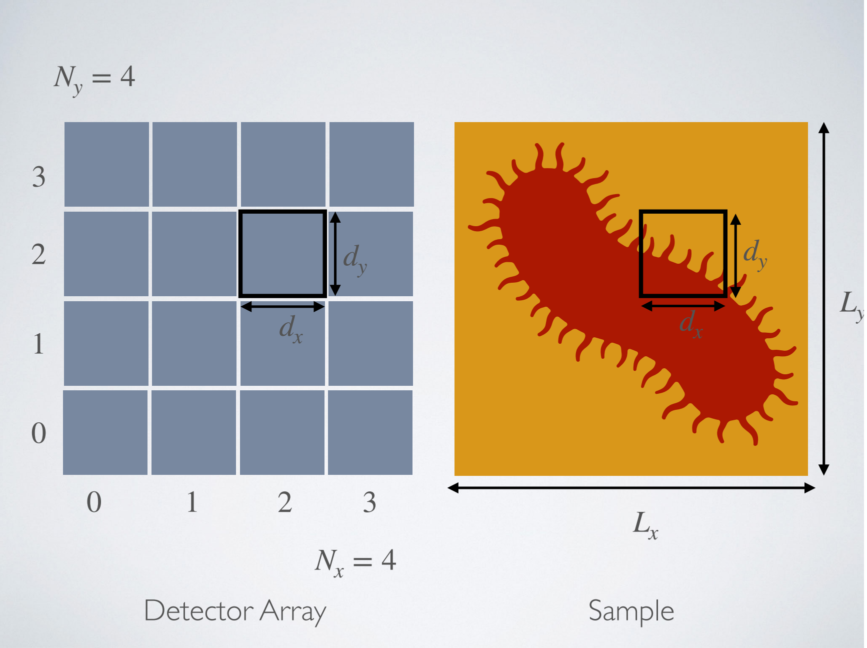


Figure 3: Left: A 4 by 4 array of sensors that can turn ON when a photon hits its surface. Associated with every sensor is an index (i, j) (row numbers, column number). Right: The image on the right is a schematic of a biological sample dyed and mounted on the microscope. The black box on the right is a sub-region othe sample. Light emitted from this region is captured by the sensor with index (2,2).

Assume that the sample is of size L_x microns by L_y microns. The microscope typically is equipped with an array of $m \times n$ sensors and each sensor is of size d_x microns by d_y. The sensors are arranged in a grid and light that comes from a certain position (x, y) in the sample uniquely hits the sensor indexed by $int(x/d_y), int(y/d_y)$. Here $int(a)$ refers to rounding down a to its nearest integer. See the schematic in figure 3 for more details.

Each sensor can remain in one of two possible states ON or OFF and by default is set to OFF. Assume that even if one photon hits a sensor it switches to the ON state. Since the sensors are only in ON or OFF states the image can be represented by black or white pixels; white corresponding to ON state and black to OFF state.

Figure 4 shows a sample image obtained from the microscope. Clearly there is a lot of noise in the image. Assume that the process of taking a picture using the sensors was repeated M times. Here is how MLE can be put to use to reconstruct the image from the library of M different noisy images.

Let us just reconstruct the image one pixel at a time. As mentioned above each pixel corresponds to the ON/OFF state of one sensor. Consider the sensor indexed by (i, j). This sensor switches ON if at-least one photon hits. The probability that no photon hits the sensor even if the region in the sample has some dye molecules in it can be written as (using equation 12)

$$\pi(0) = \frac{\lambda^0 e^{-\lambda_{ij}}}{0!} = e^{-\lambda_{ij}} \tag{13}$$

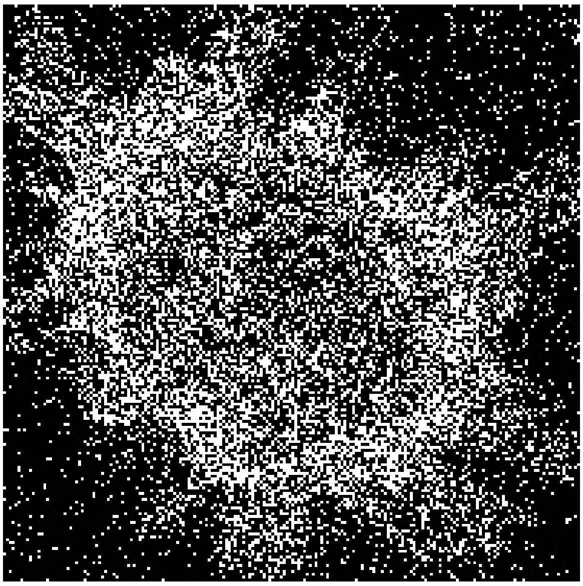


Figure 4: A noisy image imaged by a microscope fitted with a 200 by 200 sensor array. The image is therefore of size 200 by 200 pixels. Each pixel is either black or white corresponding to the ON / OFF state of the sensor.

And the probability that at-least one photon hits in $1 - \pi(0) = 1 - e^{-\lambda_{ij}}$. A pixel in one image being in ON or OFF state is therefore a Bernoulli process with parameter $p = 1 - e^{-\lambda_{ij}}$. It is white (ON state) with probability $1 - e^{-\lambda_{ij}}$ and black (OFF state) with probability $e^{-\lambda_{ij}}$.

Note that λ_{ij} is an unknown. And we are going to infer this quantity using MLE by looking at this pixel in each of the M different images collected from the microscope. Let $Y_1, Y_2, Y_3 \ldots Y_M$ be the state of the pixel in each of the M images. Each of these variables Y_i is an independent Bernoulli trial. Let the pixel be in the ON state in N images and in OFF state in $M - N$ images. The likelihood of making this observation is given by

$$\mathcal{L} = \left(1 - e^{-\lambda_{ij}}\right)^N \left(e^{-\lambda_{ij}}\right)^{M-N} \tag{14}$$

And the log-likelihood is given by

$$\log \mathcal{L} = N \log\left(1 - e^{-\lambda_{ij}}\right) + (M - N) \log\left(e^{-\lambda_{ij}}\right) \tag{15}$$

Maximizing the above equation and setting it to zero we get

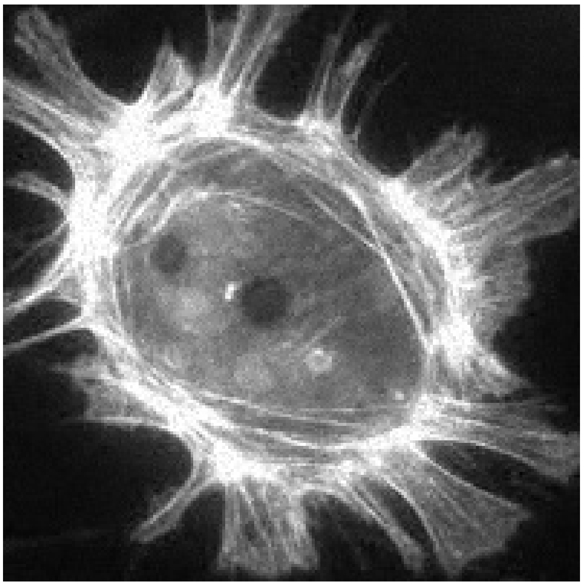


Figure 5: A library of 1000 noisy images were used to reconstruct the image above. For each pixel we applied the MLE formula dervied in equation 16. The reconstructed image is also of size 200 by 200 pixels.

$$\begin{aligned}
\frac{d\log\mathcal{L}}{d\lambda_{ij}} &= 0 \\
\frac{N}{1-e^{-\lambda_{ij}}}e^{-\lambda_{ij}} - \frac{M-N}{e^{-\lambda_{ij}}}e^{-\lambda_{ij}} &= 0 \\
\frac{N}{1-e^{-\lambda_{ij}}}e^{-\lambda_{ij}} &= M-N \\
Ne^{-\lambda_{ij}} &= M - Me^{-\lambda_{ij}} - N + Ne^{-\lambda_{ij}} \\
Me^{-\lambda_{ij}} &= M-N \\
e^{-\lambda_{ij}} &= 1-\frac{N}{M} \\
\lambda_{ij} &= -\log\left(1-\frac{N}{M}\right) \qquad (16)
\end{aligned}$$

The MLE estimator for λ_{ij} is $-\log\left(1-\frac{N}{M}\right)$, where N is the number of images in which pixel i, j is in the ON state and M is the total number of images collected. Applying this equation to every single pixel can help us infer λ_{ij} a quantity proportional to the concentration of the dye for all pixels, thereby reconstructing the image. In figure 5 I show an image obtained by reconstructing from the data of $M = 1000$ different noisy images. The Python code that was used to reconstruct this image is given below. To run this code below you need a library called `opencv` and if you do not have it installed you can install it via the command `pip install opencv`. This code has also been uploaded to Canvas together with the library of noisy images.

```
import cv2
from scipy.stats import poisson
import numpy as np

# Total number of noisy images
mfiles = 1000

# No of pixels in each images is 200 x 200
nx     = 200
ny     = 200

recon    = np.zeros((200,200))
scounts  = np.zeros((200,200))

for f in range(mfiles):

    colorimage = cv2.imread('noisy_images/image_'+str(f)+'.jpg')
    img        = cv2.cvtColor(colorimage, cv2.COLOR_BGR2GRAY)

    for i in range(nx):
        for j in range(nx):

            # ON correpsonds to 255 and OFF to zero
            # Counting number of ON states
            scounts[i,j] += img[i,j]/255

for i in range(nx):
    for j in range(nx):

        # Applying the MLE estimator formula equation 16
        recon[i,j] = -np.log(1 - (scounts[i,j]/mfiles))

# Show image
cv2.imshow('Noise', recon)

# Reset white to 255
cv2.imwrite('reconstructed.jpg', 255*recon)

cv2.waitKey(0)

# Destroying image
cv2.destroyAllWindows()
```

3.4 Properties of mean and variance of random variables

Next we will learn about estimators of mean and variance. But before that let us remind ourselves of a few definitions and in the process learn a few properties of mean and variance which will be used later in this document.

The mean of a random variable X, which is often denoted by the symbol μ, is calculated using the equation $E(X) = \mu = \sum x_i \pi(x_i)$, if X is discrete and $E(X) = \mu = \int_{-\infty}^{\infty} x f(x) dx$, if X is continuous. The first property concerns the expected value of the random variable X, multiplied by a constant a.

Mean : Property I

If X is a random variable with $E(X) = \mu$, for any constant a, $E(aX) = aE(X) = a\mu$

The next property concerns the mean of a sum of random variables $X_1, X_2 \dots X_n$. Note that the mean is computed with respect to the joint pdf if X_i s are continuous, or with respect to the joint pmf if X_i s are discrete.

Mean : Property II

The expectation (mean) of $\sum X_i$ with respect to the joint probability distribution of $\{X_i\}$ is simply the sum of the expectation of individual random variables. $E(\sum X_i) = \sum E(X_i)$.
Let us prove this for the case of two continuous random variables X_1 and X_2.

$$\begin{aligned}
E(\sum X_i) &= E(X_1 + X_2) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_1 + x_2) f(x_1, x_2) dx_1 dx_2 \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 f(x_1, x_2) dx_1 dx_2 + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_2 f(x_1, x_2) dx_1 dx_2 \\
&= E(X_1) + E(X_2) \\
&= \sum E(X_i) \qquad (17)
\end{aligned}$$

In general if we have functions of random variables $h(X_1), h(X_2) \dots h(X_n)$, then $E(\sum h(X_i)) = \sum E(h(X_i)$

The variance of a random variable X, which is denoted by the symbol σ^2, is calculated using the equation $V(X) = \sigma^2 = \sum (x_i - \mu)^2 \pi(x_i)$, if X is discrete and $V(X) = \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$. Here μ is the mean of the population. The first property simply connects the variance of a random variable X to the expectations $E(X)$ and $E(X^2)$.

Variance : Property I

Let X be a random variable with mean $E(X) = \mu$. The variance of a random variable X, $V(X) = E(X^2) - (E(X))^2$. Here is the proof for this statement for the case where X is a continuous random variable.

$$\begin{aligned} V(X) &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \\ &= \int_{-\infty}^{\infty} (x^2 - 2x\mu + \mu)^2) f(x)dx \\ &= \int_{-\infty}^{\infty} x^2 f(x)dx - 2\mu \int_{-\infty}^{\infty} xf(x)dx + \mu^2 \int_{-\infty}^{\infty} f(x)dx \end{aligned} \tag{18}$$

$$\begin{aligned} V(X) &= \int_{-\infty}^{\infty} x^2 f(x)dx - 2\mu \int_{-\infty}^{\infty} xf(x)dx + \mu^2 \int_{-\infty}^{\infty} f(x)dx \\ &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - \mu^2 \\ &= E(X^2) - (E(X))^2 \end{aligned} \tag{19}$$

The above equation (eqn. 10) can also be rewritten as

$$E(X^2) = V(X) + (E(X))^2 \tag{20}$$

3.5 Sample mean

Let us now learn how to estimate the mean value of a random variable. Assume that a company manufactures 100,000 PVC pipes each day. You are interested in knowing the mean diameter of the population. If a statistical model for the manufacturing process in available, the mean of the population can simply be calculated from the pdf using the equation $\mu = \int_{-\infty}^{\infty} xf(x)dx$. Since, we do not know a statistical model for this process, we have to settle for an *estimate* of the population mean μ, obtained from a small sample.

Here are the diameters of a sample of 20 pipes chosen at random from the warehouse. Any quantity computed from such sampled data is called a statistic.

24.46	27.15	30.88	25.61	29.5	27.31	28.04	28.28	27.94	26.66
29.13	27.54	26.42	28.5	27.98	29.11	28.49	28.87	26.25	27.74

An obvious choice for the estimator of the population's mean diameter μ is the statistic $\overline{X}_n = \frac{\sum X_i}{n} = 27.793$. Here X_i is the random variable which gives the diameter of pipe i in centimeters. Note that the X_i s are independent and identically distributed random variables. The law of large of large numbers guarantees that as the sample size n increases

$\overline{X}_n$ approaches the population mean μ. Here are a few more statistics that can potentially be considered as estimates of the population mean.

- Let us define a statistic X_e as the average of the minimum and maximum value of the observed data.

$$X_e = \frac{min(\{X_i\}) + max(\{X_i\})}{2} = \frac{24.46 + 30.88}{2} = 27.670 \tag{21}$$

- Here is another statistic $X_{tr10\%}$, called the 10 % trimmed mean. It is the arithmetic mean of the observed data after eliminating the top 10% and the bottom 10% of data.

$$X_{tr10\%} = \frac{\sum X_i - 24.46 - 25.51 - 29.5 - 30.88}{20 - 4} = 27.838 \tag{22}$$

Note that all three statistics have similar numerical values, and can be considered as good estimates of the population mean. We can also cook up more statistics that can serve as estimates of μ. *How does one decide if a statistic is a good estimator of the population mean ?*

One of the criteria that statisticians employ to judge an estimator, is the criterion of unbiasedness. If $\hat{\theta}$ is an estimator of the quantity θ, it is said to be unbiased if $E(\hat{\theta}) = \theta$. In other words the estimates produced by $\hat{\theta}$ will be centered around θ. The difference between $E(\hat{\theta})$ and θ is called as the bias of the estimator.

Let us check if the sample mean $\overline{X}_n$ is an unbiased estimator of the population mean μ. We will do this by verifying if the expectation of $\overline{X}_n$, equals the population mean μ.

$$\begin{aligned} E(\overline{X}_n) &= E(\frac{\sum X_i}{n}) \\ &= \frac{1}{n}E(\sum X_i) \end{aligned} \tag{23}$$

In the second line of the above equation we have used property I of the mean of random variables, described above. Using property II of the mean we know that the expectation of $\sum X_i$, is simply $n\mu$, where n is the sample size.

$$\begin{aligned} E(\overline{X}_n) &= \frac{1}{n}E(\sum X_i) \\ &= \frac{n\mu}{n} \\ &= \mu \end{aligned} \tag{24}$$

Since $E(\overline{X}_n) = \mu$, we conclude that it is an unbiased estimator of the population mean. The data from some samples might overestimate μ, while others may underestimate μ, however the estimate $\overline{X}_n$, will always be centered around the true population mean μ.

3.6 Sample variance

Next let us estimate the variance of a population. Some of you might have guessed that a good starting point is the statistic

$$s_a^2 = \sum \frac{(X_i - \overline{X}_n)^2}{n} \tag{25}$$

This is almost correct, however a minor change is required to the above statistic to turn it into an unbiased estimator of the population variance σ^2. Here is another statistic that we can consider to be an estimator of the population variance.

$$s^2 = \sum \frac{(X_i - \overline{X}_n)^2}{n-1} \tag{26}$$

Let us verify if s^2 is an unbiased estimator of the population variance σ^2. To do that, we have to check if $E(s^2) = \sigma^2$, but before let us simplify the expression for s^2.

$$\begin{aligned}
s^2 =& \frac{\sum(X_i - \overline{X}_n)^2}{n-1} \\
=& \frac{1}{n-1}\left[\sum(X_i^2 - 2X_i\overline{X}_n + \overline{X}_n^2)\right] \\
=& \frac{1}{n-1}\left[\sum X_i^2 - \sum 2X_i\overline{X}_n + \sum \overline{X}_n^2)\right] \\
=& \frac{1}{n-1}\left[\sum X_i^2 - 2\overline{X}_n \sum X_i + \overline{X}_n^2 \sum(1))\right] \\
=& \frac{1}{n-1}\left[\sum X_i^2 - 2\frac{\sum X_i}{n} \sum X_i + (\frac{\sum X_i}{n})^2 n)\right] \\
=& \frac{1}{n-1}\left[\sum X_i^2 - \frac{1}{n} \sum (X_i)^2\right]
\end{aligned} \tag{27}$$

Let us now compute the expectation of s^2.

$$\begin{aligned}
E(s^2) =& \frac{1}{n-1}\left[E(\sum X_i^2) - \frac{1}{n}E(\sum(X_i)^2)\right] \\
=& \frac{1}{n-1}\left[\sum E(X_i^2) - \frac{1}{n}E(\sum(X_i)^2)\right]
\end{aligned} \tag{28}$$

In the second line of the above equations, I have applied property II of the mean of variables to rewrite $E(\sum X_i^2)$ as $\sum E(X_i^2)$. Next let us apply property I of variance derived above, to both the terms in equation 28. This gives use

$$E(s^2) = \frac{1}{n-1}\left[\sum V(X_i) + \sum(E(X_i))^2 - \frac{1}{n}(V(\sum X_i) + (E(\sum X_i))^2)\right] \tag{29}$$

Since X_i s are i.i.ds obtained from a distribution with mean μ and variance σ^2, $V(X_i) = \sigma^2$ and $E(X_i) = \mu$. Furthermore in the appendix of chapter we have derived the result that $V(\sum X_i) = n\sigma^2$ and $E(\sum X_i) = n\mu$. Putting all these pieces together we have,

$$\begin{aligned} E(s^2) =& \frac{1}{n-1}\left[\sum V(X_i) + \sum (E(X_i))^2 - \frac{1}{n}(V(\sum X_i) + (E(\sum X_i))^2)\right] \\ =& \frac{1}{n-1}\left[n\sigma^2 + n\mu^2 - \frac{1}{n}(n\sigma^2 + n^2\mu^2)\right] \\ =& \frac{1}{n-1}\left[n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2)\right] \\ =& \sigma^2 \end{aligned} \tag{30}$$

Therefore s^2 is an unbiased estimator of the variance of a population. Let us briefly revisit our first choice s_a^2, and check if it is an unbiased estimator.

$$\begin{aligned} E(s_a^2) &= E(\sum \frac{(X_i - \overline{X}_n)^2}{n}) \\ &= \frac{1}{n}E\left(\sum_i (X_i - \overline{X}_n)^2\right) \end{aligned} \tag{31}$$

Let us multiply and divide by $n-1$.

$$\begin{aligned} E(s_a^2) =& \frac{n-1}{n(n-1)}E\left(\sum (X_i - \overline{X}_n)^2\right) \\ =& \frac{n-1}{n}E\left(\frac{\sum (X_i - \overline{X}_n)^2}{n-1}\right) \end{aligned} \tag{32}$$

The second term in the above equation is simply $E(s^2) = \sigma^2$. Therefore

$$E(s_a^2) = \frac{n-1}{n}\sigma^2 \tag{33}$$

As you can see from the above equation, the estimator s_a^2, carries a bias. It is not centered around the population variance σ^2, but around $(n-1/n)\sigma^2$. The bias present in this estimator is $(n-1/n)\sigma^2 - \sigma^2 = -\sigma^2/n$. For small sample sizes s_a^2 has a significant bias, and it is better to employ s^2 as the estimator of the population variance.