Mathematics Notes

Dinesh Thogulua
dinesh.thogulua@gmail.com

September 24, 2020

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 STATISTICS

1.1 PROBABILITY OVERVIEW

1.1.1 Parameters Describing Distributions

Mean and Median represent the **central tendency** of a distribution. **Mean** is essentially the center of gravity of a distribution: If we consider the values that a random variable takes as distances on a plank and their probabilities as weights, then the mean point would be the distance point on the plank below which, if a fulcrum is placed, the plank will be perfectly balanced. Hence the formula of mean μ is derived as below:

$$\int_{-\infty}^{\mu} (\mu - x) f_X(x) dx = \int_{\mu}^{\infty} (x - \mu) f_X(x) dx$$

Since μ is a constant, we can pull it out of the integrals on both sides,

$$\mu \int_{-\infty}^{\mu} f_X(x) dx - \int_{-\infty}^{\mu} x f_X(x) dx = \int_{\mu}^{\infty} x f_X(x) dx - \mu \int_{\mu}^{\infty} f_X(x) dx$$

Rearranging the equation, we get,

$$\mu \int_{-\infty}^{\mu} f_X(x) dx + \mu \int_{\mu}^{\infty} f_X(x) dx = \int_{-\infty}^{\mu} x f_X(x) dx + \int_{\mu}^{\infty} x f_X(x) dx$$

$$\mu \int_{-\infty}^{\infty} f_X(x) dx = \int_{-\infty}^{\infty} x f_X(x) dx$$

Recognizing that the area under the probability curve for entire range of x values is 1,

$$\mu = \int_{-\infty}^{\infty} x f_X(x) dx$$

Note that the above equation is only for continuous random variables which take real values from $-\infty$ to ∞. For all other cases, the formula and the derivation are similar.

Median represents the midpoint value of the random variable, X, if the values of X are ordered. It is the value, above and below which, there are equal *number* of X values. In other words, while mean is the central tendency if both X values and their probabilities, median is the central tendency if only the probabilities are considered. Median is a more useful parameter in some contexts: Take, for instance, the wealth distribution among the people of India. Since most wealth is accumulated in the hands of a very small number of people, if one were to calculate the mean wealth it will be pulled in one direction due to the few people who have enormous wealth and hence may paint a rosy picture about the economic condition of the population in general. However, median wealth would inform that half the population is below that value and half the people have more wealth than that. So median would be a better indication of the economic condition of the population.

Apart from Mean and Median, **Mode** is also an important parameter related to the "signal part" of a random variable (Explained in the next paragraph). It is simply the value of X that occurs most frequently. Mode is more important than Mean or Median is some contexts: For ex., when we use the colloquial term, "Average Indian", we actually (technically) are talking about the Modal Indian. Because, with "Average Indian", we are talking about the "Common Man" that we are mostly likely to encounter in the streets. The "Common Man" has a certain income level, certain age, height, weight etc. If one thinks of a common man as a value the multi-dimensional random variable "Indian" can take, then the Mean value of that random variable would represent have the mean values of every dimension, i.e., income level, age etc., and that may be far away from the "Common Man". In fact, the "Mean Indian" may be a completely unrealistic person. But a "Modal Indian" is simply the multi-dimensional value of X that occurs most frequently and, hence, by definition, a realistic person.

Variance (σ^2) and, it's square root Standard Deviation (σ) represent the **spread** in the population. **Variance** is defined as $E[(X-\mu)^2]$. It is a common practice, while making measurements of a quantity, to take multiple measurements and then average out the resulting values. This is done because in most measurement instruments have tolerances defined as some $\pm k\%$ - If the error in the measurement (i.e., noise) has equal probability of taking slightly higher or lower values as compared to the true value (i.e., signal), then taking an average would zero out the noise. Hence, if we think of the measured value as a random variable, then mean would be the true value of a measurement and Variance, the probablistic version of the mean square error in the measurement. We need **Standard Deviation** because Variance is a square value and hence has squared units (ex. km^2) as compared to the measured quantity (ex. km) - Taking the square root of the mean square error gives us an error quantity with comparable units to the signal (Think of Signal-to-Noise ratio). This is why the Standard Deviation is also called the **standard error**.

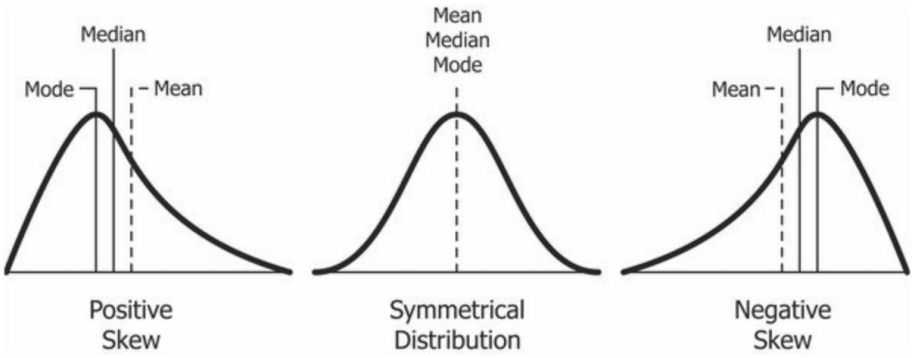


Figure 1.1: Relationships between different parameters

Skewness and Kurtosis represent the **shape** of a distribution. **Skewness** (γ_1)of a dsitribution, defined as $E[(\frac{X-\mu}{\sigma})^3]$, is the measure of asymmetry. A perfectly symmetric distribution has Skewness equal to 0. Skewness will have a positive value if a random variable has higher probability of having low values. Skewness will have a positive value if a random variable has higher probability of having high values. Skewness is also called "Pearson Skewness".

Figure 1.1 shows the relationship between skewness, mean, median and mode. Remember how we talked about how a few wealthy Indians could pull the mean and give a false story about the economic status of the population? The positively skewed (skewness is a positive value) distribution in the picture represents such a scenario: A lot of people (represented by y-axis) have low values (concentrated on the left side of the x-axis).

Kurtosis (γ_2), is defined as $E[(\frac{X-\mu}{\sigma})^4]$, is a measure of the peakedness or flatness of a distribution as compared to normal distribution (refer to Central Limit Theorem). Univariate Normal distribution has a Kurtosis of 3. If a distribution looks "pulled up" version of a normal distribution, its Kurtosis will be greater than 3, and if it looks "stretched out", the Kurtosis will be less than 3. Kurtosis is also called "Pearson Kurtosis"

Note that, for a Normal distribution, Mean = Median = Mode, and Skewness is 0, all of which establish a convenient baseline with which to compare other distributions (abNormal!). But Kurtosis alone, for a Normal distribution is 3! So, some people use an "Adjusted Pearson Kurtosis", which is simply $Kurtosis - 3$. In literature, if we find mentioning of "Postive" or "Negative" Kurtosis, it means they are talking about the Adjusted Pearson Kurtosis. In other contexts, one should be careful to find out whether a piece of literature uses Pearson Kurtosis or the adjusted version before interpreting the text.

Figure 1.2 shows what positive and negetive (Adjusted Peason) Kurtosis looks like when compared with the normal distribution.

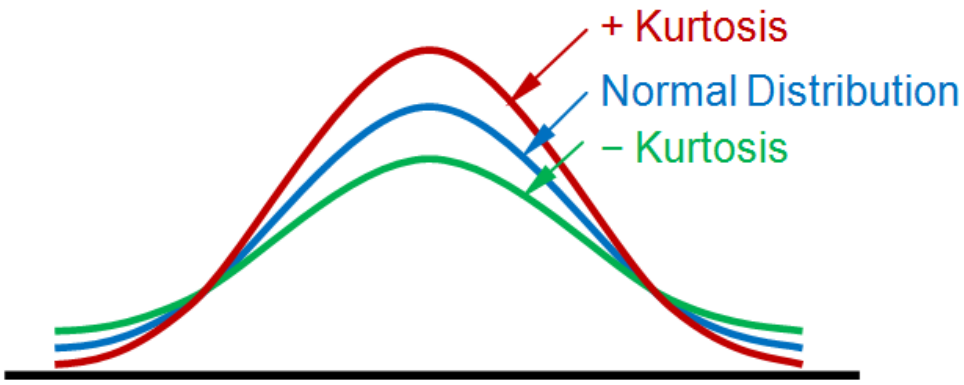


Figure 1.2: Adjusted Pearson Kurtosis

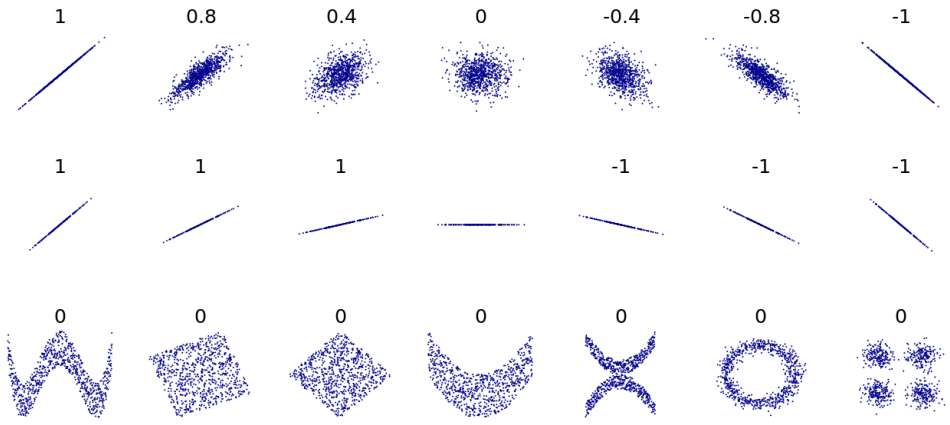


Figure 1.3: Correlation Coefficient of Various Joint Distributions

So far, we have talked about parameters related to univariate distributions, where one random variable X has a certain probability distribution associated with it. But sometimes a probability distribution is associated with a set of two random variables. Ex: We can associate probabilities to (height, weight) tuples for all the girls of a certain age. For such cases, some new parameters arise. **Covariance** (σ_{XY}), defined as $Cov[X,Y] = E[(X-\mu_X)(Y-\mu_Y)]$. The usefulness of Covariance is apparent when you look at its normalized variant, namely the **Correlation Coefficient** (ρ), defined as $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$. The values of the Correlation Coefficient can vary from -1 to 1. When X and Y are independent, σ_{XY} will be 0 and hence ρ_{XY} will also be 0. If Y is linearly dependent on X, then ρ_{XY} will be 1 or -1. If Y is loosely dependent on X, then ρ values will be close to zero (on the positive or negative side).

Figure 1.3 shows the graphical representation of distribution of (X,Y) tuples and the correlation coefficients in each case. One can see how the Correlation

Coefficient gives an idea about the tightness and direction of the relationship between X and Y.

1.1.2 Common Probability Distributions

The most basic distribution is the **Bernoulli distribution**. The Bernoulli random variable (R.V.) takes two values, 1/0 corresponding to Yes/No, On/Off etc. Coin flips, Gender of a person, bit error, Yes/No answers to survey questions etc., can be modeled as Bernoulli R.V.s. It has only one parameter p, the probability of getting a 1 - knowing p, one can derive the only other unknown required to entirely describe the Bernoulli distribution, namely the probability of getting a 0, because it is simply $1-p$, which is sometimes referred to as q[1]. Most distributions of interest in probability theory and in statistics arise as extensions of Bernoulli distribution or closely associated with it. The following paragraphs talk about these distributions.

One can generalize a Bernoulli distribution to a case where the result can take a certain number of fixed values. This distribution is called a **Categorical distribution**. A six sided die, the colour of a candy coated chocolate etc., follow Categorical distributions. A Catergorical distribution is defined by $K-1$ parameters, where K is the number of possible values the R.V. can take, These parameters are simply $p_1, p_2, ..., p_K - 1$, i.e, the probabilities associated with $K-1$ possibilities. It is not uncommon to have all the K probabilities mentioned as parameters of a Categorical distribution for ease of calculations.

Suppose one where to conduct n independent trials of a Bernoulli R.V. with parameter p, then the number of 1s (or 1s or heads etc.), in those n trials follow a **Binomial Distribution**. Later, when we deal with concepts in Statistics, the Binomial distribution will come handy when we talk about proportions - proportions of females in a population of IT workers (as compared to non-females, i.e., males), proportion of certain species of whales living above a certain age etc. This is because, when we conduct surveys, and, say, ask people's gender, at the end, the total number of females is essentially a count of one of the two possible value, this Bernoulli R.V. namely gender, can take - and n will be the number of people surveyed. The same goes to a research study that counts whales. It is important to remember, before modeling an R.V. as a Binomial R.V., that the individual Bernoulli trials should all be independent and have the same probability of success p. And that the number of trials n is predecided. So, for example, if one draws 5 cards out of a stack, *without* replacement and counts cards above a threshold value, that count won't follow a Binomial distribution - This is because, since he is not replacing the cards drawn back into the stack, the p would change from one Bernoulli trial to another. Similarly, if someone where to take a gender

[1] When a researcher models a survey question with a Yes/No answer, while looking at its statistical analysis, one should not automatically assume that "Yes" is considered 1 and "No" is considered 0. Same thing goes for gender. One should always check how the researcher has assigned R.V. values to the two possibilities. If you are the researcher, it is important to explicitly mention how you have done the assignment

survey and stop when, say, 10 females are reached, the count of males in that survey is not a Binomial R.V., as n was not pre-decided.

Whereas a Binomial distribution comes from conducting n independent Bernoulli trials and counting the successes, a **Multinomial distribution** comes from doing the same, but with a Categorical R.V. So, just as a Categorical distribution is a generalization of the Bernoulli distribution, a Multinomial distribution is a generalization of the Binomial distribution. If we use k to denote the number of successes, i.e., the x-axis of a Binomial distribution, it is implicit that for every value of k, the number of failures is $n-k$. So, in a way, every point on the PMF shows the probability of getting k successes and $n-k$ failures in n Bernoulli trials. In a Multinomial distribution, we could have more than two possible values for the R.V. Hence the PMF is specified for every possible combination of $x_1, x_2, ..., x_k$. where the x_1 is the number of 1s in n trials, x_2 the number of 2s in n trials and so on. And k is basically the total number of values the underlying Categorical R.V. can assume.

Binomial distribution

Notation: $X \sim B(n,p)$

Mean: np

Variance: $np(1-p)$

PMF: $\Pr(X=k) = \binom{n}{k} p^k q^{n-k}$

Multinomial distribution

Notation: $X \sim C(p_1, p_2, ..., p_k)$

Mean: np_i for $X_i, i \in 1,k$

Variance: $np_i(1-p_i)$, for $X_i, i \in 1,k$

PMF: $\Pr(X_1 = x_1 \text{ and } \ldots \text{ and } X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$

If one were to measure the number of Bernoulli trials to get a success, that measure will be an R.V. that follows a **Geometric Distribution**. Examples are, the number of lottery tickets one has to buy before winning the lottery, the number of times one needs to throw a dart before landing it on the board etc. A **Poisson** R.V. counts the number of events of a certain nature within a give time period, when the smallest time-slice used for counting an event tends to zero. Instead of using the time period T as a parameter, the Poisson distribution uses λ, the mean number of desired events, as the parameter. Poisson distribution is basically counting successes in n trials, just like a Binomial distribution, exept n is 1 unit of

time and counting is done in infinitesimal time slices. Examples of Poisson R.V. are number of customers arriving at a restaurant in 1 hour, number of cars crossing an intersection in 2.5 minutes etc. Note that whether it is an hour or 2.5 minutes, in a given context, it is taken as one unit of time and λ is calculated accordingly. Also note that, the concept of counting something within an interval of time can be extended to other things, such as space. For example, number of cars parked between intersection to another. Although the number of events that can occur in an interval is always an integer, the mean number of events need not be - So λ can take all positive real values.

Geometric distribution

Notation: $X \sim G(p)$

Mean: $\frac{1}{p}$

Variance: $\frac{1-p}{p_2}$

PMF: $\Pr(X = k) = (1-p)^{k-1}p$

Poisson distribution

Notation: $X \sim Poiss(\lambda), \lambda \in \mathbb{R}^+$

Mean: λ

Variance: λ

PMF: $\Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$

Until now, we have only talked about discrete distributions. Now let us look at some continuous distributions. In a Poisson distribution, we were looking at, say, the number of customers arriving at a restaurant. If one were to model the waiting time until the arrival of the first customer as a random variable, it would follow an **Exponential distribution**. It is entirely defined by one parameter θ, the average waiting period. Note that, θ is nothing but the inverse of Poisson λ.

Exponential distribution

Notation: $X \sim Exp(\theta), \theta \in \mathbb{R}^+$

Mean: θ

Variance: θ^2

PDF: $f_W(w) = \frac{1}{\theta} e^{-w/\theta}$

Suppose one were to generalize the exponential random variable to the average waiting period until the arrival of the kth customer, that R.V. would follow a **Gamma distribution**. Although α in this context is an integer, Gamma distribution is used in some other contexts (they tell me), where k is not an integer. The PDF of the Gamma distribution, when k is an integer can be written as:

$$F_W(w; k, \theta) = \frac{1}{(k-1)!\theta^k} w^{k-1} e^{-w/\theta}, \ k \in \mathbb{Z}^+, \ \theta \in \mathbb{R}^+$$

But to generalize this PDF for when k is not an integer, one needs a continuous function equalent of the factorial function. This is where the Gamma function comes in, and hence the name "Gamma distribution". Gamma function is a extension of the factorial function to complex numbers when the real part of the complex number is positive.

Gamma Function
Definition:

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \, dx, \qquad \Re(z) > 0$$

Properties:

$$\Gamma(z) = (z-1)\Gamma(z-1)$$

Hence, when z in an integer number

$$\Gamma(z) = (z-1)!$$

With this definition of the Gamma function, one can write the PDF of the Gamma distribution as,

$$F_W(w; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \ \alpha \in \mathbb{R}^+, \beta \in \mathbb{R}^+$$

where, α is called the "shape" parameter, and β, the "rate" parameter. β is simply $1/\theta$, and hence the equivalent of the Poisson λ parameter.

A **Chi-Square distribution** is a special case of a (integer) Gamma distribution, when $\theta = 2$ and $k = r/2$, where r is any positive integer, and is called "The degrees of freedom". The importance of Chi-Square distribution will be apparent later when we deal with statistical estimation and hypothesis testing.

Gamma distribution

Notation: $X \sim Gamma(\alpha, \beta),\ \alpha \in \mathbb{R}^+, \beta \in \mathbb{R}^+$

Mean: $\dfrac{\alpha}{\beta}$

Variance: $\dfrac{\alpha}{\beta^2}$

PDF: $F_W(w; \alpha, \beta) = \dfrac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$

Chi-Square distribution

Notation: $X \sim \chi^2(r)$ or χ^2_r

Mean: r

Variance: r

PDF: $F_X(x; r) = \dfrac{1}{\Gamma(r/2)2^{r/2}}\ x^{r/2-1} e^{-x/2}$

When several independent and identically distribution random variables are added together, the resulting random variable follows a **Normal distribution** when the number of random variables added together is very large. This is irrespective what probability distribution the random variables individually followed (This is called the "Central Limit Theorem" and will be explained further in the next section). So, for example, if one were to sum a lot of Binomail R.V.s or Poisson R.V.s together, the sum would follow a Normal distribution. A Normal distribution is entirely defined by two parameters namely the mean, μ and the standard deviation σ.

Normal distribution

Notation: $X \sim \mathcal{N}(\mu, \sigma^2)$

Mean: μ

Variance: σ^2

PDF: $\dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

1.2 MOTIVATION FOR STATISTICAL ANALYSIS

It is often the case that facts about things that vary in nature are impossible to find: For example, the average weight of oranges coming from the Nagpur region, or the median height of 10 year olds in India etc. It is impossible to weigh every single orange ever grown in the Nagpur region and then find the average weight; It is impossible to measure the height of every single 10 year old in India. Yet, if you are starting an orange juice business or if you are going make uniforms for 10 year olds, you need to know these facts. If we cannot find the facts, we have to find out good estimates of those facts. Statistics is the field that deals with these estiamtes: What is a good estimate, What is the best way of getting that estimate, What can we say about how close our estimate is to the actual fact etc. The basic idea is that we take samples of oranges or students - however many samples as it is practical - and see what we can say about the facts we are after from these sample data points. The measure we obtain from these samples is called ***'a statistic'***. So the average weight of 20 oranges that we sampled or the median height of 100 ten-year olds, is a statistic.

TERMINOLOGY:
Population Mean : The average weight of all oranges in Nagpur region. *i.e.*, the fact
Sample Mean: The average weight of a sample set, of say, 20, oranges. *i.e.*, the statistic

The same terminology applies to other measures such as median, mode etc.

1.3 QUANTIFYING CONFIDENCE

Is the average weight of n oranges a good estimate for the average weight of all the oranges grown in the Nagpur region? Our intuition says that this should be the case. But let us verify using mathematics. Let us say that the population mean is μ and let us call the average weight of 'n' oranges as A_n, i.e.,

$$A_n = \frac{X_1 + X_2 + ... + X_n}{n}$$

where, $X1, X2$ etc. are placeholder random variables for the first orange, second orange and so on. For A_n is a good estimate of μ, it should satifsfy the following criteria:

1. A_n should be unbiased. i.e., $E[A_n] = \mu$
2. The error, say, ϵ, between the estimate and the population mean should decrease with increasing n, eventually reaching 0, i.e., $\lim_{n\to\infty} \epsilon = 0$

If the estimation error is unbiased, it means that the estimation is as much likely to be erroneous on the higher side as it will be on the lower side of the μ. Otherwise, if the estimate is biased, we need to then determine the bias - and we would still be confused as to how to apply the bias to a single trial of A_n, as there is no way to tell which side of μ this sample of A_n falls into! If A_n is unbiased, then we don't have to worry about extra steps or confusions.

Anyways, it is very easy to verify this for A_n. Since $X1, X2, ..$ are independent of one another[2],

$$E[A_n] = \frac{E[X_1] + E[X_2] + ... + E[X_n]}{n} = \frac{n * \mu}{n} = \mu$$

Figure 1.4 shows multiple trials for every value of n from 10 to 1000. Different colours represent different samples of A_n for the same n value. We can clearly see that the samples of A_n are spread evenly on the positive and negative sides of the population mean, μ.

As for the second condition, $\lim_{n\to\infty} \epsilon = 0$, we need to first define the error: It is a distance metric between A_n and μ. Since A_n itself is a random variable, a suitable distance metric is the probablistic variant of euclidean distance:

$$d(A_n, \mu) = \sqrt{E[(A_n - \mu)^2]}$$

. Since $E[A_n] = \mu$, this distance metric is nothing but the Standard Deviation of A_n! So, let us calculate the Standard Deviation of A_n:

$$\begin{aligned} Var[A_n] &= \frac{1}{n^2} * Var[X_1 + X_2 + ... + X_n] \\ &= \frac{1}{n^2} * E[(X_1 + X_2 + ... + X_n)^2 - n\mu^2] \\ &= \frac{1}{n^2} * E[n E[{X_k}^2] + n(n-1)\mu^2 - n\mu^2],\ k \in [1, n] \\ &= \frac{\sigma^2}{n} \\ \therefore, SD[A_n] &= \frac{\sigma}{\sqrt{n}} \end{aligned}$$

As n increases, the Standard Deviation of A_n decreases as $\sqrt{n}$. And, as n approaches infinity, the Standard Deviation will approach zero. Figure 1.5 shows how the spread of A_n decreases with increasing n value.

1.3.1 Law of Large Numbers

We have established that the arithmetic average function is a good estimator of the population mean. But the arithmetic average of random variables itself is a

[2]The weight of the second orange won't magically change because the first orange we selected has a certain weight

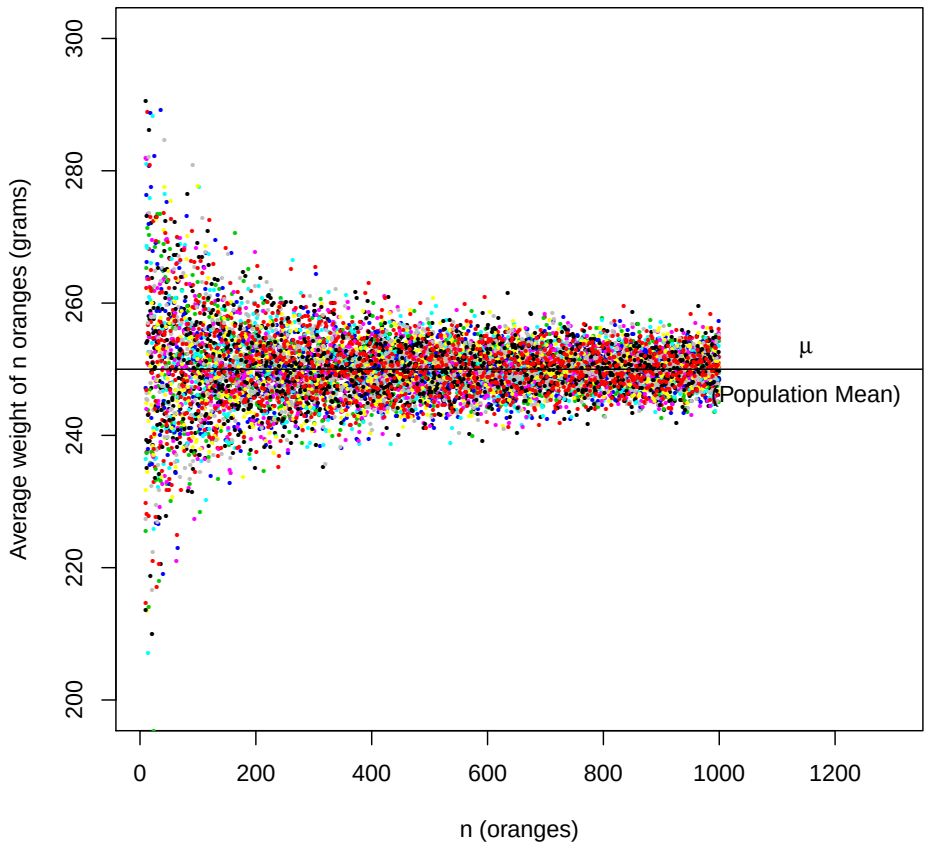


Figure 1.4: Demonstration of unbiasedness of A_n

random variable: Knowing its mean and Standard Deviation still doesn't guarantee that, if we conduct one trial and make one sample of A_n, it will not be wildly far away from the population mean. We won't be able to conduct several trials and come up with several samples of A_n either: The very thing that got us think about finding an estimate for the population mean is that finding the exact population mean would require resources we cannot muster and gathering n samples instead is affordable. If we can afford conduct serveral trials, then one might as well make an arithmetic average of all the samples of all the trials and call the new n value as the total number of samples used for that average. In that case, we will be back starting at the fact that we have a single sample of A_n and that could be widely far away from the population mean. But it would be such a shame if we can't say *anything* about the size of the estimation error and, well, the field of statistics won't exist. This is where the law of large numbers comes into picture.

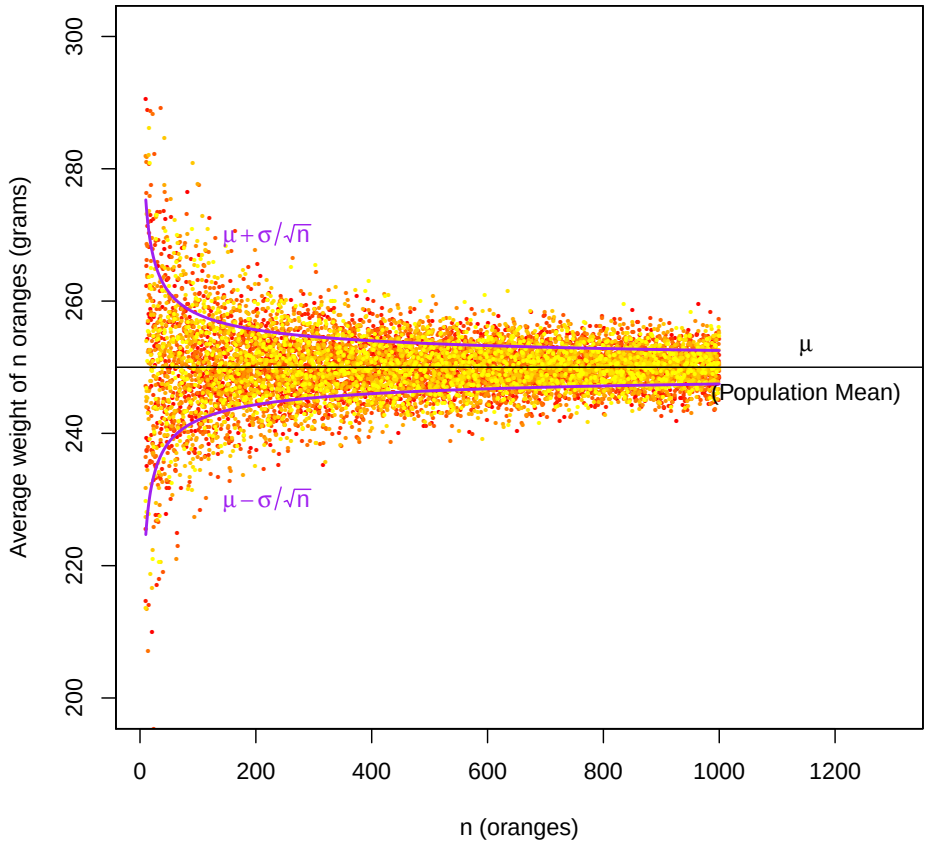


Figure 1.5: Reduction in the spread of A_n values with increasing n

There are two such laws, one called WLLN - the weak law of large numbers, and the other SLLN - the strong law of large numbers. The weak law states that the probability that the estimation error is greater than a certain bound will be lesser than a certain bound. And that, as n becomes infinity, that probability will be zero.

Theorem 3.1: Weak law of large numbers

$$P(|A_n - \mu| > \xi) \leq \frac{\sigma^2}{n\xi^2}$$
$$\therefore, \lim_{n\to\infty} P(|A_n - \mu| > \xi) = 0$$

where, ξ is any arbitrary non-negative value

The proof is simple: We expand the Variance formula for A_n and segregate all items above ξ as shown below:

$$\begin{aligned} Var[A_n] = \ldots &+ (-\xi-\delta 1)^2 P(A_n - \mu = -\xi - \delta 1) + (-\xi - \delta 2)^2 P(A_n - \mu = -\xi - \delta 2) + \ldots. \\ &+ (-\xi)^2 P(A_n - \mu = -\xi) \\ &+ (-\xi + \alpha 1)^2 P(A_n - \mu = -\xi + \alpha 1) + (-\xi + \alpha 2)^2 P(A_n - \mu = -\xi + \alpha 2) + \ldots. \\ &+ (\xi)^2 P(A_n - \mu = \xi) \\ &+ (\xi + \beta 1)^2 P(A_n - \mu = \xi + \beta 1) + (\xi + \beta 2)^2 P(A_n - \mu = -\xi + \beta 2) + \ldots. \\ &+ (\xi + \eta 1)^2 P(A_n - \mu = \xi + \eta 1) + (\xi + \eta 2)^2 P(A_n - \mu = \xi + \eta 2) + \ldots. \end{aligned}$$

The terms in blue have probabilities of $|A_n - \mu| > \xi$. Noting that all probabilities are non-negative and the terms in front of the probabilities are also non-negative, the sum of all the terms in black will also be non-negative. Using this fact, and the fact that $Var[A_n] = \frac{\sigma^2}{n}$, we have,

$$\begin{aligned} \frac{\sigma^2}{n} \geq \ldots &+ (-\xi-\delta 1)^2 P(A_n - \mu = -\xi - \delta 1) + (-\xi - \delta 2)^2 P(A_n - \mu = -\xi - \delta 2) + \ldots. \\ &+ (\xi + \eta 1)^2 P(A_n - \mu = \xi + \eta 1) + (\xi + \eta 2)^2 P(A_n - \mu = \xi + \eta 2) + \ldots. \end{aligned}$$

Now, using the same logic that we used to eliminate the terms that were black in colour, we have substitute ξ in place of every term in front of the probabilities and we get,

$$\begin{aligned} \frac{\sigma^2}{n} \geq \ldots &+ (\xi)^2 P(A_n - \mu = -\xi - \delta 1) + (\xi)^2 P(A_n - \mu = -\xi - \delta 2) + \ldots. \\ &+ (\xi)^2 P(A_n - \mu = \xi + \eta 1) + (\xi)^2 P(A_n - \mu = \xi + \eta 2) + \ldots. \end{aligned}$$

Notice how we have removed the minus signs in the first line of the equation above? Well, it is because all the terms are squares - So it doesn't matter. Bringing all the ξ^2 to the left hand side of the equation, we get,

$$\begin{aligned} \frac{\sigma^2}{n\xi^2} \geq \ldots &+ P(A_n - \mu = -\xi - \delta 1) + P(A_n - \mu = -\xi - \delta 2) + \ldots. \\ &+ P(A_n - \mu = \xi + \eta 1) + P(A_n - \mu = \xi + \eta 2) + \ldots. \end{aligned}$$

The terms on the right side of the equation, when added together, essentially boil down to $P(|An-\mu|>\xi)$. Hence we get,

$$\frac{\sigma^2}{n\xi^2} \geq P(|A_n-\mu|>\xi)$$

Since Standard Deviation of A_n is $\frac{\sigma}{\sqrt{n}}$, we can also rewrite the equation as,

$$P(|A_n-\mu|>\xi) \leq \frac{SD[A_n]^2}{\xi^2}$$

From the WLLN, we see that, knowing the number of samples and population variance, we could express our confidence that our estimate of the population mean is *probably* accurate. In other words, we are quantifying our confidence. Just as Information Theory quantified information and built several useful constructs on top of that, Statistics builds many useful constructs based on the idea that confidence, and indirectly, chaos, can be quantified. In the next sections, we formally describe various methods in Statistical analysis for estimating population unknowns (**Estimation**), validating our assumptions about the population (**Hypothesis Testing**) etc.

Note that, so far, we have only assumed that $X_1, X_2, ..., X_n$ are independent. When we formally introduce Estimation Theory, we can see that, depending upon how much more we know about these random variables (such as their probability distribution), we can get much tighter bounds on the error.

1.4 ESTIMATION

The population distribution and its parameters (mean, median, variance etc.) are all facts. There is no concept of likelihood associated with them. So the population is not a random variable and each possibility just has a frequency associated with it, and not a "probability". However, when you draw a random sample, $X_1, X_2, ..., X_n$, each X_k is a random variable. And each X_i has a probability distribution the same as the frequency distribution of the population, with the relative frequency as the probability [3]. The reason of this is that the X_is are just placeholders for the ith draw and hence the probabilities of drawing a certain values is same as the relative frequency of it in the population. **Bear in mind, though, that this is true, strictly, only when we put back a drawn sample member back into the population before the next draw**. This is an important concept. Because if we don't put back what we drew from the population, the relative frequencies of all events/measurements/samples change because now there is one less member in that population. For, ex., suppose we are trying to estimate the mean weight of an

[3] In fact, one way of looking at probability is that it is the relative frequency of a sample as the sample size approaches infinity, or, in our case, the size of the population

orange in a crate, after randomly taking one orange and measuring its weight, we should put it back into the crate and draw the next random orange.

The reality is a little bit different though. Often once we take one member out of the population, we don't put it back. Take the case of estimating the height of ten year olds in the state of Tamil Nadu - We will proceed by first choosing a sample size (which will depend upon our budget and other practical considerations [4]). Let us say, that we decide on a sample size of 50. We are then likely to randomly select 50 ten year olds, measure their heights and call it our sample. This goes against the idea of putting back a selected individual back into the population before we select the next individual. But this is considered fine, as often, the population is very large compared to the sample size, and not putting back an ten year old into a population of 10 million ten year olds in Tamil Nadu, doesn't significantly alter the probabilities of any individual ten year old in the population from being selected in the next draw. Some Statisticians are OK with not putting individuals back, as long as the sample size is less than 1% of the population. Some are even OK with it when the sample size is less than 10% of the population. It is really up to the researcher at the end how much of an approximation error in the estimate is OK.

1.4.1 Maximum Likelihood Estimation

One way to find an appropriate estimator for a population parameter, say θ, for a given distribution, is to look at the random sample $X_1, X_2, ..., X_n$ and ask, "what θ value increases the joint probability of getting $X_1, X_2, ..., X_n$?". In other words, suppose we define a function, $L(\theta)$ as,

$$L(\theta) = f(x_1;\theta).f(x_2;\theta)....f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

then, then θ value that maximizes $L(\theta)$ is the **Maximum Likelihood Estimate** (MLE) $\hat{\theta}$ of θ. $L(\theta)$ is called the *likelihood function*.

Let us take the example of estimating a population proportion. The measurements X_is in this case, will (logically) follow a Bernoulli distribution. In other words,

$$f(x_i;p) = p^{x_i}(1-p)^{1-x_i}$$

So the likelihood function is,

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}$$

To find the p that maximizes $L(p)$, we have first take the derivative of $L(p)$ with respect to p and equate it to zero, i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}p}(p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}) = 0$$

[4] After all, Statistics is all about what is practically possible

Differentiating this is difficult. But we can use a neat trick: Since the likelihood function is a product of several probability values, it will always be a non-negative value. And we know that the logarithm function $log(\alpha)$ is an increasing function of α. So taking the derivative of the logarithm of $L(p)$ and equating it to 0 will have the same effect of taking the derivative of $L(p)$ and equating it to 0. So, we have,

$$\frac{\mathrm{d}}{\mathrm{d}p} log(p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}) = 0$$

This, after some more derivation will yield,

$$\begin{aligned} \Sigma x_i(1-p) - (n-\Sigma x_i)p &= 0 \\ \Sigma x_i - np &= 0 \\ p &= \frac{\Sigma x_i}{n} \end{aligned}$$

To ensure that this extrema is actually the maxima, we need to take double derivative of $L(p)$ and evaluate it at $p = \frac{\Sigma x_i}{n}$ and make sure it is indeed a negative value. That is beyond the scope of this notes. Assuming that this is indeed the maxima, all we have to do now, to state the MLE for p, is use the random variable X_i in place of the sample x_i in the above result:

$$\hat{p}_{MLE} = \frac{\Sigma X_i}{n}$$

In the case of proportions, we defined the likelihood function in terms of one parameter, p. But it can be generalized to multiple parameters as,

$$L(\theta_1, \theta_2, \ldots, \theta_m) = \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2, \ldots, \theta_m)$$

Let us look at an example where we derive the Maximum Likelihood Estimate of the mean and the variance of a Normally distributed population, from a random sample. The likelihood function in this case is:

$$L(\mu, \sigma) = \sigma^{-n}(2\pi)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

As usual, we will take the log of the likelihood function before we move on to finding the extremas. But before that, for notational convenience, let us rewrite the likelihood function with generic parameters θ_1 and θ_2, in the place of μ and σ^2 respectively:

$$L(\theta_1, \theta_2) = \theta_2^{-n/2}(2\pi)^{-n/2} \exp\left[-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i-\theta_1)^2\right]$$

Now we take the log of the likelihood function:

$$\log L(\theta_1, \theta_2) = -\frac{n}{2}\log\theta_2 - \frac{n}{2}\log(2\pi) - \frac{\sum(x_i - \theta_1)^2}{2\theta_2}$$

Since we have two parameters, we need to do partial differentiations in order to find the extremas of the likelihood function with respect to one parameter at a time. First we do it for θ_1:

$$\frac{\partial}{\partial\theta_1} L(\theta_1, \theta_2) = \frac{-2\Sigma(x_i - \theta_1)(-1)}{2\theta_2} = 0$$

Resolving the equation, we get the estimate for θ_1 as,

$$\hat{\theta}_1 = \hat{\mu} = \frac{\sum x_i}{n} = \bar{x}$$

Now, doing the same for θ_2, we get,

$$\frac{\partial}{\partial\theta_2} L(\theta_1, \theta_2) = \frac{-n}{2\theta_2} + \frac{\Sigma(x_i - \theta_1)^2}{2\theta_2^2} = 0$$

Multiplying both sides of the above equation by $2\theta_2^2$ and solving for θ_2, we get,

$$-n\theta_2 + \sum(x_i - \theta_1)^2 = 0$$

$$\hat{\theta}_2 = \hat{\sigma}^2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

Again, taking the second partial derivatives to confirm that the extremas are indeed maximas are beyond the scope of this notes. Assuming that this are indeed the maximas, all we have to do now, to state the MLE for the mean and the variance, is by using the random variable X_i in place of the sample x_i in the above results:

$$\hat{\mu}_{MLE} = \frac{\sum X_i}{n} = \bar{X}$$

$$\hat{\sigma}^2_{MLE} = \frac{\sum(X_i - \bar{X})^2}{n}$$

1.4.2 Method of Moments

The method of moments involves equating sample moments with theoretical moments. So, let's start with their definitions

Definition 4.1: Moments

$E(X^k)$is the kth theoretical moment about the origin

$E\left[(X-\mu)^k\right]$is the kth theoretical moment about the mean

$M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$is the kth sample moment about the origin

$M_k^* = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$is the kth sample moment about the mean

The idea of the method of moments is that, we create as many simultaneous equations as there are parameters to be estimated, by equating sample moments to theoretical moments. One can use moments about the origin or about the mean or a combination of both. For Bernoulli distribution's p or Normal distribution's μ and σ, the **Method of Moments Estimate** (MM) ends up being trivial, as the paramters to be estimated themselves are theoretical moments - so we just use their sample moment equivalent as their corresponding estimates. But, let us look at the example of a Gamma distribution,

$$f(x_i) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-x/\theta}$$

Here, the parameters α and θ are not the moments of the Gamma distributions. We use the MM Estimate for α, θ as shown below:

$$E(X) = \alpha\theta = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

$$Var(X) = \alpha\theta^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

From the first equation, we get α in terms of θ:

$$\alpha = \frac{\bar{X}}{\theta}$$

We substitute this in the second equation and get,

$$\alpha\theta^2 = \left(\frac{\bar{X}}{\theta}\right)\theta^2 = \bar{X}\theta = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Now, we can solve for θ and then use it to calculate α :

$$\hat{\theta}_{MM} = \frac{1}{n\bar{X}} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

$$\hat{\alpha}_{MM} = \frac{\bar{X}}{\hat{\theta}_{MM}} = \frac{\bar{X}}{(1/n\bar{X}) \sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{n\bar{X}^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

1.4.3 Unbiased Estimate

We have so far seen two ways of coming up with estimates for distribution parameters. There may be even more ways of coming up with estimates. We need a way to evaluate how good these estimates are before we use them. One criterion is whether or not they are unbiased estimates.

Definition 4.2

n estimate, $\hat{\theta}$ is said to be **Unbiased Estimate** of the parameter θ if:

$$E[\hat{\theta}] = \theta$$

For example, let us look at the MLE for the variance of the normal distribution:

$$\hat{\sigma}^2_{MLE} = \left(\frac{1}{n} \sum_{i=1}^{n} X_i^2 \right) - \bar{X}^2$$

Taking the expectation of this Estimate, we get:

$$\begin{aligned} E(\hat{\sigma}^2_{MLE}) &= E\left[\frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}^2 \right] = \left[\frac{1}{n} \sum_{i=1}^{n} E(X_i^2) \right] - E(\bar{X}^2) \\ &= \frac{1}{n} \sum_{i=1}^{n} (\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2 \right) \\ &= \frac{1}{n} (n\sigma^2 + n\mu^2) - \frac{\sigma^2}{n} - \mu^2 \\ &= \sigma^2 - \frac{\sigma^2}{n} = \frac{n\sigma^2 - \sigma^2}{n} = \frac{(n-1)\sigma^2}{n} \end{aligned}$$

Since it doesn't meet the criterion, the MLE for the variance of the normal distribution is *not* an unbiased estimate. So, what do we do now? Well, we can simply massage the last equation above as:

$$\frac{n}{n-1} E(\hat{\sigma}^2_{MLE}) = \sigma^2$$

$$E(\frac{n}{n-1}\hat{\sigma}^2_{MLE}) = \sigma^2$$

In other words, we can get the unbiased estimate of variance by multiplying the MLE for variance with $\frac{n}{n-1}$. This will result in:

$$\hat{\sigma}^2 = \frac{\left(\sum_{i=1}^{n} X_i - \bar{X}\right)^2}{n-1}$$

1.4.4 Estimating Mean When Variance is Known

Suppose we are interested in estimating the mean of a population which is Normally distributed, where we know the Variance. This happens in situations where we are manufacturing, say resistors, and the machine used already has a known error spread.

We already know that the sample average is a good estimate (It is the MLE, MM and unbiased estimate of μ

$$\bar{X} = \frac{\sum X_i}{n}$$

While we know that $\bar{X}$ is a good estimate of μ, we also know that there will be some error in the estimation. WLLN gives us a way of quantifying the error with quantifying our level of confidence in that error. We recall it below for convenience:

$$P(|\bar{x} - \mu| > \xi) \leq \frac{\sigma^2}{n\xi^2}$$

Now when we derived WLLN, we mentioned how, knowing more information about the X_i values could give us tighter confidence bounds on the error. So let us see what all we know in this situation.

It is a fair assumption that the X_i are themselves Normally distributed with mean μ and variance σ^2. They are also independent of one another. We know that the sum of two Normally distributed, independent, random variables, will result in a Normally distributed random variable, with mean and variance equal to the sum of the means and sum of the variances of those two random variables (Proof using convolution property). We also know any linear transformation of a Normal R.V. will result in a Normal R.V. (see proof)[5]. So, considering these facts, we can say that $\bar{X}$ will also be a Normal random variable. And we also know how to calculate its mean and variance. Here it is:

$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$$

[5]The trick to deriving the pdf of a transformation of a continuous R.V. is through the CDF route - This is because, one will have to calculate the probability of the transformed variable, and for a continuous R.V. probabilities have to be derived from the CDF, as the probability of continuous R.V. to have an exact value is zero

Knowing this, instead of an upper bound on the error, we can actually find the exact probability, $P(|\bar{X}-\mu| > \xi)$ - In Estimation Theory, the actual practice is to instead find $P(|\bar{X}-\mu| \leq \xi)$, which basically conveys the same meaning. Let us call this probibility value p_v. We can derive p_v from the probability density funciton of $\bar{X}$:

$$p_v = P(|\bar{X}-\mu| \leq \xi) = P(\mu-\xi \leq \bar{X} \leq \mu+\xi) = \int_{\mu-\xi}^{\mu+\xi} \frac{1}{\sqrt{2\pi}(\sigma/\sqrt{n})} e^{-\frac{1}{2}\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)^2} \mathrm{d}\bar{x}$$

Note that,

$$P(|\bar{X}-\mu| \leq \xi) = P(|\mu-\bar{X}| \leq \xi)$$

. This also means,

$$P(\mu-\xi \leq \bar{X} \leq \mu+\xi) = P(\bar{X}-\xi \leq \mu \leq \bar{X}+\xi)$$

In other words, the result of that integral can be interpreted as the probability that a calculated sample mean value $\bar{x}$ is within a certain distance from the population mean, or as the probability that the population mean μ is within a certain distance from the sample mean value $\bar{x}$. The latter statement is considered somewhat sacrilegious by statisticians, because they think no one should be making probability statement about a fact - i.e., here, the population mean. But nevertheless, this equivalence concept is handy because, what we know is $\bar{x}$, so it makes sense to say that we are certain cofident with a certain probability that μ lies within a certain distance of $\bar{x}$. In other words, instead of thinking of the distribution of the random variable $\bar{X}$ as a Normal distribution with population μ at its centre, we can think of the distribution of population μ with $\bar{x}$ (current sample mean, not the R.V), at the centre. Although μ is a constant, when it comes to the probability of the error, these two scenarios are equivalent.

Anyway, once we calculate p_v from that integral, suppose it is say, 0.75, then we can make a statement, that, *"we are confident that, 75 out of 100 times we conduct a trial (observe a $\bar{X}$ value), the population mean will be within $\pm\xi$ of the observed value of $\bar{X}$"*. Or as statisticians prefer to say, *"we are 75% confident that the population mean will be within $\pm\xi$ of the observed value of $\bar{X}$"*.

There is however one big problem: It turns out that the integral for calculating p_v is a pain in the posterior to evaluate analytically. So one has to use numerical methods to find out the result. If we are gong to calculate numerically, then it made sense that we do the exercise of calculating the integral for various limit values once and store the results in a lookup table - The next time we need to do this integral we could just look up the result on this table. But, unfortunately, this doesn't solve the problem either: The lookup table thus calculated would contain p_vs for one specific Gaussian - with one specific mean and standard deviation. If we never want to calculate the integral, then we would have to create one look up

table for every possible value of mean and standard deviation. This is impossible! Enter the **Z statistic**:

We know that a linear transformation of a Normally distributed R.V. results in another Normally distributed R.V. Suppose we have an R.V. $X \sim \mathcal{N}(\mu, \sigma^2)$, then we can derive another R.V. such that it has a mean of 0 and a standard deviation of 1. Let us call this new R.V., Z. The transform for calculating Z from X is as below:

$$Z = \frac{X - \mu}{\sigma} \mid Z \sim \mathcal{N}(0, 1)$$

Z is called the **Standard Normal distribution**

In our case, since the random variable we are concerned with is $\bar{X}$, we can derive $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \mid Z \sim \mathcal{N}(0, 1)$. When we do this, we can translate $P(\bar{X} - \xi \leq \mu \leq \bar{X} + \xi)$ into a probabilty on Z lying between new limits - let us call them, $(-z_{\alpha/2}, z_{\alpha/2})$ - such that:

$$\begin{aligned}
-z_{\alpha/2} &\leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \\
-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) &\leq \bar{X} - \mu \leq +z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \\
-\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) &\leq -\mu \leq -\bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \\
\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) &\leq \mu \leq \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)
\end{aligned}$$

Now, we originally set out to find $P(\bar{X} - \xi \leq \mu \leq \bar{X} + \xi)$ - If we compare this with the above inequality, if we set

$$z_{\alpha/2} = \xi \frac{\sigma}{\sqrt{n}}$$

then,

$$P(\bar{X} - \xi \leq \mu \leq \bar{X} + \xi) \equiv P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2})$$

So we could create just one lookup table, called **Z table**, that has $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2})$ values calculated for various $z_{\alpha/2}$ (at some granularity) and then use this table to calculate $P(\bar{X} - \xi \leq \mu \leq \bar{X} + \xi)$ using the above relationship.

A note: The terminology of $z_{\alpha/2}$ is not coincidental. It is common among Statisticians to refer to the Z value above which area under the pdf, i.e., probability, is $\alpha/2$ as $z_{\alpha/2}$. So if we look at the probability between $-z_{\alpha/2}$ and $z_{\alpha/2}$, it will be $1 - \alpha$. i.e., One typical way statisticians ask about our confidence on our estimate $\bar{X}$, is *"What is your 90% confidence interval?"*. Then one would translate it to an α

of 10%, look up on the Z-table what $-z_{0.05}, z_{0.05}$, and finally come with the answer as, $\bar{X} - z_{0.05}\left(\frac{\sigma}{\sqrt{n}}\right), \bar{X} + z_{0.05}\left(\frac{\sigma}{\sqrt{n}}\right)$.

$$\int_{-z_{\alpha/2}}^{z_{\alpha/2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \mathrm{d}z = (1-\alpha)$$

Actually the normal practice among Statisticians, when it comes to estimation theory is not to start with an error threshold value and then answer the question of how confident we are that our estimation error is within that threshold value[6]. Instead they define a **Confidence Interval** $[-z_{\alpha/2}, z_{\alpha/2}]$, such that,

When we are using the Z-statistic, the $(1-\alpha)$% Confidence Interval on $\bar{x}$, i.e., $[\bar{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right),\ \bar{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)]$. is called the **z-interval for the mean**. The most often used confidence intervals are 68%, 95% and 99.7%, because these correspond to $z_{\alpha/2}$ values of 1, 2 and 3 respectively, or euivalently the error being one, two, three sampling standard deviations away, respectively. The standard deviation (or its estimate) of a sampling distribution is called the **standard error**[7].

Remember that z-interval works only if the samples are Normally distributed. What if they follow some other distribution? One way to tackle it is by using a large sample size n, due to Central Limit Theorem, $\bar{X}$ would look close to a Normal distribution. Then we can go ahead and use the z-interval. Of course, one could instead derive, from scratch, the appropriate test for a sample that doesn't have Normally distributed random variables, but that would also involve deriving area under the curve of whatever is the distribution of $\bar{X}$ - Most statistical softwares have tools for calculating z-interval. So, it is better to use, instead, a large n and take advantage of tools in statistical softwares.

1.4.5 Estimating Mean When Variance is Unknown

In most cases of social studies, we don't know the mean and the variance. To estimate the mean in this situation, we cannot use Z-table on the R.V., $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$. The logical thing to do is to substitute σ with Sample Standard Deviation, S, where,

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

. But when we substitute S, which is a Chi-squared distribution, in place of σ in $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ the result will no longer be a Gaussian. It will instead be a Student's T

[6] In Hypothesis Testing, they use both ways

[7] In the *Probability Overview* section, when we talked about Variance, we discussed the idea of standard error

distribution with $n-1$ degrees of freedom:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim T(n-1)$$

The complete proof can be found here., but the basic idea is to recognize that:

- $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$
- $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$
- $\bar{X}$ and S^2 are independent

Anyways, instead of the confidence interval derived based on $z_{\alpha/2}$, we derive the interval based on the equivalent value on the **T-Table** called the **t-interval for mean**:

$$\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$$

Note that, due to the Central Limit Theorem, as n becomes large, the difference between z-interval and t-interval shrinks. Figure 1.6 shows how, as the sample size increases, the t-distribution approaches the Z-distribution. When $n = 30$, there is hardly any difference between the two distributions. So, some statisticians use the $z_{\alpha/2}$ in place of $t_{\alpha/2,n-1}$ in the above equation for the t-interval, if the sample size is greater than 30 [8]. It is not fully clear why they would do that. It may have had some significance before the advent of computers when statisticians where using physical z-table, t-table charts - It is possible that carrying a single page of z-table is more convenient than carrying an entire book of t-tables, as there is one t-table for every n.

1.4.6 Estimating Difference in Means

Quite often, we come across a situation, where we are interested in estimating the difference between two populations: For instance, the difference in response times between sobre and drunk people, the difference between the growth levels of the same species of plants under two different lighting conditions (with all else being equal) etc. In such situations, we can derive confidence intervals for $\mu_X - \mu_Y$, where X and Y are used to describe the two different populations. We won't do the derivations here. One can find them in academic literature such as this one. We will simply mention the intervals in Appendix Statistics - At A Glance. But

[8] In general, the z-interval is tighter than a t-interval as the t-distribution would always have fatter tails than the Z-distribution. So a z-interval will be an underestimation of the estimation error in cases where t-inveral should have been used

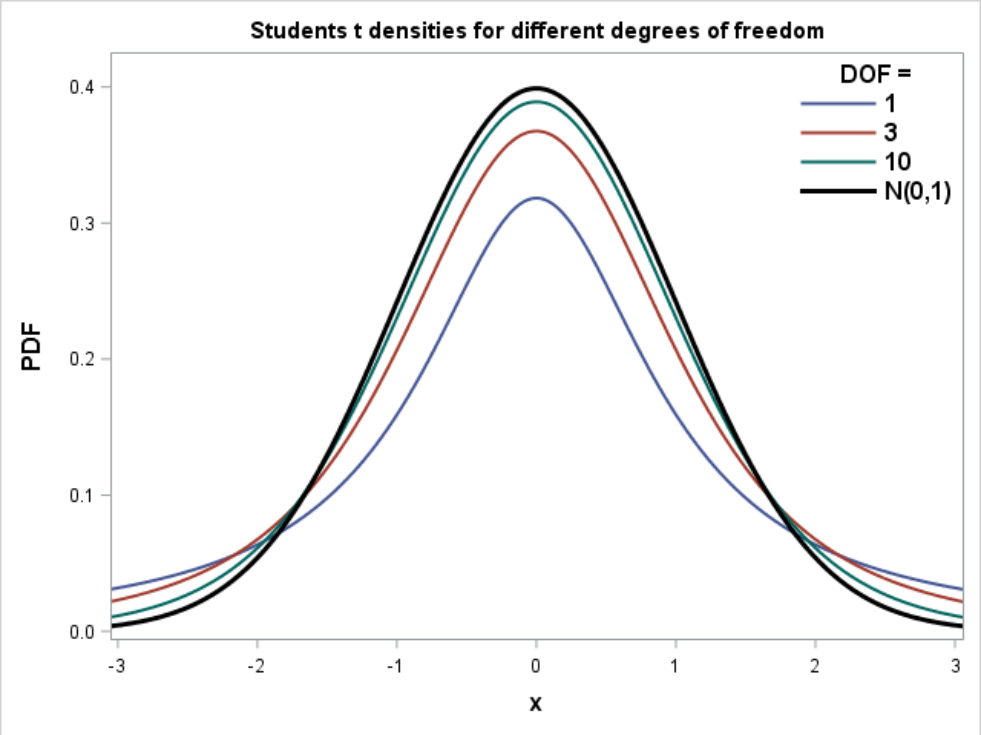


Figure 1.6: T-distribution approaching Z-distribution as n increases

we will describe here when to use which of the different tests when it comes to confidence intervals for differences between means. In all cases, the populations are assumed to be Normally distributed.

We start with the simplest case where we know that the variance of both populations X and Y are the same and are known to us. And if we derive two random samples from X and Y, with sample sizes n and m respectively, then,

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right)$$

We can call $\bar{X} - \bar{Y}$ as $\bar{D}$, with $\mu_D = \mu_X - \mu_Y$ and $\sigma_D^2 = \left(\frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right)$. Then it is a case of a simple z-interval for a Normally distributed random variable as before:

$$\bar{d} \pm z_{\alpha/2} * \sigma_D$$

Th next case is when X and Y are considered to have the same variance, but that variance is unknown. So we need to instead use an unbiased estimate of this unknown variance. It turns out, "pooled sample variance", S_P^2 is an unbiased estimate of the population variance:

$$S_P^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

And, with this, we can come up with a **Pooled t-interval** for difference between means. $\mu_X - \mu_Y$ as:

$$(\bar{X} - \bar{Y}) \pm (t_{\alpha/2,n+m-2}) S_P \sqrt{\frac{1}{n} + \frac{1}{m}}$$

But what if we know that X and Y have different variances and they are unknown? Then we have to use a **Welch's t-interval** as:

$$\bar{X} - \bar{Y} \pm t_{\alpha/2,r} \sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$$

where, r degrees of freedom is approximated by:

$$r = \frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{(s_X^2/n)^2}{n-1} + \frac{(s_Y^2/m)^2}{m-1}}$$

The general practice is to use the Welch's t-interval when $\frac{s_X^2}{s_Y^2} > 4$, or when $\frac{s_Y^2}{s_X^2} > 4$. Otherwise, Pooled t-interval is used.

There is one more interesting case left, which is when, we are interested in estimating the mean of the differences, instead of the differences of means. As an example, suppose we are interested in knowing the effect of a certain fertilizer on the yield of many types of vegetable plants: Here it doesn't make sense to take a random sample from a group of vegetables where the fertilizer was not used (say, Y) and another from a group where the fertilizer was used (say, X). This is because, there may already be inherent differences in yields between different vegetables. So, for instance, if a tomato plant, under any condition, is expected to give more fruits per plant, than a brinjal plant under the same condition, then it doesn't make sense to take a random sample where there may be 5 tomato plants and 3 brinjal plants in the group of vegetables where the fertilizer wasn't used and 2 tomato plants and 6 brinjal plants in the group of vegetables where the fertilizer was used. Instead it may make sense to pair plants - For every measurement X_i, we measure a Y_i where both measurements are taken from the same plant type. That way, we can literally compare "apples to apples". So in this case, we find the differences between paired measurements, i.e., $D_i = X_i - Y_i$ and then estimate the mean of D, the population difference. We can come up with a **Paired t-interval** for μ_D as:

$$\bar{d} \pm t_{\alpha/2,n-1} \left(\frac{s_d}{\sqrt{n}}\right)$$

There are many instances when both X_i and Y_i are the same person in an experiment: Some examples are, the weight of persons before and after a diet regime, the effect of a medicine on sugar levels of diabetic persons etc. In all these cases, the above Paired t-interval is used.

1.4.7 Confidence Intervals for Variances

Estimation of variance is very important in manufacturing. Most goods manufactured are spec'ed out both for their nominal values (1/4" screw, 3.3K Ohm resistor etc.), as well as their tolerances often specified as $pmx\%$. If the sample measurements are normally distributed, then, $(1-\alpha)\%$ **confidernce intervals for variance and standard deviation** are:

$$\left(\frac{(n-1)}{\chi^2_{\alpha/2,n-1}} S^2 \leq \sigma^2 \leq S^2 \frac{(n-1)}{\chi^2_{1-\alpha/2,n-1}} \right)$$

$$\left(\frac{\sqrt{(n-1)}}{\sqrt{\chi^2_{\alpha/2,n-1}}} S \leq \sigma \leq S \frac{\sqrt{(n-1)}}{\sqrt{\chi^2_{1-\alpha/2,n-1}}} \right)$$

We won't derive the proof here, but it starts with realising that,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Sometimes we need to estimate the ratio of two variances. There is a reason why we want to estimate the ratio of variances as opposed to difference in variances (like we do for means): For example, if you are a shirt manufacturer manufacturing clothese for Chinese as well as Indians, the differences between the mean height of Chinese and the mean height of Indians will tell you how much more cloth will be used on an average to make a shirt for an Indian compared to that for a Chinese. But variance will tell you how many SKUs need to be made for Indians vs. Chinese. A ratio makes sense here, because suppose variance in the height of Indians is twice as much as that of Chinese, then No. of SKUs for Indians = No. of SKUs for Chinese $* \frac{\sigma^2_I ndians}{\sigma^2_C hinese}$. Another reason for using the ratio is that, one can make sense of the square root of estimated ratio of variances as an estimate of the ratio of stardard deviations - and standard deviations are important as they have the same unit of measure as the measured quantity. However we can't say much about the difference in standard deviations between populations by looking at the estimate of the difference in their variances. For these reason, you can see that, even while estimating one variance, we start off by looking at the ratio of unbiased estimate of the variance to the variance.

To find the confidence interval for ratio of variances, we first start with the following Chi-squared distributions related to the population and sample variances

of R.V.s X and Y in question:

$$\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi^2_{n-1} \text{ and } \frac{(m-1)S_Y^2}{\sigma_Y^2} \sim \chi^2_{m-1}$$

Here n and m are the number of samples used for calculating sample variances of X and Y respectively. Then we observed that, when X and Y are independent, the following ratio is a F-distribution of m, n degrees of freedom:

$$\frac{\chi^2_{m-1}/(m-1)}{\chi^2_{n-1}/(n-1)} \sim F(m-1, n-1)$$

i.e.,

$$F = \frac{\frac{(m-1)S_Y^2}{\sigma_Y^2}/(m-1)}{\frac{(n-1)S_X^2}{\sigma_X^2}/(n-1)} = \frac{\sigma_X^2}{\sigma_Y^2} \cdot \frac{S_Y^2}{S_X^2} \sim F(m-1, n-1)$$

Therefore, the following probability statement holds:

$$P\left[F_{1-\frac{\alpha}{2}}(m-1, n-1) \leq \frac{\sigma_X^2}{\sigma_Y^2} \cdot \frac{S_Y^2}{S_X^2} \leq F_{\frac{\alpha}{2}}(m-1, n-1)\right] = 1-\alpha$$

From this, we can device confidence interval for ratio of variances as:

$$\left(\frac{1}{F_{\alpha/2}(n-1, m-1)} \frac{s_X^2}{s_Y^2} \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq F_{\alpha/2}(m-1, n-1) \frac{s_X^2}{s_Y^2}\right)$$

1.4.8 Confidence Intervals for Proportions

If one were to draw a random sample from a population that has a Bernoulli distribution, and count the number of time one of the two possible values occurred (no. of males or no. of heads etc.), the resulting random variable will have, a Binomial distribution. So one can derive confidence intervals based on Binomial distribution PMF. However the PMF becomes difficult to calculate[9] when the sample size increases. At the same time, the count also starts looking more and more like a Normal distribution. This is because, if one assigned a value of 1 for the one of the two possible values and 0 for the other, then $X_1 + X_2 + ... + Xn$ will result in a Normally distributed R.V. as the X_is are independent and identitically distributed (C.L.T). And if we also realize that the sample proportion $\hat{p} = \sum_{i=0}^{n} \frac{X_i}{n}$ is also the sample mean. So taking these two facts together, if the sample size is large, one could use a z-interval for means as the confidence interval for proportions, i.e.,

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

[9] It is a question of how much computational power one has

But the problem in the above interval is that it is unreasonable that we would know the population variance $p(1-p)$ directly, not could we have found it from the value of p, the value of which is the very thing we are trying to estimate. And we cannot use a t-interval like we did while estimating the mean of a Normally distributed population with the help of the sample standard deviation: For a Normally distributed population, the sample standard deviation is Chi-square distributed and hence the t-statistic is T-distributed. But in our present case of Bernoulli distributed R.V.'s, the sample standard deviation won't be Chi-square distributed. So, what do we do now? Well, the logical thing to do is to simply substitute p with $\hat{p}$, but keep using the z-interval:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Of course, this is an approximation, but we can't do any better.

So, we have been assuming that, when we draw a sample member from the population, we put it back before we make the next draw. But we often take surveys with "Yes" or "No" answers, where there is no concept of putting back an earlier draw before the next draw. In that case, $X_1 + X_2 + ... + X_n$ will not be a Binomial distriution, as Binomial requires that we put back a sample before the next draw. It is still fine to use Binomial distribution (extended to Normal R.V) in many cases because the sample size is typically very small when compared with the size of the population. But what about finite populations? For ex., if we are surveying the number of people in a village who approve a recent panchayat decision. Suppose the population of the village is only 2000, and we want to take a quick survey of just 50 people, and from them, gauge the approval rating for the panchayat. In this case, we cannot use a binomial distribution. Instead, it would follow a hyper-geometric distribution, that has mean np and variance $np(1-p)\left(\frac{N-n}{N-1}\right)$, where N is the size of our finite population (in our example, $N = 2000$, $n = 50$). While we could derive a confidence interval based on the hyper-geometric PMF, it is cumbersome[10]. And even though the X_is aren't identitically distributed, (because of non-replacement,) the sample mean, $\hat{p} = \frac{X_1 + X_2 + ... + X_n}{n}$, still approaches a Normal distribution. So, it is a normal practice among statisticians to use a z-interval for $\hat{p}$:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} \cdot \frac{N-n}{N-1}}$$

Note how this is identical to the confidence interval for $\hat{p}$ when the population is large, except for a "correction factor" to account for the finite population size. As one can easily see, if $n << N$, then $\frac{N-n}{N-1}$ will work out approximately to unity. In

[10]Computation power is cheap now-a-days. So it may be possible to compute a confidence interval based on the hyper-geometric PMF

other words, when the population is very large compared to the sample size, there is no difference between the confidence interval for $\hat{p}$ with or without replacement.

The confidence intervals for two proportions can use pooled t-interval or Welch's t-interval in the exact same way as we did with means, using the same logic that $\hat{p}$ is normally distributed and is also an estimate of the population proportion.

1.4.9 Determining Sample Size

If we are trying to unearth facts about the population from data that is already available, then we are constrained by the sample size of that data set. All confidence intervals we have calculated so far involve sample size - The smaller the n values, the wider, or loser, the interval. What if we are proavtively going to make measurements of some quantity to estimate the population fact? Then we have control over the sample size - We can decide how much max error $\xi = |\theta - \hat{\theta}|$ is acceptable. Mostly the sample size will be constrained by the cost of making measurements, such as the budget allocated for the effort in dollars, the time we have in our hands to make the measurements, the man-hours of effort we are willing to put in etc. Here we will see some nuances about determining the sample size for our choice of error.

We start with the confidence interval for the mean. We will assume that the variance of the population is unknown. When we define the confidence interval with a certain α value, what we mean is, we are $(1-\alpha)*100\%$ sure that,

$$\mu = \bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$$

So the upper bound of the error is,

$$\xi = |\mu - \bar{x}| = t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$$

If we rearrange this equation, we can come up with the formula for n:

$$n = t^2_{\alpha/2,n-1}\left(\frac{s^2}{\xi^2}\right)$$

There are a couple of problems in this formula though: The first is that the t-distribution on the right side depends on the very n that we are trying to find using it! The second is that the sample standard deviation s is not known before we have a sample! The first problem can be taken care of by substituting the t-distribution with the z-distribution, because, while the t-distribution changes with n, the standard normal doesn't. However, recall that the t-distribution always has fatter tails than z-distribution (as shown in Figure 1.6. This means that substituting $t^2_{\alpha/2,n-1}$ with $z^2_{\alpha/2}$ will result in a smaller n that what is right value for that α . One

way to take care of this is by simply ensuring that n is greater than 30 - because as one can see from Figure 1.6, when $n = 30$, already there is very little difference between a t-distribution and the standard normal.

The problems with not knowing the sample standard deviation s can be taken care of using several methods. One of them is to use a smaller pilot survey/measurements and then using the s from that sample to find the n value for the actual survey/measurements. Another method is to use s from a different study which involved taking sample from the same population as the one we are interested in. Yet another method is to judge the minimum and maximum values of the measurements and then use them as the end points of a 95% confidence interval, and then, using $s = \frac{Max - Min}{4}$, because a 95% confidence interval is $\pm 2\sigma$ on either side of the mean. An example case where this is possible, is when we are trying to find out, say, the number of cups of tea drunk by Indians: We know that the minimum number of cups will be 0. And we personally may have never come across a person who drinks more than 10 cups of tea a day. If we are 95% confident about this fact, then we can use $s = 2.5$ for estimating the sample size.

Determining the sample size required in the case of proportions of a large population, is very similar: We just substitute s^2 with $\hat{p}(1-\hat{p})$ in the last equation. However, calculating the sample size required for a finite population, when we don't replace draws, requires us to deal with the confidence interval we derived for this case earlier:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} . \frac{N-n}{N-1}}$$

It can shown easily that this will result in the sample size equation:

$$n = \frac{m}{1 + \frac{m-1}{N}}$$

where, N is the size of the finite population and m is sample size *if* we were dealing with a large population, i.e:

$$m = \frac{z^2_{\alpha/2} \hat{p}(1-\hat{p})}{\xi^2}$$

This is like we calculate the sample size as if we are doing it for a large population (or for a small population, but with replacement), but then we account for a correction factor of $1 + \frac{m-1}{N}$. As expected, this correction factor will work out to unity when N is very large, and we will get $n = m$.

1.5 REFERENCES

[1] PSU STAT 414: Introduction to Probability Theory [2] PSU STAT 415: Introduction to Mathematical Statistics

2 LINEAR ALGEBRA

2.1 INTRODUCTION

Linear Algebra is all about solving a system of linear equations. It has a general form as shown below.

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

One can interpret this geometrically, where each equation represents a line or a plane or a hyperplane. The xs can be thought of as the coordinates of the point where the lines or planes or hyperplanes intersect. For example,

$$\begin{aligned} x - y &= -1 \\ 3x + y &= 9 \end{aligned}$$

represent two lines intersecting in a 2 dimensional space at $(2,3)$ as shown in figure 2.1. [1] Similarly, the equations,

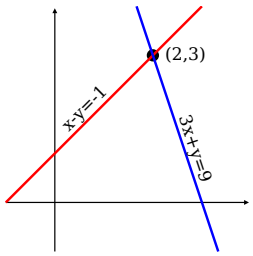


Figure 2.1: Intersecting Lines

[1] An alternative geometric interpretation is that x, y are slope and y intercept of a line that contains the coordinates -1,1 and 3,9 on it. So solving the equations is akin to estimating the (best) line that would transform input values of -1 and 3 to 1 and 9 respectively. Note: One needs to ensure the equations are translated so that they look like mx+c

$$
\begin{aligned}
3x + 2y - z &= 1 \\
2x - 2y + 4z &= -2 \\
-x + 0.5y - z &= 0
\end{aligned}
$$

represent three planes intersecting at $(1, -2, -2)$ in a 3 dimensional space as shown in figure 2.2 [2]

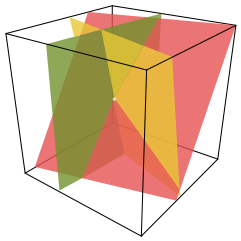
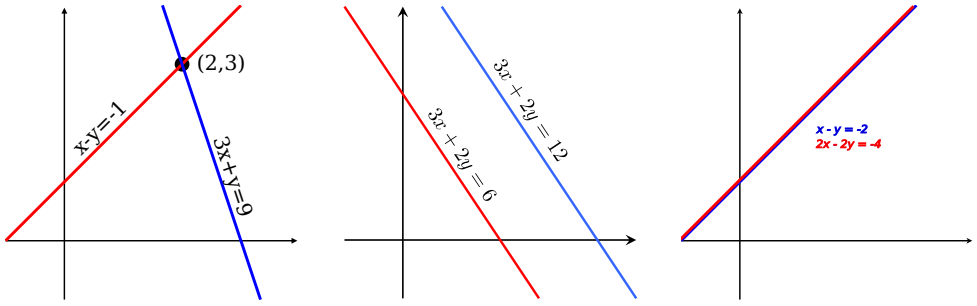


Figure 2.2: Intersecting Planes

With any given system of linear equations, for the unknowns, there could be a unique solution or multiple solutions or no solution at all. Let us look at these possibilities in a 2 dimensional space with two line equations. If the two lines are at different angles, then there is a unique solution. If the two lines are completely overlapping with one another, then there are multiple (infinite) solutions. If the two lines are parallel to each other, then there is no solution. Figure **??** shows these options.



(a) Unique solution (b) No solution (c) Infinite solutions

One can notice, that, when two lines overlap and result in infinite solutions, the corresponding equations are linearly related to one another - It is just one information masquerading as two. Linear dependency reduces the number of

[2] Again, we can think of solving for x,y,z as trying to estimate the parameters of a plane that would have transformed a bunch of input values to a bunch of output values (The equations must be translated so that, one of the parameters always has a coefficient of +1). We will deal with this "geometric dual" in a later chapter

equations we have and we could end up with less independent equations than we have unknowns. This is called an **underdetermined system**. The opposite of that - where we have more independent equations than unknown - is called an **overdetermined system**. An example of an overdetermined system is shown in figure 2.4. An underdetermined system may have infinite solutions, as we have

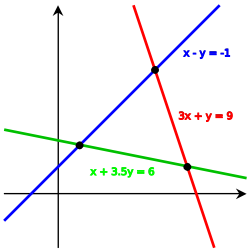


Figure 2.4: Overdetermined System

seen in the overlapping lines example, or no solution - one can think of two parallel planes. An overdetermined system (in general) has no solution.

We have seen, through a geometric interpretation of lines, a unique solution is possible only if the number of unknowns is the same as the number of independent equations. But when we go beyond lines in 2D space and planes in 3D space, generalizing results, such as what kind of a system results in a unique solution, becomes difficult and non-intuitive (how many can imagine even a 4D space, let alone a much higher dimensional space?). For this, we should go follow a vector spaces approach - The rest of the document deals only with this approach. Vector spaces approach also allows us to go beyond just lines and planes to functions, polynomials etc.

2.2 VECTOR SPACES APPROACH

This can be written in vector algebra notation as:

$$A\mathbf{x} = \mathbf{y}$$

where A is an mxn matrix, $\mathbf{x}$ is a column vector with n entries, and $\mathbf{y}$ is a column vector with m entries.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

One can interpret $A\mathbf{x} = \mathbf{y}$ as a linear transformation of $\mathbf{x}$ vector to $\mathbf{y}$ by the linear operator A. Or one could think of it as $\mathbf{y}$ being represented as a linear combination

of the columns of A weighted by rows of $\mathbf{x}$ as shown below.

$$x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} + \ldots + x_n \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

In this case, columns of A can be thought of as vectors themselves. So we could say that we are trying to represent a vector $\mathbf{y}$ as a linear weighted comination of n vectors. We will see later that this idea of representing one vector in terms of a set of other vectors has its advantages. But first, we introduce the some important concepts about vector spaces:

Definition A **linear vector space** S over a set of scalars R is a collection of objects known as vectors, together with an additive operation $+$ and a scalar multiplication operation . [a], that satisfy the following properties

- S forms a *group* under addition. i.e.,
  1. Addition operation is closed.
     $$\forall \mathbf{x}, \mathbf{y} \in S, \mathbf{x} + \mathbf{y} \in S$$
  2. Identity element, denoted as $mathbf\,0$ exists, such that
     $$\mathbf{x} + \mathbf{0} = \mathbf{0} + \mathbf{x} = \mathbf{x}$$
  3. Additive inverse exists
     $$\forall \mathbf{x} \in S, \exists \mathbf{y} \mid \mathbf{x} + \mathbf{y} = \mathbf{0}$$
     [b]
  4. Addition is associative
     $$\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in S, (\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$$
- $\forall a, b \in R,\ for\ any\ \mathbf{x}, \mathbf{y} \in S,$
  1. $a\mathbf{x} \in S$
  2. $a(b\mathbf{x}) = (ab)\mathbf{x}$
  3. $(a + b)\mathbf{x} = a\mathbf{x} + b\mathbf{x}$
  4. $a(\mathbf{x} + \mathbf{y}) = a\mathbf{x} + a\mathbf{y}$
- Multiplicative identity, denoted by $1 \in R$, exists, such that $1\mathbf{x} = \mathbf{x}$
- An element $0 \in R$ exists, such that $0\mathbf{x} = \mathbf{0}$ [c]

[a] Vector multiplication need not be defined

[b] This is just a sneaky way of introducing $-$ within the definition of the additive operator $+$.

[c] Note than multiplicative inverse need not be defined

Before we continue, it is important to understand that the scalars in set R need not be real or complex numbers as one usually imagines them to be. They can be anything that satisfy operations of additions and multiplications as defined for them. For example, R can be numbers modulo 256, or even polynomials. Similarly, for vectors, they need not be a collection of real or complex numbers. For example,

$x(t) = c_1 + c_2 t + c_3 t^2$ can be thought of as a vector $x(t)$ being represented as a linear combination of vectors $1, t, t^2$.

Once a vector space is defined with multiplication and addition, one can select a some vectors and linear combination of those will produce numerous vectors in S.

Definition If we select a set of vectors $T \subset S$ then the set of vectors that can be produced by (finite) linear combinations of vectors in T is called the **span** of T. It is denoted as $V = span(T)$. [a]

[a] T can have infinite number of vectors, but V will contain only finite linear combinations (which are infinitely many)

Note that V itself is a vector space! Since $V \subset S$, V is called a subspace of S. In other words one could just wildly pick a set of vectors from S and form a subspace. It can be proved that V is also the smallest subspace that contains all the vectors in T. Picking a T and then coming up with a V is used in some areas like digital communication (out of the scope of this notes), but what about the other way around? If we start with a V and we want to find a T whose vectors can span V, we will find several candidates. Then the question arises: "Which candidate is the best?". One way to answer it is to find the T that is the smallest. We will see soon that even in this case we will find several candidates. But let us first focus on what would make a T the smallest possible set that spans V.

Let us take an example $T1$ that has four vectors namely $\{\mathbf{p_1}, \mathbf{p_2}, \mathbf{p_3}, \mathbf{p_4}\}$. Let us take a vector in $\mathbf{v_1} \in V$. This can be represented as a linear combination of the vectors in $T1$ as below.

$$\mathbf{v_1} = c_1\mathbf{p_1} + c_2\mathbf{p_2} + c_3\mathbf{p_3} + c_4\mathbf{p_4}$$

. Now, let us say that **p4** itself can be represented as a linear combination of the remaining three vectors in $T1$ as:

$$\mathbf{p_4} = d_1\mathbf{p_1} + d_2\mathbf{p_2} + d_3\mathbf{p_3}$$

We can substitute this equation in the previous one and get:

$$\mathbf{v_1} = (c_1 + d_1)\mathbf{p_1} + (c_2 + d_2)\mathbf{p_2} + (c_3 + d_3)\mathbf{p_3}$$

Since the vector $\mathbf{v_1}$ is a placeholder for *any* vector in V, this means that the subset $\{\mathbf{p_1}, \mathbf{p_2}, \mathbf{p_3}\}$ of $T1$ itself is a spanning set of V. One could thus come with an initial T and eliminate vectors in it that can be represented as linear combinations of other vectors and eventually end up with the smallest possible set that can span V.

Definition A **Hamel Basis** is a smallest subset of vectors in vector space that can span that vector space.

There need not be a unique Hamel Basis for a vector space. But all the Hamel Bases have the following properties.

1. The vectors that comprise a Hamel Basis are linearly independent (of one another), i.e., no vector can be represented as a linear combination of other vectors
2. The Hamel Bases all have the same cardinality

We will not go into the proofs of the above points, but they can be found in many text books. So far, we have talked about dimensions casually. But we define it formally now.

Definition The **dimension** of a vector space is the cardinality of a Hamel Basis of that vector space. In other words, the dimension of a vector space is the same as the the smallest number of vectors whose linear combination can create any vector in that vector space

The idea of linear independence at the root of the definition of a Hamel basis is of high importance and hence requires more attention. One way to check if a set of vectors is linearly independent is to see if we can come up with a set of coefficients $\{c_1, c_2, ...c_n\}$, not all zeros such that,$c_1\mathbf{p_1} + c_2\mathbf{p_2} + \ldots + c_n\mathbf{p_n} = \mathbf{0}$, then the vectors$\{\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_n}\}$ are *not* linearly independent. It is easy to see why: We can readily rewrite $c_1\mathbf{p_1} + c_2\mathbf{p_2} + \ldots + c_n\mathbf{p_n} = \mathbf{0}$ as:

$$c_1\mathbf{p_1} = (-c_2)\mathbf{p_2} + (-c_3)\mathbf{p_3} + \ldots + (-c_n)\mathbf{p_n}$$

which, by definition, means $\mathbf{p1}$ is linearly dependent on other $\mathbf{p}$ vectors. Now, we will use this property to prove that every $\mathbf{x} \in V$ has a unique representation as a linear combination of vectors in its Hamel basis. To prove this, let us assume that there is an $\mathbf{x} \in V$ that can be represented by two different set of coefficients $\{c_1, c_2, \ldots, c_n\}$ and $\{d_1, d_2, \ldots, d_n\}$ of the vectors $\{\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_n}\}$. i.e.,

$$c_1\mathbf{p_1} + c_2\mathbf{p_2} + \ldots + c_n\mathbf{p_n} = \mathbf{x} = d_1\mathbf{p_1} + d_2\mathbf{p_2} + \ldots + d_n\mathbf{p_n}$$

then,

$$(c_1 - d_1)\mathbf{p_1} + (c_2 - d_2)\mathbf{p_2} + \ldots + (c_n - d_n)\mathbf{p_n} = \mathbf{0}$$

Since $\{\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_n}\}$ are linearly independent, this means that the above equation is only possible if all $c_i - d_i$ terms are zeroes, or in other words, $c_i = d_i, \forall i$. Now

we can similarly prove that there cannot be two subsets of the Hamel basis, namely $\{\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_n}\}$, $\{\mathbf{q_1}, \mathbf{q_2}, \ldots, \mathbf{q_m}\}$, where n may or may not be the same as m and the two sets may or may not have overlaps, i.e, some $\mathbf{q}$s same as some $\mathbf{p}$s), such that,

$$c_1\mathbf{p_1} + c_2\mathbf{p_2} + \ldots + c_n\mathbf{p_n} = \mathbf{x} = d_1\mathbf{q_1} + d_2\mathbf{q_2} + \ldots + d_n\mathbf{q_m}$$

then,

$$c_1\mathbf{p_1} + c_2\mathbf{p_2} + \ldots + c_n\mathbf{p_n} - d_1\mathbf{q_1} - d_2\mathbf{q_2} - \ldots - d_n\mathbf{q_m} = \mathbf{0}$$

By definition of linear independence, the above equation should only be possible if all the coefficients are 0s, or if the terms with non-zero coefficients all, somehow, cancel each other out. Note that a single term $c_i\mathbf{p_i}$ or $d_i\mathbf{q_i}$ itself cannot evaluate to $\mathbf{0}$ for non-zero coefficient value. So there has to be at least one $c_i\mathbf{p_i} - d_i\mathbf{q_i}$, that evaluates to $\mathbf{0}$, which is only possible, under the assumption of linear independence, if $c_i = d_i$ and $\mathbf{p_i} = \mathbf{q_i}$. In other words,

$$c_1\mathbf{p_1} + c_2\mathbf{p_2} + \ldots + c_n\mathbf{p_n} - d_1\mathbf{q_1} - d_2\mathbf{q_2} - \ldots - d_n\mathbf{q_m} = \mathbf{0}$$

implies,

- $n = m$
- $c_i = d_i$
- $\mathbf{p_i} = \mathbf{q_i}$

which in turn implies that the representation of $\mathbf{x}$ as a linear combination of the vectors in the Hamel basis of V is unique.

With the above concepts in linear algebra, we are almost ready to turn our attention back to solving a system of linear equations: Just one more concept remains as stated in the following Lemma.

Lemma If we pick any arbitrary set B of linearly independent vectors that has the same cardinality as the dimension of the vector space Y of which B is a subset, then B is a Hamel basis of Y

The proof is straightforward: Let us assume a condradiction, where there are some vectors in Y namely, $\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_k}$ that are outside $span\{B\}$. Then it means that a new set $B \cup \{\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_k}\}$ will span Y and hence a Hamel basis of Y. However, this new set's cardinality is greater than that of the dimensionality of the vector space, which is impossible. Hence there cannot be any vector $\mathbf{y} \in Y$ that is outside $span\{B\}$.

Alright, now we are ready to go back to our system of linear equations, $A\mathbf{x} = \mathbf{y}$, where,

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

If we represent the columns of the matrix A as $\mathbf{a_1}, \mathbf{a_1}, \ldots \mathbf{a_n}$, then, the system of linear equations boils down to the familiar representation of a vector as a linear combination of other vectors:

$$x_1\mathbf{a_1} + x_2\mathbf{a_2} + \ldots + x_n\mathbf{a_n} = \mathbf{y}$$

This is simply an attempt to represent *any* $\mathbf{y} \in Y$ as a linear combination of some vectors $\{\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_n}\}$. For a unique solution to exist for this system of linear equations, the following conditions must be true about $\{\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_n}\}$.

1. They must reside in Y
2. They should be linearly independent
3. Their cardinality, n should be the same as that of any Hamel basis of Y

To evaluate the first condition, let us look at the concept of **natural basis** for vectors that are represented as a collection of scalars (as in our case). Any $\mathbf{y} \in Y$ can be represented as:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_m \end{bmatrix} = y_1 \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + y_1 \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \ldots + y_m \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

It can be shown that the set $\{[1, 0, \ldots, 0]^T, [0, 1, \ldots, 0]^T, \ldots, [0, \ldots, 0, 1]^T\}$ is linearly independent and hence a Hamel basis of Y. With this definition, we can say that any vector $\mathbf{a}$ resides in Y as long as it can be represented by the natural basis above. For the vectors $\{\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_n}\}$, this means that, as long as $n = m$, they will belong to Y. Note that this also means that A above has to be a square matrix.

The second condition can be evaluated quite easily using our previous method of finding if there are any set of coefficients that will make the linear combination of $\mathbf{a_i}$ vectors equal to $\mathbf{0}$.

As for the third condition, $n = m$, implied by the first condition already ensures that the cardinality of $\{\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_n}\}$ is same as that of a Hamel basis (ex. the natural basis). In other words, for a unique solution to exist, all we need is:

1. n should be the same as m,
2. The vectors $\{\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_n}\}$ should be linearly independent

But what about the existance of *any* solution at all (whether or not unique)? It can be shown that, if the following conditions are true, a solution and multiple solution will exist.

1. $n \geq m$, where m is the cardinality of a Hamel basis of Y
2. There are at least m linearly independent vectors in $\{\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_n}\}$

To understand these conditions, imagine that $n = m + 1$. When there are at least m linearly indenpendent vectors in $\{\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_n}\}$, we have:

$$\mathbf{a_n} = c_1\mathbf{a_1} + c_2\mathbf{a_2} + \ldots + c_m\mathbf{a_m}$$

Now, using this we can show the following:

$$\mathbf{y} = x_1\mathbf{a_1} + x_2\mathbf{a_2} + \ldots + x_n\mathbf{a_n}$$

$$\mathbf{y} = x_1\mathbf{a_1} + x_2\mathbf{a_2} + \ldots + x_m\mathbf{a_m} + c_1\mathbf{a_1} + c_2\mathbf{a_2} + \ldots + c_m\mathbf{a_m}$$

$$\mathbf{y} = (x_1 + c_1)\mathbf{a_1} + (x_2 + c_2)\mathbf{a_2} + \ldots + (x_2 + c_m)\mathbf{a_m}$$

Since $\{\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_n}\}$ is a Hamel basis of Y, this means that one can find unique coefficents, say, $\{k_1, k_2, \ldots, k_m\}$ that satisfy the above equation, where,

$$k_i = x_i + c_i, \forall i \in [1, m]$$

So, whereas we can find unique k_is, and hence at least one solution for x_is exists, multiple possible values of c_is and x_is exists, that will give these k_is - So we have indeed multiple solutions for x_is (i.e., we have multiple solutions for the linear system of equations). One can see that, when multiple solutions exist, $n > m$ makes the system of equations an **underdetermined** one.

Needless to say, if the conditions for the existance of at least one solution are not satisfied, then no solution would exist. For ex., if the number of linearly independent vectors in $\{\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_n}\}$ is less than m: say, m-1, then, $\{\mathbf{a_1}, \mathbf{a_2}, \ldots, \mathbf{a_n}\}$ will span a space that is one dimension smaller than Y, i.e, they will span a subspace of Y. In that case, there is a way of finding the vector $\hat{\mathbf{y}}$ in this subspace of Y that is closest to $\mathbf{y}$. This is called an approximation problem [3]. Exploring this approximation problem requires that we define some measure of length of a vector (called, "norm") and a measure of projection of one vector over another (called, "inner product").

2.2.1 Norm and inner product

[3] where we have (possibly with some manipulations) an **overdetermined** system of equations.

Definition A real valued function of a vector $\mathbf{x} \in S$, denoted as $\|\mathbf{x}\|$.is said to be a **norm** if $\|\mathbf{x}\|$ satisfies the following properties.

1. $\|x\| \geq 0 \ \forall \ \mathbf{x} \in S$
2. $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = 0$
3. $\|\alpha\mathbf{x}\| = |\alpha| \ \|\mathbf{x}\|$, where α is an arbitrary scalar
4. $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$. This is called triangule inequality.

   if $\|\mathbf{x}\|$ is a norm then $d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|$ is a metric.

The point about the norm being used as a metric is intuitive: one can think of $\|\mathbf{x}\|$ as $d(\mathbf{x},\mathbf{0})$. And the distance between a vector $\mathbf{y}$ and its approximation, $\hat{\mathbf{y}}$ can be thought of as the length of the approximation "error vector" $\mathbf{e}$, i.e., $\|\mathbf{e}\| = d(\mathbf{y}-\hat{\mathbf{y}}) = \|\mathbf{y}-\hat{\mathbf{y}}\|$.

The most popular norms are (n is the dimension of the vector):

1. The l_1 norm: $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$
2. The l_p norm: $\|\mathbf{x}\|_p = (\sum_{i=1}^{n} |x_i|^p)^{1/p}$
3. The l_∞ norm: $\|\mathbf{x}\|_\infty = max_{i=1,2,\ldots,n}|x_i|$

The l_2 norm of the "error vector" is basically the Euclidean distance between $\mathbf{y}$ and $\hat{\mathbf{y}}$. Note that, although the exact values of a norm will differ from one definition of the norm to another, a vector that is small with respect to one norm is also small with respect to another norm.

Definition An **inner product** in a vector space S is a function that operates on two vectors and returns a scalar, i.e., $\langle\cdot,\cdot\rangle : S \times S \rightarrow R$. It has the following properties:

1. $\langle\mathbf{x},\mathbf{y}\rangle = \langle\mathbf{y}^*,\mathbf{x}^*\rangle$
2. $\langle\alpha\mathbf{x},\mathbf{y}\rangle = \alpha\langle\mathbf{x},\mathbf{y}\rangle$
3. $\langle\mathbf{x}+\mathbf{y},\mathbf{z}\rangle = \langle\mathbf{x},\mathbf{z}\rangle + \langle\mathbf{y},\mathbf{z}\rangle$
4. $\langle\mathbf{x},\mathbf{x}\rangle > 0$ if $x \neq 0$, and $\langle\mathbf{x},\mathbf{x}\rangle = 0$, if and only if, $\mathbf{x} = \mathbf{0}$

The most popular inner product is $\langle\mathbf{x},\mathbf{y}\rangle = \mathbf{y}^H\mathbf{x}$ (For real vectors, the hermitian becomes just a transpose). One can define a norm in terms of the inner product. Such a norm is called an **induced norm**. For example, we can defined the l_2 norm as $\|\mathbf{x}\| = \langle\mathbf{x},\mathbf{x}\rangle^{1/2}$.

One can geometrically interpret the inner product as a function of the angle between two vectors. If one were to represent the two vectors using polar coordinates as $\mathbf{x} = [r_1 cos\theta_1, r_1 sin\theta_1]$ and $\mathbf{y} = [r_2 cos\theta_2, r_2 sin\theta_2]$, then if we compute the inner product $\mathbf{y}^T\mathbf{x}$, it will result in $r_1 r_2 cos(\theta_1 - \theta_2)$. Also the induced norms of $\mathbf{x}$ and $\mathbf{y}$ will end up as r_1 and r_2 respectively. So one can infer that,

$$cos(\theta_1 - \theta_2) = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{||\mathbf{x}||||\mathbf{y}||}$$

Although we have used $\mathbb{R}^2$ in this example, one can extend this geometric interpretation to higher dimensions, because, when we are considering two vectors, we can always come up with a 2D subspace in which both vectors reside and then the maths will follow.

An important implication of this geometric interpretation is that, if two vectors are perpendicular to each other, then their inner product is 0 . This fact is very useful for our next geometric interpration of the inner product: inner product as a projection of one vector on another. Check out figure 2.5. We show two vectors $\mathbf{y}$ and $\mathbf{a}$. One can imagine $\mathbf{a}$ being parallel to the ground (if you will). Now imagine the sun at 12'O' clock. It will create a shadow of $\mathbf{y}$ on $\mathbf{a}$. Let us call the length of this shadow as x. Now one can imagine a third vector $\mathbf{e}$ connecting the end of the shadow to the end of $\mathbf{y}$. We can show that,

In fact, in a normed vector space, no matter how the norm is defined, two vectors in that space are said to be orthogonal if their inner product is 0. *So even when the vectors are functions or polynomials, and the inner product has integrals and what not, the idea of orthogonality can be extended using this definition*

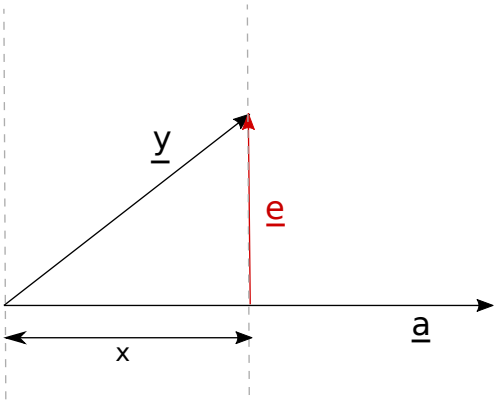


Figure 2.5: Inner product as a projection

$$\mathbf{y} = x\mathbf{a} + \mathbf{e}$$
$$\langle \mathbf{y}, \mathbf{a} \rangle = x^2 \langle \mathbf{a}, \mathbf{a} \rangle + \langle \mathbf{e}, \mathbf{a} \rangle$$

Now, using our previous understanding of inner product of two vectors orthogonal to each other being 0

$$\langle \mathbf{y}, \mathbf{a} \rangle = x^2 ||\mathbf{a}||^2 + 0$$

If $\mathbf{a}$ is a unit vector, then,

$$\langle \mathbf{y}, \mathbf{a} \rangle = x^2$$

In other words, if $\mathbf{a}$ is a unit vector, then the inner product $\langle \mathbf{y}, \mathbf{a} \rangle$ is square of the length of the shadow of $\mathbf{y}$ falling on $\mathbf{a}$.

Now, recall our linear system of equations, represented as

$$\mathbf{y} = x_1\mathbf{a_1} + x_2\mathbf{a_2} + \ldots + x_n\mathbf{a_n}$$

One can generate an alternative system of linear equations by calculating the norm of $\mathbf{y}$ with each of the $\mathbf{a_i}$ as shown below:

$$\begin{aligned}
\langle \mathbf{y}, \mathbf{a_1} \rangle &= x_1\langle \mathbf{a_1}, \mathbf{a_1} \rangle + x_2\langle \mathbf{a_2}, \mathbf{a_1} \rangle + \ldots + x_n\langle \mathbf{a_n}, \mathbf{a_1} \rangle \\
\langle \mathbf{y}, \mathbf{a_2} \rangle &= x_1\langle \mathbf{a_1}, \mathbf{a_2} \rangle + x_2\langle \mathbf{a_2}, \mathbf{a_2} \rangle + \ldots + x_n\langle \mathbf{a_n}, \mathbf{a_2} \rangle \\
&\vdots \\
\langle \mathbf{y}, \mathbf{a_n} \rangle &= x_1\langle \mathbf{a_1}, \mathbf{a_n} \rangle + x_2\langle \mathbf{a_2}, \mathbf{a_n} \rangle + \ldots + x_n\langle \mathbf{a_n}, \mathbf{a_n} \rangle
\end{aligned}$$

In matrix notation, we can represent this as:

$$R\mathbf{x} = \mathbf{p}$$

where,

$$R = \begin{bmatrix}
\langle \mathbf{a_1}, \mathbf{a_1} \rangle & \langle \mathbf{a_2}, \mathbf{a_1} \rangle & \cdots & \langle \mathbf{a_n}, \mathbf{a_1} \rangle \\
\langle \mathbf{a_1}, \mathbf{a_2} \rangle & \langle \mathbf{a_2}, \mathbf{a_2} \rangle & \cdots & \langle \mathbf{a_n}, \mathbf{a_2} \rangle \\
\vdots & \vdots & \ddots & \vdots \\
\langle \mathbf{a_1}, \mathbf{a_n} \rangle & \langle \mathbf{a_2}, \mathbf{a_n} \rangle & \cdots & \langle \mathbf{a_n}, \mathbf{a_n} \rangle
\end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{p} = \begin{bmatrix} \langle \mathbf{y}, \mathbf{a_1} \rangle \\ \langle \mathbf{y}, \mathbf{a_2} \rangle \\ \vdots \\ \langle \mathbf{y}, \mathbf{a_n} \rangle \end{bmatrix}$$

So, one could use $\mathbf{x} = R^{-1}\mathbf{p}$ as much as $\mathbf{x} = A^{-1}\mathbf{y}$. It turns out that, just as we found out that A^{-1} would exist only if $\mathbf{a_i}$ are linearly independent, the same is true for the existence of the R^{-1}. The matrix R is called the **Grammian matrix**. The Grammian matrix is always a square matrix and hermitian symmetric, i.e., $R^H = R$.

There are three advantages of using $\mathbf{x} = R^{-1}\mathbf{p}$ over $\mathbf{x} = A^{-1}\mathbf{y}$:

1. R is always a matrix whether or not A is a matrix (Recall the note about how a vector need not a tuple of scalars...how a function or a polynomial can be a vector)

2. When it comes to finding a solution for overdetermined system of equations where no solutions exist, the former method can be used with some tricks. We explain this in the next section.

3. Because R is hermitian symmetric, it is easier to invert than A that, in general, has no structure.

Before we move on, we make a note about the relationship between these two equivalent solutions. One can easily see that R is nothing but $A^H A$, and $\mathbf{p}$ is nothing but $A^H\mathbf{y}$. So one can rewrite $\mathbf{x} = R^{-1}\mathbf{p}$ as:

$$\mathbf{x} = (A^H A)^{-1} A^H \mathbf{y}$$

The usefulness of the above representation will be more apparent in the next section.

2.3 APPROXIMATION

In the previous chapter, we saw the conditions for the existence of a unique solution to $A\mathbf{x} = \mathbf{y}$, and for the existence of at least once (and many) solutions to the same. In this chapter we focus on the an overdetermined system of equations where no solutions exist. Below is such a system of equations:

$$x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} + \ldots + x_n \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

We saw in the previous chapter that a system is overdetermined if $n < m$. When there exists no solution to $\mathbf{x}$, one can find the best candidate for $\mathbf{x}$ that will result in a $\hat{\mathbf{y}}$ that is closest to $\mathbf{y}$. It turns out that this best candidate will be unique. Let us see how this happens by first representing this concept mathematically as:

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$$

If we were to visualize this relationship, it is intuitive that we can minimize ivece when $\mathbf{e}$ is orthogonal to $\hat{\mathbf{y}}$. This is shown in figure 2.6 Substituting $\hat{\mathbf{y}}$ in terms of $\mathbf{x}$, we get,

$$\hat{\mathbf{y}} = x_1\mathbf{a_1} + x_2\mathbf{a_2} + \ldots + x_n\mathbf{a_n}$$

In other words,

$$\mathbf{y} = x_1\mathbf{a_1} + x_2\mathbf{a_2} + \ldots + x_n\mathbf{a_n} + \mathbf{e}$$

Taking the inner product of $\mathbf{y}$ with every $\mathbf{a_i}$, we establish our familiar alternative system of linear equations:

$$\begin{aligned}
\langle \mathbf{y}, \mathbf{a_1} \rangle &= x_1 \langle \mathbf{a_1}, \mathbf{a_1} \rangle + x_2 \langle \mathbf{a_2}, \mathbf{a_1} \rangle + \ldots + x_n \langle \mathbf{a_n}, \mathbf{a_1} \rangle + \langle \mathbf{e}, \mathbf{a_1} \rangle \\
\langle \mathbf{y}, \mathbf{a_2} \rangle &= x_1 \langle \mathbf{a_1}, \mathbf{a_2} \rangle + x_2 \langle \mathbf{a_2}, \mathbf{a_2} \rangle + \ldots + x_n \langle \mathbf{a_n}, \mathbf{a_2} \rangle + \langle \mathbf{e}, \mathbf{a_1} \rangle \\
&\vdots \\
\langle \mathbf{y}, \mathbf{a_n} \rangle &= x_1 \langle \mathbf{a_1}, \mathbf{a_n} \rangle + x_2 \langle \mathbf{a_2}, \mathbf{a_n} \rangle + \ldots + x_n \langle \mathbf{a_n}, \mathbf{a_n} \rangle + \langle \mathbf{e}, \mathbf{a_1} \rangle
\end{aligned}$$

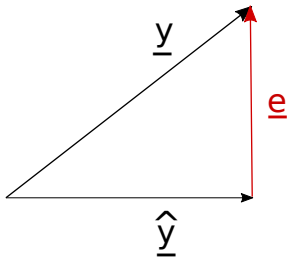


Figure 2.6: Minimum error

If $\mathbf{e}$ minimized when it is orthogonal to $\hat{\mathbf{y}}$, it will be orthogonal to every $\mathbf{a_i}$. So this system of equations becomes:

$$\begin{aligned}
\langle \mathbf{y}, \mathbf{a_1} \rangle &= x_1 \langle \mathbf{a_1}, \mathbf{a_1} \rangle + x_2 \langle \mathbf{a_2}, \mathbf{a_1} \rangle + \ldots + x_n \langle \mathbf{a_n}, \mathbf{a_1} \rangle \\
\langle \mathbf{y}, \mathbf{a_2} \rangle &= x_1 \langle \mathbf{a_1}, \mathbf{a_2} \rangle + x_2 \langle \mathbf{a_2}, \mathbf{a_2} \rangle + \ldots + x_n \langle \mathbf{a_n}, \mathbf{a_2} \rangle \\
&\vdots \\
\langle \mathbf{y}, \mathbf{a_n} \rangle &= x_1 \langle \mathbf{a_1}, \mathbf{a_n} \rangle + x_2 \langle \mathbf{a_2}, \mathbf{a_n} \rangle + \ldots + x_n \langle \mathbf{a_n}, \mathbf{a_n} \rangle
\end{aligned}$$

And this is exactly the same as the alternative system of equations $R\mathbf{x} = \mathbf{p}$ we obtained in the previous chapter for solving $\mathbf{x}$, when $\mathbf{a_i}$ are linearly independent! So the solution $\mathbf{x} = R^{-1}\mathbf{p}$ for finding the unique solution for $\mathbf{x}$ also works for finding the best solution for $\mathbf{x}$ when $\mathbf{y}$ isn't reachable by $\mathbf{x}$! As before we can write the solution entirely in terms of A as:

$$\mathbf{x} = (A^H A)^{-1} A^H \mathbf{y}$$

This is called the **least squares solution**. Compare this equation to $\mathbf{x} = A^{-1}y$, which will work only in the specfic case of $n = m$. the above equation will work for the general case of $n \leq m$ as long as $\mathbf{a_i}$ are linearly independent. Hence, $(A^H A)^{-1} A^H$ is called a **pseudo inverse** of A. So, in the case of a properly-determined system, the pseudo inverse based solution works out to be the same as that of the regular inverse, in the case of an over-determined system of equations This also means that, in the case of linearly *dependent* set of equations, while multiple solutions exists, one way to find a solution is to simply remove $\mathbf{a_i}$ until the remaining ones are linearly independent - and whether $n = m$ or $n < m$, pseudo-inverse of the new A can be used to arrive at *a* solution[4].

[4]There is a better way of finding a solution that is more meaningful in such cases. That solution is called the minimum-norm solution. We will talk about this later

If we use the least squares solution of $\mathbf{x}$, in $\hat{\mathbf{y}} = A\mathbf{x}$, we get,

$$\hat{\mathbf{y}} = A(A^H A)^{-1} A^H \mathbf{y}$$

This is called the **least squares approximation** of $\mathbf{y}$ in the vector space spanned by $\mathbf{a_i}$. And the matrix $P = A(A^H A)^{-1} A^H$ is called the **projection matrix** as it "projects" $\mathbf{y}$ into the column vector space of A.

2.3.1 Choosing a basis function

In the earlier section we talked about representing vectors in given vectors space S as a linear combination of a set of vectors forming a basis, and how a Hamel basis is smallest such a basis that can span S. But we did not talk about how to pick one Hamel basis among several Hamel bases? We are now eqipped to answer that question. One meaningful way is to choose a Hamel basis that makes finding coefficients of the linear combination easy for a given vector $\mathbf{y}$. Since this involves nothing but solving $A\mathbf{x} = \mathbf{y}$, where the column vectors $\mathbf{a_i}$ of A are the vectors that form the basis, and $\mathbf{x}$ the set of coefficients. We now know that finding $\mathbf{x}$ involves inverting A or inverting the Grammian R. Inverting matrices is generally painful. We can make it easy or completely avoid it by cleverly choosing a Hamel basis. We observe from the Grammian matrix that, if we choose the $\mathbf{a_i}$ such that they are all orthogonal to one another and they are unit vectors, then the Grammian matrix reduces to an identity matrix!! And hence solving $R\mathbf{x} = \mathbf{p}$ becomes trivial - We simply get:

$$x_i = \langle \mathbf{y}, \mathbf{a_i} \rangle$$

Such $\mathbf{a_i}$ are said to be **orthonormal**. Generally inverting matrices is painful, but it is increasingly so when the number of vectors in the basis become very large - or even infinite. In such cases, choosing orthonormal vectors is pretty much the only choice. There is a process called the **Gram-Schmidt orthogonalization** by which one can start with any basis and manipulate the vectors in it to come with an alternative basis that has orthonormal vectors.

2.3.2 Least squares approximation and fitting noisy data

A

STATISTICS - AT A GLANCE

Whatever