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 2: Univariate Random Variables

Models of Univariate RVs

- A wide range of RV models can be used to capture trends observed in real-life phenomena (e.g., diffusion, mixing, extreme events, particle sizes, failures).
- Type of model used depends on nature of phenomenon observed.
- Some of these models are "universal", in the sense that they capture limiting behavior for many systems.
- We discuss properties of univariate RVs and connections between them.
- Many RVs result from algebraic transformations of other RVs (e.g., summations, logarithmic, exponential, quadratic).
- Later we explore estimation techniques to determine if an RV model matches data.
- Selection of an appropriate RV model is analogous to how we select a physical model that is appropriate for a specific application.

Discrete Uniform

The (discrete) uniform RV is simplest model that one can think of to model phenomena.

Model inspired by the fact that, in many applications, we often have limited data to characterize phenomenon.

- Imagine all we know is that realizations of RV can take discrete values $k = a, a+1, a+2, ..., b-1, b$ and that these values are equally likely.

- Accordingly, pdf is given by:

$$\mathbb{P}(X = x) = f(x) = \frac{1}{n}, \; x \in \mathcal{D}.$$

  with domain $\mathcal{D} = \{a, a+1, a+2, ..., b-1, b\}$ and $n = b - a + 1$.

- We express fact that X follows a uniform model as $X \sim \mathcal{U}(a, b)$, where $a, b \in \mathbb{Z}$ are model parameters (adjusted to match data).

- Cdf for uniform RV model is given by:

$$\mathbb{P}(X \leq x) = F(x) = \frac{x - a + 1}{n}, \; x \in \mathcal{D}.$$

Discrete Uniform

- Mean of uniform RV is:

$$\begin{aligned}\mathbb{E}_X &= \sum_{x \in D} f(x) \cdot x = \sum_{x=a}^{b} \frac{1}{n} x \\ &= \frac{1}{n}\left(a + (a+1) + (a+2) + \cdots + (b-2) + (b-1) + b\right) \\ &= \frac{1}{2}(a+b)\end{aligned}$$

- Median of RV is value at which $F(x) = 1/2$ and thus:

$$\frac{1}{2} = \frac{x - a + 1}{n}.$$

  This occurs at $\mathbb{M}_X = \frac{1}{2}(a+b)$.

- Variance of RV is:

$$\mathbb{V}_X = \sum_{x \in D} f(x) \cdot (x - \mathbb{E}_X)^2 = \frac{1}{n} \sum_{x=a}^{b} \left(x - \frac{1}{2}(a+b)\right)^2 = \frac{1}{12}(n^2 - 1).$$

- Uniform model has many applications and is used in sampling (data collection) schemes (e.g., want equally-likely samples).

Discrete Uniform

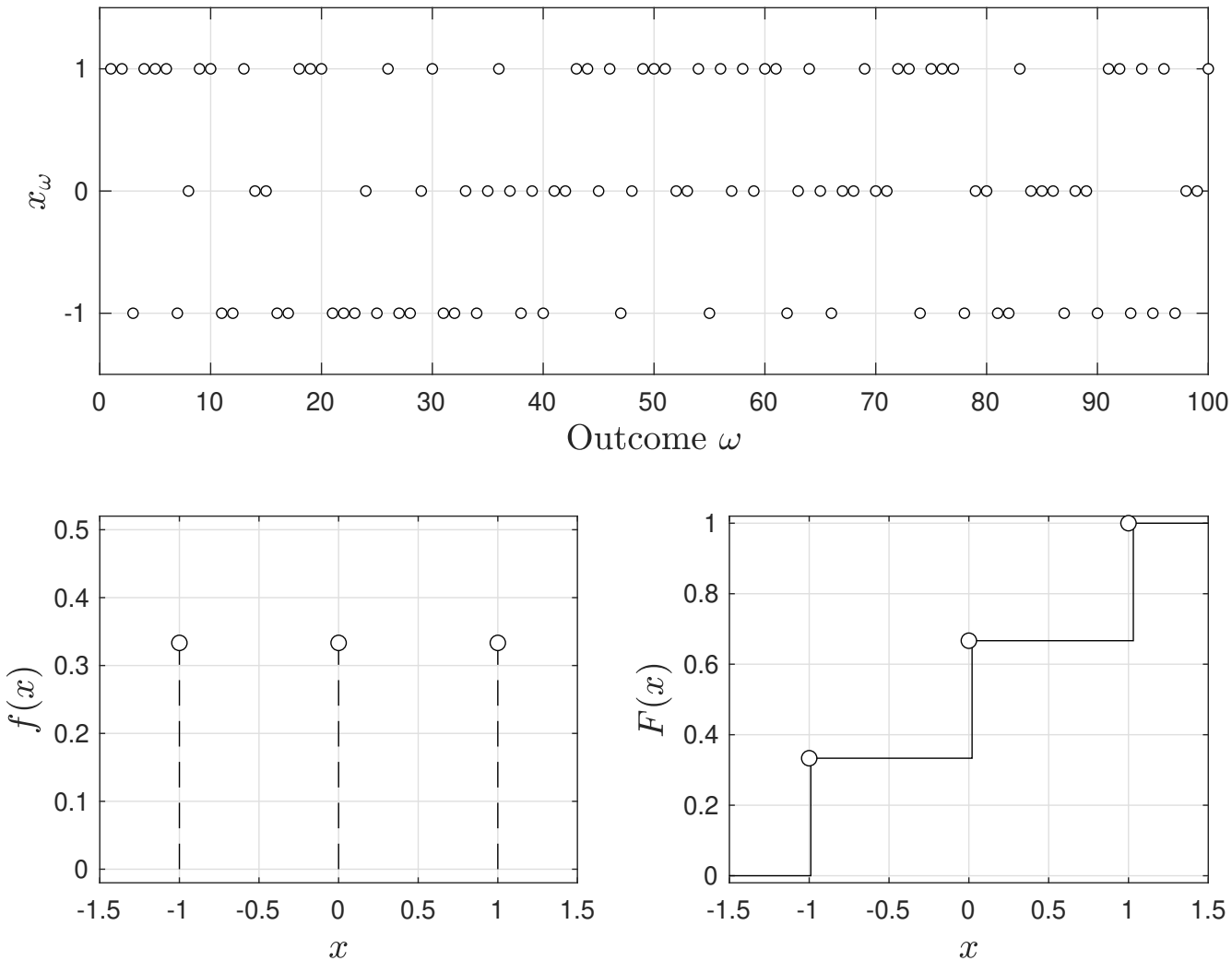


Figure: Pdf and cdf for a uniform random variable with $a = -1, b = 1$.

.

Example Discrete Uniform `ch2_example_uniform.m`

- Have machine that appears to fail randomly and you have collected weekly data for over 3 years (156 weeks).
- You have not identified any pattern but you have noticed that the machine has not failed 79 out of 156 weeks.
  - Is it suitable to model machine status as a uniform RV?
  - If so, what would be parameters?

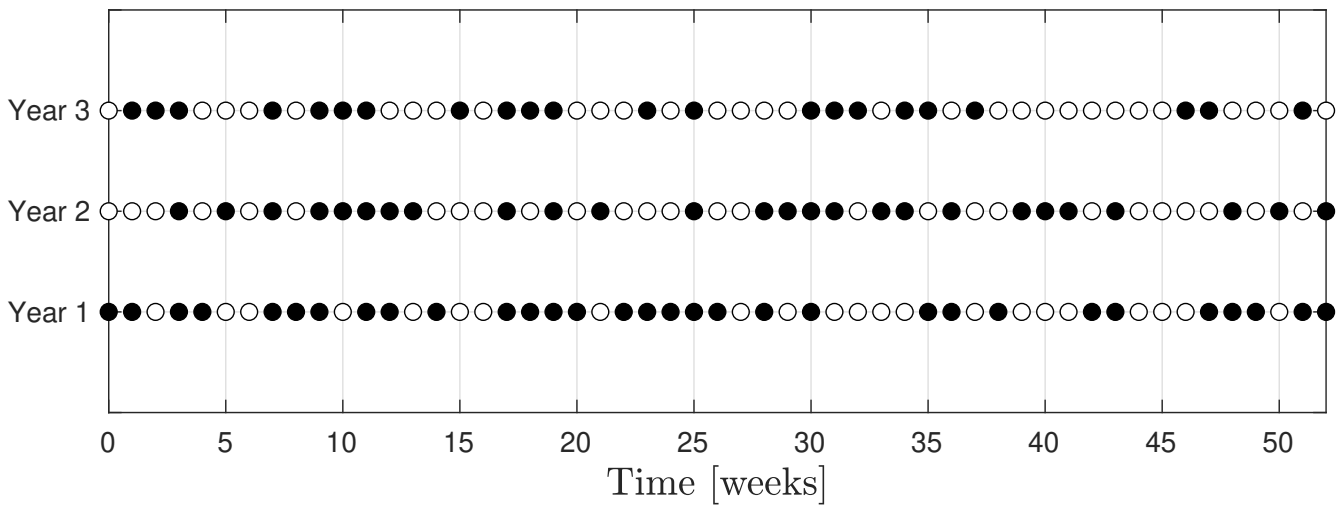


Figure: Failures (dark dots) recorded over three years.

.

Example Discrete Uniform `ch2_example_uniform.m`

- Define machine status as $X = 0$ when it fails and $X = 1$ when it works.
- We thus have $\mathcal{D} = \{0, 1\}$ and thus $a = 0$, $b = 1$ and $n = 2$; i.e., $X \sim \mathcal{U}(0, 1)$.
- For uniform RV model we have that $\mathbb{P}(X = x) = 1/n$ for all $x \in \mathcal{D}$ and thus $\mathbb{P}(X = 0) = 1/2$ and $\mathbb{P}(X = 1) = 1/2$.
- From data we see that empirical pdf is:

$$\hat{\mathbb{P}}(X = 0) = 79/156 = 0.5064$$
$$\hat{\mathbb{P}}(X = 1) = (156 - 79)/156 = 0.4936.$$

- Data suggests that machine status follows behavior of $\mathcal{U}(0, 1)$ (it is equally likely that it fails or works).

Bernoulli

- Another simple RV model is the Bernoulli RV.
- You run an experiment with two possible outcomes Y (yes) or N (no).
- We have $X = 1$ if outcome is Y and $X = 0$ if outcome is N.
- Define $p \in [0, 1]$ as probability of outcome Y and thus probability of N is $(1 - p)$.
- Domain is thus $\mathcal{D} = \{0, 1\}$ and pdf is:

$$f(x) = \begin{cases} (1-p) & \text{if } x = 0 \\ p & \text{if } x = 1 \end{cases}$$

Pdf can be written in compact form:

$$f(x) = p^{\mathbf{1}[x=1]}(1-p)^{\mathbf{1}[x=0]}.$$

Bernoulli

- We say X follows a Bernoulli model as $X \sim \mathcal{B}(p)$, where p is the parameter.
- Mean and variance of this RV are:

$$\mathbb{E}_X = \sum_{x \in \mathcal{D}} x \cdot f(x) = 0 \cdot (1-p) + 1 \cdot p = p$$

$$\mathbb{V}_X = \sum_{x \in \mathcal{D}} (x - \mathbb{E}_X)^2 \cdot f(x) = p \cdot (1-p).$$

- Bernoulli RV is used in sampling and forms basis of more sophisticated RVs.

Bernoulli

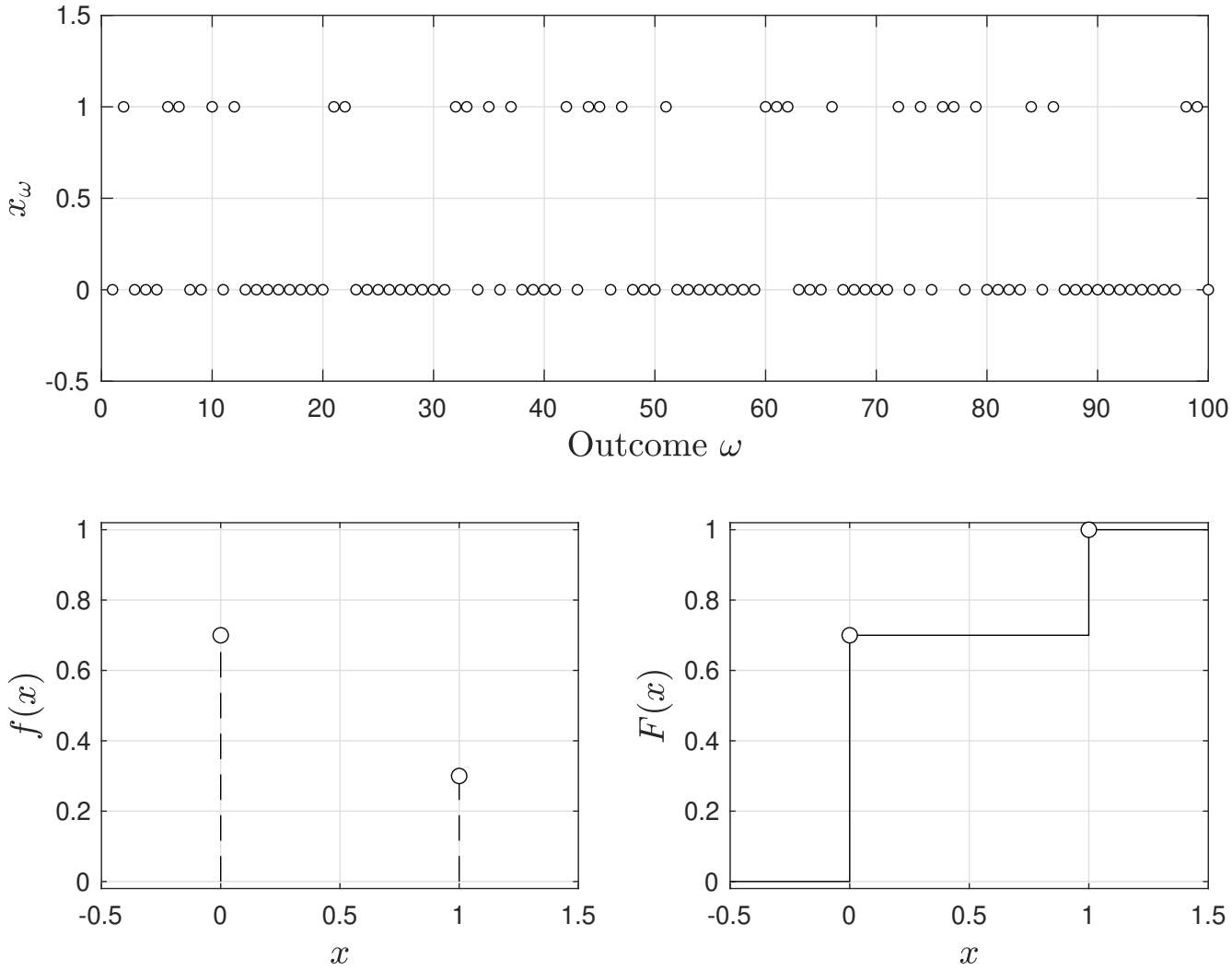


Figure: Pdf and cdf for Bernoulli random variable with $p = 0.3$.

.

Example Bernoulli `example_bernoulli`

- Consider the previous machine status problem and the following questions:
  - Is it suitable to model the machine status as a Bernoulli RV?
  - If so, what would be its parameters?
  - What makes this different than a uniform RV?
- Define status as $X = 0$ when fails and $X = 1$ when works (domain is $\mathcal{D} = \{0, 1\}$).
- For a Bernoulli RV, we must have that $\mathbb{P}(X = 0) = p$ and $\mathbb{P}(X = 1) = 1 - p$.
- From our data we have:

$$\hat{\mathbb{P}}(X = 0) = 79/156 = 0.5064$$
$$\hat{\mathbb{P}}(X = 1) = (156 - 79)/156 = 0.4936.$$

- Machine status thus seems to behave as a Bernoulli RV $\mathcal{B}(0.5)$.
- In this example, uniform and Bernoulli models are identical.
- Uniform model is a generalization of Berrnoulli model.

Hypergeometric

- Arises in applications in which one draws random samples from a population.
- For instance, in a manufacturing, we choose a set of products to inspect for quality control.
- Main issue is that population can be large and we cannot afford to inspect all.
- We want to select a subset and infer from this behavior of general population.
- Assume population is composed of mutually exclusive groups Y or N.
- Total units in population is N_t and define total number of units of type Y as N_y. We thus have that $(N_t - N_y)$ are of type N.
- The fraction of Y items is thus $p = N_y/N_t$.
- We draw sample of n items and check how many items are of type Y.
- The RV X is the number of items of type Y in our sample.

Hypergeometric

- Total number of distinct ways of choosing n items from the population is:

$$N_\Omega = \begin{pmatrix} N_t \\ n \end{pmatrix} = \frac{N_t!}{n! \cdot (N_t - n)!}.$$

  This is the size of sample set Ω (in general an extremely large number).

- For instance, for $N_t = 100$ and $n = 10$ we have $N_\Omega = 1.73 \times 10^{13}$.

- Probability of obtaining x items of type Y in sample is given by pdf:

$$f(x) = \frac{\begin{pmatrix} N_y \\ x \end{pmatrix} \cdot \begin{pmatrix} N_t - N_y \\ n - x \end{pmatrix}}{\begin{pmatrix} N_t \\ n \end{pmatrix}}, \quad x \in \mathcal{D}.$$

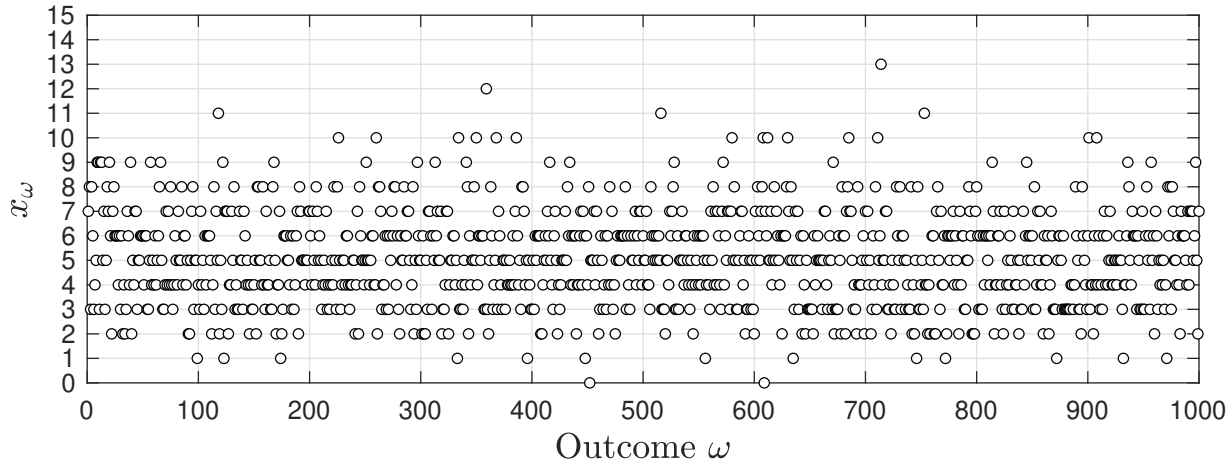
  where the domain is $\mathcal{D} = \{\max(0, n + N_y - N_t), ..., \min(n, N_y)\}$.

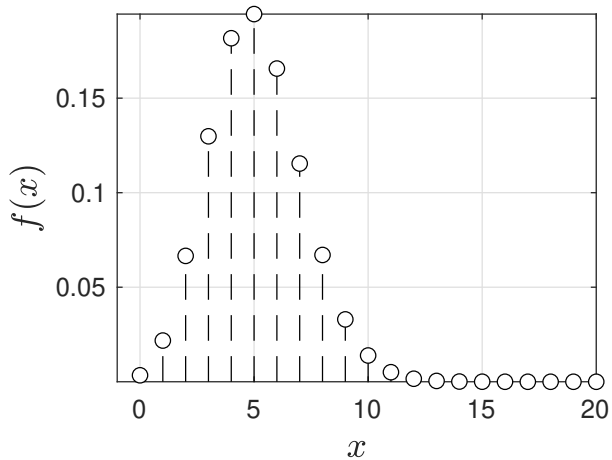
- Hypergeometric RV is denoted as $X \sim \mathcal{H}(N_y, N_t, n, p)$.

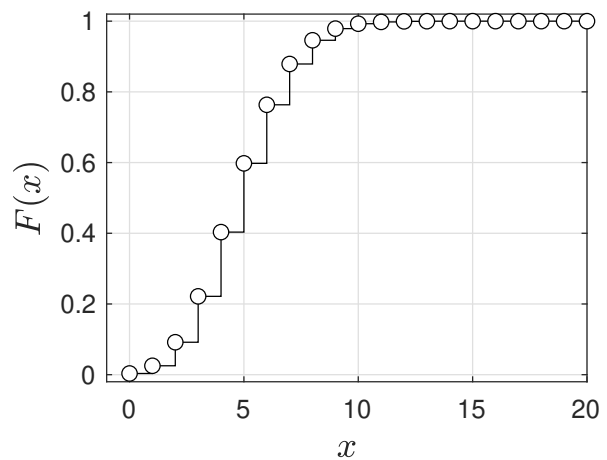
Hypergeometric

- Below is pdf and cdf of a hypergeometric RV for $N_t = 500$, $N_y = 50$, and $n = 50$
- Fraction of items of type Y in population is $p = N_y/N_t = 0.1$.
- Pdf grows quickly with x, reaches peak (mode) at $x = 5$ and then decays rapidly.
- This indicates that it is most likely to find exactly 5 items of type Y.
- More likely to find 5 items of type Y than 1 item of type Y in our sample.

Hypergeometric







.

Hypergeometric

- Expected value of X (average number of items Y in our sample) is:

$$\mathbb{E}_X = n \cdot p.$$

- Value n is known (we control this) but p is usually not known (need to sample the entire population to determine N_y, this would be expensive).
- Parameter p can be estimated from empirical estimates of $\mathbb{E}_X$.
- For instance, assume you pick different samples S (each of size n) and record the number of items of type Y in each sample (given by x_ω, $\omega \in \mathcal{S}$).
- Empirical mean is $\hat{\mathbb{E}}_X = \frac{1}{S} \sum_{\omega \in \mathcal{S}} x_\omega$ and thus we can estimate p as:

$$\hat{p} = \hat{\mathbb{E}}_X / n.$$

Hypergeometric

- Variance of the hypergeometric RV is given by:

$$\mathbb{V}_X = n \cdot p \cdot (1-p) \cdot \frac{N_t - n}{n-1}$$

- Note that variance is zero if $n = N_t$, why?
- If we pick all items $n = N_t$ then there is no variability in the number of items of type Y that we observe.
- In general, variance will decrease as n approaches N_t.
- Can you estimate p from the empirical variance $\hat{\mathbb{V}}_X$?

Example Hypergeometric `ch2_example_hypergeometric.m`

Cannot afford to monitor machine status every week for 3 years (time-consuming)

- If you inspect the machine 5 times (selected at random)... What is the probability that the machine will be in failure mode 5 out of 5 times that you inspected?
- if you inspect the machine 10 times (selected at random)... What is the probability that the machine will be in failure mode 10 out of 10 times that you inspected?

We inspect at random, it is thus appropriate to model number of failures as $X \sim \mathcal{H}(N_y, N_t, p, n)$.

- We define $N_t = 176$, $N_y = 79$, $p = N_y/N_t \approx 0.50$, and $n = 5$.
- Probability that machine is in failure mode 5 out of 5 times inspected is:

$$\mathbb{P}(X = 5) = f(5) = 0.1312.$$

- Probability of observing 10 out of 10 failures is:

$$\mathbb{P}(X = 10) = f(10) = 8 \times 10^{-4}.$$

Binomial

- In Bernoulli setting, a trial has possible outcomes Y (yes) or N (no).
- Imagine we perform a *sequence* of n Bernoulli trials and record number of times that outcomes are of type Y. Each trial is *independent* of the others.
- Binomial RV X counts number of trials of type Y in our n trials.
- Domain is $\mathcal{D} = \{0, 1, ..., n\}$ and probability that number of Y trials is x is:

$$f(x) = \left(\begin{array}{c} n \\ x \end{array} \right) \cdot p^x \cdot (1-p)^{(n-x)}, \quad x \in \mathcal{D}.$$

- We denote Binomial RV as $X \sim \text{Bi}(n, p)$.
- Expected value and variance are:

$$\mathbb{E}_X = n \cdot p, \qquad \mathbb{V}_X = n \cdot p \cdot (1-p).$$

Binomial

- Binomial RV is a generalization of Bernoulli RV; $\mathrm{Bi}(1,p) = \mathcal{B}(p)$.
- Consider sequence of independent Bernoulli RVs $X_1, X_2, ..., X_n$ (each with probability p), then $X = \sum_{i=1}^{n} X_i$ is a binomial RV $X \sim \mathrm{Bi}(n,p)$.
- Binomial RV is limiting case of a hypergeometric RV in the sense that:

$$\lim_{N_t \to \infty} \mathcal{H}(N_y, N_t, n, p) = \mathrm{Bi}(n,p).$$

  i.e., binomial setting is equivalent to hypergeometric setting when size of the population is large.

- In hypergeometric setting, we choose n items at once while in binomial setting we choose n items by choosing one item at a time (resetting system each time).
- Intuitively, when population N_t is large, size of the sample does not matter as much and picking n items at once or as an independent sequence is equivalent.

Binomial

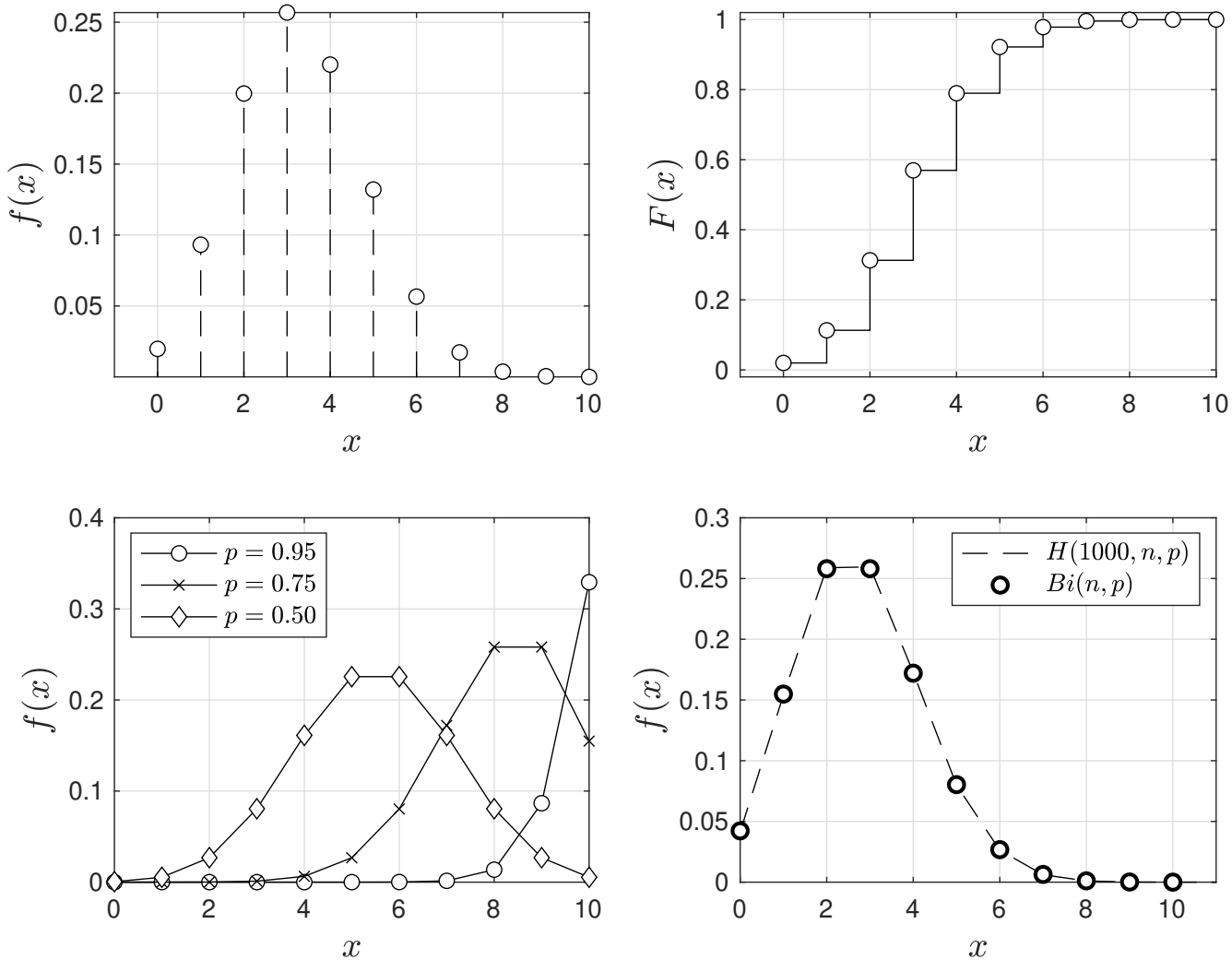


Figure: Pdf and cdf for binomial RV with $n = 10$ and $p = 0.3$ (top). Binomial pdf for different p (bottom-left). Convergence of hypergeometric pdf to binomial pdf for $N = 1000$ (bottom-right).

Example Binomial `ch2_example_binomial.m`

Consider machine status problem; recall we have established that status is $\mathcal{B}(0.5)$.

- What is probability of finding 2 failures in a sequence of 10 times?
- What is probability of finding 10 failures in a sequence of 10 times?
- What is average number of failures in a sequence of 10 times?

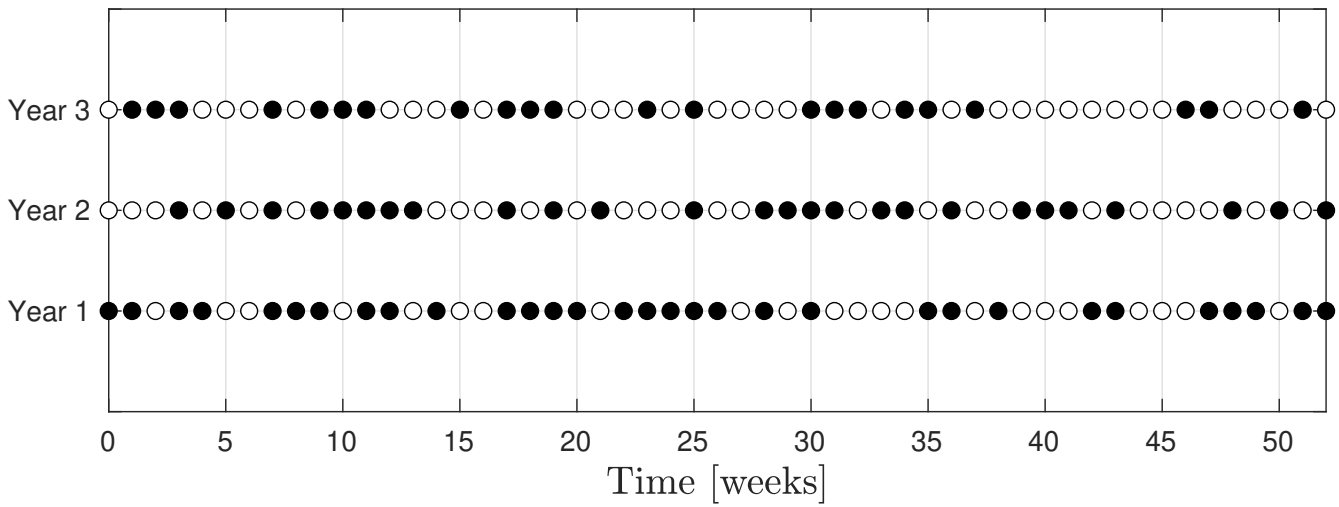


Figure: Failures (dark dots) recorded over three years.

.

Example Binomial `ch2_example_binomial.m`

Status is a Benoulli RV with $p = 0.5$ and thus number of failures n-sequence is $\mathrm{Bi}(n, p)$.

- Probability of finding 2 failures in a sequence of $n = 10$ times is:

$$\mathbb{P}(X = 2) = f(2) = 0.0439.$$

- Probability of finding 10 failures is:

$$\mathbb{P}(X = 10) = f(10) = 9.76 \times 10^{-4}$$

- Average number of failures is:

$$\mathbb{E}_X = n \cdot p = 10 \cdot 0.5 = 5$$

Poisson

- Poisson RV model is important generalization of binomial RV.
- In binomial setting, we consider sequence of n Bernoulli trials and record number of occurrences of type Y.
- Imagine trials are taken at consecutive times $z_1, z_2, ..., z_n$ and define Δz as time between trials (sampling interval).
- If we have continuous time interval $[0, z]$ (of total length z), then number of trials in interval is $n = z/\Delta z$.

Poisson

- Poisson setting is limiting case of $\Delta z \to 0$ ($n \to \infty$); specifically, we seek to understand probability of occurrences over infinitesimally small sampling intervals.
- Understanding behavior over small intervals is important because we might miss occurrences if sampling interval Δz is long.
- Poisson RV gives number of occurrences of type Y in interval of length z.
- Domain is $\mathcal{D} = \{0, 1, 2, ..., \infty\}$ and pdf:

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x \in \mathcal{D}.$$

Denote Poisson RV as $X \sim \mathcal{P}(\lambda)$ with parameter $\lambda \in \mathbb{R}_+$.

Poisson

- Expected value and variance are:

$$\mathbb{E}_X = \lambda$$
$$\mathbb{V}_X = \lambda.$$

- Parameter λ is average number of occurrences found in interval.

- Parameter λ is also often expressed as:

$$\lambda = \eta \cdot z$$

  where $\eta \in \mathbb{R}_+$ is intensity (average occurrences per unit of interval length).

- Poisson RV has a property known as self-reproducibility. If $X_i \sim \mathcal{P}(\lambda_i)$, $i = 1, 2, ..., n$ are independent, their sum is also a Poisson RV:

$$\sum_{i=1}^{n} X_i \sim \mathcal{P}(\lambda)$$

  with $\lambda = \sum_{i=1}^{n} \lambda_i$.

Poisson

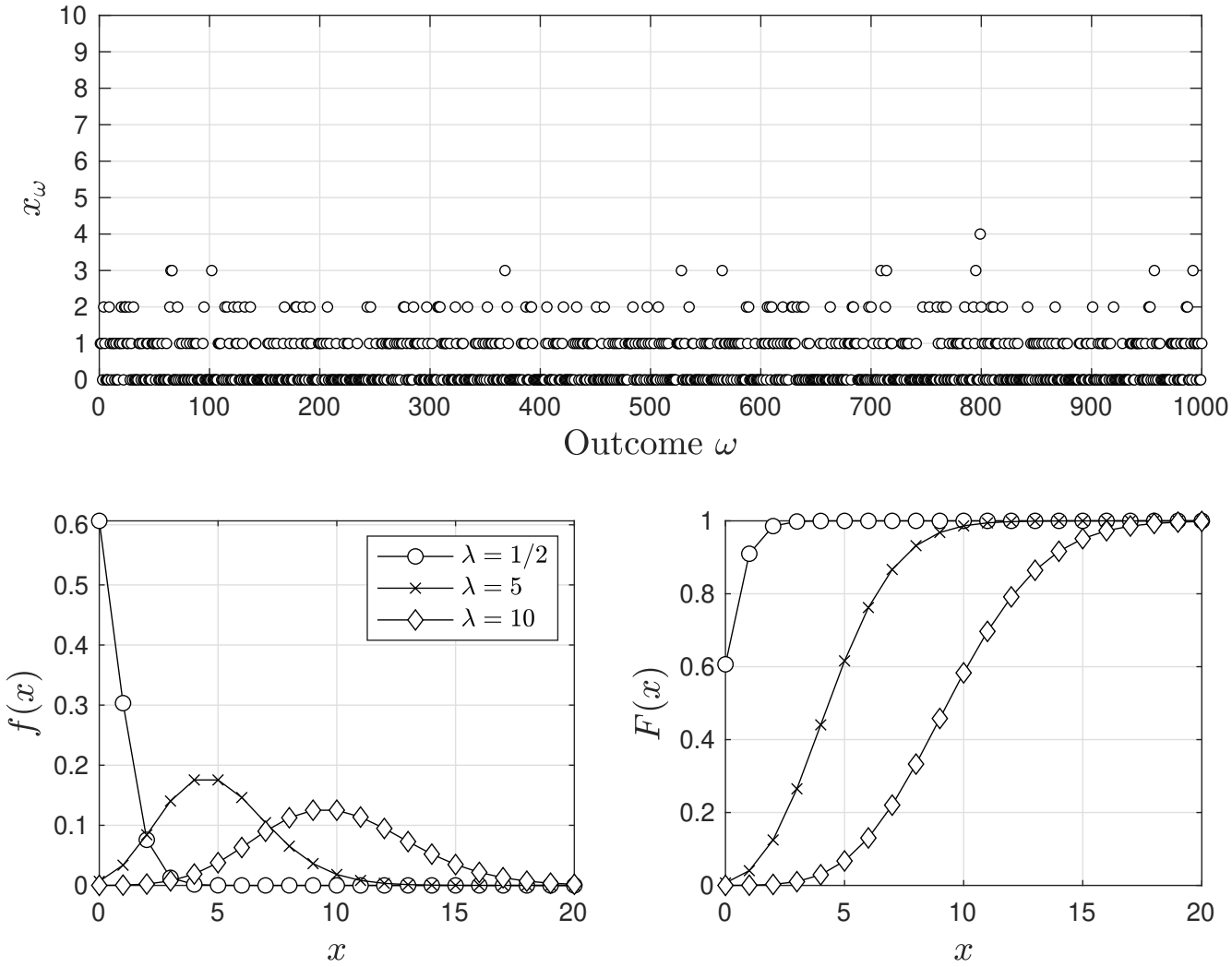


Figure: Pdf and cdf for Poisson random variable.

.

Example Poisson `ch2_example_poisson.m`

- Experimental reactor is equipped with automated cooling system.
- Pump of cooling agent is not reliable (cavitates); failures have been recorded daily for last three months.
- Assuming number of failures per month is Poisson, determine:
  - Is it more probable that the pump does not fail in a month or that the pump fails daily?
  - Probability that the pump fails an average number of times.

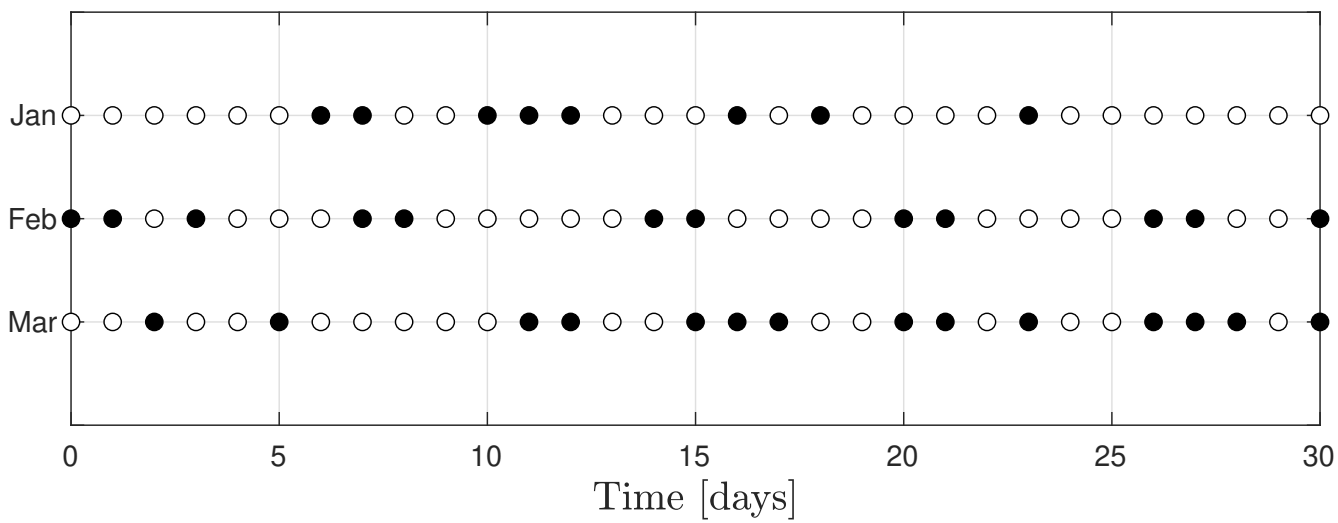


Figure: Pump failures (dark dots) recorded over three months.

Example Poisson `ch2_example_poisson.m`

- From data, number of failures per month are $x_{Jan} = 8$,$x_{Feb} = 12$, $x_{Mar} = 14$.
- From these values we obtain empirical mean:

$$\hat{\mathbb{E}}_X = \frac{1}{3}(8 + 12 + 14) = 11.3$$

- For Poisson RV, we know that $\mathbb{E}_X = \lambda$ and we thus we estimate $\hat{\lambda} = 11.3$.
- Use Poisson RV model to determine that:

$$\mathbb{P}(X = 0) = \frac{\lambda^0 e^{-\lambda}}{0!} = 1.23 \times 10^{-5}$$

$$\mathbb{P}(X = 30) = \frac{\lambda^{30} e^{-\lambda}}{30!} = 1.82 \times 10^{-6}.$$

  More probable that pump does not fail in a month (than failing daily).

- Probability that the pump fails average number of times is:

$$\mathbb{P}(X = 11) = \frac{\lambda^{11} e^{-\lambda}}{11!} = 0.12.$$

  We evaluated at $x = 11$ (and not $x = 11.3$) because domain $\mathcal{D}$ contains integers.

Uniform (Continuous)

- A continuous uniform RV was a pdf of the form:

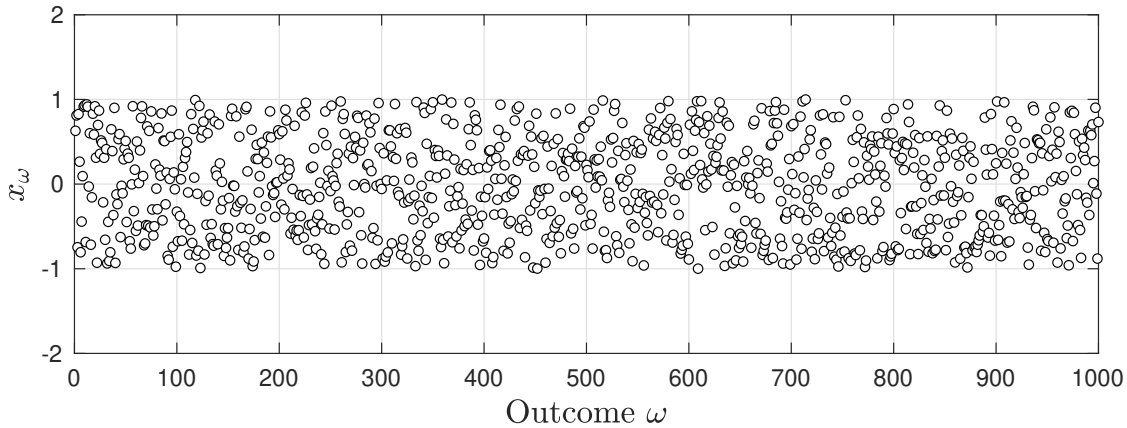
$$f(x) = \frac{1}{b-a}, \; x \in \mathcal{D}$$

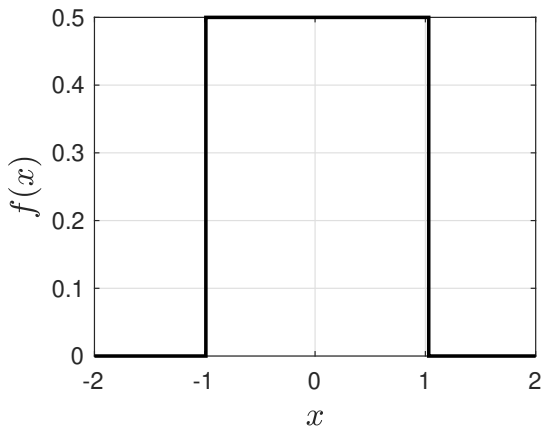
  where domain $\mathcal{D} = [a, b]$ is continuous.
- RV is denoted as $X \sim \mathcal{U}_c(a, b)$.
- Cdf is:

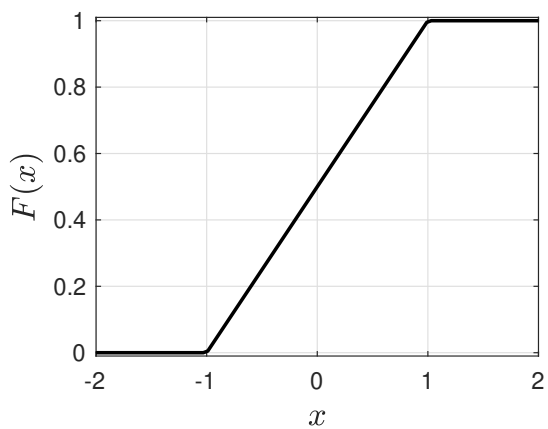
$$F(x) = \frac{x-a}{b-a}, \; x \in \mathcal{D}$$

- Recall that pdf is derivative of cdf (it is the slope of the cdf); consequently, cdf has constant slope $f(x) = 1/(b-a)$.
- Below we show an example of a pdf/cdf of a uniform RV along with realizations.

Uniform (Continuous)







.

Uniform (Continuous)

- Mean of the RV is:

$$\begin{aligned}\mathbb{E}[X] &= \int_a^b \frac{1}{b-a} x dx \\ &= \frac{1}{2}\frac{1}{b-a}(b^2 - a^2) \\ &= \frac{1}{2}(a+b)\end{aligned}$$

- From cdf we note that α-quantiles are values x satisfying:

$$\alpha = F(x) = \frac{x-a}{b-a}$$

  we thus have that $\mathbb{Q}_X(\alpha) = a + \alpha\,(b-a)$.

- Mode (point maximizing $f(x)$) is not unique because pdf is flat function.

- Variance for this RV is:

$$\mathbb{V}[X] = \frac{1}{b-a}\int_a^b \left(x - \frac{1}{2}(a+b)\right)^2 dx = \frac{1}{12}(b-a)^2.$$

  When range $[b, a]$ is wide we have that variance (uncertainty) is large.

Example: Setting up an Alarm

Consider Gibbs reactor and consider situation in which there is no pressure data recorded; all that you know is that, historically, pressure has remained in the range $63 - 234$ bar.

- Is it adequate to model pressure as a uniform RV?
- Thermodynamics tells us that we will have unacceptably low conversions at pressures below 100 bar. What is probability that reactor will exhibit low conversions?
- Reactor vessel is 30-yr old and you are unsure if it can tolerate high pressures for much longer. To be safe, you want to design an alarm system; you want to set a threshold that triggers alarm only 1% of the time. What is such threshold value?

Example: Setting up an Alarm

- We only know that pressure is in continuous range $[63, 234]$ bar, we thus have no reason to doubt that it is equally likely to find pressure anywhere in the range. Consequently, uniform RV is an adequate model with $b = 234$ and $a = 63$.

- Probability that reactor exhibits low conversion is:

$$\mathbb{P}(X \leq 100) = \frac{100 - 63}{234 - 63} = 0.21.$$

- Want to find value x such that $\mathbb{P}(X > x) = 0.01$. This is equivalent to find x such that $\mathbb{P}(X \leq x) = 0.99$ (find x such that $F(x) = 0.99$):

$$0.99 = \frac{x - 63}{234 - 63}$$

  Solving for x we get that a threshold of 232.3 bar activates alarm 1% of the time.

- Note that threshold value is quantile:

$$\mathbb{Q}_X(0.99) = 63 + 0.99 \cdot (234 - 63) = 232.3.$$

Gaussian

- Many random phenomena follow the behavior of a normal RV (a.k.a. Gaussian RV).
- A Gaussian RV is continuous and has an associated pdf:

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

  with domain $\mathcal{D} = (-\infty, \infty)$.

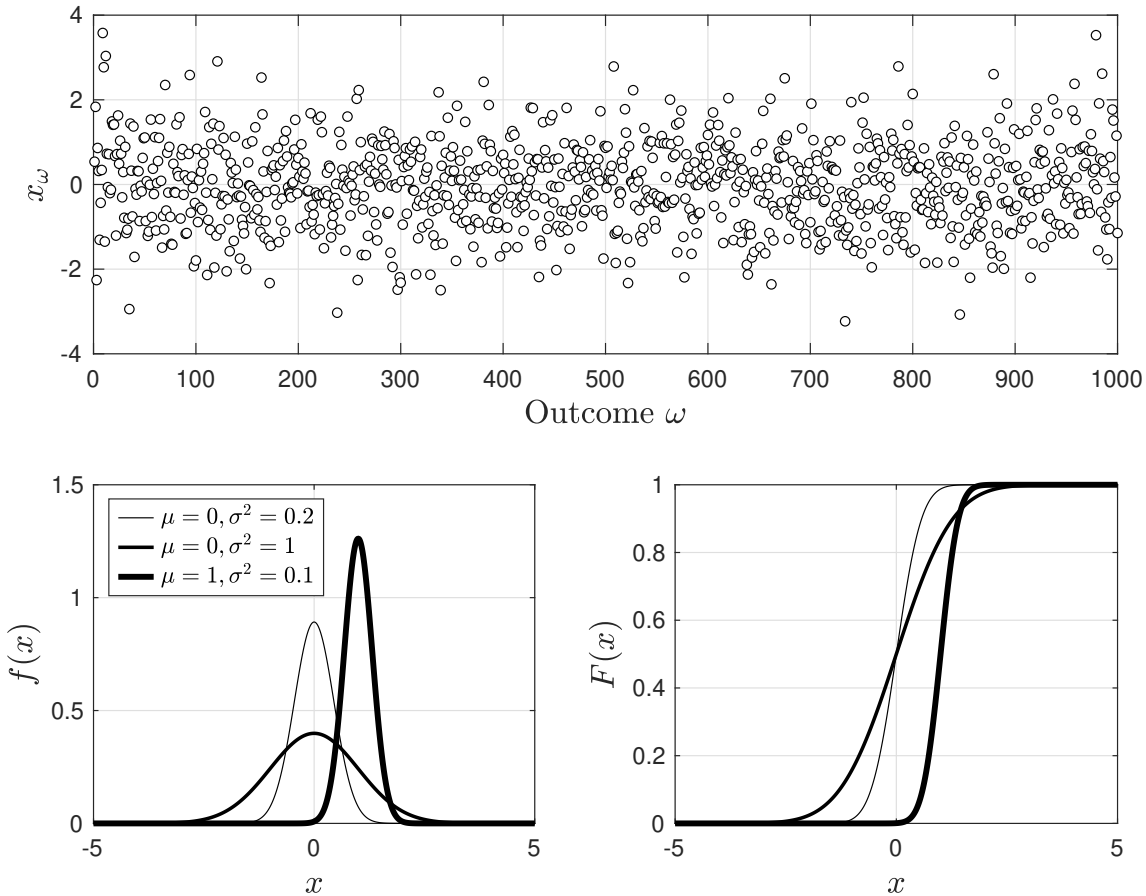
- Parameters $\mu \in \mathbb{R}, \sigma \in \mathbb{R}_+$ are specific to application of interest.
- We express the fact that X is Gaussian as $X \sim \mathcal{N}(\mu, \sigma^2)$.

Gaussian

- Pdf tells us the behavior captured by a Gaussian RV:
  - Probability of an outcome x decays exponentially fast as we move from μ
  - Outcome of maximum probability (most likely outcome) is μ
  - Speed of the decay is dictated by σ
  - Decay in probability is symmetric around μ
- Gaussian model assumes that an outcome x can take any value in domain $(-\infty, \infty)$.
- This introduces complications, as many phenomena involve 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).
- Many phenomena have Gaussian RV as limiting case (we will show this later).

Gaussian RV `ch2_compare_gaussians.m`

- Here are the pdfs for $\mathcal{N}(\mu, \sigma)$ for different values of μ and σ.
- What do you observe?
- What happens when $\sigma^2 \to 0$?



Example: Diffusion Phenomena

- Gaussian RV naturally emerges in "diffusion-like" phenomena
- What is density of particles at location x in domain $(-\infty, \infty)$ and at a given time $t \in [0, T]$?
- One can show that such density, denoted as $f(x,t)$, solves the diffusion equation:

$$\frac{\partial f(x,t)}{\partial t} = D \frac{\partial^2 f(x,t)}{\partial x^2}$$

with boundary conditions:

$$f(x,t) = 0, \quad x \in \{-\infty, \infty\}$$

$$\int_{-\infty}^{\infty} f(x,t)dx = 1, \quad t \in [0, T]$$

- Solution is:

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

with$\sigma^2 = 2Dt$.

- Particles position at any time is random and given by $X(t) \sim \mathcal{N}(0, \sigma^2)$
- How does uncertainty change with diffusivity D and time t?

Gaussian

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

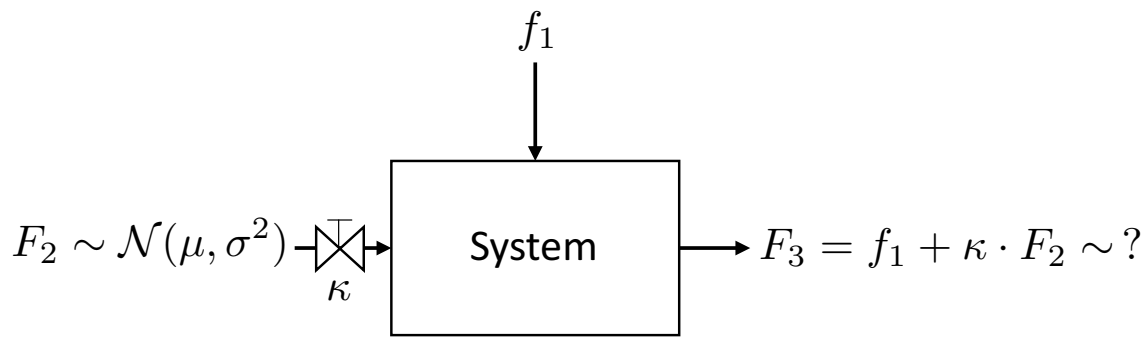
- It expected value and variance are $\mathbb{E}_X = \mu$ and $\mathbb{V}_X = \sigma^2$.
- Any linear transformation $Y = a + bX$ yields a Gaussian $Y \sim \mathcal{N}(a + b\mu, b^2\sigma^2)$.
- This implies that $\mathbb{E}_Y = a + b\mathbb{E}_X$ and $\mathbb{V}_Y = b^2\mathbb{V}_X$.
- Cdf of $Y = a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

Think about implications from an estimation and uncertainty propagation perspective:

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

Example: Gaussian Mixing Problem

- We have input flow $f_1 = 10$ (gpm) that can be measured with high accuracy so it is OK to assume this to be deterministic.
- We have another input flow F_2 (gpm) that cannot be measured with high accuracy and is thus modeled as an RV $\mathcal{N}(20, 1)$.
- Uncertain flow F_2 can be controlled using a valve with coefficient $\kappa \in [0, 1]$.



- What type of RV is output flow $F_3 = f_1 + \kappa \cdot F_2$? What is its mean and std dev?
- How does uncertainty in F_3 change $\kappa \to 0$ and $\kappa \to 1$? Why?

- We have that $F_3 = 10 + \kappa F_2$ and is thus F_3 a linear transformation of F_2
- We thus have that $F_3 \sim \mathcal{N}(10 + \kappa \cdot 20, \kappa^2 \cdot 1)$

Gaussian

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

Gaussian

- Fortunately, one can exploit properties of Gaussian RVs to avoid this issue.
- $Z = (X - \mu)/\sigma$ is a linear transformation of $X \sim \mathcal{N}(\mu, \sigma^2)$ and thus $Z \sim \mathcal{N}(0, 1)$.
- Scaling a Gaussian X using its mean μ and SD σ is known as *standardization*.
- Now note that pdf and cdf of Z are simply:

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

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

  which do not depend on parameters.

- Z is known as the standard normal RV $\mathcal{N}(0, 1)$.

Gaussian

- From linear transformation we have that $F_X(x) = F_Z(z)$ holds for any $z = (x - \mu)/\sigma$ and we can thus evaluate $F_X(x)$ at 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).
- We can compute $\mathbb{Q}_X(\alpha)$ by using pre-computed quantiles of Z, $z_\alpha := \mathbb{Q}_Z(\alpha)$.
- Relationship between quantiles of X and Z is obtained directly from the linear transformation $x = \mu + \sigma z$:

$$\mathbb{Q}_X(\alpha) = \mu + \sigma \cdot \mathbb{Q}_Z(\alpha).$$

- Values z_α are known as the critical values of the standard normal.
- As with cdf, critical values are available in software packages.

Example: Mixing Problem `ch2_mixing_gaussians.m`

- Recall $F_3 \sim \mathcal{N}(10 + \kappa \cdot 20, \kappa^2 \cdot 1)$
- Consider $\kappa = 1$ and thus flow $F_3 \sim \mathcal{N}(30, 1)$

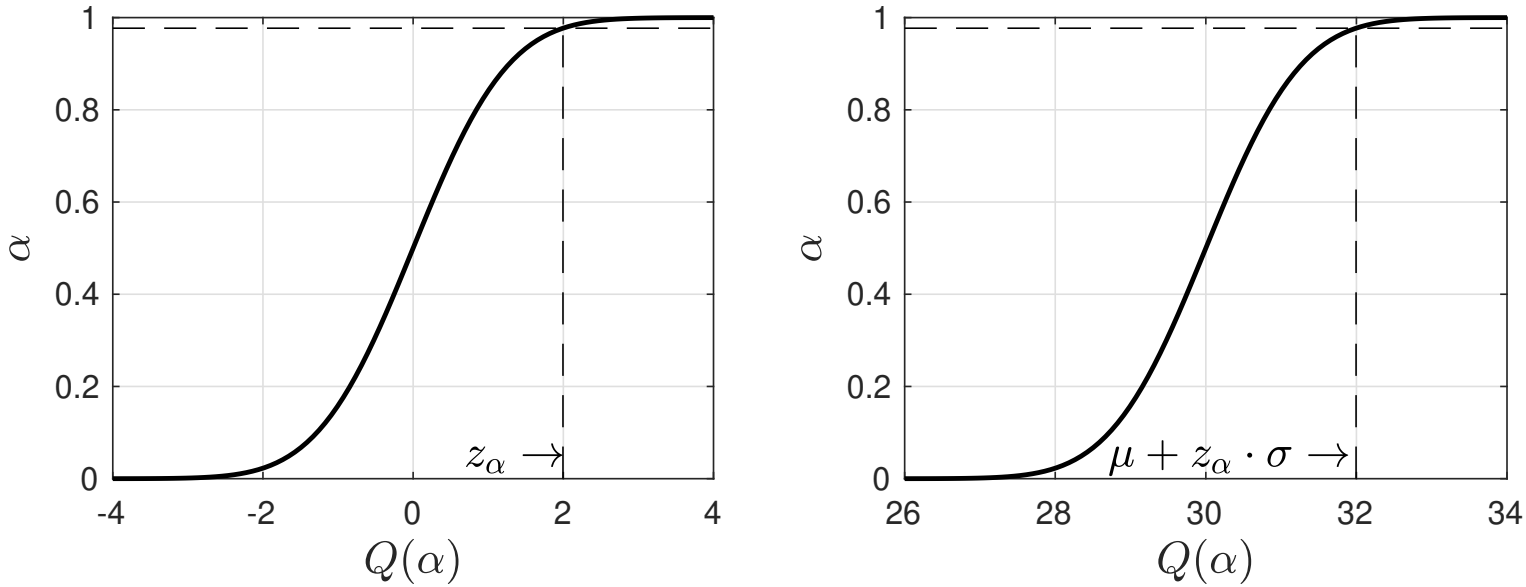


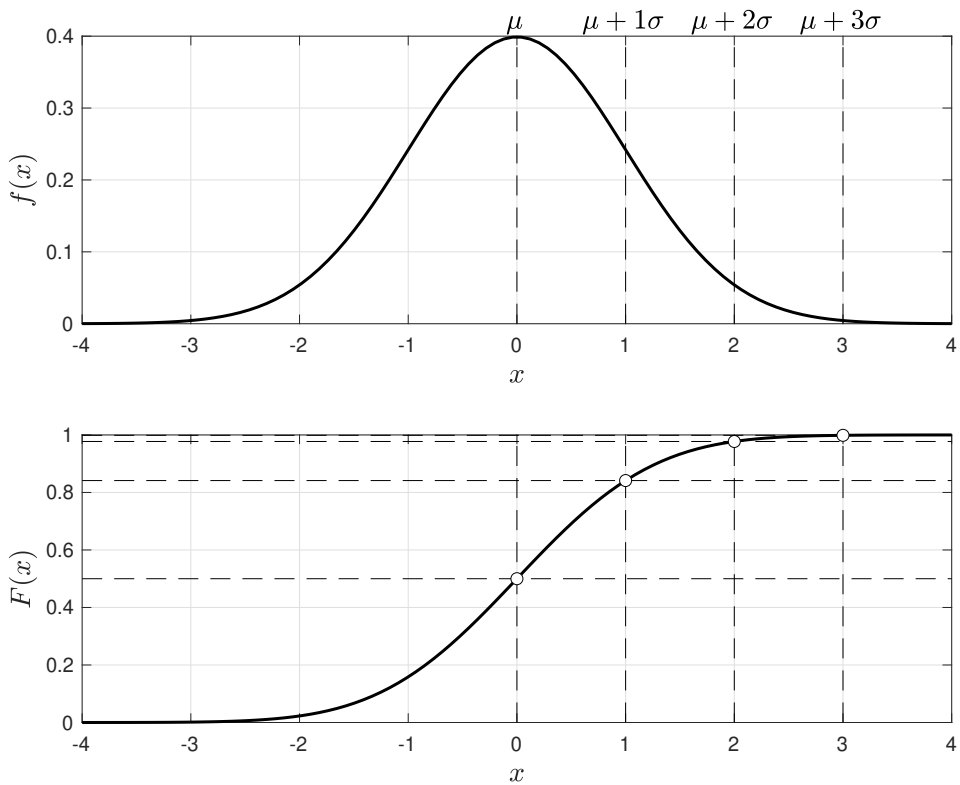
Figure: Quantile functions for $\mathcal{N}(0, 1)$ (left) and $\mathcal{N}(30, 1)$ (right).

- Quantile of $\mathcal{N}(0, 1)$ at $\alpha = 0.977$ is $z_\alpha = 2$
- Quantile of $\mathcal{N}(\mu, \sigma)$ should be $\mu + 2 \cdot \sigma = 32$. Confirm this is true from plot

Gaussian

- Imagine that you precomputed $F_Z(k) = \mathbb{P}(Z \leq k)$ for $k = 0, 1, 2, 3, \ldots$. We have:

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



- Standarization allows us to easily determine probability that X is in specific ranges.

Gaussian

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

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

  is $z = \mathbb{Q}_Z(1 - \frac{\alpha}{2})$; for simplicity, we denote this quantile as $z_{\frac{\alpha}{2}}$.
- In other words, we have that:

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

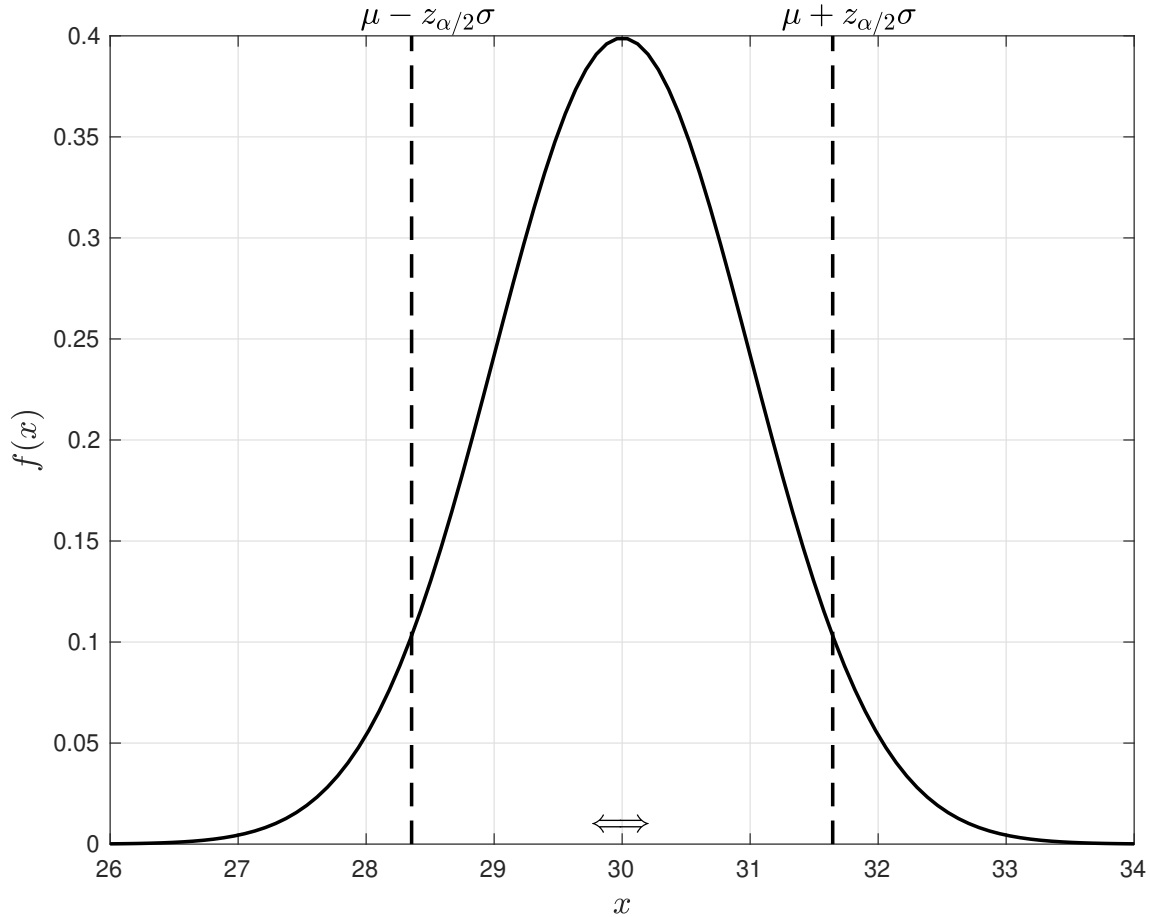
- Using linear transformation property we obtain:

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

- i.e.; probability of finding $X \sim \mathcal{N}(\mu, \sigma^2)$ in region $[\mu \pm z_{\frac{\alpha}{2}} \cdot \sigma]$ is $1 - \alpha$.
- This gives idea of how confident we are of finding X in a region around mean.
- Confidence region is important in many topics (e.g., hypothesis testing).

Example: Mixing Problem `ch2_mixing_gaussians.m`

- In what region (around mean) do we expect $F_3 \sim \mathcal{N}(30, 1)$ to be with 90% prob?



- We have $1 - \alpha = 0.9$ and thus $\alpha = 0.1$ and $\mathbb{Q}_Z(1 - \frac{\alpha}{2}) = 1.644$.
- We thus have $F_3 \in [30 - 1.64 \cdot 1, 30 + 1.64 \cdot 1] = [28.36, 31.64]$ with 90% prob.

Log-Normal

- Log-normal RV helps describe phenomena that span *multiple scales*.
- RV is related to normal RV (via a logarithmic transformation) and has pdf:

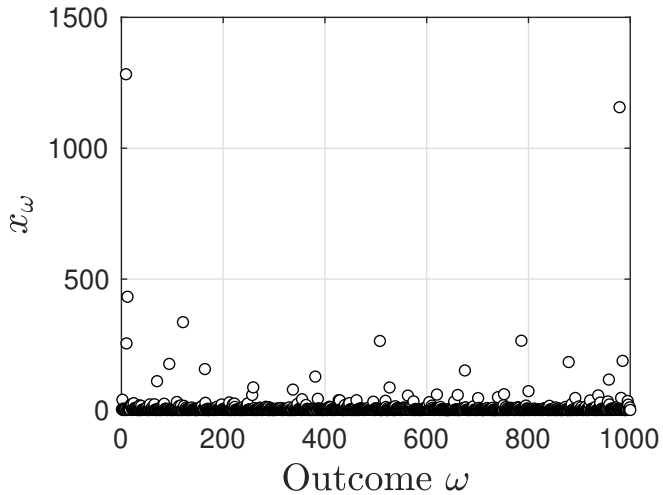
$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\log(x)-\mu)^2}{2\sigma^2}\right], \; x \in \mathcal{D}.$$

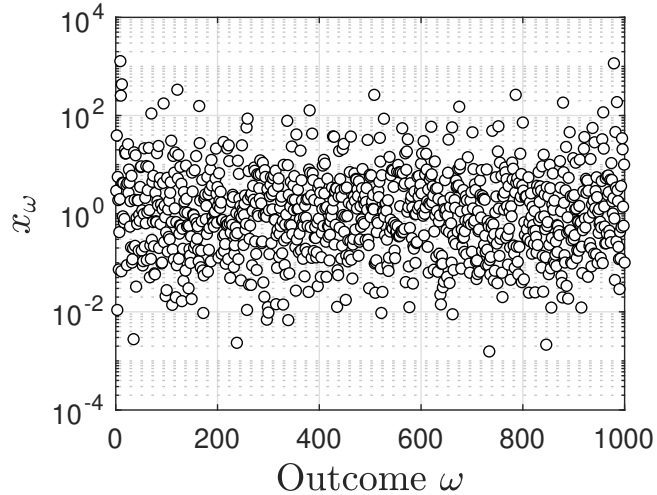
  with hyperprameters $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$.

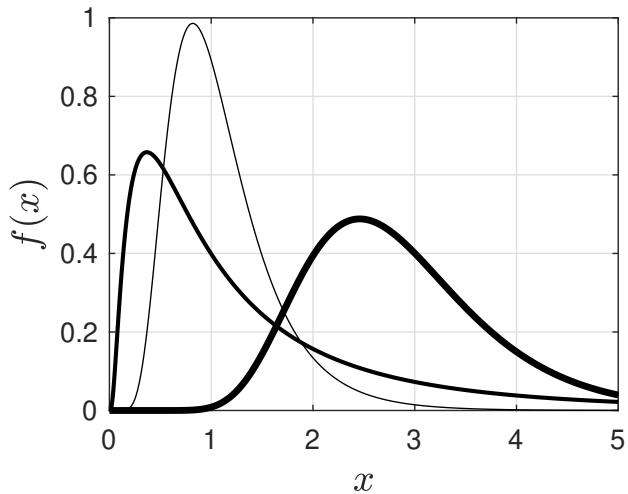
- Domainis $\mathcal{D} = (0, \infty)$ (cannot take negative values)
- Median and quantiles are:

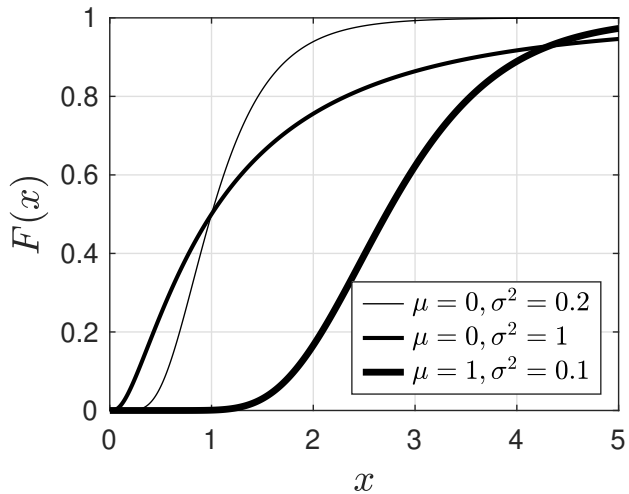
$$\begin{aligned} \mathbb{M}_X &= \exp(\mu) \\ \mathbb{Q}_X(\alpha) &= \exp\left(\mu + \sigma \cdot \sqrt{2} \cdot \text{erf}^{-1}(2 \cdot \alpha - 1)\right). \end{aligned}$$

Log-Normal


1500
1000
500
0
x_ω
0
200
400
600
800
1000
Outcome ω

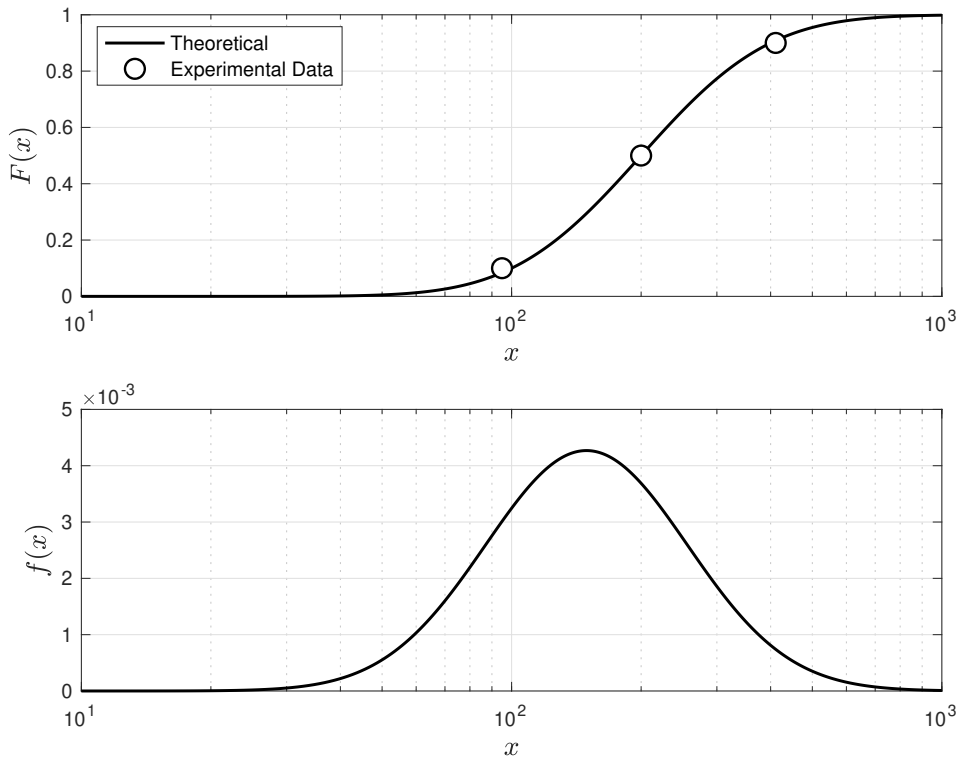

10^4
10^2
10^0
10^{-2}
10^{-4}
x_ω
0
200
400
600
800
1000
Outcome ω


1
0.8
0.6
0.4
0.2
0
$f(x)$
0
1
2
3
4
5
x


1
0.8
0.6
0.4
0.2
0
$F(x)$
$\mu = 0, \sigma^2 = 0.2$
$\mu = 0, \sigma^2 = 1$
$\mu = 1, \sigma^2 = 0.1$
0
1
2
3
4
5
x

Example: Particle Size Distribution in Pharma Products `ch2_example_lognormal.m`

- In pharma industry, particle characterization of powder materials is crucial in drug product development and quality control of solid oral dosage forms.
- Particle size distribution (PSD) affects product performance (e.g., dissolution and bioavailability).
- Laser diffraction was used to characterize PSD for 1000 kg of product



Example: Particle Size Distribution in Pharma Products `ch2_example_lognormal.m`

- Address the following questions:
  - Does PSD follow a lognormal RV?
  - If so, what are the parameters of RV? What is PSD?
  - Product is considered to be acceptable if particle sizes are in $[250, 400]$ μm.
  - How many kg of the batch are in the acceptable range?

Example: Particle Size Distribution in Pharma Products `ch2_example_lognormal.m`

- From data at size 200 μm, we have that probability of finding a particle with size below this value is 50%. This is empirical median (obtained from experimental data).
- If PSD is lognormal, median should be $\mathbb{M}_X = \exp(\mu) = 200$ and from here we estimate that $\mu = \log 200$.
- From data we observe that probability of finding particle size below 400 μm is 90%. Under our model hypothesis, we thus have that:

$$\mathbb{Q}_X(\alpha) = \exp\left(\mu + \sigma \cdot \sqrt{2} \cdot \text{erf}^{-1}(2 \cdot \alpha - 1)\right).$$

Solving for σ we have that

$$\sigma = \frac{\log \mathbb{Q}_X(\alpha) - \mu}{\sqrt{2} \cdot \text{erf}^{-1}(2 \cdot \alpha - 1)}$$

Substituting terms we find estimate $\sigma = 0.54$.

Example: Particle Size Distribution in Pharma Products `ch2_example_lognormal.m`

- Use parameters to obtain the cdf of the PSD and we see this matches data well.
- Having parameters, we can also obtain pdf (note that the particle size spans two orders of magnitude)
- Probability that the particle size is in the acceptable range is:

$$\mathbb{P}(250 \leq X \leq 400) = F(400) - F(250) = 0.24.$$

  i.e., 24% of our product is acceptable ($1000 \cdot 0.24 = 240$ kg is acceptable).

Exponential

- Exponential RV has domain $\mathcal{D} = (0, \infty)$ and pdf:

$$f(x) = \frac{1}{\beta} e^{-x/\beta}, \quad x \in \mathcal{D}.$$

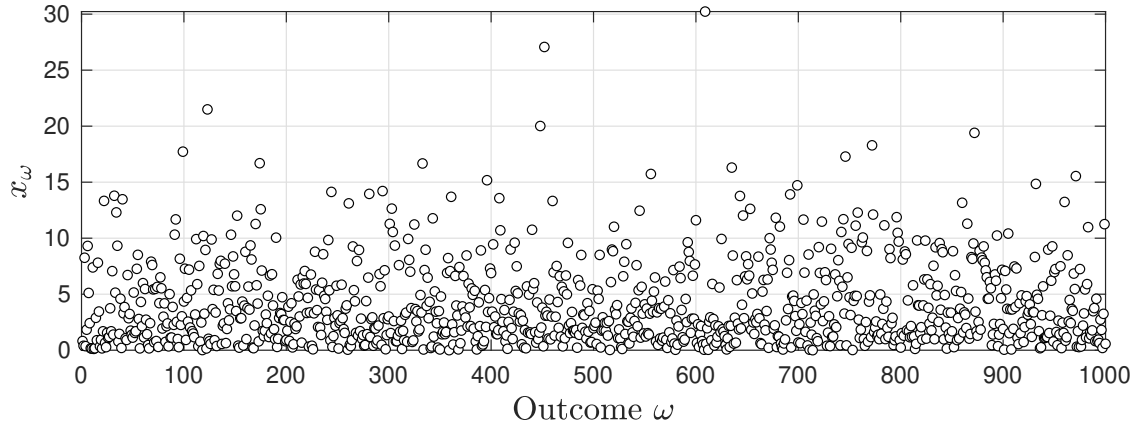
- Only parameter is $\beta \in \mathbb{R}_+$ (a.k.a. scale value).
- Reciprocal $\eta = 1/\beta$ is known as intensity and thus pdf can also be written as $f(x) = \eta e^{-\eta \cdot x}$.
- Express fact that X is exponential as $X \sim \text{Exp}(\beta)$.

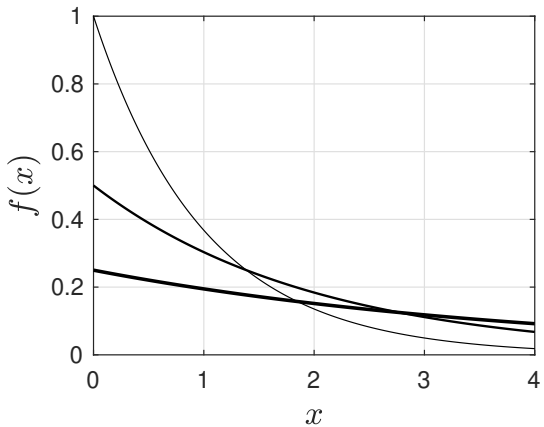
Exponential RV

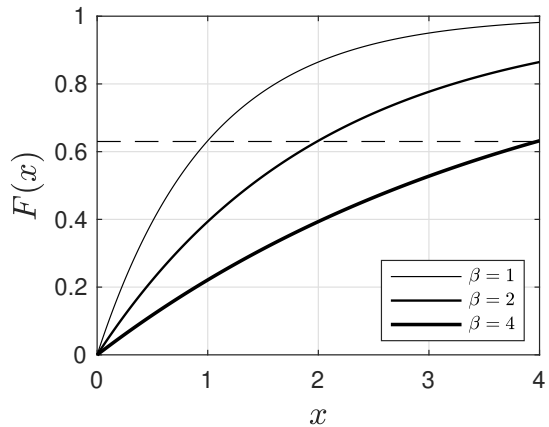
- Pdf of exponential RV tells us that:
  - Probability of finding X away from zero decays exponentially fast at rate η
  - There is *zero* probability of finding X below zero (pdf is asymmetric).
- Cdf is $F(x) = 1 - e^{-x/\beta}$ and note $1 - F(x) = e^{-x/\beta}$.
- Expected value and variance are $\mathbb{E}_X = \beta$ and $\mathbb{V}_X = \beta^2$.
- This RV is often used to model time-dependent phenomena (e.g., *failures*).
- For instance, X can be used to model time that we have to wait until we observe first occurrence of an event (e.g., engine fails).
- In this context, we know average waiting time ($\mathbb{E}_X = \beta$) but actual time is unknown.
- Parameter $\eta = 1/\beta$ can be interpreted as event rate (how frequently they occur).

Exponential `ch2_compare_exponentials.m`

- Pdfs and cdfs for $\text{Exp}(\beta)$ for different values of β.







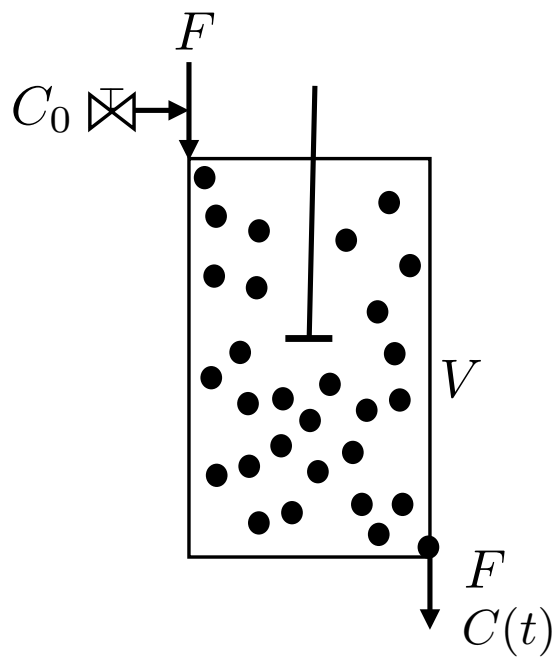
How to determine β from cdf?

- Note that $F(x) = 1 - e^{-1} = 0.63$ when $x = \beta$. This is known as characteristic time.

Example: Residence Time in Mixing System

- Mixed system with volume V and input/output flow F
- At time $t = 0$ we inject tracer so that input concentration is C_0.
- Tracer concentration in system at time t is $C(t)$

For how long will a tracer particle reside in the system? What factors influence this time?



Example: Residence Time in Mixing System

- Material balance reveals that:

$$f_T(t) = \frac{C(t)}{C_0} = \frac{1}{\tau} e^{-t/\tau} \quad \text{with} \quad \tau = V/F$$

- Interpret T as random time required for a particle to exit system (residence time)
- Pdf of T is $f_T(t)$ is interpreted as fraction of injected particles exiting at time $T = t$
- Balance suggests that $T \sim \text{Exp}(\tau)$

Example: Residence Time in Mixing System

- Mean residence time is $\mathbb{E}[T] = \tau = V/F$ and variance is $\mathbb{V}[T] = (V/F)^2$.
- What is effect of V and F on uncertainty?
- Fraction of particles that have exited up to time t is $\mathbb{P}(T \leq t) = F_T(t)$.
- $F_T(t) = \int_0^t f_T(t')dt' = (1 - e^{-t/\tau})$ and thus $F(0) = 0$ and $F(\infty) = 1$.
- Fraction of particles that remain in system at time t is $\mathbb{P}(T > t) = 1 - F_T(t) = e^{-t/\tau}$
- Function $\mathbb{P}(T > t)$ is known as the "survival function".

Gamma

Gamma RV is a generalization of exponential RV that has domain $\mathcal{D} = [0, \infty)$ and pdf:

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} e^{-x/\beta} x^{\alpha-1}, \; x \in \mathcal{D}$$

- Pdf has parameters $\alpha, \beta \in \mathbb{R}_+$
- $\Gamma(\alpha)$ is the gamma function; for integer $\alpha \geq 1$ we have $\Gamma(\alpha) = (\alpha - 1)!$.
- We express fact that X is a Gamma RV as $X \sim \text{Gamma}(\alpha, \beta)$
- For integer α, the cdf of the Gamma RV takes the form:

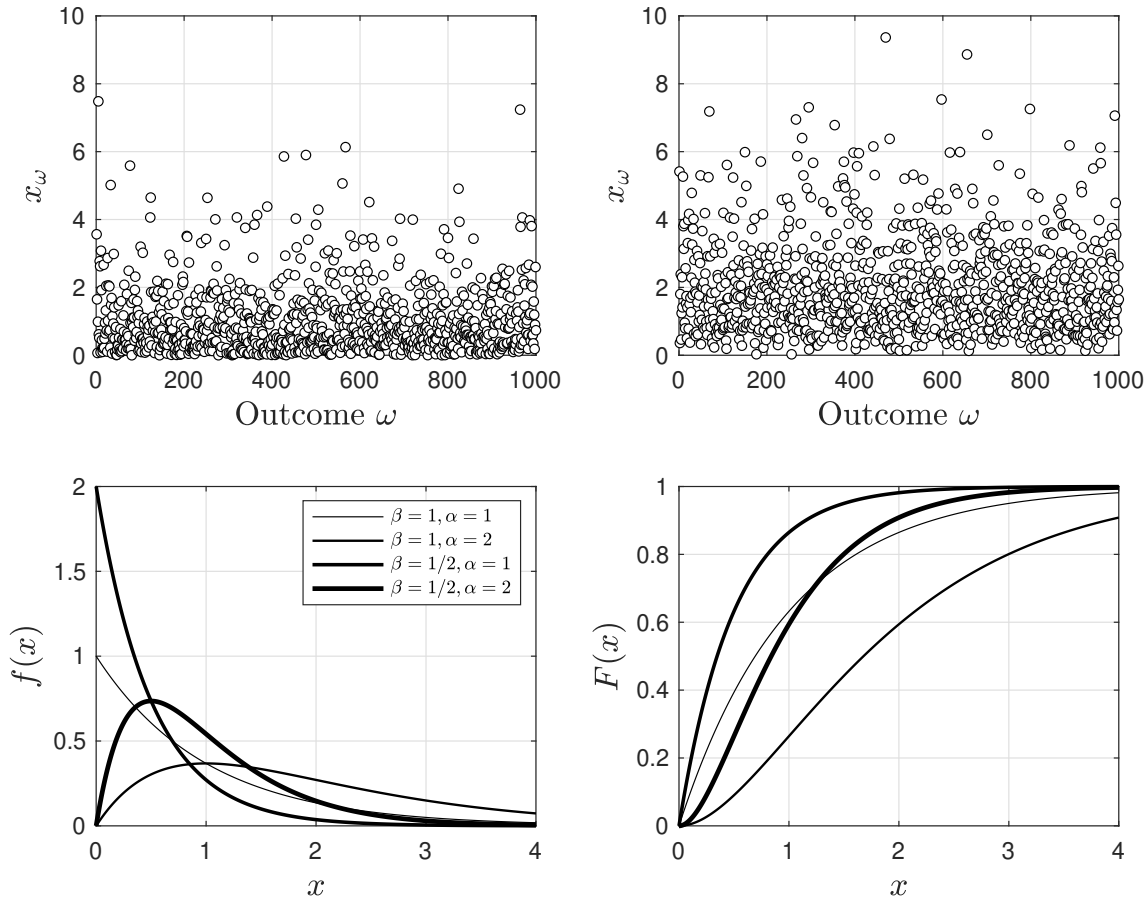
$$F(x) = 1 - \sum_{k=0}^{\alpha-1} \frac{1}{k!} \left(\frac{x}{\beta}\right)^k e^{-x/\beta} \tag{1}$$

Gamma

- In context of time phenomena, RV generalizes exponential in that it models amount of time that we have to wait until we observe the α-th occurrence of an event.
- Consequently, α=1 means first event (as in exponential RV).
- This RV has applications not only in temporal but also in spatial phenomena.
- For instance, it can be used to model distance traveled until we find α-th occurrence of certain type of atom in a chain or crack in a pipeline.

Gamma `ch2_compare_gammas.m`

- Here are pdfs and cdfs for $\text{Gamma}(\alpha, \beta)$ for different values of α, β.



- Note same as exponential for $\alpha = 1$.
- Note emergence of peaks due to competing effects for $\alpha = 2$.

Example: Service Time

- You run warehouse logistics for your company and are responsible for making sure that queue of electric forklift trucks at charging stations remains under control.
- Average charging rate at a station is 0.25 trucks/hr (average charge time is 4 hours).
- Charge time is random because there is variability in state-of-charge and age of batteries and average charge time is quite long because you need to protect them.

You are interested in addressing the following questions:

- If queue has 2 trucks, what is prob that next truck will wait more than 3 hours?
- What is prob if queue has 3 trucks?
- What is prob of waiting more than 4 hours if queue has 1 truck?
- You buy brand new batteries and this allows you to cut down average charging time to 2 hours. What is prob of waiting more than 3 hours if queue has 1 truck?

Example: Service Time

- Use X to denote waiting time of next truck and let Poisson event represent completion of truck charging.
- Prob that waiting time for next truck is more than 3 hours is $\mathbb{P}(X > 3)$.
- This is the prob that the charging of the 2 trucks in the queue *cannot* be completed in the next 3 hours.
- X is thus time to wait until we observe 2 Poisson events (two charge completions)

Example: Service Time

- Average waiting time is $\beta = 1/0.25 = 4$ hr. We use:

$$P(X > x) = 1 - F(x) = \sum_{k=0}^{\alpha-1} \frac{1}{k!} \left(\frac{x}{\beta}\right)^k e^{-x/\beta}$$

with $x = 3$ (hr) that $\alpha = 2$ and $\beta = 4$ (hr) and thus:

$$\begin{aligned} P(X > 3) &= \sum_{k=0}^{1} \frac{1}{k!} \left(\frac{3}{4}\right)^k e^{-3/4} \\ &= \frac{1}{0!} \left(\frac{3}{4}\right)^0 e^{-3/4} + \frac{1}{1!} \left(\frac{3}{4}\right)^1 e^{-3/4} = e^{-3/4} + (3/4)e^{-3/4} \\ &= 0.826. \end{aligned}$$

- If there are 3 trucks in queue we expect a higher probability. Indeed, set $\alpha = 3$:

$$P(X > 3) = \sum_{k=0}^{2} \frac{1}{k!} \left(\frac{3}{4}\right)^k e^{-3/4} = 0.959.$$

Chi-Squared

Chi-Square RV has pdf of the form:

$$f(x) = \frac{1}{2^{r/2}\Gamma(r/2)} e^{-x/2} x^{r/2-1}, \; x \in \mathcal{D}.$$

with domain $\mathcal{D} = [0, \infty)$.

- Pdf has only one parameter $r \in \mathbb{Z}_+$ (a.k.a degrees of freedom).
- We express fact that X is a chi-squared RV as $X \sim \chi^2(r)$
- This is a Gamma RV with $\beta = 2$ and $\alpha = r/2$ (for a positive integer r).

Chi-Squared

- Expected value and variance can be derived from those of the Gamma RV.
- Crucial property of a chi-squared RV is that it is related to standard normal RV. In particular, one can show that:

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

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

- From relationship with $\mathcal{N}(0,1)$ we can show that $\mathbb{Q}_Y(1-\alpha)$ of $Y \sim \chi^2(1)$ is equal to $\mathbb{Q}_Z(1-\alpha/2)^2$ of $Z \sim \mathcal{N}(0,1)$.
- This is because $\mathbb{P}(-z \leq Z \leq z) = \mathbb{P}(Y \leq z)$ with $Y = Z^2$.

Example: Computing Confidence Intervals

- Consider output flow $F_3 \sim \mathcal{N}(30, 1)$.
- Compute 90% confidence interval for F_3 using quantile of Chi-Squared RV.
- Confidence interval is $\mu \pm z_{\alpha/2}\sigma$ with $z_{\alpha/2} = \mathbb{Q}_Z(1 - \alpha/2)$ and $Z \sim \mathcal{N}(0, 1)$.
- Instead of using $Z \sim \mathcal{N}(0, 1)$, here we compute $z_{\alpha/2}$ using $Y \sim \chi^2(1)$ as:

$$\mathbb{Q}_Z(1 - \frac{1}{2}) = z_{\alpha/2} = \sqrt{\mathbb{Q}_Y(1 - \alpha)}$$

- We use numerical package to compute $\mathbb{Q}_Y(1 - 0.1) = 2.71$ and thus $z_{\alpha/2} = \sqrt{2.71} = 1.64$.
- Confidence interval is $[30 \pm 1 \cdot 1.64] = [28.36, 31.64]$ gpm. This is the same confidence interval computed previously using the Gaussian RV model.

Weibull

Weibull RV is a generalization of exponential RV that has a domain $\mathcal{D} = [0, \infty)$ and pdf:

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

- Pdf has parameters $\beta, \xi \in \mathbb{R}_+$ (a.k.a scale and shape).
- We express fact that X is a Weibull RV as $X \sim \text{Weibull}(\xi, \beta)$.
- One recovers an exponential RV when $\xi = 1$.

Weibull

- Expected value and variance are:

$$\mathbb{E}_X = \beta\Gamma(1 + 1/\xi)$$
$$\mathbb{V}_X = \beta^2 \left(\Gamma(1 + 2/\xi) + \Gamma(1 + 1/\xi)^2\right)$$

- Cdf has nice form $F(x) = 1 - e^{-(x/\beta)^\xi}$ (ξ, β are often inferred from cdf).

- Note also that $\mathbb{P}(X \leq \beta) = 1 - e^{-1} = 0.63$ *for any* ξ. This property is useful for estimating β from the empiricial cdf.

- Because cdf has a nice exponential form, quantile function is explicit:

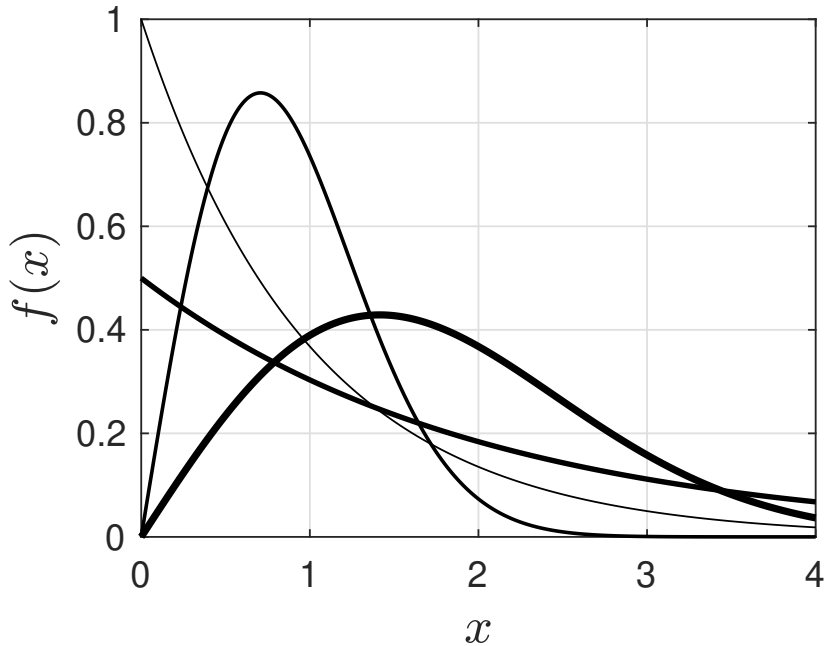
$$\mathbb{Q}_X(\alpha) = \beta \left(-\log(1 - \alpha)\right)^{1/\xi}.$$

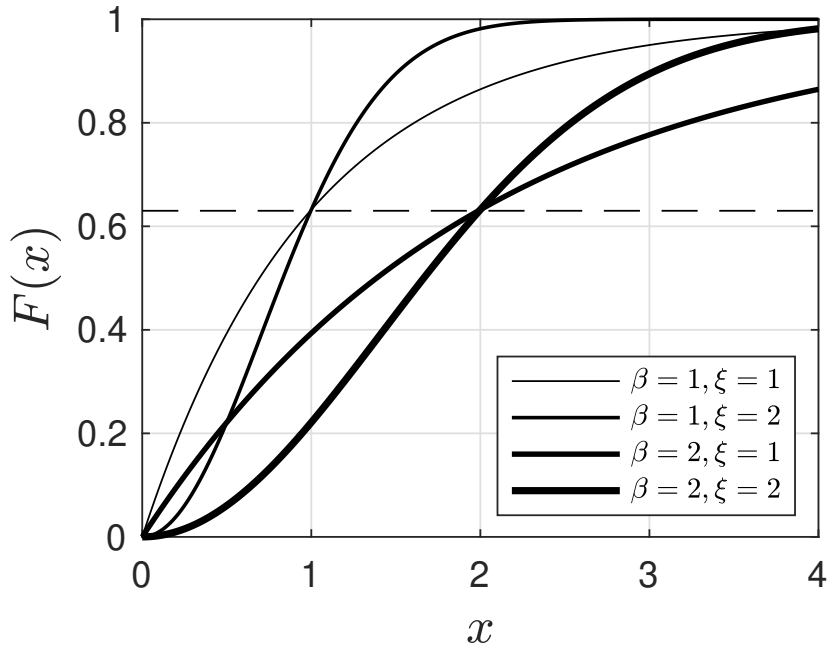
  From here, we determine that median is $\mathbb{M}_X = \beta(\log 2)^{1/\xi}$.

- Weibull is *de facto* model used in failure analysis and many phenomena have Weibull-like RVs as limiting case.

Weibull `ch2_compare_weibull.m`

- Pdfs and cdfs for $\text{Weibull}(\beta, \xi)$ for different values of β, ξ.



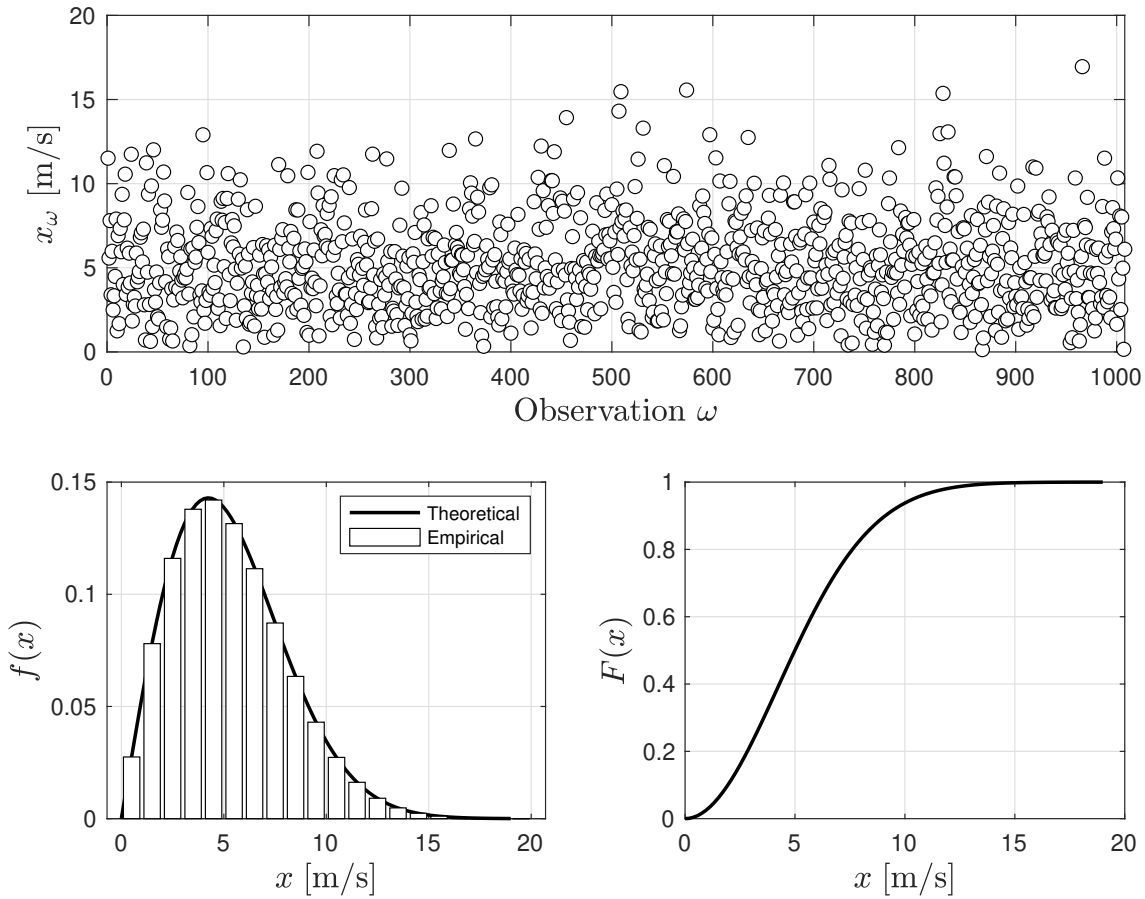


- Same as exponential for $\xi = 1$.
- Note emergence of peaks due to competing effects for $\xi = 2$.
- Note $\mathbb{P}(X \leq \beta) = 0.63$ for any ξ (this differentiates it from Gamma RV).

Example: Modeling Wind Conditions `ch2_example_weibull.m`

- Wind is rather unpredictable but it is known that the 10-min wind speed (time average over 10 minutes) follows a Weibull distribution (see Figure).

- Information on wind speed distribution is used in planning studies to determine how much wind power can be extracted using wind turbines at a given location. .



Example: Modeling Wind Conditions `ch2_example_weibull.m`

Use the pdf and cdf to determine the following:

- Scale and shape parameters for 10-min wind speed.
- Average and most likely wind speed at this location. Which of these statistics would you use to recommend?
- Wind turbines extract high power when wind speed is [10-20 m/s]. Over what fraction of the time will the turbine operate in this high-power range at this location?

Example: Modeling Wind Conditions `ch2_example_weibull.m`

- We have that $F(x) = 0.63$ occurs at $x =$6 m/s (this gives β).
- $F(x) = 0.5$ occurs at $x = 5$ m/s; we thus have $\mathbb{M}_X = 5$ and $5 = \beta(\log 2)^{1/\xi}$.
- Applying logarithms we obtain:

$$\log 5 = \log(\beta) + (1/\xi)\log(\log 2)$$

Solving for ξ we obtain $\xi = 2$.

Example: Modeling Wind Conditions `ch2_example_weibull.m`

- The average 10-min windspeed (in m/s) is

$$\mathbb{E}_X = 6 \cdot \Gamma(1 + 1/2) = 5.3.$$

  while the most likely value (the mode, also in m/s) is:

$$\mathbb{MO}_X = 6 \cdot \left(\frac{2-1}{2} \right)^{1/2} = 4.2.$$

  Mode has a more intuitive interpretation (most likely wind speed value).

- Probability that 10-min wind speed is in [10-20] m/s is given by:

$$\mathbb{P}(10 \leq X \leq 20) = (1 - e^{-(20/\beta)^{\xi}}) - (1 - e^{-(10/\beta)^{\xi}}) = 0.06$$

  Wind turbine will only operate 6% of the time in the high-power range.

Families of Random Variables

- Exponential, Gamma, Chi-squared, and Weibull RVs are interrelated.
- In fact, these are captured by generalized Gamma model with pdf:

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

- These RVs are known as the Gamma family.
- There are three major families of continuous RVs:
  - Gaussian family (includes Gaussian, LogNormal, and Raleigh)
  - Gamma family (includes exponential, Gamma, Chi-Squared, and Weibull)
  - Ratio family (includes Cauchy, Uniform, Beta, Fisher, Student)

Families of Random Variables

- One family for discrete RVs: includes Uniform (discrete), Bernoulli, Binomial, and Poisson.
- Each family models different phenomena, important to understand their origin (just like with physical models).
- Some RVs are limiting cases, generalizations, or transformations of others.
- A detailed discussion of all RV types and properties is beyond our scope.
- Here, we have only discussed some RVs to highlight their rich behavior.

Important: Pay attention to the origin of these RV models. As with mechanistic models, understanding their origin gives insights into what phenomena might exhibit similar characteristics (e.g., do not model a diffusion-like system using a Gamma RV).

Families of Random Variables