MTH211A: Theory of Statistics

Module 1: Introduction to Inference

January 30, 2024

Week 0

I. What is statistical inference?

The basic idea of statistical inference is to learn to infer about an underlying truth from the data. We will understand this using some examples.

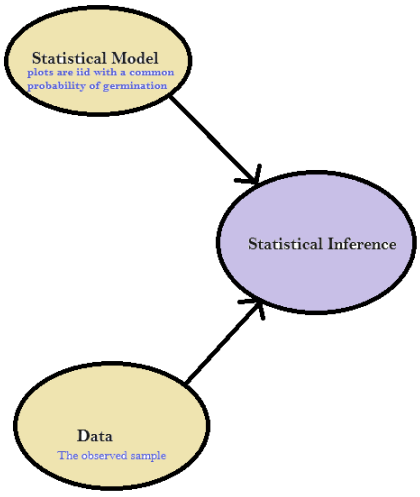
Example 1. Suppose a land of 25 acre is split into 10,000 plots of equal area each, and a seed of type A is planted in each plot on 1st August 2023. After one month 15 plots are surveyed at random, and the following germination status is observed (see Table 1).

sample plot	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Germination status	1	0	1	0	1	0	1	1	1	0	1	1	0	1	1

Table 1: Germination status of 15 sample plots

Here germination status 1 indicates that the seed is germinated, while 0 indicates that the seed is not germinated. We are interested in the proportion of germination θ.

To answer the above statistical problem we need to build an appropriate model and insert the observed data in the model.



First note that, as we could not enumerate the entire 25 acre land, we need to infer the proportion θ using a randomly selected sub-population, called a sample.

How the data is related to the sample?

The word 'sample' is typically used to mean a part of a totality. Here the totality is called the population. The population in this example is the collection of all the plots in the 25 acre land. The fact that some units are randomly picked up from the population induces two characteristics:
(i) The germination status of the sampled units are not known in advance, hence they are random. The data in Table 1 is the realization of the random sample.
(ii) The sampled units are expected to have same characteristic as that of the population. Thus, if the population has a high rate of germination (θ is large), then the sample would also have a high *probability* of germination.

To answer the above problem, we have to make a few assumptions on the underlying truth. The set of assumptions together is called the model, or statistical model. In this problem we may assume that (a1) the soil condition and other environment related factors remain nearly identical over the entire 25 acre land. Further, we may assume that (a2) the sampled plots are independent in the sense that the germination status of one plot does not affect that of other plots. The above two assumptions together build the following model:

Define X_i : Germination status of the i-th sampled plot, $i = 1, \ldots, 15$. Then $X_i \overset{iid}{\sim} \texttt{Bernoulli}(\theta)$ for some $0 < \theta < 1$. Due to (ii) θ appears to be the probability that the seed in the i-th sample plot germinates. Further, due to assumption (a1) the probability is same across plots (identical distribution). The unknown probability θ is called a parameter of the model.

As we don't know θ, we need to approximate θ using the sample realizations or data. If the model $X_i \overset{iid}{\sim} \texttt{Bernoulli}(\theta)$ is true, then the realizations of $X_1, \cdots, X_{15}$ provide an idea about θ. For example, from Table 1 we see that a total of $\sum_i X_i = 10$ plants are germinated out of 15. Thus, $\bar{X} = 2/3$ seems to be a reasonable estimate for the rate of germination. Statistical inference is the procedure to come up with an appropriate method of estimating the parameter θ or a function of the parameter $g(\theta)$ based on a sample.

Example 2. Suppose we are interested in estimating the gravitational constant g. Usual way to estimate g is by the pendulum experiment, where $g \approx 4\pi^2 l/T^2$, where l is the length of a simple pendulum and T is time period of the pendulum for an oscillation. Suppose length of a simple pendulum is 75 cm. Due to variation which depends on several factors such as skill of the experimentor and measurement error, T can not be measured exactly. Instead 10 repeated measurements are taken which are

Repetitions	1	2	3	4	5	6	7	8	9	10
Measurements (s)	1.822	1.684	1.688	1.758	1.739	1.805	1.809	1.702	1.614	1.764

Table 2: Results of 10 repeated measurements

Set your assumptions to build a model. Identify the parameter(s). Can you express the quantity of interest g as a function of the parameter(s)? Can you interpret the data in Table 2 as realization of a sample from some population? How would you apply the data into the model and estimate g?

II. Now we formally define the different ingredients of statistical inference.

(a) **Population.** In statistical inference we seek information about some numerical characteristic of a collection of units, called population. A population can be finite or infinite.

Examples. (i) Suppose you are interested in the average grade obtained by students in MTH211. Then the population consists of all students who took this course in the past years, this year and will be taking this course in future years. (ii) Suppose you are interested in the proportion of homeless people who live in Kanpur, then the population consists of every individual who live in Kanpur. (iii) Suppose you are interested in the lifetime of Samsung mobiles of a particular model A, then the population consists of all Samsung mobiles of model A.

(b) **Sample.** It is often not possible to enumerate the entire population due to several constraints, for e.g., time constraint, cost, inaccessibility, etc. In such cases, one examines a part of the population, called sample. The sample must be a good representative of the population. How does one select a representative sample? One way is to select sample units at random.

(c) **Random sample.** Suppose we are interested in the proportion of students in IITK whose weekly study time is between 40 to 60 hours on an average in 2023. If we had enumerated all students of IITK on each week through out 2023 and prepared a grand list of average study time, then it would be possible to calculate the proportion exactly. Suppose the histogram in Figure 1 represent the completely enumerated data. Let the true proportion of interest be μ, then μ is the relative frequency of the purple portion in Figure 1.

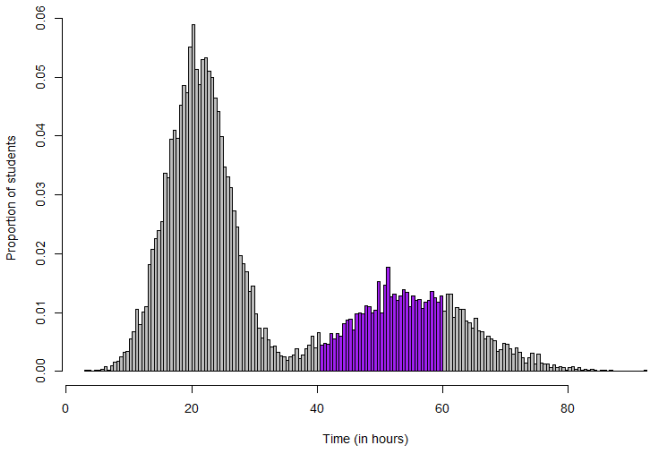


Figure 1: Histogram of average weekly study time (in hours) of students of IITK

Now, suppose we take a random element from this population (i.e., randomly enquire a student of IITK). Prior to observing the sample, its realization is a random phenomenon to us, and we expect the features of the population to be present in the sample. Let X denotes the random variable indicating the average weekly study time of the randomly selected student. Then the probability $P(X \in (40, 60))$ is expected to be same as μ. Further, this phenomenon remains unaltered if we replace the numbers 40 and 60 by any arbitrary $a, b \in \mathbb{R}$ with $a < b$. Therefore, the probability distribution of the random sample is expected to be same as the relative frequency distribution of the population. Suppose we denote the distribution by F. Then we say that the random sample X is distributed as F.

Now, suppose instead of considering a random sample of size 1, we consider a random sample of size n (i.e., we randomly select n students and enquire). For each of these samples, we have an associated random variable. Let us denote those by $X_1, \ldots, X_n$. Then we expect that each X_i has the same underlying probability distribution (marginally). Thus X_i is identically distributed as the distribution F, for each $i = 1, \ldots, n$.

Finally, unless otherwise mentioned, there is no reason to believe that the samples are dependent. Thus, we assume that the random samples are mutually independent. Combining the

above facts, the term "$X_1, \ldots, X_n$ is a random sample from F" indicates $X_i \overset{iid}{\sim} F$. The F here can be probability distribution of some random variable, for example, normal, binomial, etc. It depends on the nature of the underlying population. In the above example (see Figure 1), it is natural to consider F to be a mixture of two normal distributions, one is centered around 20 with a narrow standard deviation, and the other is centered around 55 with a large standard deviation.

(d) **Statistical Model.** Practically, it is not feasible to see the relative frequency distribution of the population phenomenon, and choose F accordingly. In practice, statisticians take some assumptions on F based on their experience, past record, or pilot survey. These set of assumptions (including the assumptions on the samples, for e.g., the sample units are iid) together is called the statistical model. The statistical model must be a reasonable one.

Examples. Throughout 2020-2022, many mathematicians, statisticians, computer scientists, epidemiologists have tried to model the growth of COVID-19. Although COVID-19 was a newly discovered disease, the growth modelling was done using past knowledge on pandemic growth mechanism. See this article for a comprehensive review. Based on the assumed model, the scientists tried to infer the number of affected person on a future date, or time for eventual abolition of the disease, etc.

(e) **Parameter.** Sometimes the form of the underlying probability distribution F is assumed to be known except for some constants. For instance, in the previous example, a statistician may suspect the bimodal nature of study time of students and the bell shaped nature of each of the sub-populations. However, s/he may not know that the location of the modes, the spreads, and the proportion of students in each population. Then s/he may set F to be the mixture of two normal distributions, i.e., $\pi \times \texttt{normal}(\mu_1, \sigma_1^2) + (1-\pi) \times \texttt{normal}(\mu_2, \sigma_2^2)$. Thus, F is completely specified except the unknown constants, π (proportion of students in first sub-population), μ_i (locations of the sub-populations) and σ_i^2 (scales of the sub-populations), $i = 1, 2$. These unknown quantities are called parameters.

To answer different questions related to this population, one needs to make inference on these parameters only. This type of inference is called parametric inference. An other section of inference problems are solved without assuming a particular form of F. This section is called non-parametric inference. In MTH211, we will consider parametric inference only.

(f) **Statistic.** In parametric, as well as non-parametric inference, the ultimate goal is to infer a population feature, say μ, based on sample realizations. Statisticians often put forward a summary function of the samples as an estimate for μ. For instance, in germination status example (see Example 1), $\bar{X}$ was used to estimate θ, in the study time example the sample proportion of students having average weekly study time between 40 to 60 hours can be used to estimate μ. Such functions of observable samples are called sample statistics.

Some notations.

1. The random quantities, including random samples will be indicated by capital letters, $X_1, X_2, \cdots$. The realizations will be indicated by small letters, $x_1, x_2, \ldots$, for e.g., $P(X = x)$.

2. Boldface will be used to indicate vectors, for e.g., $\mathbf{X}$ indicates random vector, $\mathbf{x}$ indicates a vector of realizations.

3. The parameters of a distribution are treated as unknown fixed quantities in frequentist inference, and will be indicated in Greek letters, for example, μ, σ, etc. Here also, boldface will be used to indicate parameter vectors, for example, $\boldsymbol{\mu}$.

III **More on statistics and sampling distributions.** Let $X_1, \ldots, X_n$ be a random sample of size n from a population F. The collection of all possible values of $(X_1, \ldots, X_n)$ is called the sample space. As random variables are measurable functions on $\mathbb{R}$, the sample space is a subset of $\mathbb{R}^n$. Let $T(\cdot)$ be a real (or vector) valued function whose domain includes the sample space of $(X_1, \ldots, X_n)$. Then the random variable (or vector) $Y = T(X_1, \ldots, X_n)$ is called a statistic.

As $(X_1, \ldots, X_n)$ is random, so is a statistic $Y = T(X_1, \ldots, X_n)$. The probability distribution of a statistic is called the sampling distribution of the statistic.

Note that, a statistic does NOT involve a parameter. However, the sampling distribution of a statistic may involve parameters. For example, let $X_1, \ldots, X_n$ be a random sample of size n from $\texttt{normal}(\mu, \sigma^2)$ distribution. Then $\sum_{i=1} X_i$ is a statistic, and its distribution is $\texttt{normal}(\mu, \sigma^2/n)$. However, $\sum_{i=1}^n (X_i - \mu)/S$ is not a statistic, as it involves μ.

Support of a random variable

Let X be a *discrete* random variable. Then the support of X, say $\mathcal{S}_X$, is the collection of points x in $\mathbb{R}$ such that $P(X = x) > 0$, i.e., $\mathcal{S}_X = \{x \in \mathbb{R} : P(X = x) > 0\}$.

Let X be a *continuous* random variable with CDF F_X. Then the support of X, say $\mathcal{S}$, is the collection of points x in $\mathbb{R}$ such that X has a probability mass at each non-trivial neighborhood of x, i.e., $\mathcal{S}_X = \{x \in \mathbb{R} : F_X(x + h) - F_X(x - h) > 0, \quad \text{for all } h > 0\}$.

Thus, if X is a discrete (or, continuous) random variable with pmf (or, pdf) f_X and support $\mathcal{S}_X$. Then $x \notin \mathcal{S}_X$, then $f_X(x) = 0$. (WHY? Is the converse of the above statement true?)

Some important distributions

1. Discrete distributions

(a) **Bernoulli.** Let $X \sim \texttt{Bernoulli}(p)$ distribution. Then

$$P(X = x) = p^x (1 - p)^{1-x}, \qquad \mathcal{S}_X = \{0, 1\}, \quad 0 < p < 1.$$

(b) **Binomial.** Let $X \sim \texttt{binomial}(n, p)$ distribution. Then

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \qquad \mathcal{S}_X = \{0, 1, \ldots, n\}, \quad 0 < p < 1, \quad n \in \mathbb{N}.$$

(c) **Poisson.** Let $X \sim \texttt{Poisson}(\lambda)$ distribution. Then

$$P(X = x) = \exp\{-\lambda\} \lambda^x / x!, \qquad \mathcal{S}_X = \{0, 1, \ldots\}, \quad \lambda > 0.$$

(d) **Geometric.** Let $X \sim \texttt{geometric}(p)$ distribution. Then

$$P(X = x) = (1 - p)^{x-1} p, \qquad \mathcal{S}_X = \{1, \ldots\}, \quad 0 < p < 1.$$

2. Continuous distributions

(a) **Uniform.** Let $X \sim \texttt{uniform}(\alpha, \beta)$ distribution. Then the probability density function of X, $f_X(\cdot)$, is given by

$$f_X(x) = \frac{1}{\beta - \alpha}, \qquad \alpha < x < \beta, \quad \mathcal{S}_X = [\alpha, \beta], \quad \alpha, \beta \in \mathbb{R}, \quad \beta > \alpha.$$

(b) **Gamma.** Let $X \sim \texttt{Gamma}(\alpha, \beta)$ distribution. Then the probability density function of X, $f_X(\cdot)$, is given by

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \qquad x > 0, \quad \mathcal{S}_X = [0, \infty), \quad \alpha > 0, \ \beta > 0.$$

(c) **Exponential.** Let $X \sim \texttt{exponential}(\lambda)$ distribution. Then the probability density function of X, $f_X(\cdot)$, is given by

$$f_X(x) = \lambda e^{-\lambda x}, \qquad x > 0, \quad \mathcal{S}_X = [0, \infty), \quad \lambda > 0.$$

Clearly, exponential is a special case of Gamma distribution with parameters 1 and λ.

(d) **Normal.** Let $X \sim \texttt{normal}(\mu, \sigma^2)$ distribution. Then the probability density function of X, $f_X(\cdot)$, is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \qquad x \in \mathcal{S}_X = \mathbb{R}, \quad \mu \in \mathbb{R}, \quad \sigma > 0.$$

(e) **Beta.** Let $X \sim \texttt{beta}(\alpha, \beta)$ distribution. Then the probability density function of X, $f_X(\cdot)$, is given by

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 < x < 1, \quad \mathcal{S}_X = [0, 1], \quad \alpha > 0, \ \beta > 0.$$

(f) **Cauchy.** Let $X \sim \texttt{Cauchy}(\mu, \sigma)$ distribution. Then the probability density function of X, $f_X(\cdot)$, is given by

$$f_X(x) = \frac{1}{\pi\sigma}\left[1 + \left(\frac{x-\mu}{\sigma}\right)^2\right]^{-1}, \qquad x \in \mathcal{S}_X = \mathbb{R}, \quad \mu \in \mathbb{R}, \quad \sigma > 0.$$

(g) **Chi-squared distribution.** Let $X_i \overset{iid}{\sim} \texttt{normal}(0, 1)$ distribution, $i = 1, \ldots, n$. Then $T = \sum_{i=1}^n X_i^2$ follows a `Chi-squared` distribution with degrees of freedom (d.f.) n, notationally $T \sim \chi^2_n$, and the pdf of T, f_T, is given by

$$f_T(t) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, \qquad x > 0, \quad \mathcal{S}_X = [0, \infty), \quad n \in \mathbb{N}.$$

Chi-squared distribution with d.f. n is a special case of Gamma distribution with parameters $n/2$ and $1/2$.

(h) F **distribution.** Let $X \sim \chi^2_{n_1}$, $Y \sim \chi^2_{n_2}$ and X and Y are independently distributed, then $F = n_2 X/(n_1 Y)$ follows an F distribution with d.f. n_1 and n_2, notationally $F \sim F_{n_1, n_2}$, and the pdf of F, f_F, is given by

$$f_F(x) = \frac{\Gamma((n_1+n_2)/2)}{\Gamma(n_1/2)\Gamma(n_2/2)} \left(\frac{n_1}{n_2}\right)^{n_1/2} x^{n_1/2-1} \left(1 + \frac{n_1}{n_2}x\right)^{-(n_1+n_2)/2}, \quad x > 0, \quad \mathcal{S}_X = [0, \infty).$$

(i) t **distribution.** Let $X \sim \texttt{normal}(0, 1)$, $Y \sim \chi^2_n$ and X and Y are independently distributed. Then $W = X/\sqrt{Y/n}$ follows a t-distribution with d.f. n, notationally $W \sim t_n$, and the pdf of W, f_W is given by

$$f_W(x) = \frac{\Gamma((n+1)/2)}{\sqrt{\pi n}\Gamma(n/2)} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}, \qquad x \in \mathcal{S}_X = \mathbb{R}.$$

Properties (Homework):

1. Find the expectations and variances of each of the above distributions (if exist).
2. (**Additive properties**) *Prove the following statements using moment generating functions, or characteristic functions.*
   (a) Let $X_i \overset{ind}{\sim} \texttt{binomial}(n_i, p)$, for $i = 1, \ldots, k$, then $T = \sum_{i=1}^{k} X_i$ follows $\texttt{binomial}(\sum_i n_i, p)$.
   (b) Let $X_i \overset{ind}{\sim} \texttt{Poisson}(\lambda_i)$, for $i = 1, \ldots, n$, then $T = \sum_{i=1}^{n} X_i$ follows $\texttt{Poisson}(\sum_i \lambda_i)$.
   (c) Let $X_i \overset{ind}{\sim} \texttt{normal}(\mu_i, \sigma_i^2)$, for $i = 1, \ldots, n$, then $T = \sum_{i=1}^{n} X_i$ follows $\texttt{normal}\left(\sum_i \mu_i, \sum_i \sigma_i^2\right)$.
   (d) Let $X_i \overset{ind}{\sim} \texttt{Gamma}(\alpha_i, \beta)$, for $i = 1, \ldots, n$, then $T = \sum_{i=1}^{n} X_i$ follows $\texttt{Gamma}\left(\sum_i \alpha_i, \beta\right)$.
   (e) Let $X_i \overset{ind}{\sim} \chi^2_{n_i}$, for $i = 1, \ldots, k$, then $T = \sum_{i=1}^{k} X_i$ follows χ^2_N where $N = \sum_i n_i$.
3. Let $X \sim \texttt{normal}(\mu, \sigma^2)$ distribution, then $T = aX + b \sim \texttt{normal}(a\mu + b, a^2\sigma^2)$.
4. Let $X \sim \texttt{Gamma}(\alpha, \beta)$ distribution, then $T = aX \sim \texttt{Gamma}(\alpha, \beta/a)$.
5. Let $X \sim \texttt{beta}(n/2, m/2)$ distribution, then $T = mX/\{n(1-X)\} \sim F_{n,m}$.
6. Let $X \sim \texttt{uniform}(0, 1)$ distribution, and $\alpha > 0$ then $T = X^{1/\alpha} \sim \texttt{beta}(\alpha, 1)$.
7. Let $X \sim \texttt{Cauchy}(0, 1)$ distribution, then $T = 1/(1 + X^2) \sim \texttt{beta}(0.5, 0.5)$.
8. Let $X \sim \texttt{uniform}(0, 1)$ distribution, then $T = -2 \log X \sim \chi^2_2$.
9. Let X be distributed as some absolutely continuous distribution with cdf G_X, then $T = G_X(X) \sim \texttt{uniform}(0, 1)$.

3. Multivariate distributions

Suppose X_i is an absolutely continuous (or, discrete) random variable with pdf (or, pmf) f_i, $i = 1, \ldots, n$, and $X_1, \ldots, X_n$ are mutually independent, then the multivariate distribution of the random vector $\mathbf{X} = (X_1, \ldots, X_n)$ has the joint pdf (or, pmf) $f_{\mathbf{X}}$, where

$$f_{\mathbf{X}}(x_1, \ldots, x_n) = f_1(x_1) \times \cdots \times f_n(x_n), \qquad \text{for each} \quad \mathbf{x} \in \mathbb{R}^n.$$

Thus, if $X_1, \ldots, X_n$ is a random sample from some distribution with pdf (or, pmf) f_X. Then the joint pdf (or, pmf) of $(X_1, \ldots, X_n)$ evaluated at the point $\mathbf{x} \in \mathbb{R}^n$ is $\prod_{i=1}^n f_X(x_i)$.

However, if $X_1, \ldots, X_n$ are not mutually independent, then the above generalization is NOT possible. In that case the joint distribution of $X_1, \ldots, X_n$ can NOT be expressed in terms marginal distributions. In order to infer about the joint behavior of $X_1, \ldots, X_n$, one needs to know the joint distribution.

A k-dimensional random vector $\mathbf{X}$ can be characterized by the joint CDF

$$F_{\mathbf{X}}(\mathbf{x}) = P(X_1 \leq x_1, \cdots, X_k \leq x_k),$$

or the moment generating function (MGF, if exists) $E(\exp\{\mathbf{X}'\mathbf{t}\})$, or the characteristic function $E(\exp\{i\mathbf{X}'\mathbf{t}\})$.

A k-dimensional discrete random vector $\mathbf{X}$ can also be characterized by its joint pmf

$$f_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}) = P(X_1 = x_1, \cdots, X_k = x_k).$$

A k-dimensional absolutely continuous random vector $\mathbf{X}$ can also be characterized by its joint pdf $f_{\mathbf{X}}$, where $f_{\mathbf{X}}$ satisfies

$$F_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{\mathbf{X}}(\mathbf{t}) d\mathbf{t}, \qquad \text{and} \qquad \frac{\partial^k}{\partial x_1 \cdots \partial x_k} F_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}).$$

Marginal distribution Let $\mathbf{X}$ be a k-dimensional random vector with CDF $F_{\mathbf{X}}$, then the marginal CDF of j-th component of $\mathbf{X}$, X_j is

$$F_{X_j}(x) = F_{\mathbf{X}}(\infty, \cdots, \infty, \underbrace{x}_{\text{j-th}}, \infty, \cdots, \infty), \qquad x \in \mathbb{R}.$$

It can be shown that F_{X_j} is a CDF of some distribution, and the corresponding distribution is called the marginal distribution of X_j.

Conditional distribution Let $(\mathbf{X}, \mathbf{Y})'$ be a discrete random variable. Then the conditional distribution of $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$ where $\mathbf{y} \in \mathcal{S}_{\mathbf{Y}}$, is the discrete random variable with pmf

$$f_{\mathbf{X}|\mathbf{Y}=\mathbf{y}}(\mathbf{x}) = P(\mathbf{X} = \mathbf{x} \mid \mathbf{Y} = \mathbf{y}) = \frac{P(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y})}{P(\mathbf{Y} = \mathbf{y})} = \frac{f_{(\mathbf{X},\mathbf{Y})}(\mathbf{x}, \mathbf{y})}{f_{\mathbf{Y}}(\mathbf{y})}.$$

Let $(X, Y)'$ be an absolutely continuous random variable with joint CDF $F_{(X,Y)}$, joint pdf $f_{(X,Y)}$ and y be such that $f_Y(y) > 0$. Then the CDF and pdf of X given $Y = y$, denoted by $F_{X|Y=y}$ and $f_{X|Y=y}$, respectively, are defined as

$$F_{X|Y=y}(x) = \lim_{h \downarrow 0} P(X \leq x \mid y - h < Y \leq y), \qquad \text{and} \qquad f_{X|Y=y}(x) = \frac{f_{(X,Y)}(x, y)}{f_Y(y)}.$$

Similarly, if $(\mathbf{X}, \mathbf{Y})'$ is an absolutely continuous random variable and $\mathbf{y}$ be such that $f_{\mathbf{Y}}(\mathbf{y}) > 0$, then the pdf of $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$ is $f_{\mathbf{X}|\mathbf{Y}=\mathbf{y}}(\mathbf{x}) = f_{(\mathbf{X},\mathbf{Y})}(\mathbf{x}, \mathbf{y}) / f_{\mathbf{Y}}(\mathbf{y})$.

Some multivariate distributions are often of interest to MTH211. One such distribution is multivariate normal distribution.

Multivariate normal distribution. The random vector $\mathbf{X} = (X_1, \ldots, X_k)'$ is distributed as multivariate normal distribution with parameters $\boldsymbol{\mu} \in \mathbb{R}^k$ and Σ, where Σ is a $k \times k$ symmetric and positive semi-definite matrix, if for any vector $\mathbf{a} \in \mathbb{R}^k$, $T_a = \mathbf{a}'\mathbf{X} \sim \texttt{normal}\,(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\Sigma\mathbf{a})$. Notationally, $\mathbf{X} \sim N_k\,(\boldsymbol{\mu}, \Sigma)$.

Special case: The two-dimensional normal distribution is called bivariate normal distribution, and it has 5 parameters $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$. If $\mathbf{X} \sim N_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$, then the pdf of $\mathbf{X}$, $f_{\mathbf{X}}$, is given by

$$f_{\mathbf{X}}(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right\},$$

$$\begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2, \quad \mu_x, \mu_y \in \mathbb{R}, \quad \sigma_x, \sigma_y > 0, \quad 0 \le \rho \le 1.$$

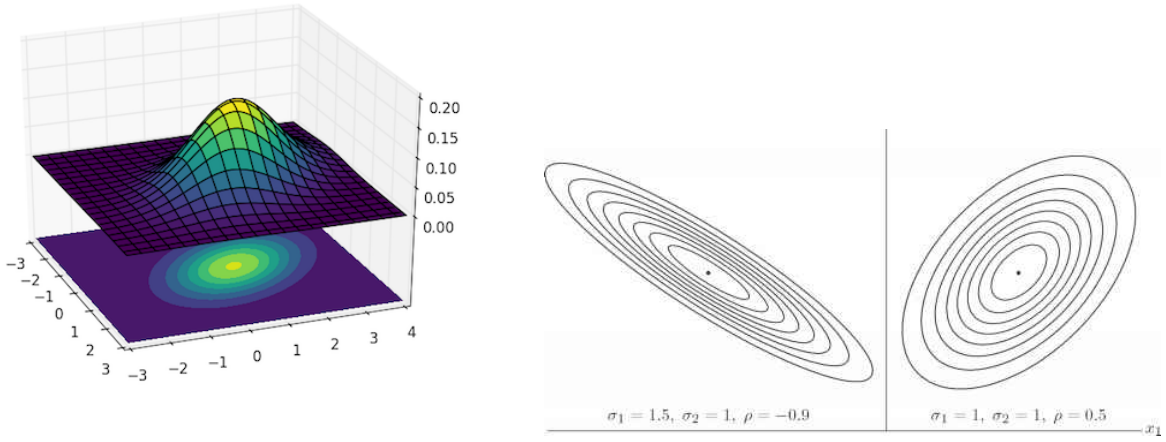


Figure 2: Contour plot of bivariate normal distribution

Properties (Homework):

1. Let $\mathbf{X} \sim N_k(\boldsymbol{\mu}, \Sigma)$. Then the expectation of $\mathbf{X}$ is $E(\mathbf{X}) = \boldsymbol{\mu}$, variance-covariance matrix is $E\left[\{\mathbf{X} - E(\mathbf{X})\}\{\mathbf{X} - E(\mathbf{X})\}'\right] = \Sigma$, and the moment generating function is $\exp\{\boldsymbol{\mu}'\mathbf{t} + \mathbf{t}'\Sigma\mathbf{t}/2\}$.

2. (Box-Muller transformation) Let $U_i \overset{iid}{\sim} \texttt{uniform}(0, 1)$, $i = 1, 2$. Consider the transformations $Z_1 = \sqrt{-2\log U_1}\cos(2\pi U_2)$, and $Z_2 = \sqrt{-2\log U_1}\sin(2\pi U_2)$. Then $(Z_1, Z_2)' \sim N_2(\mathbf{0}, I)$, where I is the identity matrix.

3. Let $X_i \overset{iid}{\sim} \texttt{Gamma}(\alpha_i, \beta)$, $i = 1, 2$. Consider the transformation $Z = X_1/(X_1 + X_2)$. Then $Z \sim \texttt{beta}(\alpha_1, \alpha_2)$.

4. Let (X, Y) is jointly distributed as $N_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$. Then the marginal distribution of X is $\texttt{normal}(\mu_x, \sigma_x^2)$. Also, the conditional distribution of Y given $X = x$ is $\texttt{normal}(\mu_y + \rho\sigma_y(x - \mu_x)/\sigma_x, \sigma_y^2(1-\rho^2))$.

5. Let $X_i \overset{iid}{\sim} \texttt{Bernoulli}(p)$, $i = 1, \ldots, n$. Then the conditional distribution of $\mathbf{X}$ given $\bar{X}_n = y$ is free of p.

6. Let $X_i \overset{iid}{\sim} \texttt{Poisson}(\lambda)$, $i = 1, \ldots, n$. Then the conditional distribution of $\mathbf{X}$ given $\bar{X}_n = y$ is free of λ.

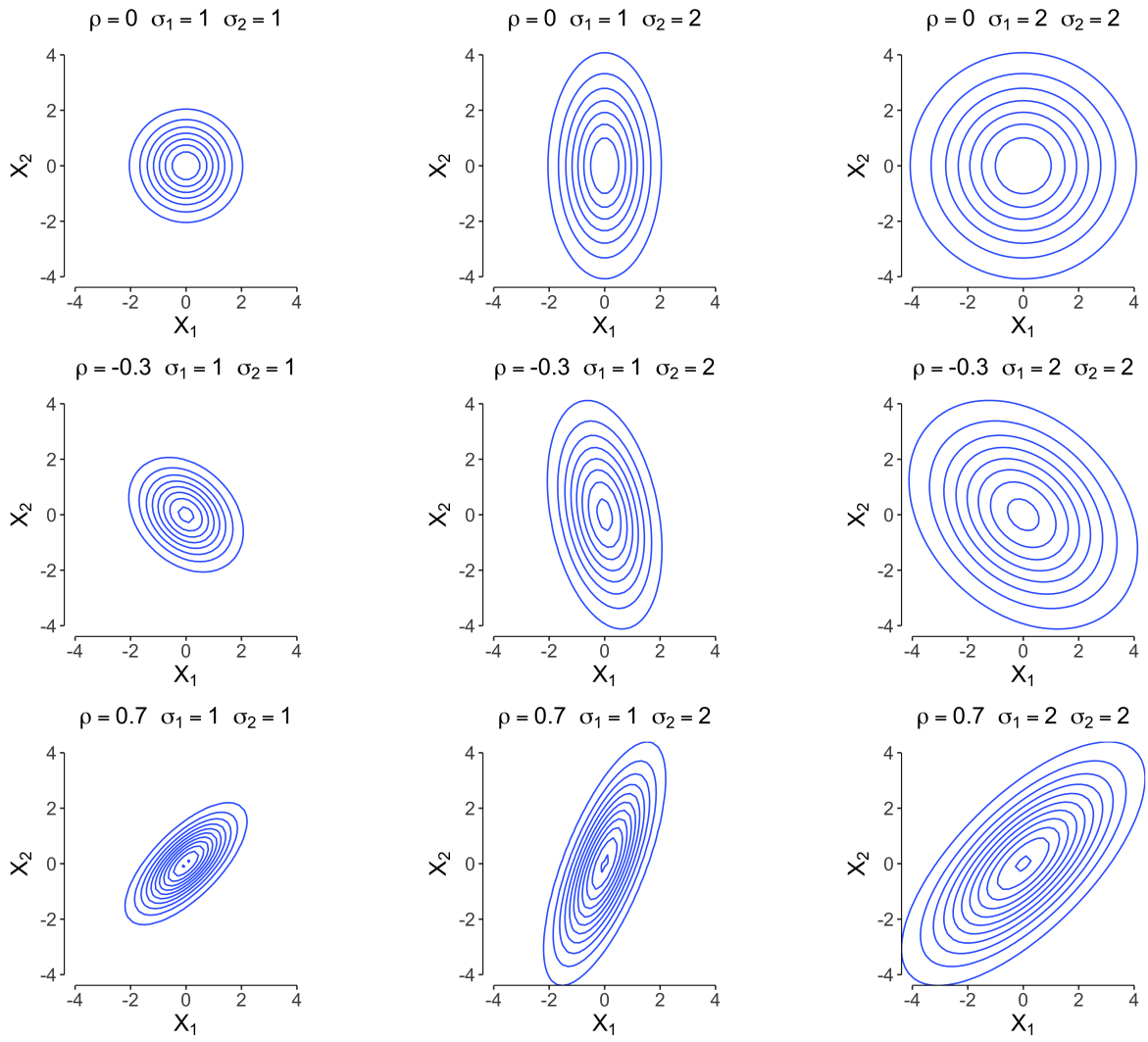


Figure 3: Contour plots of bivariate normal distribution with different parameters

Some important statistics and their sampling distribution

1. **Sample mean.** Let $X_1, \ldots, X_n$ be a random sample from some distribution F. Then $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ is called the sample mean.

   Properties of sample mean.

   (a) Let $X_1, \ldots, X_n$ be a random sample from some distribution F with expectation μ and finite variance σ^2. Then $E(\bar{X}_n) = \mu$ and $\text{var}(\bar{X}_n) = \sigma^2/n$. [Proof]

   (b) Let $X_1, \ldots, X_n$ be a random sample from $\texttt{normal}(\mu, \sigma^2)$ distribution. Then the sampling distribution of $\bar{X}_n$ is $\texttt{normal}(\mu, \sigma^2/n)$. [Proof]

   (c) Let $X_1, \ldots, X_n$ be a random sample from some distribution F with expectation μ, and finite variance σ^2. Then $\bar{X}_n$ is the best linear unbiased estimator (BLUE) of μ. [Proof]

2. **Sample variance.** Let $X_1, \ldots, X_n$ be a random sample from some distribution F. Then $S_n^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = n^{-1}\sum_{i=1}^n X_i^2 - \bar{X}_n^2$ is called the sample variance.

   The positive square root of sample variance is called sample standard deviation, and is denoted by S_n.

   Properties of sample variance.

   (a) Let $X_1, \ldots, X_n$ be a random sample from some distribution F with expectation μ, and finite variance σ^2. Then $E\left(nS_n^2/(n-1)\right) = \sigma^2$. [Proof]

   The statistics $S_n^{\star 2} = nS_n^2/(n-1)$ is sometimes also referred to as the sample variance. We will term this statistic as unbiased sample variance.

   (b) Let $X_1, \ldots, X_n$ be a random sample from $\texttt{normal}(\mu, \sigma^2)$ distribution. Then the sampling distribution (SD) of nS_n^2 is $\sigma^2\chi^2_{n-1}$. [Proof]

   (c) (**Joint distribution of sample mean and variance for normal samples**) Let $X_1, \ldots, X_n$ be a random sample from $\texttt{normal}(\mu, \sigma^2)$ distribution. Then $\bar{X}_n$ and S_n^2 are independently distributed. [Proof]

   Thus, $T_\mu = \sqrt{n}(\bar{X}_n - \mu)/S_n^\star \sim t_{n-1}$. (Why?)

3. **Correlation coefficient.** Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a bivariate random sample from some distribution F. Then the sample correlation coefficient

$$r_{x,y} = \frac{n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{S_X S_Y} = \frac{n^{-1}\sum_{i=1}^n X_i Y_i - \bar{X}_n\bar{Y}_n}{S_X S_Y},$$

   where S_X and S_Y are SDs of X and Y, respectively.

   The numerator in the expression of $r_{x,y}$ is called sample covariance.

4. **Multivariate extensions of sample mean and variance.** Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from some multivariate distribution F. Then the sample mean $\bar{\mathbf{X}}_n$ and sample variance covariance matrix is defined as

$$\bar{\mathbf{X}}_n = n^{-1}\sum_{i=1}^n \mathbf{X}_i, \qquad \text{and} \qquad S_n = n^{-1}\sum_{i=1}^n \left(\mathbf{X}_i - \bar{\mathbf{X}}_n\right)\left(\mathbf{X}_i - \bar{\mathbf{X}}_n\right)' = n^{-1}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i' - \bar{\mathbf{X}}_n\bar{\mathbf{X}}_n'.$$

Properties of sample mean and variance.

(a) If the samples are k-dimensional then $\bar{\mathbf{X}}_n$ is a k-vector, and S_n is a $k \times k$ positive semi definite matrix. The j-th diagonal of S_n is the sample variance of the j-th component of $\mathbf{X}$, and the (i,j)-th component of S_n is the covariance between the i-th and j-th element of $\mathbf{X}$. (WHY?)

(b) Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from some multivariate distribution F with expectation $\boldsymbol{\mu}$ and variance covariance matrix Σ having finite components. Then $E(\bar{\mathbf{X}}_n) = \boldsymbol{\mu}$, $\text{var}(\bar{\mathbf{X}}_n) = n^{-1}\Sigma$ and $E(nS_n/(n-1)) = \Sigma$. [Proof]

5. **Sample moments.** Let $X_1, \ldots, X_n$ be a random sample from some distribution F. Then the r-th order raw moment m'_r and r-th order central moment m_r are defined as

$$m'_r = n^{-1}\sum_i X_i^r, \qquad \text{and} \qquad m_r = n^{-1}\sum_i \left(X_i - \bar{X}_n\right)^r, \quad r > 0.$$

Properties of sample moments.

(a) The sample central moments can be derived from the sample raw moments and vice versa.

(b) Let $X_1, \ldots, X_n$ be a random sample from some distribution F with r-th population moment $\mu'_r = E\left(X^r\right) < \infty$. Then $E\left(m'_r\right) = \mu'_r$. [Proof]

6. **Order statistics.** Let $X_1, \ldots, X_n$ be a random sample from some distribution F. Then the r-th order statistics $X_{(r)}$ is the r-th smallest of $X_1, \ldots, X_n$, $r = 1, \ldots, n$. Therefore, $X_{(1)} = \min\{X_1, \ldots, X_n\}$ and $X_{(n)} = \max\{X_1, \ldots, X_n\}$.

Properties of sample variance.

(a) Let $X_1, \ldots, X_n$ be a random sample from some distribution F with pdf f_X. Then

$$f_{(X_{(1)},\cdots,X_{(n)})}(\mathbf{x}) = \begin{cases} n!\prod_{i=1}^n f_X(x_i) & \text{if} \quad x_1 < x_2 < \cdots < x_n, \\ 0 & \text{otherwise.} \end{cases} \quad \text{[Proof]}$$

(b) Let $X_1, \ldots, X_n$ be a random sample from some distribution with CDF F_X. Then the CDF of $X_{(n)}$, say G_n, is given by

$$G_n(t) = P(X_{(n)} \le t) = \{P(X \le t)\}^n = F_X^n(t). \quad \text{[Proof]}$$

(c) Let $X_1, \ldots, X_n$ be a random sample from some distribution with CDF F_X. Then the CDF of $X_{(1)}$, say H_n, is given by

$$H_n(t) = P(X_{(1)} \le t) = 1 - \{1 - P(X \le t)\}^n = 1 - \{1 - F_X(t)\}^n.$$

(d) Let $X_1, \ldots, X_n$ be a random sample from some distribution with CDF F_X and pdf f_X. Then the pdf of $X_{(r)}$, say g_r, is given by

$$g_r(t) = \frac{n!}{(r-1)!(n-r)!} F_X(t)^{r-1} f_X(t)\{1 - F_X(t)\}^{n-r}.$$

(e) Let $X_1, \ldots, X_n$ be a random sample from some distribution with CDF F_x and pdf f_X. Then the joint pdf of $X_{(r)}$ and $X_{(s)}$ with $r < s$, say $g_{r,s}$, is given by

$$g_{r,s}(w,t) = \begin{cases} \dfrac{n!}{(r-1)!(s-r-1)!(n-s)!} F_X(w)^{r-1}\{F_X(t) - F_X(w)\}^{s-r-1} f_X(w) f_X(t)\{1 - F_X(t)\}^{n-s} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if} \quad w < t, \\ 0 \qquad \text{otherwise.} \end{cases}$$

7. **Sample median.** Let $X_1, \ldots, X_n$ be a random sample from some distribution F. Then the sample median, $\tilde{X}_{me}$ is given by

$$\tilde{X}_{me} = \begin{cases} X_{((n+1)/2)} & \text{if } n \text{ is odd,} \\ \{X_{(n)} + X_{(n+1)}\}/2 & \text{if } n \text{ is even.} \end{cases}$$

IV **Large Sample Results.** Two large sample results would be useful in MTH211.

1. (**Weak Law of Large Numbers, WLLN**) Let $X_1, \ldots, X_n$ be a random sample from a population with $E(g(X)) = \eta < \infty$, where $g : \mathbb{R} \to \mathbb{R}$ is a function. Then $n^{-1} \sum_{i=1}^{n} g(X_i) \xrightarrow{p} \eta$ as $n \to \infty$, i.e., for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left| n^{-1} \sum_{i=1}^{n} g(X_i) - \eta \right| > \epsilon \right) = 0.$$

2. (**Central Limit Theorem, CLT**) Let $X_1, \ldots, X_n$ be a random sample from some distribution with expectation μ and finite variance σ^2. Then $T_n^{\mu,\sigma} = \sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{d} \texttt{normal}(0, 1)$, that is for any $x \in \mathbb{R}$, the CDF of $T_n^{\mu,\sigma}$, say G_n, satisfies

$$G_n(x) \to \Phi(x), \qquad \text{as} \qquad n \to \infty,$$

where Φ is the CDF of $\texttt{normal}(0, 1)$ distribution.

Corollary:

(a) For any $r > 0$ such that $\mu'_{2r} < \infty$, $m'_r \xrightarrow{p} \mu'_r$.

(b) Let $X_1, \cdots, X_n$ be a random from $\texttt{unoform}(0, \theta)$, $\theta > 0$. Then $X_{(n)} \xrightarrow{d} \delta_{\{\theta\}}$, where $\delta_{\{\theta\}}$ is the degenerate distribution with non-zero probability mass at θ.