Fundamentals of Statistics:
From Data to Knowledge to Decision-Making

Victor M. Zavala

Department of Chemical and Biological Engineering
University of Wisconsin-Madison

victor.zavala@wisc.edu

Motivation

As engineers, we often use *laws* of physics and chemistry to make *decisions*:

- The discovery of these governing laws has been the result of extensive collection and analysis of observations (data)
- A governing law is often expressed in the form of a mechanistic model
- A mechanistic model provides a concise summary of observations (knowledge) that allow us to predict and generalize

These laws are powerful but only provide limited descriptions of phenomena:

- Laws are applicable under specific settings (e.g., continuum vs. atomistic)
- Discovering laws and new mechanistic models might be challenging or cost-prohibitive (e.g., climate)

Mechanistic predictions will *always* face a certain degree of *uncertainty* (due to our limited knowledge of the world). Despite these limitations, we still want to be able to *make decisions*. In fact, we (as humans) make many decisions in our daily lives without using any mechanistic models and (somehow) by accounting for uncertainty.

Motivation

Statistics is the branch of mathematics that offers tools to:

- Collect, analyze, and extract knowledge (models) from data
- Characterize and model the unknown (uncertainty)
- Systematically make decisions in the face of uncertainty

For an engineering perspective, *statistics* aids the discovery and development of mechanistic models and provides complementary (data-driven) modeling capabilities.

From a scientific perspective, *statistics* provides a way of thinking about the world that can help us understand how humans naturally process data to extract knowledge and to ultimately make decisions.

Random Variables

In statistics, we use random variables (RVs) to *model* uncertain phenomena. An RV (denoted as X) does not have a known value and exhibits *variability* and has the following properties:

- An RV is characterized by a realization set $\omega \in \Omega$ with associated values $x_\omega \in \mathcal{D}_X$ (a.k.a. realizations of X). Here, $\mathcal{D}_X$ is the domain of X.
- An RV is characterized by a measure $\mathbb{P} : \Omega \to [0, 1]$, which assigns probability to events (combinations of realizations); e.g., $\mathbb{P}(X \in \mathcal{A})$ for some $\mathcal{A} \subseteq \mathcal{D}_X$.

Random Variables

- The measure $\mathbb{P}$ has an associated cumulative density function (cdf) $F_X : \mathcal{D}_X \to [0,1]$. The cdf assigns a probability to the event that X is below a certain threshold value x; i.e., $F_X(x) = \mathbb{P}(X \leq x)$.
- The cdf has an associated probability density function (pdf) $f_X : \mathcal{D}_X \to [0,1]$. The pdf assigns a probability to the event that X takes a specific value x; i.e., $f_X(x) = \mathbb{P}(X = x)$.

An RV that has a unique and known value (exhibits no variability) is called a *deterministic variable*.

Random Variables

Don’t Forget: A random variable is a *model* of a unknown phenomenon.

Types of Random Variables

RVs are categorized as multivariate vs. univariate and continuous vs. discrete:

- A *multivariate* RV $X = (X_1, X_2, ..., X_n)$ has realizations that are vector values $x_\omega = (x_{\omega,1}, x_{\omega,2}, ..., x_{\omega,n}) \in \mathbb{R}^n$; e.g., temperature and reaction conversion.
- A *univariate* RV X is a multivariate with $n = 1$ and has realizations that are scalar values $x_\omega \in \mathbb{R}$; e.g., temperature.
- A *continuous* RV X is that in which the domain $\mathcal{D}_X$ is continuous; e.g., $X = (X_1, X_2)$ has realizations satisfying $0 \leq x_{\omega,1} \leq 1$ and $0 \leq x_{\omega,2} \leq 1$.
- A *discrete* RV X is that in which the domain $\mathcal{D}_X$ is discrete; e.g., $X = (X_1, X_2)$ has realizations satisfying $x_{\omega,1} \in \{0, 1\}$ and $x_{\omega,2} \in \{0, 1\}$.

There is a wide range of models of random variables that apply to the different categories (e.g., Gaussian is for continuous and Poisson for discrete). We will explore these later.

Probability Density of Discrete and Continuous RVs

A discrete RV X has a discrete domain $\mathcal{D}_X$ and, as such, its pdf $f_X(x)$ is not a continuous function. Consequently:

$$
\begin{aligned}
f(x) &\geq 0, \quad x \in \mathcal{D}_X \\
\sum_{x \in D_X} f(x) &= 1 \\
\mathbb{P}(X \in \mathcal{A}) &= \sum_{x \in \mathcal{A}} f(x), \quad \mathcal{A} \subseteq \mathcal{D}_X.
\end{aligned}
$$

A continuous RV X has a continuous domain $\mathcal{D}_X$ and its pdf $f_X(x)$ is a continuous function. Consequently:

$$
\begin{aligned}
f(x) &\geq 0, \quad x \in \mathcal{D}_X \\
\int_{x \in \mathcal{D}_X} f(x) dx &= 1 \\
\mathbb{P}(X \in \mathcal{A}) &= \int_{x \in \mathcal{A}} f(x) dx, \quad \mathcal{A} \subseteq \mathcal{D}_X.
\end{aligned}
$$

Probability Density of Discrete and Continuous RVs

- A discrete RV is easy to handle computationally (involves summations):
  - If $\mathcal{D}_X$ is discrete then $\mathbb{P}(X \leq a) = F_X(a) = \sum_{x \in \mathcal{D}_X} f_X(x)\mathbf{1}[x \leq a]$.

  Here, we use the indicator function $\mathbf{1}[x \leq a] = 1$ if $x \leq a$ and $\mathbf{1}[x \leq a] = 0$ if $x > a$.
- A continuous RV is difficult to handle computationally (involves integrals) but has useful properties that facilitate analysis:
  - If $\mathcal{D}_X$ continuous then $\mathbb{P}(X \leq a) = F_X(a) = \int_{x \in \mathcal{D}_X} f_X(x)dx$.
  - The cdf and pdf are related as $\frac{dF_X(x)}{dx} = f_X(x)$ and thus $\int_{x \in \mathcal{A}} f_X(x)dx = \int_{x \in \mathcal{A}} dF_X(x)$.
- Continuous RVs are often approximated using discrete RVs (discretization). This is analogous to discrete time and continuous time.

From Data to Random Variables

We will begin our discussion by considering RVs that are *univariate*.

- In practice, we often count with observations (data) for a given variable of interest $x_\omega,\, \omega \in \mathcal{S}$. Our objective is to use the data to model this variable as an RV.
- If we assume that $\mathcal{S}$ (a.k.a. sample set) is a subset of the realization set Ω (which is usually extremely large), we can construct a data-driven approximation (a.k.a. empirical or sample approximation) of the domain, cdf, and pdf of X.
- The empirical domain $\hat{D}_X$ is the domain covered by the observations $x_\omega,\, \omega in \Omega$.
- The pdf is approximated using the empirical pdf:

$$\hat{f}_X(x) = \frac{1}{S} \sum_{\omega \in \mathcal{S}} [x_\omega = x],\; x \in \hat{D}_X$$

  i.e., this is the frequency at which X takes a value x normalized by $S = |\mathcal{S}|$.
- The cdf is approximated using the empirical cdf:

$$\hat{F}_X(x) = \frac{1}{S} \sum_{\omega \in \mathcal{S}} \mathbf{1}[x_\omega \leq x],\; x \in \hat{D}_X$$

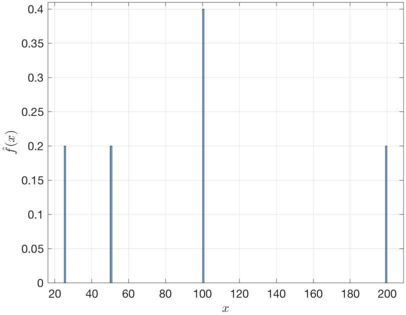
  i.e., this is the frequency at which X takes a value below x normalized by $S = |\mathcal{S}|$.

From Data to Random Variables

Example `rainfall.m`: Consider RV X representing the uncertain input flow of a system (e.g., rainfall). The set of available observations is $\mathcal{S} = \{1, 2, ..., 10\}$ with associated values x_ω is given by $(100, 200, 100, 200, 100, 50, 50, 25, 100, 25)$ gpm.

- The empirical domain is: $\hat{D}_X = \{25, 50, 100, 200\}$
- The empirical pdf is visualized using a histogram:

$$\hat{f}_X(x) = \begin{cases} 2/10 & \text{if} \quad x = 200 \\ 4/10 & \text{if} \quad x = 100 \\ 2/10 & \text{if} \quad x = 50 \\ 2/10 & \text{if} \quad x = 25 \\ 0 & \text{if} \quad \text{otherwise} \end{cases}$$



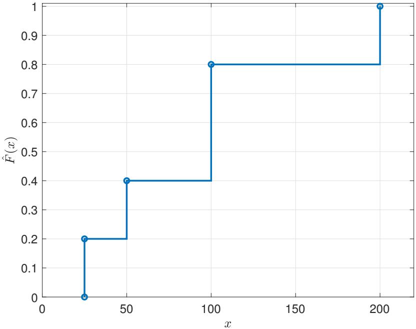
The empirical pdf is a discontinuous function defined by a finite number of points.

From Data to Random Variables

Example `rainfall.m`: Consider RV X representing the uncertain input flow of a system (e.g., rainfall). The set of available observations is $\mathcal{S} = \{1, 2, ..., 10\}$ with associated values x_ω is given by $(100, 200, 100, 200, 100, 50, 50, 25, 100, 25)$ gpm.

The empirical cdf is:

$$\hat{F}_X(x) = \begin{cases} 2/10 & \text{if} \quad x \leq 25 \\ 4/10 & \text{if} \quad x \leq 50 \\ 8/10 & \text{if} \quad x \leq 100 \\ 10/10 & \text{if} \quad x \leq 200 \end{cases}$$



The empirical cdf is a discontinuous function with jumps corresponding to the points defining the empirical pdf.

Summarizing Statistics (Basic)

The pdf and cdf are *functions* that fully characterize an RV X. However, in practice, we might be interested in using scalar quantities and not functions. This can be done by using summarizing statistics.

For a discrete RV we have:

- *Expected Value (measure of magnitude):* $\mathbb{E}_X := \sum_{x \in \mathcal{D}_X} x f_X(x)$
- *Variance and Standard Deviation (measure of variability):*

$$\mathbb{V}_X := \sum_{x \in \mathcal{D}_X} f_X(x)(x - \mathbb{E}_X)^2, \qquad \mathrm{SD}_X := \sqrt{\mathbb{V}_X}$$

For a continuous RV we have:

- *Expected Value (measure of magnitude):* $\mathbb{E}_X := \int_{x \in \mathcal{D}_X} x f_X(x) dx$
- *Variance and Standard Deviation (measure of variability):*

$$\mathbb{V}_X := \int_{x \in \mathcal{D}_X} (x - \mathbb{E}_X)^2 f_X(x) dx, \qquad \mathrm{SD}_X := \sqrt{\mathbb{V}_X}$$

When convenient, we will use notation $\mathbb{E}[X]$ and $\mathbb{V}[X]$ to define summarizing statistics.

Summarizing Statistics (Sample Approximations)

If we only have a finite set of data for X available, we can approximate the summarizing statistics using their sample approximations:

- *Sample Mean (measure of magnitude):*

$$\hat{\mathbb{E}}_X := \sum_{x\in\hat{\mathcal{D}}_X} x\hat{f}_X(x) = \frac{1}{S}\sum_{\omega\in\mathcal{S}} x_\omega$$

- *Sample Variance and Standard Deviation (measure of variability):*

$$\hat{\mathbb{V}}_X := \sum_{x\in\hat{\mathcal{D}}_X} (x-\hat{E}_X)^2\hat{f}_X(x) = \frac{1}{S}\sum_{\omega\in\mathcal{S}}(x_\omega-\hat{E}_X)^2, \qquad \hat{\mathrm{SD}}_X = \sqrt{\hat{\mathbb{V}}_X}$$

Intuition tells us that these become better approximations as we accumulate data of X (i.e., as $\mathcal{S}$ approaches Ω). We will see later on that this is indeed the case.

Summarizing Statistics (Quantiles)

An important family of summarizing statistics are the quantiles (a.k.a. percentiles).

- The quantile is the inverse function of the cdf and, as such, it might be easier to explain it from this perspective. Consider the following equation for some $\alpha \in [0, 1]$:

$$F_X(x) = \mathbb{P}(X \leq x) = \alpha$$

- A value x that satisfies this equation is the α-quantile of the random variable X and is denoted as $\mathbb{Q}_X(\alpha)$. This means that we can express the quantile as:

$$\mathbb{Q}_X(\alpha) = F_X^{-1}(\alpha)$$

Summarizing Statistics (Quantiles)

Some important observations about quantiles:

- Since the cdf can have a “staircase” form, there might be multiple values of x satisfying $F_X(x) = \alpha$. Consequently, the α-quantile might be not be unique. Typically, the definition of the quantile is refined by looking for the smallest or center values of x satisfying $F_X(x) \geq \alpha$.
- Quantiles are related to other summarizing statistics for interest. For instance:
  - $\mathbb{Q}_X(1/2)$ is the *center value* of X (a.k.a. the median and denoted as $\mathbb{M}_X$)
  - $\mathbb{Q}_X(1) = \max_{x \in \mathcal{D}_X} x$ is the maximum value of X
  - $\mathbb{Q}_X(0) = \min_{x \in \mathcal{D}_X} x$ is the minimum value of X
- We can use the empirical cdf $\hat{F}_X(x)$ to estimate empirical quantiles $\hat{\mathbb{Q}}(\alpha)$.

Summarizing Statistic (Moments)

- An important family of summarizing statistics are central moments.
- The moments of X with pdf $f(x, \theta)$ are given by:

$$m_k := \mathbb{E}[(X - \mathbb{E}[X])^k], \qquad k = 1, 2, 3, 4, ...,$$

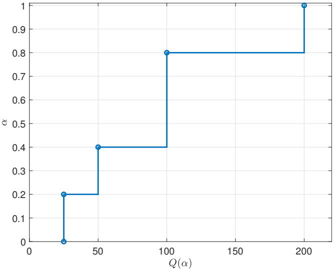
- The first moment is simply $m_1 = 0$ while the second moment $m_2 = \mathbb{V}[X]$ is the variance, the third moment m_3 is known as skewness, and the fourth moment m_4 as kurtosis.
- As with the expectation and variance, we can use data to construct sample approximations for the moments.

Summarizing Statistics

Example `rainfall.m`: Consider RV X representing the uncertain input flow of a system (e.g., rainfall). The set of available observations is $\mathcal{S} = \{1, 2, ..., 10\}$ with associated values x_ω is given by $(100, 200, 100, 200, 100, 50, 50, 25, 100, 25)$ gpm.

- Sample mean is: $\hat{\mathbb{E}}_X = \sum_{x \in \hat{D}_X} x \hat{f}_X(x) = 95$
- Sample variance is: $\hat{\mathbb{V}}_X = \sum_{x \in \hat{D}_X} (x - \hat{\mathbb{E}}_X)^2 \hat{f}_X(x) = 4000$
- Empirical quantiles (assuming center values in flat portions) are:

$$\mathbb{Q}(0) = 25$$
$$\mathbb{Q}(0.5) = 100$$
$$\mathbb{Q}(0.8) = 150$$
$$\mathbb{Q}(1) = 200$$



Uncertainty Propagation and Mitigation

The uncertainty (and variability) associated with random variables propagate through systems in complex ways. Fortunately, we often have the ability to manipulate a system (through design or control) in order to mitigate the effects of this variability.

Consider the propagation of X through a system described by a function $\varphi(X, u)$, where $u \in \mathcal{U}$ is an action (decision):

$$Y = \varphi(X, u)$$

We make the following observations:

- The output Y is an RV if the input X is an RV.
- The nature of Y (its cdf, pdf, and domain) depends on the system function φ. Some systems might magnify uncertainty and variability while others might damp it.
- The nature of Y depends on the decision u. We can implement actions on the system φ to control the uncertainty and variability of Y.

Uncertainty Propagation and Mitigation

Having data x_ω, $\omega \in \mathcal{S}$ and a system φ, we can characterize the cdf, pdf, domain, and summarizing statistics of Y using the following procedure.

- For a given u, perform simulations of the form:

$$y_\omega = \varphi(x_\omega, u), \ \omega \in \mathcal{S}$$

- We use y_ω to compute approximations of quantities of interest for Y such as:
  - Sample mean:

$$\hat{\mathbb{E}}_Y = \frac{1}{S} \sum_{\omega \in \mathcal{S}} y_\omega = \frac{1}{S} \sum_{\omega \in \mathcal{S}} \varphi(x_\omega, u)$$

  - Sample variance:

$$\hat{\mathbb{V}}_Y = \frac{1}{S} \sum_{\omega \in \mathcal{S}} (y_\omega - \hat{\mathbb{E}}_Y)^2$$

  - Empirical cdf:

$$\hat{F}_Y(y) = \frac{1}{S} \sum_{\omega \in \mathcal{S}} \mathbf{1}[y_\omega \leq y]$$

  and quantiles $\hat{Q}_Y(\alpha)$.

Uncertainty Propagation and Mitigation

VZ: Mixer tank

Decision-Making under Uncertainty

Consider now that we would like to find a decision $u \in \mathcal{U}$ that controls $Y(u) = \varphi(X, u)$ in some desirable way. This gives rise to a couple of questions:

- If we have a couple of competing decisions u and u' giving rise to random outputs $Y(u)$ and $Y(u')$. How can we tell which one is better?
- How can we find the best possible decision u?

Some observations:

- If we assume a *deterministic setting* with no uncertainty, then $Y(u)$ and $Y(u')$ will each take a single value and one would select, unambigously, the one with the larger (or smaller) value. Mathematically, one would select u if $Y(u) \geq Y(u')$.
- In a *setting under uncertainty* this is no longer possible because $Y(u)$ and $Y'(u)$ have multiple possible outcomes and with different probabilities.
- The concept of "better" under uncertainty is ambiguous and the mathematical statement $Y(u) \leq Y(u')$ does not even make sense. Does $Y(u) \leq Y(u')$ mean that all the outcomes of $Y(u)$ are lower than those $Y(u')$? Does it mean that only a subset of outcomes is lower?

Decision-Making under Uncertainty

VZ: Which area is better based on rainfall?

Estimation

Given that we have data $x_\omega, y_\omega,\ \omega \in \mathcal{S}$ available, we now place our attention to the following questions:

Is the data following a particular pattern (trend)? If there is a pattern, can we model it?

In this context, by a model, we mean two things:

- If we have empirical statistics (e.g., cdf, pdf, mean, variance) obtained from data x_ω, y_ω, do these match the statistics of a *known* RV?
- If we do not know the system model φ that relates x_ω and y_ω, can we determine this by using input-output data?

The task of determining models from data is known as *estimation*.
Having a model will allow us:

- Determine if the available data is sufficient to say something meaningful about events that have not been observed (e.g., need more data to make a decision?).
- Make predictions about other possible events and their respective probabilities (e.g., how likely is an extreme event from happening?)
- Extract trends that help us summarize the data available (from data to knowledge).

Our first step is to postulate an RV model and see if this fits the data.

Model of a Gaussian RV

- A wide range of RV models have been developed over the years based on identification of common patterns that emerge in real-life phenomena.
- Many phenomena follow the behavior of a normal RV (a.k.a. Gaussian RV).

A Gaussian RV is continuous and has an associated pdf:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathcal{D}_X = \{-\infty \leq x \leq \infty\}$$

- The scalar values $\mu \in \mathbb{R}, \sigma \in \mathbb{R}$ are hyperparameters that are specific to the application of interest. These are the values that we will seek to tune to match the model to the available data.
- We express the fact that an RV X is Gaussian as $X \sim \mathcal{N}(\mu, \sigma^2)$.

Model of a Gaussian RV

- The pdf tells us the behavior of the Gaussian RV:
  - the probability of an outcome x decays exponentially fast as we move from μ
  - the outcome of maximum probability (most likely outcome) is μ
  - the speed of the decay is dictated by σ
  - the decay in probability is symmetric around μ
- Gaussian model assumes that an outcome x can take any value in $(-\infty, \infty)$. This introduces complications, as many phenomena involves variables that cannot take negative values (e.g., mass) or infinite values (e.g., temperatures).
- Gaussian RVs can model a wide range of phenomena (e.g., diffusion). Moreover, many phenomena have the Gaussian RV as a limiting case (we will show this later).

Model of a Gaussian RV

VZ: Analyze behavior of Gaussian RV

Properties of a Gaussian RV

A Gaussian RV $X \sim \mathcal{N}(\mu, \sigma^2)$ has many useful properties. For instance:

- It expected value and variance are $\mathbb{E}_X = \mu$ and $\mathbb{V}[X] = \sigma^2$.
- Any linear transformation $Y = a + bX$ yields a Gaussian RV $Y \sim \mathcal{N}(a + b\mu, b^2\sigma^2)$. This implies that $\mathbb{E}_Y = a + b\mathbb{E}_X$ and $\mathbb{V}_Y = b^2\mathbb{V}_Y$.
- The cdf of $Y = a + bX$ satisfies $F_Y(y) = F_X(x)$ for all $y = a + bx$.

Think about the implications of the above properties from an estimation and uncertainty propagation perspective:

- We can estimate μ and σ from data simply as $\mu = \hat{\mathbb{E}}_X$ and $\sigma^2 = \hat{\mathbb{V}}[X]$. This is sufficient to create our empirical Gaussian model.
- Any linear system $\varphi(X) = a + bX$ will generate a Gaussian RV as output if the input X is linear. Moreover, the system will shrink the variability of X if $b < 1$ and will magnify if $b > 1$.

Properties of a Gaussian RV

VZ: Show fit as data is accumulated and show linear transformation

Properties of a Gaussian RV

The cdf of a Gaussian RV is given by:

$$F_X(x) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{(x-\mu)/\sigma}{\sqrt{2}}\right)\right)$$

where $\operatorname{erf} : \mathbb{R} \to \mathbb{R}$ is the error function:

$$\operatorname{erf}\left(\frac{(x-\mu)/\sigma}{\sqrt{2}}\right) = \frac{2}{\sqrt{\pi}} \int_0^{\frac{(x-\mu)/\sigma}{\sqrt{2}}} e^{-t^2} dt$$

Computing the cdf involves evaluating an integral that depends on μ and σ.

Properties of a Gaussian RV

VZ: Show side by side Gaussian RV and its cdf

Properties of a Gaussian RV

- Fortunately, one can exploit properties of Gaussian RVs to avoid this issue.
- Note that $Z = (X - \mu)/\sigma$ is a linear transformation of $X \sim \mathcal{N}(\mu, \sigma^2)$ and and thus $Z \sim \mathcal{N}(0, 1)$.
- Now note that the pdf and cdf of Z are simply:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

$$F_Z\left(z\right) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right)\right), \quad \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right) = \frac{2}{\sqrt{\pi}} \int_0^{\frac{z}{\sqrt{2}}} e^{-t^2} dt$$

which do not depend on hyperparameters; Z is known as the standard normal RV.

Properties of a Gaussian RV

- One can show that $F_X(x) = F_Z(z)$ holds for any $z = (x - \mu)/\sigma$ and thus we can evaluate $F_X(x)$ at a given value x by transforming x into z and then evaluate $F_Z(z)$.
- Since $F_Z(z)$ does not depend on any parameters, it can be precomputed (values of $F_Z(z)$ are available in software packages).
- Now imagine that we want to compute $Q_X(\alpha) = F_X^{-1}(\alpha)$. As with the cdf, we can compute this by using pre-computed quantiles of Z, which we denote as $z_\alpha := F_Z^{-1}(\alpha)$.
- The relationship between the quantiles of X and Z is obtained directly from the linear transformation $x = \mu + \sigma z$:

  $$Q_X(\alpha) = \mu + \sigma z_\alpha.$$

  The values z_α are known as the critical values of the standard normal. As with the cdf, these values have been precomputed and are available in software packages.

Properties of a Gaussian RV

VZ: Show calculation of critical values

Properties of a Gaussian RV

- Standarization allows us to easily determine probability that X is in specific ranges.
- Imagine that you precomputed $F_Z(k) = \mathbb{P}(Z \leq k)$ for $k = 0, 1, 2, 3, \dots$. We have:

$$\begin{aligned}
\mathbb{P}(Z \leq 0) = 50.0\% &\iff \mathbb{P}(X \leq \mu) = 50.0\% \\
\mathbb{P}(Z \leq 1) = 84.1\% &\iff \mathbb{P}(X \leq \mu + \sigma) = 84.1\% \\
\mathbb{P}(Z \leq 2) = 97.7\% &\iff \mathbb{P}(X \leq \mu + 2\sigma) = 97.7\% \\
\mathbb{P}(Z \leq 3) = 99.9\% &\iff \mathbb{P}(X \leq \mu + 3\sigma) = 99.9\%
\end{aligned}$$

- i.e., the probability that X is below its mean μ plus one σ is always 84.1%, the probability that it is below μ plus three σ is always 99.9% and so on.

VZ: Show pdf of standard normal

Properties of a Gaussian RV

- Standarization allows us to easily determine *confidence regions* for X.
- Assume a probability level $\alpha \in [0,1]$; we can show that critical value z satisfying

$$\mathbb{P}(-z \leq Z \leq z) = 1 - \alpha$$

  is $z = F_Z^{-1}(1-\alpha/2)$ (i.e., $z_{1-\alpha/2}$).
- In other words, we have that:

$$\mathbb{P}(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha$$

- Using the linear transformation property we obtain:

$$\mathbb{P}(\mu - z_{1-\alpha/2}\sigma \leq X \leq \mu + z_{1-\alpha/2}\sigma) = 1 - \alpha.$$

- i.e.; the probability of finding $X \sim \mathcal{N}(\mu, \sigma^2)$ in region $\mu \pm z_{1-\alpha/2}\sigma$ is $1-\alpha$.
- This gives us an idea of how confident we are of finding X in a given region; conversely, we can also determine the region under which we can find X with a desired confidence.
- The concept of the confidence region is important in many topics of statistics.

Model of an Exponential RV

Another RV that often appears in applications is the exponential random variable.

An exponential RV is continuous and has a pdf of the form:

$$f_X(x) = \frac{1}{\beta} e^{-x/\beta}, \quad x \in \mathcal{D}_X = \{0 \leq x \leq \infty\}.$$

- The only hyerparameter of this model is $\beta \in \mathbb{R}_+$ (a.k.a. scale value).
- The reciprocal $\eta = 1/\beta$ is known as the intensity and thus the pdf can also be written as $f_X(x) = \eta e^{-\eta \cdot x}$.
- We express the fact that X is exponential as $X \sim \mathrm{Exp}(\beta)$.

Model of an Exponential RV

- The pdf of the exponential RV tell us that:
  - The probability of finding x away from zero decays exponentially fast at rate η
  - There is zero probability of finding X below zero (pdf is asymmetric). One can think of an exponential RV as one side of a Gaussian RV.
- The cdf is $F_X(x) = 1 - e^{-x/\beta}$ and the expected value and variance are $\mathbb{E}_X = \beta$ and $\mathbb{V}_X = \beta^2$ (i.e., β can be easily estimated from data).
- This RV is often used to model time phenomena associated with *failures*.
- For instance, X can be used to model the amount of time that we have to wait until we observe the first occurrence of an event (e.g., engine fails). In this context, we know the average time that we have to wait ($\mathbb{E}_X = \beta$) but the actual time is random (unknown).

Model of a Gamma RV

The Gamma RV is a generalization of the exponential RV that has a pdf of the form:

$$f_X(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} e^{-x/\beta} x^{\alpha-1},\ x \in \mathcal{D}_X = \{0 \leq x \leq \infty\}.$$

- The pdf has two hyperparameters $\alpha, \beta \in \mathbb{R}_+$
- $\Gamma(\alpha)$ is the gamma function (for integer $\alpha \geq 1$ we have $\Gamma(\alpha) = (\alpha - 1)!$).
- We express the fact that X is a gamma RV as $X \sim \mathrm{Gamma}(\alpha, \beta)$

Model of a Gamma RV

- The pdf tells us that one recovers an exponential RV when $\alpha = 1$ and the term $x^{\alpha-1}$ introduces a competing (opposite) effect for the exponential decay.
- The expected value and variance are $\mathbb{E}_X = \alpha\beta$ and $\mathbb{V}_X = \alpha\beta^2$ (i.e., the hyperparameters can be estimated from data by solving a set of two equations).
- In the context of time phenomena, this RV generalizes the exponential in that it models the amount of time that we have to wait until we observe the α-th occurrence of an event. Consequently, α=1 means the first event (as in the exponential RV).
- This RV has applications not only in temporal but also in spatial phenomena. For instance, it can be used to model the distance until we find the α-th occurrence of certain type of atom in a molecule.

Model of a Chi-Squared RV

The Chi-Square RV has a pdf of the form:

$$f_X(x) = \frac{1}{2^{r/2}\Gamma(r/2)} e^{-x/2} x^{r/2-1}, \; x \in \mathcal{D}_X = \{0 \leq x \leq \infty\}.$$

- The pdf has only one hyperparameter $r \in \mathbb{Z}_+$ (a.k.a degrees of freedom).
- We express the fact that X is a gamma RV as $X \sim \chi^2(r)$
- Note that this is a Gamma RV with $\beta = 2$ and $\alpha = r/2$ (for a positive integer r).
- The expected value and variance can be derived from those of the Gamma RV.
- A crucial property of a Chi-Squared RV is that it is related to the standard normal RV. In particular, one can show that:

$$\sum_{i=1}^{r} X_i^2 \sim \chi^2(r)$$

if $X_i \sim \mathcal{N}(0,1)$. This property will be useful later on.

Model of a Weibull RV

The Weibull RV is a generalization of the exponential RV that has a pdf:

$$f_X(x) = \frac{\xi}{\beta}\left(\frac{x}{\beta}\right)^{\xi-1} \exp\left[-\left(\frac{x}{\beta}\right)^{\xi}\right], \; x \in \mathcal{D}_X = \{0 \leq x \leq \infty\}.$$

- The pdf has two hyperparameters $\xi, \beta \in \mathbb{R}_+$ (a.k.a scale and shape).
- We express the fact that X is a gamma RV as $X \sim \mathrm{Weibull}(\xi, \beta)$.
- Note that one recovers an exponential RV when $\xi = 1$.
- The expected value and variance are $\mathbb{E}_X = \beta\Gamma(1 + 1/\xi)$ and $\mathbb{V}_X = \beta^2\left(\Gamma(1 + 2/\xi) + \Gamma(1 + 1/\xi)^2\right)$. The dependence on the gamma function makes it difficult to estimate ξ, β from these relationships.
- The cdf has a nice form $F_X(x) = 1 - e^{(-x/\beta)^{\xi}}$ (ξ, β are often inferred from the cdf).
- The Weibull RV is the *de facto* model used in failure analysis and was discovered by Fischer and Tippett.
- Turns out that, as in the case of the Gaussian RV, many phenomena have the Weibull RV as a limiting case (we will show this later).

Families of Random Variables

- The exponential, Gamma, Chi-Squared, and Weibull RVs are interrelated. In fact, these are captured by the generalized gamma model with pdf:

$$F_X(x) = \frac{1}{\beta^{\alpha\xi}\Gamma(\alpha)} \exp\left[-\left(\frac{x-\delta}{\beta}\right)^{\xi}\right] \xi(x-\delta)^{\alpha\xi-1}, \quad x \in \mathcal{D}_X = \{0 \leq x \leq \infty\}.$$

- These RVs are known as the Gamma family.
- There are three major families of continuous RVs:
  - Gaussian family (includes Gaussian, LogNormal, and Raylegh)
  - Gamma family (includes exponential, Gamma, Chi-Squared, and Weibull)
  - Ratio family (includes Cauchy, Uniform, Beta, Fisher, Student)
- There is one family for discrete RVs, which includes Uniform (discrete), Bernoulli, Bionomial, and Poisson.
- Each family models different type of phenomena and there exist relationships between RVs within families and across families.

A detailed discussion of the modeling properties of RVs is beyond our scope. Here, we have only discussed the RVs that will be relevant in our subsequent discussion.

Families of Random Variables

VZ: Show diagram of RV families

Estimation Techniques

- Now that we have a basic idea of the types of RV models available we proceed to develop procedures to estimate an RV model from available data.
- We will denote an RV model as $f(x, \theta)$, where θ are the hyperparameters. By estimating an RV model we mean that we seek to find θ that best fits the data.
- We will explore a couple of estimation methods:
    - Point Estimation (Method of Moments and Least-Squares Method)
    - Maximum Likelihood Estimation
- Our first step will be to explore our data and postulate a specific RV model (e.g., Gaussian or Exponential) based on any patterns exposed by the data.
- Our second step would be to tune θ for the given postulated RV model to see if this fits the data satisfactorily. If the fit is not adequate, we postulate another model.

Method of Moments

- Recall the moments of X (with pdf $f_X(x|\theta)$ and hyperprameters θ) are given by:

$$m_k(\theta) := \mathbb{E}[(X - \mathbb{E}[X])^k], \qquad k = 1, 2, ..., N$$

- Here, we highlight the dependence of the moments on the hyperparameters θ.
- Method of moments uses data $x_\omega,\ \omega \in \mathcal{S}$ to obtain sample approximations:

$$M_k = \frac{1}{S} \sum_{\omega \in \mathcal{S}} (x_\omega - m)^k,\ k = 1, 2, ..., N$$

where $m = \frac{1}{S} \sum_{\omega \in \mathcal{S}} x_\omega$ is the sample mean.

Method of Moments

- Our objective is to find the hyperparameters θ that solve the matching equations:

$$m_k(\theta) = M_k, \ k = 1, 2, ..., N$$

- In other words, we want to find θ that matches model and data moments.
- A solution to the equations can also be found by solving the minimization problem:

$$\min_{\theta} \frac{1}{2} \sum_{k=1}^{N} (m_k(\theta) - M_k)^2.$$

  This problem is known as a least-squares minimization problem (seeks to minimize the discrepancy between the model and data moments).
- Least-squares is preferred when the matching equations do not have a solution. In this case, the minimization problem finds θ that is most compatible with the data.

Method of Moments

VZ: Example Gaussian and Weibull. Show Weibull is too difficult using this approach but is trivial under Least-Squares.

Least-Squares Method

- Functional form of moments $m_k(\theta)$ might be too complex for some RVs (e.g., Weibull).
- In such cases, we can use $F_X(t|\theta)$ to find the hyperparameters. For example, for Weibull we have $F_X(t|\theta) = (1 - e^{-(t/\beta)^{\xi}})$ with $\theta = (\xi, \beta)$ and threshold value t.
- In least-squares method, we find θ that best matches the empirical cdf (obtained from data $x_\omega,\ \omega \in \mathcal{S}$).
- Here, we propose a set of threshold values $t_k, k = 1, 2, ..., N$ and compute:

$$\hat{F}_k = \frac{1}{S} \sum_{\omega \in \mathcal{S}} \mathbf{1}[x_\omega \leq t_k],\ k = 1, 2, ..., N$$

- We use these approximations to solve the least-squares minimization problem:

$$\min_{\theta} \frac{1}{2} \sum_{k=1}^{N} (F_X(t_k|\theta) - \hat{F}_k)^2$$

- Least-squares method is general and can be used to match multiple statistics of the RV available such as moments, empirical quantiles, and empirical pdf.

Least-Squares Method

VZ: Show how to do least-squares in Matlab.

Maximum Likelihood Method

The maximum likelihood estimation (MLE) method proceeds as follows:

- Assume we have observations $x_\omega,\ \omega \in \mathcal{S}$ (selected at random).
- We postulate $f_X(x|\theta)$ for the RV X and recall that $f(x_\omega|\theta)$ is the probability (likelihood) that X takes the value of observation x_ω.
- Consequently, we find hyperparameters θ that maximize *joint* probability that X takes observations $x_\omega,\ \omega \in \mathcal{S}$. This is done by solving the maximization problem:

$$\max_{\theta}\ L(\theta) = \prod_{\omega \in \mathcal{S}} f(x_\omega|\theta).$$

  Here, $L(\theta)$ is known as the likelihood function.

- It is often convenient to solve the equivalent problem:

$$\max_{\theta}\ \log L(\theta) = \sum_{\omega \in \mathcal{S}} \log f(x_\omega|\theta).$$

  This problem can be solved by hand when the pdf is simple but requires numerical techniques when complex.

- We will soon discuss what “random observations” and “joint probability” mean.

Maximum Likelihood Method

VZ: Show how to do MLE in Matlab and derivation by hand for exponential.

Sampling and Asymptotic Properties

So far, we have assumed that we have data x_ω, $\omega \in \mathcal{S}$ but we have not discussed yet how this is being collected and generated. Moreover, we want to know how our empirical approximations behave as we keep accumulating data.

- The data sample sequence $x_\omega \in \mathcal{S}$ is a set of observations of X collected from a statistical population Ω by a defined procedure; i.e., sampling is a data collection procedure.
- A data sample sequence $x_\omega \in \mathcal{S}$ is called random if each sample x_ω is drawn from the same underlying pdf $f_X(x)$ and if it is drawn independently from the others. In other words, the samples are independent and identically distributed (i.i.d).
- If a sample x_ω is selected at random, the sample itself is an RV. Consequently, sometimes we denote the data sample sequence as a sequence of RVs $X_\omega \in \mathcal{S}$.

Sampling and Asymptotic Properties

- Random sample X_ω has same probability $1/S$ of being selected and is *unbiased*.
- The lack of bias indicates that $\mathbb{E}[X_\omega] = \mathbb{E}[X]$ (drawing sample many times and averaging the results gives same expected value of the actual RV X).
- Random samples can be used to construct approximation techniques with rather striking asymptotic properties, as we will show next.
- Collecting data at random is not as easy as it sounds, one must ensure that there is no bias in selecting a sample (i.e., there is no hidden mechanism). As humans, it is strikingly difficult to pick something randomly.

Monte Carlo Approximations

Monte Carlo (MC) is a set of computational techniques that use *random samples* to infer properties of X and derived quantities (e.g., summarizing statistics). For example, we can use the random sample $x_\omega \in \mathcal{S}$ to compute the sample approximations:

- Expectation of X: $\hat{E}_X^S := \frac{1}{S}\sum_{\omega\in\mathcal{S}} x_\omega \approx \mathbb{E}[X]$
- Variance of X: $\hat{V}_X^S := \frac{1}{S}\sum_{\omega\in\mathcal{S}}(x_\omega - \hat{E}_X^S)^2 \approx \mathbb{V}[X]$
- CDF of X: $\hat{F}_X^S(x) := \frac{1}{S}\sum_{\omega\in\mathcal{S}} \mathbf{1}[x_\omega \leq x] \approx F_X(x) = \mathbb{E}[\mathbf{1}[X \leq x]]$
- Expectation of system $\hat{E}_\varphi^S := \frac{1}{S}\sum_{\omega\in\mathcal{S}} \varphi(x_\omega, u) \approx \mathbb{E}[\varphi(X, u)]$.

Natural questions that emerge here are:

- Do the approximations become exact as $S \to \infty$?
- How accurate are these approximations for finite S?

- Most approximations use an expectation function and we thus restrict our discussion to the behavior of the approximation $\hat{E}_X^S$.
- There are approximation techniques that use systematic (biased) sampling methods (data is not collected at random).

Law of Large Numbers

Lets first assess the quality of the MC approximation as $S \to \infty$.

- Consider a random sample sequence $X_1, X_2, ..., X_S$ for RV X. Since samples are identically distributed, they have same underlying pdf with expected value $\mathbb{E}[X]$.
- This implies that $\mathbb{E}[X_1] = \mathbb{E}[X_2] = \cdots = \mathbb{E}[X_S] = \mathbb{E}[X]$.
- Consider the MC approximation of $\mathbb{E}[X]$:

$$\hat{\mathbb{E}}_X^S = \frac{1}{S} \sum_{\omega \in \mathcal{S}} X_\omega$$

The law of large numbers (LLN) states that:

$$\lim_{S \to \infty} \hat{\mathbb{E}}_X^S = \mathbb{E}[X]$$

- The LLN is a fundamental result in statistics and is important because it guarantees *stable long-run* behavior of random variables.
- In other words, if the process is truly random, samples fluctuate around $\mathbb{E}[X]$ and average out. Conversely, if there is a systematic bias, the fluctuations will accumulate and the process will drift.
- The LLN implies that Monte Carlo approximations become asymptotically exact as $S \to \infty$. Importantly, the result holds for any random variable (e.g., discrete, continuous, univariate, multivariate).

Central Limit Theorem

Now lets turn our attention to the issue of assessing quality of the MC approximation for finite sample size S.

- This can be addressed by using a powerful result in statistics known as the central limit theorem (CLT).
- Consider that the random sample sequence $X_1, X_2, ..., X_S$ is i.i.d and has *known* expected value $e = \mathbb{E}[X]$ and variance $v^2 = \mathbb{V}[X]$.
- We know that X_ω are RVs and so is $\hat{\mathbb{E}}_X^S$.

The CLT will answer the following questions:

- What is the distribution of $\hat{\mathbb{E}}_X^S$? If we know it, we can say something about how much variability the MC approximation has.
- What is the distribution of $\hat{\mathbb{E}}_X^S$ as $S \to \infty$?

Central Limit Theorem

The CLT states that:

$$\lim_{S\to\infty} \hat{\mathbb{E}}_X^S \sim \mathcal{N}(e, v/\sqrt{S})$$

This result is one of the most surprising and useful results in statistics. Let's explain why:

- CLT says that, *regardless* of the underlying nature of X (e.g., Uniform, Weibull, Exponential), its sample average approximation $\hat{\mathbb{E}}_X^S$ will *always* become a Gaussian RV as S increases.
- CLT also says that the variance of the Gaussian $\hat{\mathbb{E}}_X^S \sim \mathcal{N}(e, v/\sqrt{S})$ shrinks with S. In other words, $\hat{\mathbb{E}}_X^S$ becomes more certain as S increases.
- Implication is that, since we have a cdf and pdf for $\hat{\mathbb{E}}_X^S$, we can compute all quantities of interest for it (e.g., confidence regions):

$$\mathbb{P}\left(e - z_{1-\alpha/2}v/\sqrt{S} \leq \hat{\mathbb{E}}_X^S \leq e + z_{1-\alpha/2}v/\sqrt{S}\right) = 1 - \alpha.$$

Consequently, for given S, we can know how confident we are that the approximation $\hat{\mathbb{E}}_X^S$ is in a region.

Extreme Value Distribution

- In CLT we are interested in *sample average* $\hat{\mathbb{E}}_X^S = \frac{1}{S}\sum_{\omega \in \mathcal{S}} X_\omega$.
- But what if we are interested in a different statistic? For instance, the *sample max*:

$$X_{max}^S = \max\{X_1, X_2, \cdots, X_S\}$$

- There exists a result (analogous to CLT) that characterizes pdf of X_{max}^S as $S \to \infty$. The result is known as the extreme value theorem (EVT).
- Consider, as before, a random sample sequence $X_1, X_2, ..., X_S$ for RV X.

The EVT states that:

$$\lim_{S \to \infty} X_{max}^S \sim \mathrm{GEV}(a, b, c)$$

where $GEV(a, b, c)$ is the generalized extreme value RV with hyperparameters a, b, c.

Extreme Value Distribution

The generalized extreme value RV has a pdf of the form:

$$f(s;\xi) = \begin{cases} (1+cs)^{(-1/c)-1}\exp(-(1+cs)^{-1/c}) & c \neq 0 \\ \exp(-s)\exp(-\exp(-s)) & c = 0 \end{cases}$$

and a cdf of the form:

$$F_X(x) = \begin{cases} \exp(-(1+cs)^{-1/c}) & c \neq 0 \\ \exp(-\exp(-s)) & c = 0 \end{cases}$$

where $s = (x-a)/b$ is a standarized variable.

- GEV RV is a generalization that includes Weibull (for $c < 0$), Frechet (for $c > 0$), and Gumbel (for $c = 0$) RVs.
- GEV RV is widely used in failure analysis because max operator characterizes peak (extreme) events.
- As with CLT, EVT does not depend on the underlying nature of X.

Multivariate Statistics

- So far, we have assumed that the RV X is univariate and thus has observations x_ω that are scalar values ($\mathbb{R}$).
- We have also informally mentioned the concept of independent and joint RVs in the context of estimation and sampling. What do these mean?
- Consider a multivariate RV $X = (X_1, X_2, ..., X_n)$ with observations $x_\omega = (x_{\omega,1}, x_{\omega,2}, ..., x_{\omega,n}) \in \mathbb{R}^n$; e.g., consider input-output pair $(X, \varphi(X))$ discussed in uncertainty propagation.

Questions that we are interested in answering are:

- Are there any connections between RVs? Is there a pattern that suggests they vary together? Are they independent of one another?
- If there are connected, how strong are connections?
- If there are connected, how does knowledge of one affects uncertainty of the other?
- How to analyze connections between many RVs? (e.g., n is in the hundreds)
- How to generalize results from univariate case to multivariate case?

Joint PDFs and CDFs

For simplicity, we consider bivariate RV $X = (X_1, X_2)$. The concepts presented easily generalize to higher dimensions.

- Observation $\omega \in \Omega$ of RV X generates observation pair $x_\omega = (x_{\omega,1}, x_{\omega,2})$
- We assume that the domain of X is a 2-D box:

$$\begin{aligned}
\mathcal{D}_1 &= \{-\infty \le x_1 \le \infty\} \\
\mathcal{D}_2 &= \{-\infty \le x_2 \le \infty\} \\
\mathcal{D} &= \mathcal{D}_1 \times \mathcal{D}_2 = \{-\infty \le x_1 \le \infty,\ -\infty \le x_2 \le \infty\}
\end{aligned}$$

- The *joint* pdf and cdf of multivariate RV are:

$$\begin{aligned}
f(x_1, x_2) &= \mathbb{P}(X_1 = x_1 \,\&\, X_2 = x_2),\ (x_1, x_2) \in \mathcal{D} \\
F(x_1, x_2) &= \mathbb{P}(X_1 \le x_1 \,\&\, X_2 \le x_2),\ (x_1, x_2) \in \mathcal{D}
\end{aligned}$$

- Note sign $\&$ inside the measure. For example, $f(x_1, x_2)$ is probability of event in which X_1 takes value x_1 *and* X_2 takes value x_2.
- The pdf must satisfy $f(x_1, x_2) \ge 0,\ (x_1, x_2) \in \mathcal{D}$.

Joint PDFs and CDFs

Consider a subdomain $\mathcal{A} \subseteq \mathcal{D}$ of the form:

$$\mathcal{A} = \{a_1 \leq x_1 \leq b_1 \,\&\, a_2 \leq x_2 \leq b_2\}$$

- For discrete RV we have that pdf is discontinuous and:

$$\begin{aligned}\mathbb{P}(X \in \mathcal{A}) &= \sum_{(x_1,x_2)\in\mathcal{A}} f(x_1, x_2) \\ &= \sum_{\omega\in\Omega} f(x_1, x_2)\mathbf{1}[(x_{\omega,1}, x_{\omega,2}) \in \mathcal{A}].\end{aligned}$$

- For continuous RV we have that pdf is continuous and:

$$\begin{aligned}\mathbb{P}(X \in \mathcal{A}) &= \int_{x\in\mathcal{A}} f(x)dx \\ &= \int_{a_1}^{b_1}\int_{a_2}^{b_2} f(x_1, x_2)dx_1dx_2 \\ &= \int_{a_1}^{b_1}\int_{a_2}^{b_2} dF(x_1, x_2)\end{aligned}$$

The last expression implies that $\frac{dF(x_1,x_2)}{dx_1dx_2} = f(x_1, x_2)$.

Joint PDFs

The joint pdf $f(x_1, x_2)$ tells us probability that (X_1, X_2) takes value (x_1, x_2).

- Imagine now we want to know probability that X_1 takes x_1 *given knowledge* that X_2 takes value x_2 (or other way around).
- These probabilities are obtained from *conditional pdfs*:

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)},\ x_1 \in \mathcal{D}_1$$

$$f(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)},\ x_2 \in \mathcal{D}_2$$

- Conditionals represent $f(x_1|x_2) = \mathbb{P}(X_1 = x_1|X_2 = x_2)$ and $f(x_2|x_1) = \mathbb{P}(X_2 = x_2|X_1 = x_1)$.
- These expressions can also be written as:

$$f_2(x_2)f(x_1|x_2) = f(x_1, x_2),\ x_1 \in \mathcal{D}_1$$

$$f_1(x_1)f(x_2|x_1) = f(x_1, x_2),\ x_2 \in \mathcal{D}_2$$

- As an example, consider we want $\mathbb{P}(a_1 \leq X_1 \leq b_1|X_2 = x_2)$; we have that:

$$\mathbb{P}(a_1 \leq X_1 \leq b_1|X_2 = x_2) = \int_{a_1}^{b_1} f(x_1|x_2)dx_1$$

- Joint pdfs have associated marginal cdfs $F(x_1|x_2)$ and $F(x_2|x_1)$.

Marginal PDFs

- Imagine now we want to know probability that X_1 takes x_1 *regardless* of what value X_2 takes (or other way around).
- These probabilities are obtained from marginal pdfs. For a continuous RV:

$$f_1(x_1) = \int_{x_2 \in \mathcal{D}_2} f(x_1, x_2) dx_2, \ x_1 \in \mathcal{D}_1$$
$$f_2(x_2) = \int_{x_1 \in \mathcal{D}_1} f(x_1, x_2) dx_1, \ x_2 \in \mathcal{D}_2$$

  i.e., marginal pdfs integrate out effect of RV we ignore (for discrete RV we sum out)
- Marginals represent:

$$f_1(x_1) = \mathbb{P}(X_1 = x_1 | X_2 \in \mathcal{D}_2) = \mathbb{P}(X_1 = x_1)$$
$$f_2(x_2) = \mathbb{P}(X_2 = x_2 | X_1 \in \mathcal{D}_1) = \mathbb{P}(X_2 = x_2).$$

- As an example, consider we want $\mathbb{P}(a_1 \leq X_1 \leq b_1)$; we have that:

$$\mathbb{P}(a_1 \leq X_1 \leq b_1) = \int_{a_1}^{b_1} \int_{x_2 \in \mathcal{D}_2} f(x_1, x_2) dx_2 dx_1 = \int_{a_1}^{b_1} f_1(x_1) dx_1$$

- Marginal pdfs have associated marginal cdfs $F_1(x_1)$ and $F_2(x_2)$.

Models of Multivariate Random Variables

So the conditional pdfs are telling us know knowledge in one RV affects the uncertainty in another (i.e., how much knowledge of one is embedded in the other one).

So how about that knowledge of one does not affect uncertainty of the other?

This gives rise to the concept of *independence*.

- RVs X_1 and X_2 are said to be independent if:

$$f(x_1|x_2) = f_1(x_1),\ x_1 \in \mathcal{D}_1$$
$$f(x_2|x_1) = f_2(x_2),\ x_2 \in \mathcal{D}_2$$

- This implies that:

$$f(x_1, x_2) = f_2(x_2)f_1(x_1),\ (x_1, x_2) \in \mathcal{D}$$

- Equivalently:

$$\mathbb{P}(X_1 = x_1 \,\&\, X_2 = x_2) = \mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2)$$

Summarizing Statistics for Multivariate RVs

Computing summarizing statistics in multivariate case is similar to univariate case but there are a few key differences that we highlight:

- Joint expectation of $X = (X_1, X_2)$ is a vector $\mathbb{E}[X] = (\mathbb{E}[X_1], \mathbb{E}[X_2])$ with:

$$\mathbb{E}[X_1] = \int_{x_1 \in \mathcal{D}_1} x_1 f_1(x_1) dx_1, \qquad \mathbb{E}[X_2] = \int_{x_2 \in \mathcal{D}_2} x_2 f_2(x_2) dx_2.$$

- Joint expectation of $\varphi(X) = \varphi(X_1, X_2)$ is a scalar:

$$\mathbb{E}[\varphi(X)] = \int_{x_1 \in \mathcal{D}_1} \int_{x_2 \in \mathcal{D}_2} \varphi(x_1, x_2) f(x_1, x_2) dx_1 dx_2$$

- Conditional expectation of X_1 (given knowledge $X_2 = x_2$) is:

$$\mathbb{E}[X_1 | X_2 = x_2] = \int_{x_1 \in \mathcal{D}_1} x_1 f(x_1 | x_2) dx_1$$

- Conditional expectation of $\varphi(X) = \varphi(X_1, X_2)$ (given knowledge $X_2 = x_2$) is:

$$\mathbb{E}[\varphi(X) | X_2 = x_2] = \int_{x_1 \in \mathcal{D}_1} \varphi(x_1, x_2) f(x_1 | x_2) dx_1$$

- Expressions for discrete RVs are analogous (replace integrals for sums).

Summarizing Statistics for Multivariate RVs

In multivariate case, the concepts of *covariance* and *correlation* emerge:

- Define the marginal expectations and variances:

$$\begin{aligned}\mu_1 &= \mathbb{E}[X_1] \\ \mu_2 &= \mathbb{E}[X_2] \\ \sigma_1^2 &= \mathbb{V}[X_1] = \mathbb{E}[(X_1 - \mu_1)^2] \\ \sigma_2^2 &= \mathbb{V}[X_2] = \mathbb{E}[(X_2 - \mu_2)^2].\end{aligned}$$

- Covariance between X_1 and X_2 is:

$$\begin{aligned}\mathrm{Cov}(X_1, X_2) &= \mathbb{E}[(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= \int_{x_1 \in \mathcal{D}_1} \int_{x_2 \in \mathcal{D}_2} (X_1 - \mu_1)(X_2 - \mu_2) f(x_1, x_2) dx_1 dx_2.\end{aligned}$$

- Define $\sigma_{i,j} = \mathrm{Cov}(X_i, X_j)$ and note $\sigma_{2,1} = \sigma_{1,2}$, $\mathrm{Cov}(X_1, X_1) = \sigma_1^2$ and $\mathrm{Cov}(X_2, X_2) = \sigma_2^2$.
- Correlation between X_1 and X_2 is:

$$\mathrm{Corr}(X_1, X_2) = \frac{\sigma_{1,2}}{\sigma_1 \sigma_2}$$

- Define $\rho_{i,j} = \mathrm{Corr}(X_i, X_j)$ and note $\rho_{1,2} \in [-1, 1]$, $\rho_{2,1} = \rho_{1,2}$, $\rho_{1,1} = \rho_{2,2} = 1$.

Models of Multivariate Random Variables

The covariance and correlation tells us how, on average, X_1 varies with X_2:

- If $\sigma_{1,2} > 0$ $(\rho_{1,2} > 0)$ we have *positive correlation*: The event $X_1 > \mu_1$ is, on average, associated with $X_2 > \mu_2$ and event $X_1 < \mu_1$ is, on average, associated with $X_2 < \mu_2$. In other words, X_1 and X_2 move in the same direction (on average).
- If $\sigma_{1,2} < 0$ $(\rho_{1,2} < 0)$we have negative correlation: The event $X_1 > \mu_1$ is, on average, associated with $X_2 < \mu_2$ and event $X_1 < \mu_1$ is, on average, associated with $X_2 > \mu_2$. In other words, X_1 and X_2 move in opposite directions (on average).
- If $\sigma_{1,2} = 0$ $(\rho_{1,2} = 0)$ we have no correlation: On average, changes in $(X_1 - \mu_1)$ are not associated with changes in $(X_2 - \mu_2)$. If X_1 andX_2 are independent then $\sigma_{1,2} = 0$. However, $\sigma_{1,2} = 0$ does not imply independence because variables might move together (but not on average).

Presence of correlation reveals emergent trends. For instance, if X_1 and X_2 are related as $X_2 = \alpha X_1$ then $\mathrm{Cov}(X_1, X_2) = \alpha \mathbb{V}[X_1]$.

Models of Multivariate Random Variables

- Covariance between variables is often expressed in matrix form as:

$$\mathrm{Cov}[X] = \begin{bmatrix} \sigma_{1,1}^2 & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_{2,2}^2 \end{bmatrix}$$

  this matrix is symmetric because $\sigma_{1,2} = \sigma_{2,1}$ and has positive eigenvalues (it is positive definite).

- Covariance matrix (for any dimension n) can be computed as:

$$\mathrm{Cov}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T\right]$$

- Correlation between variables is often expressed in matrix form as:

$$\mathrm{Corr}[X] = \begin{bmatrix} 1 & \rho_{1,2} \\ \rho_{2,1} & 1 \end{bmatrix}$$

  this matrix is symmetric because $\rho_{1,2} = \rho_{2,1}$ and is positive definite.

- Correlation matrix (for any dimension n) can be computed as:

$$\mathrm{Corr}[X] = D^{-1}\mathrm{Cov}(X)D^{-1}$$

  where $D = \mathrm{diag}(\mathrm{Cov}[X])$.

Multivariate Gaussian Variables

Surprisingly enough, there are actually few established models for multivariate RVs. The most used model is that of the Gaussian RV, which has a wide range of properties:

- Consider a multivariate RV vector $X = (X_1, X_2, ..., X_n)$.
- Denote as $X \sim \mathcal{N}(\mu, \Sigma)$, where $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$ are hyperparameters.
- RV $X \sim \mathcal{N}(\mu, \Sigma)$ has joint pdf of the form:

$$f_X(x) = f_X(x_1, x_2, ..., x_n) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

- $|\Sigma|$ is determinant of Σ (product of its eigenvalues) and Σ is positive definite.
- Domain of X is $\mathcal{D} = [-\infty, \infty]^n$.
- Hyperparameters are given by expected value $\mu = \mathbb{E}[X]$ and covariance $\Sigma = \mathrm{Cov}[X]$.

Properties of Multivariate Gaussian Variables

Consider the case with $n = 2$:

- Can show that marginal pdfs of X are Gaussian:

$$f_1(x_1) = \int_{x_2 \in \mathbb{D}_2} f(x_1, x_2) dx_2 = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(\frac{-(x-\mu_1)^2}{2\sigma_1^2}\right)$$
$$f_2(x_2) = \int_{x_1 \in \mathbb{D}_1} f(x_1, x_2) dx_1 = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(\frac{-(x-\mu_2)^2}{2\sigma_2^2}\right)$$

- In other words, $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$.

Properties of Multivariate Gaussian Variables

- Can show that conditional pdfs of X are Gaussian:

$$f(x_1|x_2) = \frac{1}{\sqrt{2\pi\sigma^2_{1|2}}} \exp\left(\frac{-(x-\mu_{1|2})^2}{2\sigma^2_{1|2}}\right)$$
$$f(x_2|x_1) = \frac{1}{\sqrt{2\pi\sigma^2_{2|1}}} \exp\left(\frac{-(x-\mu_{2|1})^2}{2\sigma^2_{2|1}}\right).$$

- That is, $X_1|X_2 \sim \mathcal{N}(\mu_{1|2}, \sigma_{1|2})$ and $X_2|X_1 \sim \mathcal{N}(\mu_{2|1}, \sigma_{2|1})$ with hyperparameters:

$$\mu_{1|2} = \mu_1 + \sigma_{1,2}\sigma_{22}^{-1}(x_2 - \mu_2)$$
$$\sigma_{1|2} = \sigma_{1,1} - \sigma_{1,2}\sigma_{2,2}^{-1}\sigma_{2,1}$$
$$\mu_{2|1} = \mu_2 + \sigma_{2,1}\sigma_{11}^{-1}(x_1 - \mu_1)$$
$$\sigma_{2|1} = \sigma_{2,2} - \sigma_{2,1}\sigma_{1,1}^{-1}\sigma_{1,2}$$

Properties of Multivariate Gaussian Variables

- Linear transformation of a multivariate Gaussian is Gaussian:
  - If $X \sim \mathcal{N}(\mu, \Sigma)$, then $Y = AX + b$ is $Y \sim \mathcal{N}(A\mu + b, A\Sigma A^T)$.
- Linear transformation property is used to establish useful properties:
  - Mixture model:
    If $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, then $Y = \sum_{i=1}^n X_i$ is Gaussian with $Y \sim \mathcal{N}(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$.
  - Multivariate standard normal:
    If $X \sim \mathcal{N}(0, I)$ then $Y = \sqrt{\Sigma}X + \mu$ is Gaussian with $Y \sim \mathcal{N}(\mu, \Sigma)$.
- Standarization result used to generate samples of $\mathcal{N}(\mu, \Sigma)$ from samples of $\mathcal{N}(0, I)$.

Geometry of Multivariate Gaussian Variables

Understanding geometry of multivariate Gaussian facilitates *data visualization*.

- For $n = 2$ we write joint pdf of $X = (X_1, X_2)$ as:

$$f_X(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-U(x_1, x_2)\right)$$

  where

$$U(x_1, x_2) = \frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right]$$

- If we fix probability $f_X(x_1, x_2) = p$ with $\alpha \in [0, 1]$, pdf defines an equation in variables (x_1, x_2). This is equation of ellipse centered at μ_1 and μ_2. This ellipse is known as the p-level set of pdf.
- Correlation coefficient $\rho = \text{Corr}(X_1, X_2)$ dictates *orientation of ellipse*: if $\rho > 0$ this is tilted to right. if $\rho < 0$ this is tilted to left, and if $\rho = 0$ (X_1 and X_2 are independent) ellipse has no tilt.
- *Length of axes* are dictated by σ_1^2 and σ_2^2 (variances of X_1 and X_2).
- Maximum value of $f_X(x_1, x_2)$ is achieved at $x_1 = \mu_1$ and $x_2 = \mu_2$.

Geometry of Multivariate Gaussian Variables

- We are interested in box confidence regions $\mathcal{B}(\alpha)$ satisfying:

$$\mathbb{P}(X \in \mathcal{B}(\alpha)) = \alpha.$$

- For univariate $X \sim \mathcal{N}(\mu, \sigma^2)$, box region is:

$$\mathcal{B}(\alpha) = \{x \,|\, x \in [\mu \pm \sqrt{\mathbb{Q}(\alpha)}\sigma]\}$$

  where $\mathbb{Q}(\alpha)$ is the α-quantile of $\chi^2(1)$ (one can show that this quantile corresponds to quantile $\mathbb{Q}(1-\alpha/2)$ of $\mathcal{N}(0,1)$).

- For multivariate $X = (X_1, X_2)$ with $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ we can build a box of the form:

$$\mathcal{B}(\alpha) = \{(x_1, x_2) \,|\, x_1 \in [\mu_1 \pm \sqrt{\mathbb{Q}(\alpha)}\sigma_1] \,\&\, x_2 \in [\mu_2 \pm \sqrt{\mathbb{Q}(\alpha)}\sigma_2]\}.$$

- This box (a.k.a. marginal box) does not capture correlations in X_1 and X_2.

Geometry of Multivariate Gaussian Variables

- In multivariate Gaussians, observations concentrate in ellipses (p-level sets). We are thus interested in finding *ellipsoidal* confidence region $\mathcal{E}(\alpha)$ satisfying:

$$\mathbb{P}(X \in \mathcal{E}(\alpha)) = \alpha.$$

- For $X \sim \mathcal{N}(\mu, \Sigma)$, the ellipsoidal region is given by:

$$\mathcal{E}(\alpha) = \{x \,|\, (x-\mu)^T \Sigma^{-1} (x-\mu)^T \leq \mathbb{Q}(\alpha)\}.$$

  where $\mathbb{Q}(\alpha)$ is the α-quantile of $\chi^2(n)$.

- Interpretation of region is:
  - If draw a sample from $\mathcal{N}(\mu, \Sigma)$, there is probability α that it will land in $\mathcal{E}(\alpha)$
  - The larger the α, the larger the ellipsoid (more likely it is to land in $\mathcal{E}(\alpha)$)
- Tighest box that encloses $\mathcal{E}(\alpha)$ is:

$$\mathcal{B}(\alpha) = \{(x_1, x_2) \,|\, x_1 \in [\mu_1 \pm \sqrt{\mathbb{Q}(\alpha)}\sigma_1] \,\&\, x_2 \in [\mu_2 \pm \sqrt{\mathbb{Q}(\alpha)}\sigma_2]\}.$$

  where $\mathbb{Q}(\alpha)$ is α-quantile of $\chi^2(n)$ (note difference with the marginal box).

- Ellipsoidal and enclosing box capture correlations in X_1 and X_2.

Data-Driven Modeling

When X, Y are correlated, variations in Y align with those in X (there is a *trend*).

What if connection is deep (e.g., variations in Y are *caused* by variations in X)?

- Consider univariate RVs Y and X and postulate that any behavior in Y is due to a systematic dependence of X:

$$Y = \theta X$$

  Here, θ is a parameter that captures a *linear* dependence between X and Y .
- Use random samples pairs (y_ω, x_ω) and, based on postulated model, we assume they are related as:

$$y_\omega = \theta x_\omega + \epsilon_\omega,\ \omega \in \mathcal{S}.$$

- We introduce hidden RV $\epsilon_\omega \in \mathcal{N}(0, \sigma^2)$ with known σ^2 to capture behavior that cannot be explained by model (the unknown).
- Note y_ω is an RV because ϵ_ω is an RV. Moreover, y_ω is linear transformation of ϵ_ω and thus y_ω is Gaussian with $\mathbb{E}[y_\omega | x_\omega, \theta] = \theta x_\omega$ and $\mathbb{V}[y_\omega | x_\omega, \theta] = \sigma^2$.

Data-Driven Modeling

- Recall $f(y_\omega|x_\omega,\theta)$ is probability that $Y=y_\omega$ given that we know $X=x_\omega$ and θ. This is equivalent to assume that θ and x_ω are *deterministic* (more on this later).
- We use a maximum likelihood approach and seek to find θ that maximizes joint likelihood $\prod_{\omega\in\mathcal{S}} f(y_\omega|x_\omega,\theta)$. This gives:

$$\max_\theta \; \log L(\theta) = \sum_{\omega\in\mathcal{S}} \log f(y_\omega|x_\omega,\theta)$$

- Since $y_\omega \sim \mathcal{N}(\theta x_\omega,\sigma^2)$, we know that:

$$\log f(y_\omega|x_\omega,\theta) = -\log\sqrt{2\pi\sigma^2} - \frac{(y_\omega-\theta x_\omega)^2}{2\sigma^2}$$

- Terms $\log\sqrt{2\pi\sigma^2}$ and $2\sigma^2$ are constants and we thus obtain:

$$\min_\theta \frac{1}{2}\sum_{\omega\in\mathcal{S}}(y_\omega-\theta x_\omega)^2$$

This is a least-squares problem and aims to find the estimate θ that minimizes discrepancy between the observed output y_ω and model prediction θx_θ.

Data-Driven Modeling

We denote best parameter that is *learned* from data as $\hat{\theta}$.

- Recall $\hat{\theta}$ minimizes $S(\theta) = \frac{1}{2}\sum_{\omega\in\mathcal{S}}(y_\omega - \theta x_\omega)^2$ and thus satisfies:

$$\frac{\partial S(\hat{\theta})}{\partial \theta} = -\sum_{\omega\in\mathcal{S}} x_\omega(y_\omega - \theta x_\omega) = 0, \qquad \frac{\partial^2 S(\hat{\theta})}{\partial \theta^2} > 0$$

  These conditions lead to:

$$\hat{\theta} = \frac{\sum_{\omega\in\mathcal{S}} x_\omega y_\omega}{\sum_{\omega\in\mathcal{S}} x_\omega^2}, \qquad \sum_{\omega\in\mathcal{S}} x_\omega^2 > 0$$

- Estimate $\hat{\theta}$ captures *interactions* in data x_ω, y_ω.
- Estimate $\hat{\theta}$ is *unique* and becomes better defined as we add more data.
- Recall $y_\omega \sim \mathcal{N}(\hat{\theta}x_\omega, \sigma^2)$, implying that prediction $\hat{\theta}x_\omega$ is the most likely outcome and that the larger σ^2, the more uncertainty we have in y_ω.
- Estimate gives *residual noise estimates* $\hat{\epsilon}_\omega = y_\omega - \hat{\theta}x_\omega$. If these estimates follow our assumption $\mathcal{N}(0, \sigma^2)$ then the available data and postulated model is satisfactory. If not, more data or another model is needed (e.g., nonlinear).
- From $y_\omega = \hat{\theta}x_\omega + \epsilon_\omega$ we note that model $\hat{\theta}x_\omega$ represents what we know about y_ω while ϵ_ω represents the unknown. Estimation problem thus seeks to extract maximum knowledge from data (minimize the unknown).

Data-Driven Modeling

Now generalize linear model to account for influence of multiple inputs. We postulate:

$$Y = \theta_0 + \sum_{i=1}^{n} \theta_i X_i$$

Use samples (data) pairs (y_ω, x_ω) and, based on postulated model, we have that:

$$y_\omega = \theta_0 + \sum_{i=1}^{n} \theta_i x_{i,\omega} + \epsilon_\omega,\ \omega \in \mathcal{S}.$$

These set of equations can be expressed compactly using matrix notation:

$$\mathbf{y} = \mathbf{X}\theta + \epsilon$$

where:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_S \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,n} \\ \vdots & & & \vdots & \\ 1 & x_{S,1} & x_{S,2} & \dots & x_{S,n} \end{bmatrix} \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_S \end{bmatrix}$$

We assume unknown noise is $\epsilon \sim \mathcal{N}(0, \Sigma)$ with $\Sigma_{\omega,\omega} = \sigma^2$ for $\omega \in \mathcal{S}$.

Data-Driven Modeling

Best estimate $\hat{\theta}$ is found as the solution of max likelihood problem:

$$\min_{\theta} \quad \frac{1}{2}(\mathbf{y}-\mathbf{X}\theta)^T(\mathbf{y}-\mathbf{X}\theta)$$

Problem can also be written as:

$$\min_{\theta} \quad \frac{1}{2}\|\mathbf{y}-\mathbf{X}\theta\|^2$$

Solution of this problem must satisfy:

$$\frac{\partial S(\theta)}{\partial \theta} = -\mathbf{X}^T(\mathbf{y}-\mathbf{X}\theta) = 0$$
$$\left|\frac{\partial^2 S(\theta)}{\partial \theta^2}\right| = \left|\mathbf{X}^T\mathbf{X}\right| > 0$$

First condition yields:

$$\hat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Second condition indicates that $\hat{\theta}$ is unique if matrix $\mathbf{X}^T\mathbf{X}$ is positive definite.

Now note that $\mathbf{y} = \mathbf{X}\theta + \epsilon$ is an RV (θ is true parameter). Consequently:

$$\begin{aligned}\hat{\theta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\theta + \epsilon)\\ &= (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})\theta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\\ &= \theta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\epsilon\end{aligned}$$

Estimate $\hat{\theta}$ is thus an RV (a linear transformation of noise $\epsilon \sim \mathcal{N}(0, \Sigma)$) and thus:

$$\hat{\theta} \sim \mathcal{N}(\theta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$

Some observations:

- Expected value of estimate is $\mathbb{E}[\hat{\theta}] = \theta$ (estimate is unbiased).
- Covariance of estimate is $\mathrm{Cov}[\theta] = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$. As variance σ^2 increase so does variance of θ (covariance is sensitivity of estimates to noise).

Data-Driven Modeling

How much of the variability in the data $\mathbf{y}$ can be explained by our gained knowledge (model $\hat{\mathbf{y}} = \mathbf{X}^T\hat{\theta}$) and how much of it cannot be explained (unknown ϵ)?

This can be addressed using *analysis of variance* (ANOVA). Total variability of data is:

$$S_y = \sum_{\omega \in \mathcal{S}} (y_\omega - \bar{y})^2 \quad \text{with} \quad \bar{y} = \frac{1}{S} \sum_{\omega \in \mathcal{S}} y_\omega$$

Total variability can be decomposed into contributions as:

$$S_y = \underbrace{\sum_{\omega \in \mathcal{S}} (\hat{y}_\omega - \bar{y})^2}_{S_m} + \underbrace{\sum_{\omega \in \mathcal{S}} (y_\omega - \hat{y}_\omega)^2}_{S_e}$$

Here, S_m is known as model sum of squares and S_e is known as sum of squared errors. Based on these quantities we define index:

$$R^2 = \frac{S_m}{S_y} = 1 - \frac{S_e}{S_y}$$

This represents faction of variability captured by model. The fraction that is left unexplained is $1 - R^2 = S_e/S_y$. Consequently, $R^2 \to 1$ indicates model fully explains data variability.

Data-Driven Modeling

How confident are we in estimate $\hat{\theta}$ and in model predictions $\hat{\mathbf{y}} = \mathbf{X}\hat{\theta}$?

- We have established that $\hat{\theta} \sim \mathcal{N}(\theta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$. We thus have that true θ lies in region $\mathcal{E}(\alpha)$ defined by:

$$(\hat{\theta} - \theta)^T \left(\frac{\mathbf{X}^T\mathbf{X}}{\sigma^2}\right)(\hat{\theta} - \theta) \leq \mathbb{Q}(\alpha)$$

  where $\mathbb{Q}(\alpha)$ is α-quantile of $\chi^2(n+1)$.
- In most scientific literature, confidence in estimates is reported using marginals:

$$\hat{\theta}_i \sim \mathcal{N}(\theta_i, \sigma^2(\mathbf{X}\mathbf{X})^{-1}_{ii}),\ i = 0, ..., n$$

  as:

$$\theta_i = \hat{\theta}_i \pm m_i \qquad m_i = \sqrt{\mathbb{Q}(\alpha)\sigma^2(\mathbf{X}\mathbf{X})^{-1}_{ii}}$$

  where $\mathbb{Q}(\alpha)$ is α-quantile of $\chi^2(1)$. This disregards correlations in parameters.
- We can construct confidence intervals for model predictions by noticing that $\hat{\mathbf{y}} = \mathbf{X}\hat{\theta}$ and $\hat{\theta} \sim \mathcal{N}(\theta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$ and thus:

$$\hat{\mathbf{y}} \sim \mathcal{N}(\mathbf{y}, \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2)$$

Data-Driven Modeling

Is the available data sufficient (or too much) to construct model?

Some observations on having sufficient data:

- Matrix $\mathbf{X}^T\mathbf{X}$ plays a fundamental role as it contains all input data and determines sharpness of minimum and variance of estimate $\hat{\theta}$.
    - If one or more eigenvalues of $\mathbf{X}^T\mathbf{X}$ are large, $\hat{\theta}$ is well-defined by the data and its variance is small. This manifests as a sharp minimum $S(\hat{\theta})$.
    - If one or more eigenvalues of $\mathbf{X}^T\mathbf{X}$ are close to zero, $\hat{\theta}$ is ill-defined by the data and variance is large. This manifests as a flat minimum $S(\hat{\theta})$.
    - If one eigenvalue of $\mathbf{X}^T\mathbf{X}$ is zero, $\hat{\theta}$ cannot be resolve uniquely from data.
- Volume of data is not sufficient, we also require *quality of data*.
    - Observations do not provide information if redundant ($\mathbf{X}^T\mathbf{X}$ has dependent rows).
    - If selected inputs X do not explain output $\mathbf{y}$ estimates $\hat{\theta}$ might exhibit high variability (regardless of number of observations).
    - Using knowledge of application to select input variables is important.
- Selection of input variables and observations is a topic of *design of experiments*.
- Inputs X are a.k.a. *regressor variables*, *explanatory variables*, *features*, or *descriptors*.

Data-Driven Modeling

Is the available data sufficient (or too much) to construct model?

Some observations on using excessive data:

- A common issue is that we use too many inputs X to explain output Y. This can result in a large number of parameters and *overfitting*.
- Check that $\hat{\theta}$ obtained with observations $(y_\omega, x_\omega),\ \omega \in \mathcal{S}$ predicts well in an independent set of observations $(y_\omega, x_\omega),\ \omega \in \mathcal{T}$. This procedure is called *cross-validation* or out-of-sample testing.
- Cross validation will ensure that model is *generalizable*.
- In linear models, and adjusted R^2 index is used to account for number of parameters:

  $$R^2_{adj} = 1 - \frac{S_e}{S_y}\frac{(S-1)}{(S-n)}$$

  As number of parameters n increases, we have that R^2_{adj} decreases.
- Note that the size of confidence ellipsoid $\mathcal{E}(\theta)$ depends on n.
- A strategy to deal with many parameters is *regularization* (will be covered later). Regularization seeks to embed *prior* knowledge in estimation problem.

Data-Driven Modeling (Nonlinear)

Now generalize the linear model to a general model of the form:

$$y_\omega = g(\theta, x_\omega) + \epsilon_\omega,\ \omega \in \mathcal{S}.$$

where $g : \mathbb{R}^n \times \mathbb{R}^S \to \mathbb{R}$ is a model function of the parameters and inputs.

- The model function can be nonlinear and capture mechanistic relationships between the inputs, parameters, and outputs.
- In linear case we have $g(\theta, x_\omega) = \theta_0 + \sum_{i=1}^n \theta_i x_{i,\omega} + \epsilon_\omega$.
- Use an MLE framework to estimate θ:

  $$\min_\theta \ S(\theta) = \frac{1}{2} \sum_{\omega \in \mathcal{S}} (y_\omega - m_\omega(\theta))$$

  where $m_\omega(\theta) = g(\theta, x_\omega)$. Problem can be expressed in vector form:

  $$\min_\theta \ \frac{1}{2} \|\mathbf{y} - \mathbf{m}(\theta)\|^2$$

- Solution $\hat{\theta}$ satisfies following set of n nonlinear equations (a.k.a. score functions):

$$\nabla_\theta S(\theta) = 0 \qquad \Longleftrightarrow \qquad \nabla_\theta m(\theta)^T (\mathbf{y} - \mathbf{m}(\theta)) = 0$$

Nonlinear Estimation

- Solution $\hat{\theta}$ also satisfies $|\mathbf{H}(\theta)| > 0$ where:

$$\mathbf{H}(\theta) = \frac{\partial^2 S(\theta)}{\partial \theta^2} = \begin{bmatrix} \frac{\partial^2 S(\theta)}{\partial\theta_1 \partial\theta_1} & \frac{\partial^2 S(\theta)}{\partial\theta_1 \partial\theta_2} & \cdots & \frac{\partial^2 S(\theta)}{\partial\theta_1 \partial\theta_n} \\ \frac{\partial^2 S(\theta)}{\partial\theta_1 \partial\theta_2} & \frac{\partial^2 S(\theta)}{\partial\theta_2 \partial\theta_2} & \cdots & \frac{\partial^2 S(\theta)}{\partial\theta_n \partial\theta_2} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 S(\theta)}{\partial\theta_1 \partial\theta_n} & \frac{\partial^2 S(\theta)}{\partial\theta_2 \partial\theta_n} & \cdots & \frac{\partial^2 S(\theta)}{\partial\theta_n \partial\theta_n} \end{bmatrix}$$

  This matrix is known as the *Hessian* matrix.

- For linear models the Hessian is:

$$\mathbf{H}(\theta) = X^T X$$

  because $\nabla_\theta m(\theta) = X$.

- What is different about nonlinear estimation?
  - Difficult to obtain pdf of $\hat{\theta}$. Typically, this is approximated as:

$$\hat{\theta} \sim \mathcal{N}(\theta, \mathbf{H}(\hat{\theta})^{-1}\sigma^2)$$

    The approximation is accurate if nonlinearity of model $m_\omega(\theta)$ is not too strong.
  - Function $S(\theta)$ might have multiple points satisfying optimality conditions.
  - Problems are often solved using local search algorithms (based on Newton's method) and thus the initial guess of $\hat{\theta}$ influences estimate found. One can also resort to using global search algorithms.

Prior Knowledge

How can we incorporate prior (expert, physical) knowledge about parameters in the estimation problem?

- Prior knowledge help us eliminate spurious estimates $\hat{\theta}$ (e.g., avoid estimates with no physical meaning or large parameters).
- Prior knowledge help us reduce number of meaningful parameters or narrow space over we search on.

We can incorporate prior knowledge in estimation problem (a.k.a. regularization) by using:

- Bounds on parameters (e.g., kinetic parameters are positive):

$$\begin{aligned}&\min_{\theta} \frac{1}{2}\|\mathbf{y}-\mathbf{m}(\theta)\|^2\\ &\text{s.t. } \theta_L \leq \theta \leq \theta_U\end{aligned}$$

- Constraints to fix parameters (e.g., sum of parameters must be equal to some value):

$$\begin{aligned}&\min_{\theta} \frac{1}{2}\|\mathbf{y}-\mathbf{m}(\theta)\|^2\\ &\text{s.t. } \Pi\theta = r\end{aligned}$$

Prior Knowledge

- Penalty term to control parameter behavior (e.g., penalize movements away from reference value):

$$\min_{\theta} \frac{1}{2}\|\mathbf{y}-\mathbf{m}(\theta)\|^2 + \kappa \cdot \rho(\theta)$$

  Where $\kappa \geq 0$ is constant that trades-off fit and allowed movement.

- Common choices for penalty function $\rho(\theta)$ are:
  - ℓ-2 norm (a.k.a ridge or Tikhonov penalty): $\rho(\theta) = \frac{1}{2}\|\theta - \bar{\theta}\|_2^2 = \frac{1}{2}\sum_{i=1}^{m}(\theta_i - \bar{\theta}_i)^2$
  - ℓ-1 norm (a.k.a. lasso penalty): $\rho(\theta) = \|\theta - \bar{\theta}\|_1 = \sum_{i=1}^{m}|\theta_i - \bar{\theta}_i|$
  - Bayes penalty: $\rho(\theta) = \frac{1}{2}(\theta - \bar{\theta})^T \Sigma_{\theta}^{-1}(\theta - \bar{\theta})$
- Penalties (a.k.a. regularizers) induce different behavior and some of them can be derived from statistical principles. There are many regularizers in the literature (seeking to embed different type of knowledge).
- Statistics help us to understand the type of *prior knowledge* that regularizers convey.
- Understanding statistical principles of constraints is a more difficult question but one can often reformulate constraints as penalty functions.

Bayesian Estimation

Bayes theorem provides a solid statistical basis to derive a wide range of estimation formulations (e.g., that include prior knowledge).

- In the context of our estimation problem of interest, Bayes theorem states that:

$$f(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)f(\theta)}{f(\mathbf{y})}$$

- $f(\theta|\mathbf{y})$ is probability that parameters take value θ given knowledge that output takes value $\mathbf{y}$ (a.k.a. posterior pdf)
- $f(\mathbf{y}|\theta)$ is probability that output takes value $\mathbf{y}$ given knowledge that parameters take value θ
- $f(\theta)$ is marginal probability of parameters (a.k.a prior pdf)
- $f(y)$ is marginal probability of outputs (this is irrelevant as it does not carry knowledge on parameters)

Bayesian Estimation

- Goal in Bayesian estimation is to maximize probability of parameters $f(\theta|\mathbf{y})$ (not of outputs as in MLE). Bayes theorem tells us that:

$$f(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)f(\theta)$$

- This approach carries prior knowledge of θ. Recall that in MLE we find estimate $\hat{\theta}$ that maximizes $f(\mathbf{y}|\theta)$ (it is assumed that θ is deterministic).
- We thus find estimate $\hat{\theta}$ by solving:

$$\max_{\theta} \quad \log f(\mathbf{y}|\theta) + \log f(\theta)$$

  This problem is equivalent to that of MLE but we incorporate prior term $\log f(\theta)$.
- If we have prior knowledge that $\theta \sim \mathcal{N}(\bar{\theta}, \Sigma_\theta)$, then:

$$\min_{\theta} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{m}(\theta)\|^2 + \frac{\kappa}{2}(\theta - \bar{\theta})\Sigma_\theta^{-1}(\theta - \bar{\theta})$$

- Gaussian prior $f(\theta)$ achieves its maximum at $\theta = \bar{\theta}$ while $f(\mathbf{y}|\theta)$ achieves its maximum when θ fits data.
- Estimation problem seeks to find *balance* between what we previously knew about θ and new knowledge gained through observations $\mathbf{y}$.
- If ignore prior knowledge, all that we know about θ is through $\mathbf{y}$ (which might lead to ambiguity if data is insufficient).
- Adding prior knowledge reduces *ambiguity*.

Statistical Learning

Machine learning is fast growing field that combines techniques from diverse branches of science and engineering to perform tasks such as:

- Data Analysis (e.g., dimension reduction, clustering, visualization, signal recognition, computer vision)
- Data-Driven Modeling (e.g., neural nets, kriging, support vector machines)
- Artificial Intelligence (e.g., data collection, experimentation, learning, control)

Statistical learning is a subset of machine learning that provides tools derived from *statistical principles*.

- Some tools of machine learning are derived from other mathematical principles (e.g., geometry, topology, optimization, linear algebra).
- Our focus here is not to provide an extensive review of all tools. Instead, we focus on general statistical principles behind such tools.

Data Analysis

How can I interpret and extract knowledge from high-dimensional data? How can I reduce (compress) my data?

Analysis of Multivariate Data

Consider the following problem:

- You have multiple input RVs $X = (X_1, X_2, ..., X_n)$ entering a system.
- You want to create a product that is a *mixture* (blend) of these RVs:

$$t = \sum_{i=1}^{n} w_i X_i = w^T X$$

where $w_i \in \mathbb{R}$ are mixture proportions satisfying $\|w\| = 1$, with $w = (w_1, w_2, ..., w_n)$.

- What proportions w give a product that contains maximum information about X?
- What proportions w give a product that contains minimum information about X?

Think about analogy of this problem with a physical blending process:

We want to mix a set of input flows in a way that product is as valuable as possible.

Analysis of Multivariate Data

Data mixing problem can be solved using *principal component analysis* (PCA).

- Collect observations for $x_\omega \in \mathbb{R}^n$, $\omega in 1, ..., S$ for the RV $X = (X_1, X_2, ..., X_n)$.
- Store the observations in $S \times n$ data matrix:

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{S,1} & x_{S,2} & \cdots & x_{S,n} \end{bmatrix}$$

- Normalize columns in such a way that $\frac{1}{S}\sum_{i=1}^{S} X_{i,j} = 0$ for all $j = 1, ..., n$. This centers the data around zero.
- After normalization, the sample covariance matrix of X is $\mathrm{Cov}(X) = \mathbf{X}^T\mathbf{X}$ (denote this as Σ and note this is $n \times n$).
- Covariance matrix contains all information (variance) of inputs X.

Analysis of Multivariate Data

- For mixture $t = w^T X$ we can show that $\hat{\mathbb{V}}[t] = w^T \Sigma w$.
- Consequently, the mixtures proportions that contain maximum variance of X can be found by solving:

$$\max_{w} \; w^T \Sigma w \text{ s.t. } \|w\| = 1$$

- Solution of this problem is eigenvector w_1 of Σ associated with largest eigenvalue $\lambda_1 = w_1^T \Sigma w_1$. Consequently, proportions w_1 give mixture $t_1 = w_1^T X$ that contains maximum information about X.
- The mixture that contains minimum information about X is found by solving:

$$\min_{w} \; w^T \Sigma w \text{ s.t. } \|w\| = 1$$

  this gives eigenvector w_n associated with smallest eigenvalue $\lambda_n = w_n^T \Sigma w_n$
- We can find and rank mixtures based on information content by performing an eigendecomposition:

$$\Sigma = \lambda_1 w_1 w_1^T + \lambda_2 w_2 w_2^T + \cdots + \lambda_n w_n w_n^T$$

- Eigenvalues are arranged as $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_n$. Consequently, mixture $t_1 = w_1^X$ contains most information while $t_2 = w_2^T X$ contains second most information and so on.

Analysis of Multivariate Data

- Mixtures (a.k.a. principal components) $t_1, t_2, ..., t_n$ contain most information about X. Consequently, select a few of them (typically two or three) to visualize high-dimensional data X in a small dimensional space (PCA is a dimensionality reduction technique).
- Truncating the eigendecomposition series enables compression of data matrix Σ.
- Key property of principal components is that they retain structure of high-dimensional data. Visualizing data by dropping variables destroys original structure.
- We can show that principal components are uncorrelated and thus $\mathrm{Cov}(t_i, t_j) = 0$ for all $i \neq j$ (i.e., mixtures contain complementary types of information).
- We can use eigenvector matrix $W = [w_1 \,|\, w_2 \,|\, \cdots | w_n] \in \mathbb{R}^{n \times n}$, to can *project* data matrix $\mathbf{X}$ to the principal component (information) space as:

$$\mathbf{T} = \mathbf{X}W$$

where $\mathbf{T} \in \mathbb{R}^{S \times n}$ is a matrix with entries $t_{i,j}$, $i = 1, ..., S\; j = 1, ..., n$.

Data-Driven Modeling (Classification)

Consider following problem:

- Imagine you have input RVs $X = (X_1, X_2, ..., X_n)$ with observations $x_\omega \in \mathbb{R}^n$, $\omega in 1, ..., S$ and domain $\mathcal{D} = \mathbb{R}^n$ (X can be discrete or continuous) and an output RV Y with observations $y_\omega \in \mathbb{R}$ with domain$\mathcal{D}_Y = \{0, 1\}$ (discrete).
- We postulate that there exists a relationship between X and Y of the form:

$$Y = g(X, \theta)$$

  where $\theta \mathbb{R}^n$ are parameters and $g(X, \theta)$ is model.
- Our objective is to find model that best explains observations:

$$y_\omega = g(x_\omega, \theta) + \epsilon_\omega$$

- This estimation problem is known as a *classification* problem. In this context, inputs X are known as features or descriptors while Y are known as labels or classes.
- What is different (and difficult) here is the binary nature of output Y. In standard estimation problems, Y is assumed to be continuous.

Think about analogy of this problem with a chemical classification problem:

Given a set of features for a chemical, can we predict if this is toxic or not?

Classification

- To solve problem, we postulate a model function of the form:

$$g(x,\theta) = \frac{1}{1+e^{-\theta^T x}}$$

- Here, the mix $\theta^T x = \sum_{i=1}^{n} \theta_i x_i$ is known as *evidence*. The parameters reflect weights that we place on different features.
- Postulated model is known as the logistic function and seeks to capture 0-1 logic.
- Logistic function is a sigmoidal function that satisfies $g(x,\theta) \to 1$ as $\theta^T x \to \infty$, $g(x,\theta) \to 0$ as $\theta^T x \to -\infty$, and $g(x,\theta) = \frac{1}{2}$ if $\theta^T x = 0$.

Classification

- Importantly, $g(x, \theta)$ can be used to model probabilities $\mathbf{P}(Y = 1 \,|\, x, \theta)$ and $\mathbf{P}(Y = 0 \,|\, x, \theta) = 1 - \mathbf{P}(Y = 1 \,|\, x, \theta)$:
  - If evidence is strongly positive, there is a high probability that $Y = 1$
  - If evidence is strongly negative, there is a high probability that $Y = 0$
  - If evidence is weak, there is ambiguity (it is equally probable that $Y = 0$ or $Y = 1$)
- This logic can be captured by defining a conditional pdf of the form:

$$f(y|x, \theta) = \mathbb{P}(Y = y|x, \theta) = g(x, \theta)^y (1 - g(x, \theta))^{1-y}$$

- Goal is to find estimate $\hat{\theta}$ that maximizes joint probability $\prod_{\omega \in \mathcal{S}} f(y_\omega, |x_\omega, \theta)$:

$$\max_\theta \sum_{\omega \in \mathcal{S}} y_\omega \log(g(x_\omega, \theta)) + (1 - y_\omega) \log(1 - g(x_\omega, \theta))$$

- Having estimate $\hat{\theta}$ we use model $g(x, \hat{\theta})$ to predict class of input with features x.
- This approach does not minimize sum of squared errors (and is standard estimation). The objective function is called the *log loss*.

Data-Driven Modeling (Kernel Methods)

- A flexible way of constructing models is to *mix* different types of models.
- Assume that Y and $X = (X_1, ..., X_n)$ follow a relationship of the form:

$$Y = \sum_{j=1}^{m} \theta_j \phi_j(X) = \theta^T \phi(X)$$

- $\phi_j(X)$ $j = 1, ..., m$ is collection of basic models (a.k.a. basis functions). We define vector $\phi(X) = (\phi_1(X), ..., \phi_m(X))$.
- Basis function $\phi_j(X)$ can be nonlinear (e.g., polynomial, sigmoidal, exponential) or linear (in which case $\phi(X) = X$).
- Parameters $\theta_j \in \mathbb{R}$ are mixing coefficients of basis functions.

Data-Driven Modeling (Kernel Methods)

- We determine estimate $\hat{\theta}$ by solving regularized MLE problem:

$$\min_{\theta} S(\theta) = \frac{1}{2} \sum_{\omega \in \mathcal{S}} (y_\omega - \theta^T \phi(x_\omega))^2 + \lambda \theta^T \theta$$

In vector form:

$$\min_{\theta} S(\theta) = \frac{1}{2} (\Phi\theta - \mathbf{y})^T (\Phi\theta - \mathbf{y}) + \lambda \theta^T \theta$$

where $\Phi \in \mathbb{R}^{S \times n}$ is input data matrix with entries $\Phi_{\omega,j} = \phi_j(x_\omega)$ and $\mathbf{y}$ is the output data vector.

- If basis functions are linear then $\Phi = X$.

Kernel Methods

- Optimality conditions indicate that best estimate satisfies:

$$\theta = \frac{1}{\lambda}\Phi^T\mathbf{r}$$

  where we define residuals (mismatch errors) $\mathbf{r} = (\Phi\theta - \mathbf{y})$. By substituting in $S(\theta)$ we obtain:

$$S(\theta) = \frac{1}{2}\mathbf{r}^T\Phi\Phi^T\Phi\Phi^T\mathbf{r} - \mathbf{r}^T\Phi\Phi^T\mathbf{y} + \frac{1}{2}\mathbf{y}^T\mathbf{y} + \frac{\lambda}{2}\mathbf{r}^T\Phi\Phi^T\mathbf{r}$$

  We define matrix $K = \Phi\Phi^T \in \mathbb{R}^{S\times S}$ and simplify:

$$S(\theta) = \frac{1}{2}\mathbf{r}^TKK\mathbf{r} - \mathbf{r}^TK\mathbf{y} + \frac{1}{2}\mathbf{y}^T\mathbf{y} + \frac{\lambda}{2}\mathbf{r}^TK\mathbf{r}$$

- Note that function $S(\theta)$ can be defined entirely in terms of $\mathbf{r}$.

Kernel Methods

- This suggests an alternative strategy to find estimate. Here, we solve problem:

$$\min_{\mathbf{r}} \frac{1}{2}\mathbf{r}^T KK\mathbf{r} - \mathbf{r}^T K\mathbf{y} + \frac{1}{2}\mathbf{y}^T\mathbf{y} + \frac{\lambda}{2}\mathbf{r}^T K\mathbf{r}$$

  to find an optimal $\hat{\mathbf{r}}$ and then recover $\hat{\theta} = \Phi^T\hat{\mathbf{r}}$.

- But it turns out that parameters $\hat{\theta}$ are *not needed at all*. To see this, note that $\hat{\mathbf{r}}$ satisfies $\mathbf{r} = (K + \lambda)^{-1}\mathbf{y}$ and thus optimal model prediction is:

$$\begin{aligned}\hat{\mathbf{y}} &= K\hat{\mathbf{r}}\\ &= K(K + \lambda I)^{-1}\mathbf{y}\end{aligned}$$

- Optimal prediction only depends on input data (in K) and output data (in $\mathbf{y}$).
- This implies that parameters θ are not needed (these are just intermediary variables).

Kernel Methods

- Matrix K is known as kernel matrix, which captures interactions in input variables.
- Given input data x_ω, $\omega in \mathcal{S}$, a kernel matrix can be constructed as:

$$K_{i,j} = k(x_i, x_j)\ i, j \in \mathcal{S}$$

  where $k(x_i, x_j)$ is known as the *kernel function*. This provides a general approach for estimation.

- The case for linear estimation corresponds to defining kernel function:

$$k(x_i, x_j) = x_i^T x_j$$

  In this case, note that $K = \mathrm{Cov}(X)$ (after normalization).

- The case of basis functions corresponds to defining a kernel function:

$$k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

- Typically, kernel function is chosen to be the radial basis function (RBF):

$$k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$$

where γ is a hyperparameter. This kernel is also known as the Gaussian kernel.

Kernel Methods

The RBF kernel is simple but captures a wide range of nonlinear behavior. To see this:

Consider scalar case (x_ω is a scalar) and notice that:

$$\begin{aligned}\exp(-\gamma(x_i - x_j)^2) &= \exp(-\gamma x_i^2 + 2\gamma x_i x_j - \gamma x_j^2) \\ &= \exp(-\gamma x_i^2 - \gamma x_j^2)\left(1 + \frac{2\gamma x_i x_j}{1\,!} + \frac{(2\gamma x_i x_j)^2}{2\,!} + ...\right)\end{aligned}$$

We thus have that radial basis function can be written as:

$$\exp(-\gamma(x_i - x_j)^2) = \phi(x_i)^T \phi(x_j)$$

where $\phi(x)$ is an *infinite* collection of polynomial basis functions:

$$\phi(x) = e^{-\gamma x}(1, \sqrt{2\gamma/1\,!}, \sqrt{(2\gamma)^2/2\,!}, \sqrt{(2\gamma)^3/3\,!}, ...).$$

- In kernel methods we do not need to postulate an input-output model and estimate its parameters θ.
- Instead, we postulate a kernel function and estimate its hyperparameters (e.g., γ).
- The number of parameters of a kernel function is often small (typically less than ten). This is a key advantage over standard estimation approaches.

Data-Driven Modeling (Kriging)

How do we estimate kernel function hyperparameters?

This can be done by using a technique called kriging.

- In kriging we postulate a model of the form:

$$y_\omega = g(x_\omega) + \epsilon_\omega,\ \omega \in \mathcal{S}$$

  In vector form:

$$\mathbf{y} = \mathbf{g} + \epsilon$$

- If we assume that $\epsilon_\omega \sim \mathcal{N}(0, \sigma)$ then we have that $f(\mathbf{y}|\mathbf{g})$ corresponds to $\mathcal{N}(\mathbf{g}, \sigma\mathbb{I})$.
- Note the absence of parameters in postulated model.

Kriging

- What is unique about kriging is that $\mathbf{g}$ is treated as a random function (non-parametric).
- In the techniques that we have covered, we defined parametric functions $\mathbf{g}(\theta)$.
- We assume that random function $\mathbf{g}$ has pdf $f(\mathbf{g}|\gamma)$ corresponding to $\mathcal{N}(0, K(X, \gamma))$.
- Here, $\mathrm{Cov}(\mathbf{g}) = K(\gamma)$ is a covariance function and γ are its hyperparameters.
- Think about $K(\gamma)$ as a function that generates samples of the random model $\mathbf{g}$. This matrix has entries:

$$K_{i,j}(\gamma) = k(x_i, x_j, \gamma) \tag{1}$$

where $k(x_i, x_j, \gamma)$ is a kernel function.

Kriging

- One can can show that marginal of $\mathbf{y}$ is:

$$f(\mathbf{y}|\gamma) = \int f(\mathbf{y}|\mathbf{g})f(\mathbf{g}|\gamma)d\mathbf{g}$$

- This corresponds to pdf of $\mathcal{N}(0, C(\gamma))$ with entries:

$$C(x_i, x_j, \gamma) = K_{i,j}(\gamma) + \sigma\delta_{i,j}\ i, j \in \mathcal{S}$$

  where $\delta_{i,j} = 1$ if $x_i = x_j$ and zero otherwise.
- One approach to estimate $\hat{\gamma}$ consists of maximizing $\log f(\mathbf{y}|\gamma)$:

$$\max_{\gamma}\ -\frac{1}{2}\log|C(\gamma)| - \frac{1}{2}\mathbf{y}^T C(\gamma)\mathbf{y} - \frac{S}{2}\log 2\pi$$

- As a kernel method, in kriging one computes predictions by using kernel function (which only depends on $\hat{\gamma}$ and input data).
- This is done by using conditional pdf $f(y|\mathbf{y})$ where y is predicted output at new point x and $\mathbf{y}$ are observations used to determine $\hat{\gamma}$.

Kriging

- One can show that the conditional pdf $f(y|\mathbf{y})$ is Gaussian $\mathcal{N}(m(x), \sigma^2(x))$ with:

$$m(x) = \mathbf{k}^T C^{-1} \mathbf{y}$$
$$\sigma^2(x) = c - \mathbf{k}^T C \mathbf{k}$$

  and:

$$C = C(\hat{\gamma})$$
$$c = k(x, x, \hat{\gamma})$$
$$\mathbf{k}_i = k(x_i, x, \hat{\gamma}),\ i = 1, ..., S$$

- Mean prediction of y is $m(x)$ and confidence intervals can be constructed using $\sigma^2(x)$.
- There is a wide range of kernel functions that can be used in kriging. The generalized Gaussian kernel resembles the radial basis function and takes the form:

$$k(x_i, x_k, \gamma) = \gamma_0 \exp\left(-\frac{\gamma_1}{2}\|x_i - x_j\|^2\right) + \gamma_2$$

- It is possible to combine parametric and non-parametric estimation within kriging. This can be done by defining $f(\mathbf{y}|\mathbf{g}, \theta)$ as $\mathcal{N}(\mathbf{g} + \Phi\theta, \sigma\mathbb{I})$ where Φ is an input data matrix with entries $\Phi_{\omega,j} = \phi_j(x_\omega)$ and $\phi_j(x)$ is a collection of basis functions.

Data-Driven Modeling (Neural Networks)

How do humans learn? How do they establish corrections between variables to make predictions and decisions?

- Neural networks (NNs) are nature-inspired parametric models that seek to establish causal models of between $X = (X_1, ..., X_n)$ and Y of the form:

$$Y = g(X, \theta)$$

- What is different about NNs is that the model function $g(X, \theta)$ is automatically constructed by using a set of activation (basis) functions that are mixed (combined) in a hierarchical manner. This seeks to mimic how the brain works.
- NNs provide a flexible approach to capture virtually any type of relationship between X and Y. A key advantage of this is that we do not need to postulate a model (e.g., nonlinear, linear, logistic, mechanistic).
- As with estimation and classification the objective is to build the NN model (the parameters θ) that best explain the observations:

$$y_\omega = g(x_\omega, \theta) + \epsilon_\omega, \; \omega \in \mathcal{S}$$

- The determination of θ from data mimics *learning* process of a human.

Neural Networks

- NNs seek to mimic how brain responds when exposed to data and how we learn and accumulate knowledge.
- Brain is a highly sophisticated network that has neurons as basic processing (decision) units that interact with one another through signals.
- Mathematical description of a neuron is called a *perceptron*. Perceptron generates a signal if evidence $\sum_{i=1}^{n} \theta_i x_i + \theta_0$ is strong enough or does not generate signal if evidence is not strong.
- Binary behavior of perceptron is similar to that logistic classification. In fact, perceptrons are often modeled using logistic functions. The additional parameter θ_0 in the evidence is called the *bias*.
- At basic level, a perceptron is a decision unit that decides to “fire” or not when exposed to data (evidence). How to respond to different pieces of evidence is captured by parameters.
- The bias captures situations in which we make decisions by taking into account not only evidence but also our inherit biases.
- An NN is a hierarchical architecture of perceptrons that constructs *complex logic*.

Neural Networks

Components of NN hierarchy are: input layers, hidden layers, and output layer.

- Input layer contains perceptrons that take input data X to generate signals. This layer captures basic logic (i.e., I decided this because of that).
- Hidden layer contains perceptrons that take signals from input layer to generate signals. This layer captures abstract logic. This layer captures aspects that are difficult for humans to rationalize (i.e., why did I decide that?).
- Output layer contains perceptrons that take signals from hidden layer to generate a final output signal Y. Signal can be continuous or discrete.

We now proceed to show how to train NNs. For simplicity, we will assume a NN architecture with one input layer, one hidden layer, and one output layer.

Neural Networks

- *Input layer* is composed of $j = 1, ..., n_I$ perceptrons. Evidence in perceptron j is :

  $$a_j = \sum_{i=1}^{n} \theta^I_{j,i} x_i + \theta^I_{j,0}$$

  Here, $\theta^I_{j,i}$ are parameters of perceptron ($\theta_{j,0}$ is bias) and x_i is input data (for a given observation ω). Given evidence, perceptron j generates output signal:

  $$z_j = h(a_j)$$

  here, $h(a_j)$ is known as activation function and is often modeled using a sigmoidal function (tanh and max functions are also used).
- *Hidden layer* is composed of $k = 1, ..., n_H$ perceptrons. Evidence in perceptron k is:

  $$a_k = \sum_{j=1}^{n_I} \theta^H_{k,j} z_j + \theta^H_{k,0}$$

  Given evidence, output signal is:

  $$w_k = h(a_k)$$

Neural Networks

- *Output layer* takes signals of hidden layer and can have one or multiple perceptrons (depending on how many outputs Y we have). If there is one output, evidence takes the form:

$$a = \sum_{k=1}^{n_H} \theta_k^O w_k + \theta_0^O$$

  Final output (model prediction) is:

$$m = \sigma(a)$$

  where $\sigma(a)$ is a sigmoidal function if output Y is binary (e.g., classification) or $\sigma(a) = a$ if output is continuous (e.g., regression).

- Given parameters $\theta = (\theta^I, \theta^H, \theta^O)$, NN propagates input x_ω into output $m_\omega(\theta)$. This forward propagation can be written as:

$$m_\omega(\theta) = \sigma\left(\sum_{k=1}^{n_H} \theta_k^O h\left(\sum_{j=1} g\left(\sum_{i=1}^{n} \theta_{j,i}^I x_i^\omega + \theta_{j,0}^I\right) + \theta_{k,0}^H\right) + \theta_0^O\right), \; \omega \in \mathcal{S}$$

Neural Networks

- We can use MLE framework to estimate θ. For regression problems we solve:

$$\min_{\theta} \; S(\theta) = \frac{1}{2} \sum_{\omega \in \mathcal{S}} (y_\omega - m_\omega(\theta))^2$$

  In the context of NNs, $S(\theta)$ is known as the loss function.

- For classification problems we find θ that maximizes the log loss.
- Main advantage of NNs is that they are completely data-driven (no mechanistic understanding is needed). However, this is also the main disadvantage of NNs.
- Parameters and hidden variables often have no physical meaning and it is thus difficult to convey prior knowledge.
- This often manifests as a need for large amounts of data to determine parameters which are also many (on the order of thousands to millions). In the architecture discussed with have $n_I \cdot n + n_H \cdot n_I + n_H$ parameters.
- As a result, an issue with NNs is that of *overfitting*. Generic regularization terms (e.g., $\rho(\theta) = \lambda \theta^T \theta$) are often added to loss function but there is often no mechanistic basis to construct more sophisticated regularization strategies (e.g., Bayesian priors).
- Conducting cross-validation in NNs is particularly critical.

Decision-Making under Uncertainty

- We have learned to use data to model uncertainty as RVs and to build models to capture connections between RVs.
- We have also learned how to characterize uncertainty for functions of RVs using Monte Carlo simulations and RV transformations.
- We now shift our attention on how to use these capabilities to *make decisions*.
- Fundamental issue with making decisions under uncertainty is that humans take different attitudes towards risk and tend to severely under/overestimate uncertainty. Also, "risk" means different things to different people.
- Consider setting in which we would like to chose between decisions u and u' with associated univariate RVs $Y(u)$ and $Y(u')$ (e.g., cost).
- Another setting is that in which we want to find a decision u that manipulates RV $Y(u)$ in a particular way (e.g., it maximizes it or reduces its variability).

Decision-Making under Uncertainty

Consider now the following questions:

- What is risk? How do I measure risk?
- How can I make a decision that withstands uncertainty?
- How can I take optimal proactive actions in the face of uncertainty?

The concepts and techniques that answer these questions are studied in the area of *stochastic optimization*.

Defining and Measuring Risk

At this point, we have all fundamentals of statistics needed to properly define risk and figure out how to measure it.

- Consider an input X with pdf $f_X(x)$ and cdf $F_X(x)$ and utility function $\varphi(X, u)$ that depends on the input RV and a decision u.
- Since X is an RV, utility is also an RV that we represent as univariate RV $Y(u) = \varphi(X, u)$ with pdf $f_{Y(u)}(y)$ and cdf $F_{Y(u)}(y)$.
- The pdf and cdf of $Y(u)$ depend on the decision u; consequently, u can be used to manipulate these and to manipulate the statistics of $Y(u)$.
- In popular culture, risk is associated with a decision or situation (e.g., “this is a risky investment”, “the reactor is operating at risky conditions”) and is associated with extreme events or large/catastrophic losses.

Defining and Measuring Risk

- When making decisions, however, we want precise measures that tell us exactly what constitutes a risky decision and how risky it is.
- Unfortunately, there is no unique mathematical definition of risk because humans tend to value different aspects of uncertainty (e.g., probability of failure vs. magnitude of failure vs. worst-case failure).
- Moreover, humans tend to disagree on what probability levels are acceptable (e.g., what seems risky to me might not seem risky to you).
- These disagreements arise because decision-makers (DMs) take different attitudes towards risk.
- As a result, what we want to explore here is not what definition of risk exist but rather what *definitions* of risk exist. Ultimately, establishing a definition of risk for a particular situation at hand should be based on consensus between DMs.

Defining and Measuring Risk

To motivate potential definitions of risk that we might consider, we explore different attitudes towards risk that DMs might take:

- *Risk-Neutral*: This DM prefers $Y(u)$ over $Y(u')$ if $\mathbb{E}_{Y(u)} \leq \mathbb{E}_{Y(u')}$. This DM worries about performance on average and is not concerned with the fact that $Y(u)$ might have outcomes with large values compared to those of $Y'(u)$.
- *Risk-Conscious*: This DM prefers $Y(u)$ over $Y(u')$ if $\mathbb{P}(Y(u) \leq y) \geq \mathbb{P}(Y(u') \leq y)$. A risk-conscious DM wants a decision that will likely achieve lower outcomes. This type of DM might also prefer $\mathbb{V}_{Y(u)} \leq \mathbb{V}_{Y(u')}$ because u has lower variability than u'.
- *Risk-Averse*: This DM prefers $Y(u)$ over $Y(u')$ if $\max Y(u) \leq \max Y(u')$. This DM only worries about the worst possible outcome of u and u'.
- *Risk-Taker*: This DM prefers $Y(u)$ over $Y(u')$ if $\min Y(u) \leq \min Y(u')$. This DM only worries about the best possible outcome of u and u.

Defining and Measuring Risk

So, how do we measure risk?

A risk measure is a function $\rho(Y(u))$ that maps a univariate RV $Y(u)$ to a scalar quantity (it is a summarizing statistic). Common risk measures used in practice are:

- Expected Value: $\mathbb{E}[Y(u)]$
- Variance: $\mathbb{V}[Y(u)] = \mathbb{E}[(Y(u) - \mathbb{E}[Y(u)])^2]$
- Mean-Variance: $\mathbb{E}[Y(u)] + \kappa\mathbb{V}[Y(u)]$ (for some κ)
- Probability of Loss: $\mathbb{P}(Y(u) > y)$ or $\mathbb{P}(Y(u) \leq y)$ (a.k.a. probability of failure)
- Value-at-Risk: $\mathbb{Q}_{Y(u)}(\alpha)$ (also written as VaR_α)
- Conditional Value-at-Risk: $\mathbb{E}[Y(u)|Y(u) \geq \mathrm{VaR}_\alpha]$ (a.k.a. expected short fall)
- Mean Deviation: $\mathbb{E}[|Y(u) - \mathbb{E}[Y(u)]|]$
- Worst/Best-Case: $\max Y(u)$ or $\min Y(u)$

Defining and Measuring Risk

Some observations:

- Different measures are used to model different risk attitudes.
- Probability of loss is what is colloquially known as risk because it has a natural interpretation that DMs find useful in practice. We will see that this measure has its caveats.
- In principle, we have freedom of using any statistic $Y(u)$ of interest (e.g., moments, entropy) as risk measure but, as expected, not all measures are expected to be adequate.

Stochastic Dominance

So, what are the desirable features of a risk measure?

- To answer this question, it is necessary to first discuss what makes an RV better than another RV (how do we compare RVs?).
- An important concept that arises here is that of *stochastic dominance* (SD).

We say that Y dominates Y' (written as $Y \preceq Y'$) if:

$$\mathbb{P}(Y \leq y) \geq \mathbb{P}(Y' \leq y) \text{ for all } y \in \mathcal{D}$$

where $\mathcal{D}$ is the domain of Y and Y'.

Some observations:

- SD says that Y is better (dominates) Y' if probability of Y taking a value less than a threshold y is higher than probability of Y' being below same threshold. Moreover, the probability is higher or equal for any threshold value.
- Can also express SD as $F_Y(y) \geq F_{Y'}(y)$ for all $y \in \mathcal{D}$. We thus have that cdf curve of Y is always above that of Y'.

Stochastic Dominance

- Note how SD is different than the traditional dominance concept we are familiar with (a decision dominates another one if it is *always* better). In statistical terms, this would imply that outcomes of Y are always lower or equal than those of Y':

$$y_\omega \leq y'_\omega \; \omega \in \Omega$$

- This requirement is strict and it is unlikely to occur in practice. Consequently, SD is a more flexible requirement.
- A risk-conscious DM would typically require that $\mathbb{P}(Y \leq y) \geq \mathbb{P}(Y' \leq y)$ holds *or a single* y. Note that SD is a stricter requirement because it requires $\mathbb{P}(Y \leq y) \geq \mathbb{P}(Y' \leq y)$ to hold *for all possible* y. As such, there are weakened (relaxed) versions of SD that are also often used in practice (e.g., holds *for some* y).

Coherency Properties of Risk Measures

There are a number of fundamental properties (a.k.a. axioms) that a proper (a.k.a. coherent) risk measure should satisfy. These properties have been proposed based on their usefulness in actual practical applications and based on mathematical consistency.

- Translation Invariance: $\rho(Y + c) = \rho(Y) + c$ for $c \in \mathbb{R}$. This indicates that adding a constant value to an RV should result in adding same constant to the risk measure (i.e., adding a constant does not alter inherent properties of RV).
- Subadditivity: $\rho(Y + Y') \leq \rho(Y) + \rho(Y')$. This indicates that the risk of a combined pair of RVs cannot exceed the summation of their individual risks.
- Monotonicity: If $Y \preceq Y'$ then $\rho(Y) \leq \rho(Y')$. This indicates that, if Y stochastically dominates Y', then its risk should also be lower. This indicates that risk measure reflects dominance.
- Positive Homogeneity: $\rho(c \cdot Y) = c\rho(Y)$ for $c \in \mathbb{R}_+$. This indicates that scaling RV by a constant does not affect inherent properties of RV.

We will see later that some of these properties resemble those of vector norms.

Risk Measures as Norms

In fact, it turns out that the risk measure of an RV is analogous to a norm of a vector. To see this, lets begin with the following question:

How can we say that vector $\mathbf{y} = (y_1, y_2, ..., y_S)$ is better than $\mathbf{y}' = (y'_1, y'_2, ..., y'_S)$?

This reveals similar difficulties that arise when comparing RVs. Consider following possibilities:

- We can say that $\mathbf{y}$ is better than $\mathbf{y}'$ if the total magnitude of its entries is lower $(\sum_{i=1}^{S} y_i \leq \sum_{i=1}^{S} y'_i)$. This can be written as $(1/S)\sum_{i=1}^{S} y_i \leq (1/S)\sum_{i=1}^{S} y'_i$ and is analogous to expected value.
- We can say that $\mathbf{y}$ is better than $\mathbf{y}'$ if $\mathrm{max}_i \leq \mathrm{max}_i \;\, y'_i$. This is analogous to worst-case risk measure.
- We can say that $\mathbf{y}$ is better than $\mathbf{y}'$ if $y_i \leq y'_i$ for all $i = 1, ..., S$. This is analogous to say that Y is always better than Y'. Note, however, that the entries are not arranged in order so maybe an entry of $\mathbf{y}$ is much larger than any entry of $\mathbf{y}'$.
- Consequently, perhaps a better approach would be arranging entries in decreasing order and then compare the rearranged entries. This is analogous to stochastic dominance (we establish a threshold on the magnitude of the entries) and count how many entries of the vectors are below that threshold value.

Risk Measures as Norms

- Recall that norm $\rho(\mathbf{y})$ of a vector $\mathbf{y}$ is a measure of its size. Norms are used to compare vectors and to establish bounding properties.
- Recall that most used vector norm is ℓ-p norm:

$$\|\mathbf{y}\|_p = \left(\sum_{i=1}^{S} |y_i|^p \right)^{1/p}.$$

  This norm has the following special cases:
  - For $p = 1$ we have that $\|\mathbf{y}\|_1 = \sum_{i=1}^{S} |y_i|$ and note that this is S times average magnitude of the entries (analogous of expected value).
  - For $p = \infty$ we have that $\|\mathbf{y}\|_\infty = \max_i |y_i|$ and note this is largest value (analogous of worst-case).
- As in the case of risk measures, one can define many types of norms to measure and compare vectors in different forms.
- For instance, an interesting norm is k-max norm. This norm is sum of k-largest elements of vector $\mathbf{y}$. This norm is analogous to expected shortfall.

Risk Measures as Norms

As with risk measures, one must ensure that any norm that we define satisfy basic properties. A *proper* norm of a vector must satisfy the following properties:

- Homogeneity: $\rho(c \cdot \mathbf{y}) = c \cdot \rho(\mathbf{y})$ for $c \in \mathbb{R}_+$
- Subadditivity (Triangle Inequality): $\rho(\mathbf{y} + \mathbf{y}') \leq \rho(\mathbf{y}) + \rho(\mathbf{y}')$
- Normalization: $\rho(0) = 0$.

Homogeneity property is analogous to that of a proper risk measure. Moreover, the triangle inequality is analogous to the subadditivity property.

Expected Value Properties

We now have all elements to judge whether a risk measure is adequate or not. We begin by discussing the properties of expected value $\mathbb{E}[Y]$.

- Expected value is a measure of the magnitude of Y (i.e., $\mathbb{E}[Y] \leq \mathbb{E}[Y']$ indicates that Y is, on average, smaller than Y') and therefore has a natural intuitive interpretation. Because of this, this is a common measure used in practice.
- Expected value satisfies all properties of coherency.
- Expected value ignores outcomes that can result in high values. For instance, even if $\mathbb{E}[Y] \leq \mathbb{E}[Y']$ holds, Y can have outcomes of large magnitude that Y' does not have.

Variance Properties

We now discuss the properties of variance $\mathbb{V}[Y] = \mathbb{E}[(Y - \mathbb{E}[Y])^2]$.

- Variance is a measure of variability (i.e., $\mathbb{V}[Y] \leq \mathbb{V}[Y']$ indicates that Y has less variability than Y') and therefore has a natural and intuitive interpretation.
- An issue with the variance is that it does not capture magnitude of RV. To remedy this issue, variance is often used in conjunction with expected value by using mean-variance measure $\mathbb{E}[Y] + \kappa\mathbb{V}[Y]$ where κ is a weighting vector that helps span a range of attitudes towards risk (from risk neutral to risk conscious).
- Mean-variance was proposed by Harry Markowitz in the 1950s and became the standard in the finance industry for many years (Markowitz earned a Nobel prize).
- Variance takes into consideration outcomes of large values (in upper tail of pdf) but this connection is not direct and this results in several important inefficiencies. Specifically, variance is centered around the expected value and penalizes deviations from it symmetrically.
- Symmetry is undesirable in applications because we are often only interested in accounting for outcomes of large magnitude (and not with those of low magnitude). Moreover, in many applications, RV does not have a symmetric pdf and therefore variance can fail to properly capture tail effects.

.

Variance Properties

- Importantly, variance is not a proper risk measure. Specifically, it does not satisfy monotonicity (it is not consistent with SD).
- The variance and expected value do not have same units. Consequently, it is often preferred to use the standard deviation $\mathrm{SV}[Y] = \sqrt{\mathbb{V}[Y]}$, which has the same units as the expected value.
- Variance remains widely used in industry because it is an easily-interpretable measure of variability. This measure, however, has important deficiencies.

Expected Shortfall Properties

The expected shortfall overcomes many of deficiencies of the variance and has recently become the standard risk measure. This measure has many important properties and connections with other risk measures that are worth highlighting.

- Recall that expected shortfall of Y at probability α is $\mathbb{E}[Y|Y \geq \mathbb{Q}(\alpha)]$, where $\mathbb{Q}(\alpha)$ is the α-quantile of Y.
- Expected loss takes expected value losses above quantile $\mathbb{Q}(\alpha)$. The α-quantile is threshold value t at which $\mathbb{P}(Y \leq t) = \alpha$.
- Expected loss captures magnitude of outcomes of high value while ignoring those of small magnitude (it is an asymmetric risk measure). This becomes obvious if we write the expected loss as $\mathbb{E}[(Y - \mathbb{Q}(\alpha))_+]$.

Expected Shortfall Properties

- If we set $\alpha = 0$, quantile $\mathbb{Q}(\alpha)$ is minimum value of Y and therefore the expected loss is the expected value $\mathbb{E}[Y]$.
- If we set $\alpha = 1$, quantile is maximum value and therefore the expected loss is the worst-case value $\max Y$.
- Consequently, expected loss captures a range of risk attitudes (from risk neutral to conscious to averse).
- Expected loss is a coherent risk measure.
- A caveat of the expected loss is that it offers no direct control on the probability of loss, which is a measure of interest to many DMs. In other words, reducing the expected loss does not necessarily imply reducing the probability of loss.
- Moreover, DM needs to specify α and this selection can lead to disagreement.

Probability of Loss Properties

Probability of loss is one of most widely used measures of risk.

- Recall that probability of loss is simply $\mathbb{P}(Y > y)$ and this is also often expressed as probability of no loss as $\mathbb{P}(Y \leq y) = 1 - \mathbb{P}(Y > y)$.
- What constitutes a loss is defined by threshold value y. Consequently, it is more appropriate to call this measure the "probability of unacceptable loss".
- Meaning of probability of loss is intuitive.
- Biggest caveat of probability of loss is that it says nothing about actual magnitude of the losses. For example, even if $\mathbb{P}(Y > y) \leq \mathbb{P}(Y' > y)$, one might still prefer Y' because losses incurred are not that large compared to those of Y. In other words, Y' is less catastrophic than Y.
- Probability of loss is not coherent risk measure.
- Finally, one needs to specify a threshold value to express what constitutes and unacceptable loss and this selection often leads to disagreement of DMs.

Risk Measures

The different advantages and disadvantages encountered with risk measures highlight the difficulty in controlling and comparing RVs. In particular:

- Risk measures are often *conflicting* (e.g., reducing one measure often results in increasing another measure).
- Measuring risk leads to *ambiguity* (expressing mathematically what an DM might be looking for is challenging). This is analogous to measuring happiness or fairness.

Despite these limitations, risk measures are an essential component of decision-making under uncertainty.

Making Optimal Decisions

Instead of comparing decisions, a DM might also want to directly find best decision possible. Such a decision is also influenced by the attitude towards risk and gives rise to different optimization problems (depending on the risk measure used):

- Expected value and variance:

$$u^* = \underset{u}{\operatorname{argmin}} \; \mathbb{E}[Y(u)] + \kappa \mathbb{V}[Y(u)]$$

- Probability of loss

$$u^* = \underset{u}{\operatorname{argmin}} \; \mathbb{P}(Y(u) > y)$$

- Expected shortfall:

$$u^* = \underset{u}{\operatorname{argmin}} \; \mathbb{E}[Y(u) \,|\, Y(u) \geq \mathbb{Q}(\alpha)]$$

Another possibility is to find a decision that dominates a given benchmark. In other words, we seek u such that $Y(u) \preceq Y'$ for a given Y'. An issue with this approach is that the problem might have no solution.

Deterministic Decision-Making

DMs often make decisions in real-life by ignoring uncertainty all-together. This *deterministic* DM approach follows the logic:

- Assume input X takes a single value $x = \mathbb{E}[X]$ (i.e., average historical value). This assumes that input is deterministic and thus output Y is deterministic with value:

$$y = \varphi(x, u)$$

- Based on this assumed behavior, we can obtain a decision:

$$u_D = \underset{u}{\operatorname{argmin}}\, \varphi(x, u)$$

- The assumed deterministic behavior simplifies the decision-making process (removes ambiguity) but fails to ignore the inherent variability of X seen in real-life.
- Decision u_D might not be capable of controlling real output $Y(u_D)$, which is characterized by outcomes $y_\omega = \varphi(x_\omega, u_D),\ \omega \in \mathcal{S}$.
- As a result, decision u_D might be vulnerable to uncertainty and incur large losses.