Statistics for Chemical Engineers:
From Data to Models to Decisions

Victor M. Zavala

Department of Chemical and Biological Engineering
University of Wisconsin-Madison

victor.zavala@wisc.edu

Chapter 1: Introduction

UNIVERSITY OF WISCONSIN–MADISON

Motivation

As engineers, we often use *fundamental laws* to make *decisions*:

- Discovery of fundamental laws has been the result of extensive collection and analysis of observations (data)
- Fundamental law is often expressed in the form of a mechanistic model
- Mechanistic model provides a concise summary of observations (knowledge) that allow us to predict and generalize

Can you think of fundamental laws used in chemical engineering?

Motivation

Fundamental 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)

To account for this, we also often build predictive models based purely on observations (data); such models are known as *empirical models* and also embed knowledge.

Engineering decisions rely on a combination of mechanistic and empirical knowledge.

Can you think of fundamental laws used in chemical engineering?

Motivation

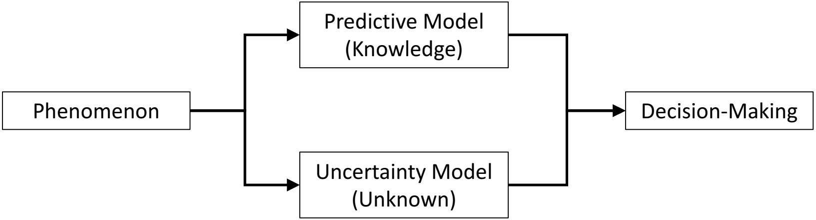
However:

- Model predictions will *always* face a certain degree of *uncertainty* (due to limited knowledge/understanding/data or due to *random* phenomena affecting systems).
- Despite these limitations, we still need to be able to *make decisions*. In fact, we (as humans), make predictions and decisions in our daily lives.
- The human brain naturally gathers knowledge (learns) from observations by building empirical models and has the ability to blend such empirical knowledge with mechanistic knowledge.
- The human brain has a natural ability to hedge against uncertainty and adapts decisions based on new knowledge.

Examples of random phenomena: time to failure of a material, "pop" time of a corn kernel, molecular fluctuations.

Example of brain learning: riding a bicycle, cooking.

Motivation



Decision-making relies on ability to characterize of what is known (predictable) and not known (not predictable).

Motivation

Statistics is the branch of mathematics that offers tools to:

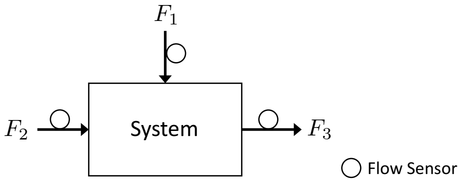
- Collect, analyze, and extract knowledge (models) from data
- Characterize and model the unknown (uncertainty)
- Systematically make decisions in the face of uncertainty

For an engineering perspective, *statistics* aids the discovery of fundamental laws and the development of mechanistic and empirical models as well as the characterization of random phenomena.

From a scientific perspective, *statistics* provides a framework for thinking about the world that can help us understand how humans extract knowledge from data and use this to make decisions in the face of uncertainty.

Example: Flow Mixer `ch1_mixer_example.m`

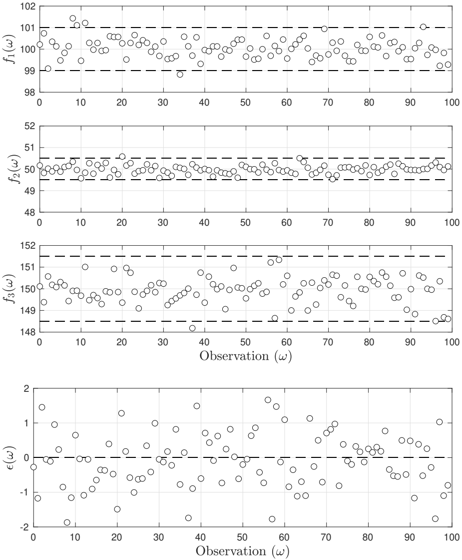
- Consider a mixing system with input flows F_1, F_2 and output flow F_3
- Sensors measure flows at time ω: $f_1(\omega) = 101.5$, $f_2(\omega) = 50.5$, $f_3(\omega) = 151$ gpm
- Conservation laws tell us that $f_3(\omega) = f_1(\omega) + f_2(\omega)$ (but this is not true). Why?



- What is known and unknown about this system?
- What is source of uncertainty?

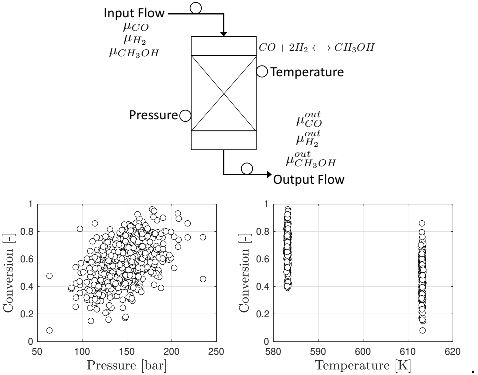
Example: Flow Mixer `ch1_mixer_example.m`

Collect 100 observations and monitor mismatch $\epsilon(\omega) = f_3(\omega) - f_1(\omega) - f_2(\omega)$



Example: Gibbs Reactor

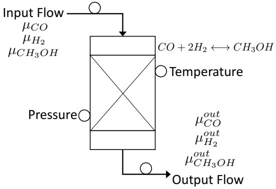
- Consider a reactor under which the reaction $CO + 2H_2 \leftrightarrow CH_3OH$ takes place



- Data seems to indicate that reaction favored (achieves higher conversion ξ) at high pressure (P) and low temperature (T) but this is blurred by variability.

- What is known and unknown about this system?
- Note that trends are blurred by uncertainty/noise.

Example: Gibbs Reactor



$$\mu_k^{out} = \mu_k + \gamma_k \cdot \xi \cdot \mu_{CO}, \; k \in K$$

$$\xi = 1 - \frac{\mu_{CO}^{out}}{\mu_{CO}}$$

$$K_{eq} = \frac{(\mu_{CH_3OH} + \xi \cdot \mu_{CO})}{(\mu_{CO} - \xi\mu_{CO})(\mu_{H_2} - 2 \cdot \xi \cdot \mu_{CO})^2} \left(\frac{\mu_{out}}{P}\right)^2$$

$$\frac{\partial \log K_{eq}}{\partial T} = -\frac{\Delta H_0}{R \cdot T}$$

- We know fundamental conservation and thermodynamic laws hold and mechanistic model allow us to predict ξ from P and T.
- This knowledge, however, is limited (e.g., makes assumptions, cannot explain random behavior).
- How to predict ξ as a function of P and T without any mechanistic knowledge?
- What are trade-offs between mechanistic and empirical knowledge?

Motivation

Don't forget:

Inherent limitation of engineering practice: no matter how sophisticated our predictive models and sensing devices are, we will *always* have a certain degree of uncertainty.

Characterizing both known and unknown aspects of a system is key.

Random Variables

- In statistics, we use random variables (RVs) to *model* unknown (random) behavior.
- An RV does not have a known value and exhibits variability.
- Statistical view of the world is significant departure from traditional deterministic viewpoint (commonly used in engineering).
- Under a deterministic view of the world, we *assume* that variables have known and unique values (ignore uncertainty and variability).

- Deterministic Thinking: Temperature is 20^oC (no uncertainty)
- Statistical Thinking: Temperature is 20 ± 0.5^oC (account for uncertainty)

Random Variables

An RV model (denoted as X) has the following elements:

- Set of possible realizations $\omega \in \Omega$ with associated values $x_\omega \in \mathcal{D}_X$.
- Domain $\mathcal{D}_X$ under which realizations x_ω of X "live".
- Probability measure $\mathbb{P} : \Omega \to [0, 1]$ that assigns probability to events.
- Cumulative density function (cdf) $F_X : \mathcal{D}_X \to [0, 1]$ that tells us $\mathbb{P}(X \leq t)$.
- Associated with cdf, there is a probability density function (pdf) $f_X : \mathcal{D}_X \to [0, \infty)$.

Random Variables

Some observations:

- Think of ω as being a pinball that carries data x_ω
- When you drop a pinball, this will "fall" in different locations of the domain $\mathcal{D}_X$
- pdf/cdf tell us how "densely" pinballs accumulate in certain locations of the domain.
- Where in the domain the pinball falls is an event and probability tell us how likely are specific event.
- An event is a subdomain $\mathcal{A} \subseteq \mathcal{D}_X$ (location or set of locations in domain).
- Probabilities are related to how densely pinballs accumulate in parts of domain.

Random Variables

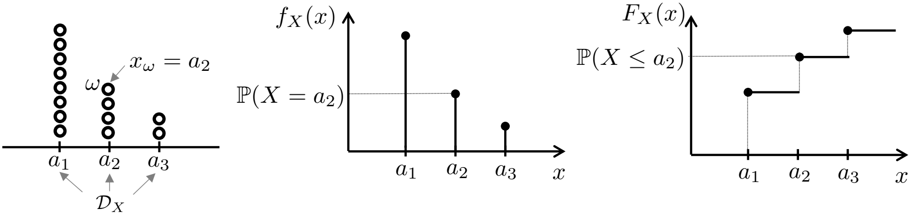


Figure: Illustration of the elements of a random variable.

.

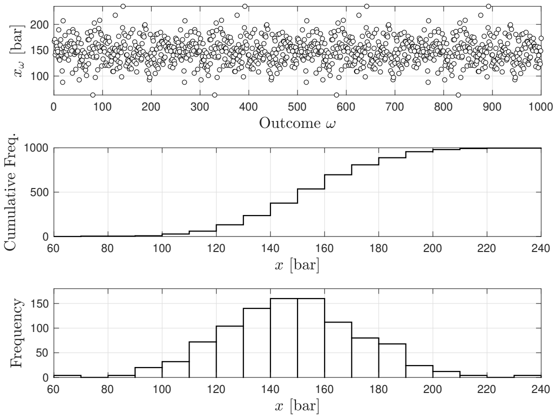
Random Variables

Don't Forget:

- A random variable is a *model* (a characterization) of a random phenomenon.
- To fully define an RV model, you need to know $\mathcal{D}_X$, f_X, and F_X.

Example: Gibbs Reactor `ch1_gibbs_example.m`

- Assume *pressure* varies due to malfunction of flow control and model this as RV X
- Collect 1000 realizations x_ω (data); visualize as cumulative frequency and frequency.
- Cumulative frequency approximates cdf, frequency approximates pdf.



Types of Random Variables

RVs are categorized as multivariate vs. univariate:

- A *multivariate* RV $X = (X_1, X_2, ..., X_n)$ has realizations that generate vector values $x_\omega = (x_{\omega,1}, x_{\omega,2}, ..., x_{\omega,n}) \in \mathbb{R}^n$.
- A *univariate* RV X is a multivariate with $n = 1$ and has realizations generate scalar values $x_\omega \in \mathbb{R}$.

- For a univariate RV, a pinball ω (observation) carries a single number x_ω (e.g., temperature).
- For a multivariate RV, a pinball ω carries a set of numbers $x_\omega = (x_{\omega,1}, x_{\omega,2}, ..., x_{\omega,n})$ (e.g., temperature, pressure, conversion)

For now, we will focus discussion on univariate RVs.

Types of Random Variables

RVs are categorized as continuous vs. discrete:

- A *continuous* RV X is that in which the domain $\mathcal{D}_X$ is continuous; e.g., X has realizations satisfying $0 \leq x_\omega \leq 1$.
- A *discrete* RV X is that in which the domain $\mathcal{D}_X$ is discrete; e.g., X has realizations satisfying $x_\omega \in \{0, 1\}$.

Many RV models are available that apply to different categories.

Type of RV used to characterize uncertainty depends on nature of random phenomenon.

Probability Density of Discrete and Continuous RVs

Discrete X has discrete domain $\mathcal{D}$ and cdf/pdf are discontinuous functions.

The pdf and cdf have the following properties:

1. $f(x) \geq 0, \quad x \in \mathcal{D}$
2. $f(x) = \mathbb{P}(X = x), \quad x \in \mathcal{D}$
3. $\sum_{x \in \mathcal{D}} f(x) = \sum_{x \in \mathcal{D}} \mathbb{P}(X = x) = 1$
4. $\mathbb{P}(X \in \mathcal{A}) = \sum_{x \in \mathcal{A}} f(x), \quad \mathcal{A} \subseteq \mathcal{D}.$
5. $F(t) = \mathbb{P}(X \leq t) = \sum_{x \in \mathcal{D} \mid x \leq t} f(x) \quad t \in \mathbb{R}.$

Probability Density of Discrete and Continuous RVs

A discrete RV is easy to handle computationally (simple summations and counting):

- Since $\mathcal{D}$ is discrete, we have that:

$$\begin{aligned}\mathbb{P}(X \leq t) &= F(t) \\ &= \sum_{x \in \mathcal{D} | x \leq t} f(x) \\ &= \sum_{x \in \mathcal{D}} \mathbf{1}[x \leq t] f(x).\end{aligned}$$

Here, we use the indicator function:

$$\mathbf{1}[x \leq t] = \begin{cases} 1 & \text{if} \quad x \leq t \\ 0 & \text{if} \quad x > t \end{cases}.$$

Probability Density of Discrete and Continuous RVs

- Can compute pdf and cdf by *counting* how many times realizations x_ω of X take a certain value (or are below a certain value):

$$f(x) = \mathbb{P}(X = x) = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \mathbf{1}[x_\omega = x]$$

$$F(t) = \mathbb{P}(X \leq t) = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \mathbf{1}[x_\omega \leq t].$$

- We can thus determine pdf and cdf of discrete RV X directly from data x_ω.

Example Discrete RV

- Consider a discrete RV with domain $\mathcal{D} = \{-1, 0, 1, 2\}$ and pdf:

$$f(x) = \begin{cases} 0 & \text{for } x = -1 \\ 0.4 & \text{for } x = 0 \\ 0.6 & \text{for } x = 1 \\ 0 & \text{for } x = 2 \end{cases}$$

- Pdf satisfies $\sum_{x \in \mathcal{D}} f(x) = 1$ and $f(x) \geq 0$ for all $x \in \mathcal{D}$.
- RV has 10 possible realizations $\Omega = \{1, 2, 3, ..., 10\}$ with associated values x_ω given by $\{0, 1, 0, 0, 1, 1, 1, 1, 1, 0\}$.
- Pdf can also be computed using the observations, for example:

$$f(0) = P(X = 0) = \frac{1}{10} \sum_{\omega \in \Omega} \mathbf{1}[x_\omega = 0] = 0.4$$

- Similarly, use observations to compute the cdf, for example:

$$F(0) = \mathbb{P}(X \leq 0) = \sum_{x | x \leq 0} f(x) = f(-1) + f(0) = 0.4$$

Example Discrete RV

Pdf and cdf of the RV are:

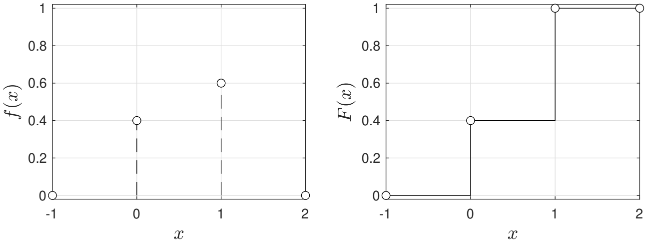


Figure: Pdf and cdf for simple discrete random variable.

.

Note discontinuous nature of these functions.

Probability Density of Discrete and Continuous RVs

Continuous X has continuous domain $\mathcal{D}$ and pdf/cdf are continuous functions.

The pdf and cdf have the following properties:

1. $f(x) \geq 0, \quad x \in \mathcal{D}.$
2. $f(x) = \frac{dF(x)}{dx}, \quad x \in \mathcal{D}.$
3. $\displaystyle\int_{x \in \mathcal{D}} f(x)dx = 1.$
4. $\displaystyle\mathbb{P}(X \in \mathcal{A}) = \int_{x \in \mathcal{A}} f(x)dx, \quad \mathcal{A} \subseteq \mathcal{D}.$
5. $F(t) = \int_{x \in \mathcal{D} | x \leq t} f(x)dx.$

Probability Density of Discrete and Continuous RVs

- The 2nd property tells us that pdf $f(x)$ for a continuous RV X is not $\mathbb{P}(X = x)$ (as in the discrete case).
- Instead, the pdf is the derivative of the cdf and thus:

$$\begin{aligned} f(x)\Delta x &\approx \mathbb{P}(X \le x + \Delta x) - \mathbb{P}(X \le x) \\ &= \mathbb{P}(x \le X \le x + \Delta x). \end{aligned}$$

  We thus have that pdf tell us the probability that X is in a *neighborhood* of x.

- Not that setting $\Delta x \to 0$ implies that $\mathbb{P}(X = x) = 0$.
- The 3rd,4th, 5th properties are analogous to the discrete case but we see that summation are replaced with integration operations.

Probability Density of Discrete and Continuous RVs

A continuous RV is more difficult to handle computationally since it involves integrals instead of summations (this prevents the use of simple counting operations).

- However, computations are analogous to those of the discrete case. For instance:

$$\begin{aligned}\mathbb{P}(X \leq t) &= \int_{x \in \mathcal{D} | x \leq t} f(x) dx \\ &= \int_{x \in \mathcal{D}} \mathbf{1}[x \leq t] f(x) dx.\end{aligned}$$

- Continuous RVs are often approximated using discretization. For instance:

$$f(x) \approx \frac{F(x + \Delta x) - F(x)}{\Delta x}.$$

- We will see that we can approximate continuous pdfs/cdfs from data x_ω.

Example Continuous RV

- Consider continuous RV with domain $\mathcal{D} = (-\infty, \infty)$ and pdf of the form:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

- It is easy to see that the pdf satisfies $f(x) \geq 0$ for any $x \in \mathcal{D}$ and one can show that $\int_{x \in \mathcal{D}} f(x) dx = 1$ (this is not that easy to show).

- The cdf associated with this pdf is given by:

$$\begin{aligned} F(t) &= \int_{x \leq t} f(x) dx = \int_{-\infty}^{t} f(x) dx \\ &= \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{t}{\sqrt{2}}\right)\right), \end{aligned}$$

- One can show $f(x) = dF(x)/dx$ and that cdf satisfies $0 \leq F(t) \leq 1$ for any $t \in \mathcal{D}$.

Example Continuous RV

Pdf and cdf of the RV are:

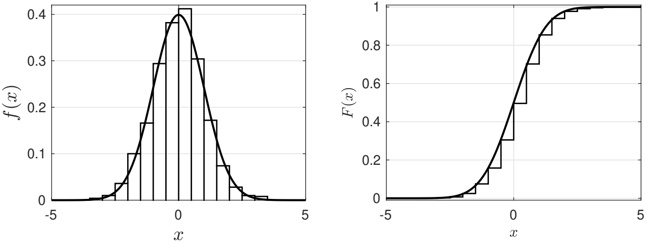


Figure: Pdf and cdf of continuous RV and corresponding discrete approximations.

Continuous pdf and cdf are approximated using discretized versions with $\Delta x = 0.5$.

From Data to Models

- In practical situations, we have access to observations (data) of our system.
- Our goal is to use such data to gain knowledge (understanding) of our system.
- Our knowledge will be extracted from data in the form of *models* of two types:
  - *Structural models*: provide a characterization of structural dependencies between variables in our system (mechanistic or empirical)
  - *Random variable (uncertainty) models*: provide a characterization of behavior that cannot be explained by structural models

Important: Both models are necessary to make proper predictions and decisions.

From Data to Models

- We will begin our discussion with random variable (RV) models.
- Assume that we have available observations (data) $x_\omega,\ \omega \in \mathcal{S}$.
- What type of an RV model are the observations following?
- How to obtain a cdf $F(x)$ and pdf $f(x)$ for an RV model from available data?

From Data to Models

- Goal is to use available data x_ω to postulate a *theoretical* RV model X.
- RV X is a *model* of a random phenomenon that generates the observed data.
- Many models are available to capture diverse phenomena seen in real life.
- A widely used model is that of a Gaussian RV.
- Model assumes that X is continuous, has domain $\mathcal{D} = (-\infty, \infty)$, and has pdf:

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

  where $\mu, \sigma \in \mathbb{R}_+$ are parameters of the model.
- Parameters define behavior of the RV and can be "tuned" to match our data.
- The cdf associated to the Gaussian model is:

$$F(x) = \frac{1}{2}\left(1 + \text{erf}\left(\frac{(x-\mu)/\sigma}{\sqrt{2}}\right)\right), \quad x \in \mathcal{D},$$

- **What other theoretical models do you know?**

From Data to Models

- Our goal is to investigate whether the available observations $x_\omega,\ \omega \in \mathcal{S}$ follow the pdf and cdf of a theoretical model that we postulate.
- To verify hypothesis, we use the data to construct *empirical* approximations for the pdf $f(x)$ and cdf $F(x)$ of the theoretical model and verify if these match.
- Empirical approximations (a.k.a data-driven or sample-based approximations) are denoted as $\hat{f}(x)$ and $\hat{F}(x)$ and these are used to *estimate* $f(x)$ and $F(x)$.

From Data to Models

The approach to construct $\hat{f}(x)$ and $\hat{F}(x)$ from data (for a continuous RV) can be summarized as follows:

1. Construct *empirical* domain $\hat{\mathcal{D}}$; this is the domain covered by observations x_ω, $\omega \in \mathcal{S}$. This gives us an approximation of the domain $\mathcal{D}$ of the RV model. Discretize the domain $\hat{\mathcal{D}}$ into bins of size Δx.

2. Construct an empirical cdf:

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

   This is the number of observations x_ω that take a value below x (normalized by S).

   The number of observations is known as the *cumulative frequency* and is given by:

$$\sum_{\omega \in \mathcal{S}} \mathbf{1}[x_\omega \leq x]$$

From Data to Models

3. Construct the empirical pdf:

$$\hat{f}(x) = \frac{\frac{1}{S}\sum_{\omega \in \mathcal{S}} \mathbf{1}[x \leq x_\omega \leq x + \Delta x]}{\Delta x}$$

where Δx is the size of the bin interval.

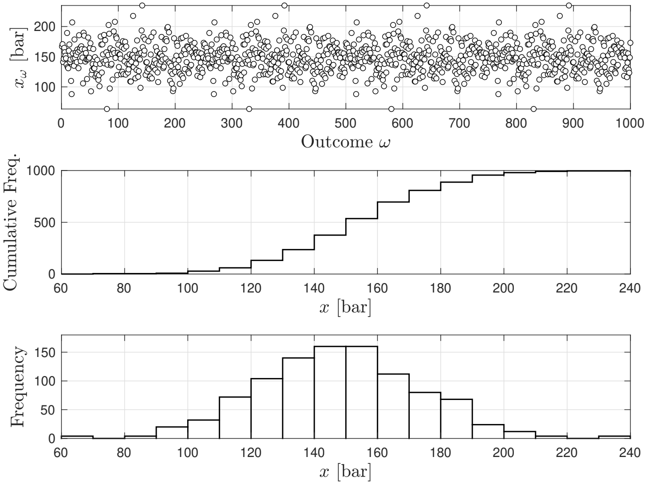
The number of observations in $[x, x + \Delta x]$ is known as the *frequency* and is given by:

$$\sum_{\omega \in \mathcal{S}} \mathbf{1}[x \leq x_\omega \leq x + \Delta x]$$

The procedure to obtain the empirical pdf/cdf of a discrete RV is easier, as the domain does not need to be discretized.

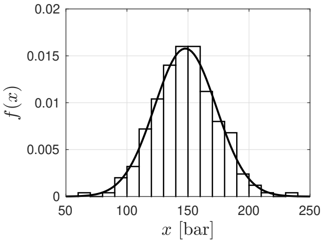
Example: Gibbs Reactor `ch1_gibbs_example.m`

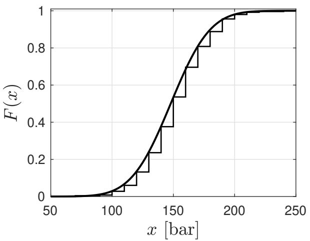
Here are S =1,000 observations x_ω along with the frequency and cumulative frequency.



Example: Gibbs Reactor `ch1_gibbs_example.m`

- Obtain empirical cdf by normalizing cumulative frequency with $S = 1000$
- Obtain empirical pdf by normalizing frequency with $S = 1000$ and $\Delta x = 10$





- Empirical pdf and cdf match the pdf and cdf of a Gaussian RV model.
- We conclude that our random phenomenon (pressure) behaves as a Gaussian RV.

Summarizing Statistics (Basic)

- Pdf and cdf are *functions* that fully characterize an RV X. However, in practice, we might be interested in using values (and not functions) to describe X.
- This is done by using *summarizing statistics* (a.k.a. descriptive statistics). Popular summarizing statistics are the expected value and variance:

For a discrete RV we have:

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

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

For a continuous RV we have:

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

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

Summarizing Statistics (Sample Approximations)

Need theoretical pdf of X to compute expected value, variance, and SD (theoretical statistics).

However, if we have data x_ω, $\omega \in \mathcal{S}$, we can approximate theoretical statistics by using empirical estimates:

- *Empirical Mean:*

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

- *Empirical Variance and Standard Deviation:*

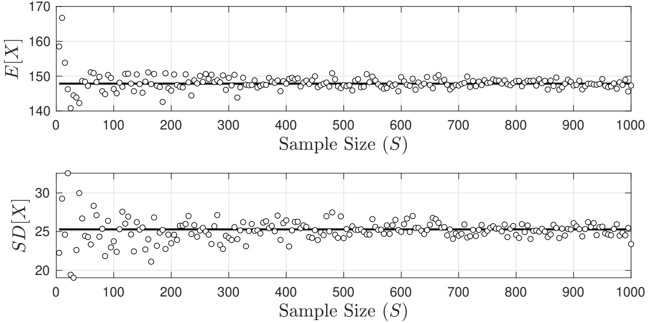
$$\hat{\mathbb{V}}_X := \frac{1}{S} \sum_{\omega \in \mathcal{S}} (x_\omega - \hat{\mathbb{E}}_X)^2, \qquad \hat{\mathbb{SD}}_X = \sqrt{\hat{\mathbb{V}}_X}$$

Intuition tells us that approx improve as we accumulate data (as S becomes large). We will see later that this is indeed the case.

Statistics can be related to model parameters (e.g., for a Gaussian: $\mathbb{E}_X = \mu$ and $\mathbb{V}_X = \sigma^2$). As such, empirical estimates of statistics can be used to obtain parameters.

Example: Gibbs Reactor `ch1_gibbs_example.m`

Use reactor pressure data to compute empirical estimates for the mean $\hat{\mathbb{E}}_X$ and standard deviation $\hat{\mathbb{SD}}_X$ (we explore effect of using increasing amounts of data S).



Empirical mean and SD converge to the theoretical values $\mathbb{E}_X = 148$ and $\mathbb{SD}_X = 25$

Estimates are not accurate for small sample sizes.

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 equation is the α-quantile of RV 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 cdf can have a "staircase" form, there might be multiple values of x satisfying $F_X(x) = \alpha$. Consequently, 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:

  - $\mathbb{Q}_X(0.5)$ 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 empirical cdf $\hat{F}_X(x)$ to estimate empirical quantiles $\hat{\mathbb{Q}}_X(\alpha)$.

Summarizing Statistics (Mode and Moments)

- The mode of an RV is the outcome of maximum probability:

$$\mathbb{MO}_X \in \underset{x \in \mathcal{D}_X}{\operatorname{argmax}}\, f_X(x).$$

Some RVs are unimodal (one peak) and some are multimodal (multiple peaks).

- *Central moments* are an important family of summarizing statistics
  - The central moments of X are given by:

$$\mathbb{CMO}_k[X] = \mathbb{E}[(X - \mathbb{E}[X])^k], \qquad k = 1, 2, 3, 4, \ldots,$$

    Note that the second ($k = 2$) central moment is the variance.
  - The standardized moments of X are given by:

$$\mathbb{SMO}_k[X] = \frac{\mathbb{E}[(X - \mathbb{E}[X])^k]}{\mathbb{SD}[X]^k}, \qquad k = 1, 2, 3, 4, \ldots,$$

    The third ($k = 3$) standardized moment is known as *skewness* and the fourth ($k = 4$) is known as *kurtosis*. Skewness is a measure of symmetry of the pdf is while kurtosis is a measure of the nature of the tails of the pdf.

Summarizing Statistics (Mode and Moments)

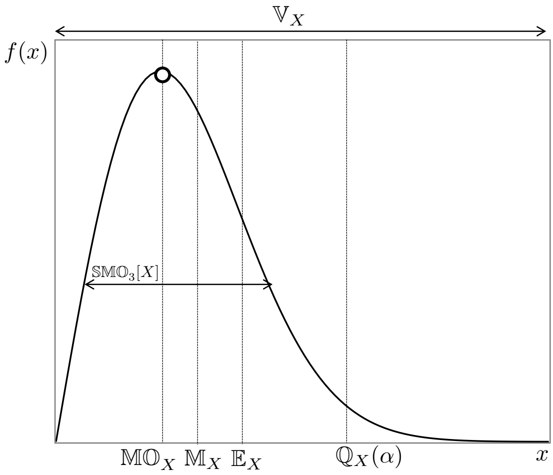


Figure: Features of the probability density function that different summarizing statistics capture.

.

From Knowledge to Decisions

We now have a characterization of a given random phenomenon (our RV model X) affecting a system of interest and we would like to exploit this to make decisions.

In this context, it is important to make a couple of important observations:

- Uncertainty *propagates* through systems in complex ways.
- Uncertainty can be *mitigated* via design or control decisions.

Uncertainty Propagation and Mitigation

- Consider propagation of X through system $\varphi(X, u)$:

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

  where u is a mitigating action (decision) and Y is the system output.

- We make the following observations:

- Output Y is an RV if the input X is an RV.
- Nature of Y (its cdf, pdf, and domain) depends on system φ. Some systems magnify uncertainty and variability while others might damp it.
- Nature of Y depends on action u. Can use action to mitigate/manipulate uncertainty of Y.

Uncertainty Propagation and Mitigation

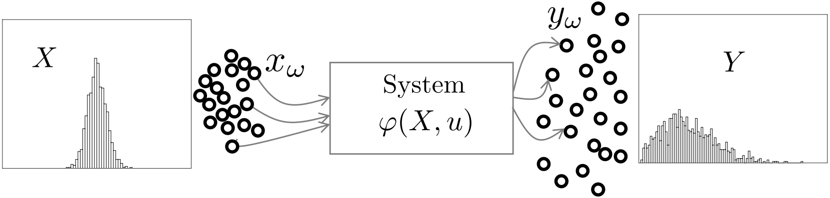


Figure: Illustration of propagation of a random variable X through a system. The propagation results in a random output Y that has different characteristics (e.g., higher variability/uncertainty.

.

Uncertainty Propagation and Mitigation

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

1. For a given decision u, perform simulations of the form:

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

2. Use y_ω to compute sample 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:

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

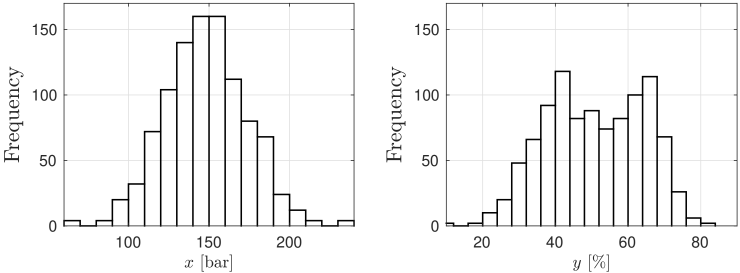
   - Empirical cdf:

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

The above procedure is known as *Monte Carlo (MC) simulation* and is widely used to estimate diverse quantities of interest for RVs.

Example: Gibbs Reactor `ch1_gibbs_example.m`

- Empirical pdf and cdf for pressure (input X) and conversion (output Y).
- Note change in behavior of Y; pdf of X is unimodal, while pdf of Y is bimodal.



- Complex behavior of the conversion pdf is the result of strong nonlinear behavior

Decision-Making under Uncertainty

We would like now to find a decision u that manipulates $Y(u) = \varphi(X, u)$ in some desirable way (e.g., minimizes uncertainty/variance).

Consider the 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? What do we mean by the "best"?

Decision-Making Under Uncertainty

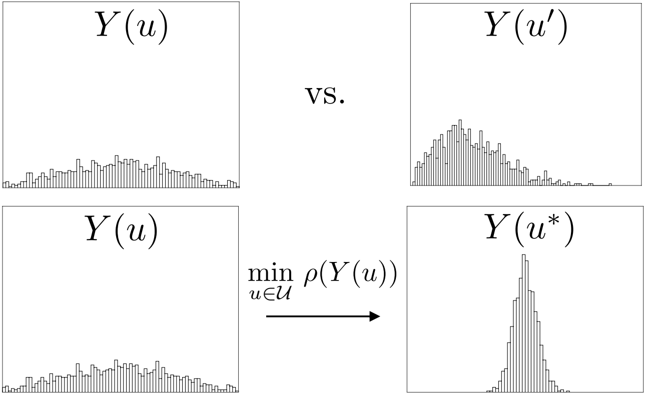


Figure: Paradigms for decision-making under uncertainty. Comparison between decisions u, u' and associated outputs $Y(u), Y(u')$ (top). Find best decision u^* that shapes $Y(u^*)$ in a desirable way (bottom).

.

Decision-Making under Uncertainty

- If we assume *deterministic setting* (no uncertainty), then $Y(u)$ and $Y(u')$ will each take a single value $y(u)$ and $y(u')$ and one would select, *unambigously*, the one with smaller (or larger) value. For instance, one would select u over u' if $y(u) \leq y(u')$.

- In a *setting under uncertainty* this is no longer possible because $Y(u)$ and $Y'(u)$ have multiple possible outcomes ($Y(u)$ and $Y'(u)$ are RV models)

- Concept of "better" under uncertainty is ambiguous and the mathematical statement $Y(u) \leq Y(u')$ does not even make sense.

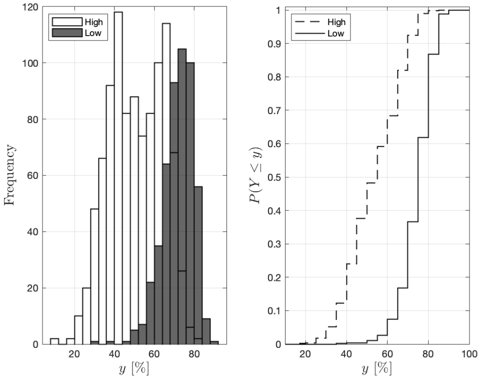
- Does $Y(u) \leq Y(u')$ mean that all outcomes of $Y(u)$ are lower than those $Y(u')$? Does it mean that a subset of outcomes are lower?

- When making a decision under uncertainty, need to capture information embedded in $Y(u)$ and $Y(u')$.

- This requires comparing RVs consistently; e.g., by using their cdfs or by using *risk measures* (summarizing statistics).

Example: Gibbs Reactor `ch1_gibbs_example.m`

- Can counteract variability in pressure X by operating at modifying temperature (u).
- We compare empirical pdf and cdf for conversion at low $Y(u)$ and high $Y(u')$ temp
- Should we operate at low or high temp?
- By comparing cdfs, we can see that operating at low temp is *consistently* more likely to achieve higher yields.



Decision-Making under Uncertainty

- We might not only be interested in comparing decisions, but we might want to find the *best* possible decision.
- To decide what is "best", we select a measure of the output $Y(u)$ (a summarizing statistic) that we denote as $\rho(Y(u))$.
- We find best decision by solving optimization problem:

$$\min_{u \in \mathcal{U}} \ \rho(Y(u)).$$

- $\mathcal{U}$ is set of possible decisions that we can choose from. In Gibbs reactor example, $\mathcal{U} = \{583, 613\}$.
- If we select $\rho(Y(u)) = \mathbb{SD}[Y(u)]/\mathbb{E}[Y(u)]$, optimization problem finds decision u such that $Y(u)$ minimizes the *coefficient of variation* (CV).
- In Gibbs reactor, operating at a low temp yields a CV of 0.07, while operating at high temp yields 0.10 (optimal decision is $u^* = 583$ K).
- We will see later that $\rho(Y(u))$ is a risk measure that aims to model attitudes towards risk by decision-makers.