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 4: Estimation for Random Variables

Estimation Techniques

- We have now a basic idea of the types of RV models available to model phenomena.
- We proceed to develop procedures to determine if a RV model fits data at hand.
- We define an RV model ($f_X(x|\theta)$ or $F_X(x|\theta)$), where θ are the model parameters.
- By estimating RV model we mean that we seek to find θ that *best* fits data.

We explore a couple of estimation methods:

- Point Estimation (Method of Moments and Least-Squares Method)
- Maximum Likelihood Estimation (MLE)

- First step will be to explore our data and application and postulate an RV model (e.g., Gaussian or Exponential) based on any patterns exposed.
- Second step will be to find θ that *best* fits data. If best fit is not adequate, we we can conclude that we need to propose a different model.

Method of Moments

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

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

- Here, we highlight dependence of moments on parameters θ.

- Method of moments uses data $x_\omega,\ \omega \in \mathcal{S}$ to obtain sample approx:

$$\mathbb{C}\hat{\mathbb{M}}\mathbb{O}_k = \frac{1}{S}\sum_{\omega \in \mathcal{S}} (x_\omega - \hat{\mathbb{E}}_X)^k,\ k = 1, 2, ..., N$$

  where $\hat{\mathbb{E}}_X = \frac{1}{S}\sum_{\omega \in \mathcal{S}} x_\omega$ is sample mean.

- In this moment matching method we can use any type of moment that we find convenient (e.g., raw, central, standardized).

Method of Moments

- Our objective is to find θ that solves matching equations:

$$\mathbb{CMO}_k(\theta) = \mathbb{C}\hat{\mathbb{M}}\mathbb{O}_k,\ k = 1, 2, ..., N$$

- In other words, we want to find θ that matches model to sample moments.
- For example, we can find the parameters by matching the model mean $\mathbb{E}[X]$ to the empirical $\hat{\mathbb{E}}[X]$ and the model variance $\mathbb{V}[X]$ to the empirical variance $\hat{\mathbb{V}}[X]$.
- We can leverage the properties of different models; for instance, for a Gaussian we know that $\mathbb{E}[X] = \mu$ and $\mathbb{V}[X] = \sigma^2$.

Example: Method of Moments for Gaussian `ch4_gaussian_moments_example.m`

- Want $\theta = (\mu, \sigma^2)$ for model $X \sim \mathcal{N}(\mu, \sigma)$ to see if this matches data $x_\omega,\ \omega \in \mathcal{S}$.

- Gaussian model has property that $\mu = \mathbb{E}_X$ and $\sigma^2 = \mathbb{V}_X$.

- Use data to obtain $\hat{\mathbb{E}}_X = \frac{1}{S}\sum_{\omega\in\mathcal{S}} x_\omega$ and $\hat{\mathbb{V}}_X = \frac{1}{S}\sum_{\omega\in\mathcal{S}} (x_\omega - \hat{\mathbb{E}}_X)^2$.

- From definition of first raw moment we have:

$$\mathbb{CMO}_1(\theta) = \mathbb{E}[X] - \mu = 0$$

  and this indicates that $\hat{\mu} = \hat{\mathbb{E}}_X$.

- From definition of second central moment we have $\mathbb{CMO}_2(\theta) = \mathbb{V}_X = \sigma^2$.

- Thus $\mathbb{CMO}_2(\theta) = \hat{\mathbb{CMO}}_2$ indicates that $\hat{\sigma}^2 = \hat{\mathbb{V}}_X$.

Example: Method Moments for Gaussian `ch4_gaussian_moments_example.m`

- We used moment matching with $S = 10$ and $S = 100$ data points to estimate parameters. If we use $S = 10$, the estimates are $\hat{\mu} = 0.45$ and $\hat{\sigma} = 1.43$; if we use $S = 100$, the estimates are $\hat{\mu} = 0.9$ and $\hat{\sigma} = 2$. The true parameters are $\mu = 1$ and $\sigma = 2$; we can thus see that we get closer as we increase S.

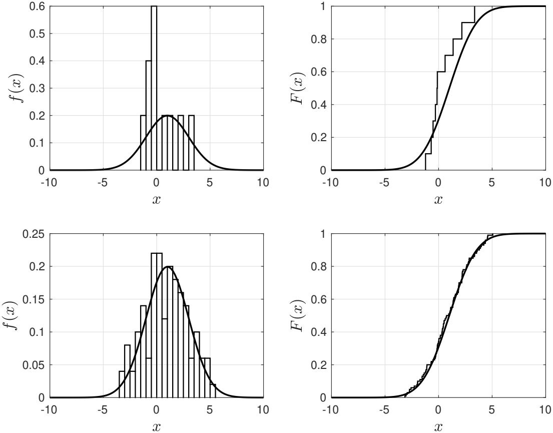


Figure: Empirical pdf/cdf for Gaussian RV obtained from data and model predictions obtained with estimated parameters. Results for $S = 10$ (top row) and for $S = 100$ (bottom row).

Example: Method of Moments for Weibull

Now want to postulate $\mathrm{Weibull}(\xi, \beta)$ and see if this matches data.

- We are given observations $x_\omega,\ \omega \in \mathcal{S}$.
- Sample mean is $\hat{\mathbb{E}}_X = \frac{1}{S}\sum_{s\in\mathcal{S}} x_\omega$ and variance $\hat{\mathbb{V}}_X = \frac{1}{S}\sum_{s\in\mathcal{S}}(x_\omega - \hat{\mathbb{E}}_X)^2$.
- We find β, ξ that matches moments $\mathbb{CMO}_1(\theta)$ and $\mathbb{CMO}_2(\theta)$ (expected value and variance) to empirical counterparts.
- This is done by solving the matching equations for β and ξ:

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

- This is challenging due to complex nature of Γ function. Is there another way?

Least-Squares Method

- Moment functions $\mathbb{CMO}_k(\theta)$ might be too complex for some RVs (e.g., Weibull).

- We can also use $F_X(t|\theta)$ (or other summarizing statistics) to find parameters.

- In LS, we find θ that best matches empirical estimates of cdf or statistics.

- Assume we use cdf and propose threshold values $t_k, k = 1, 2, ..., N$ to compute:

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

- We find parameters for the model $F_X(t_k|\theta)$ that solve LS problem:

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

- This is an optimization problem that needs to be solved numerically.

Example: Least-Squares for Weibull `ch4_leastsquares_example.m`

- Want to estimate $\theta = (\xi, \beta)$ for model $X \sim \text{Weibull}(\xi, \beta)$.

- We are given data $x_\omega,\ \omega \in \mathcal{S}$ to obtain empirical cdf $\hat{F}(t_k)$ for $t_k,\ k = 1, ..., N$.

- Recall cdf of Weibull model is $F_X(t|\theta) = (1 - e^{-(t/\beta)^\xi})$.

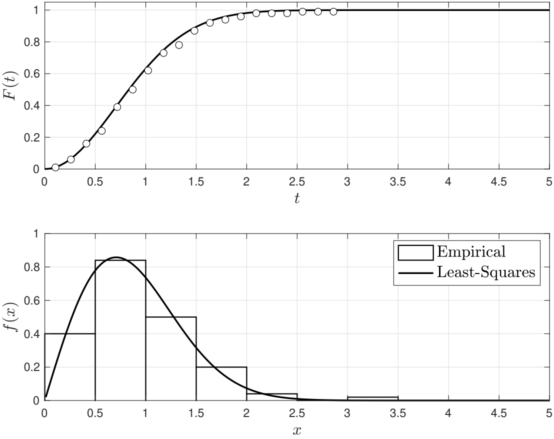
- We estimate parameters θ by solving LS problem:

$$\min_\theta \frac{1}{2} \sum_{k=1}^{N} \left((1 - e^{-(t_k/\beta)^\xi}) - \hat{F}_k \right)^2$$

- We solve this problem using a numerical software package.

Example: Least-Squares for Weibull `ch4_leastsquares_example.m`

- Weibull model using the estimated parameters fits data well.



Maximum Likelihood Estimation

- Maximum likelihood estimation (MLE) is a method that is conceptually similar LS in that tries to find the best possible set of parameters that match data.

- In MLE, by “best” we mean parameters that are most probable (likely) given data.

- In LS, by “best” we mean parameters that best fit data (minimize error).

- We thus note that MLE has a natural statistical interpretation (while LS does not).

Maximum Likelihood Estimation

- Assume observations x_ω, $\omega \in \mathcal{S}$ are collected at *random*.
- We thus have that observations are independent RVs.
- Postulate $f(x|\theta)$ and recall $f(x_\omega|\theta)$ is prob (likelihood) that X takes value x_ω.
- Find θ that maximizes *joint* probability that X takes observations x_ω, $\omega \in \mathcal{S}$.
- Joint probability is given by $\prod_{\omega \in \mathcal{S}} f(x_\omega|\theta)$.
- We thus find θ by solving maximization problem:

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

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

Maximum Likelihood Estimation

- It is often convenient to apply a log transformation to solve equivalent problem:

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

- Function $\log L(\theta)$ is known as the log likelihood function.
- Logarithmic transformation is used to avoid scaling issues (as pdf values are small).
- Maximization problem is equivalent to minimization problem:

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

- These problems can be solved by hand when pdf is simple but require numerical techniques when pdf is complex.

Example: Estimation for Exponential using MLE

- Use MLE to estimate β for model $\mathrm{Exp}(\beta)$ from data $x_\omega,\ \omega \in \mathcal{S}$.
- Pdf of exponential RV is $f(x|\beta) = \frac{1}{\beta} e^{-x/\beta}$ and thus likelihood function is:

$$\begin{aligned} L(\beta) &= \left(\frac{1}{\beta} e^{-x_1/\beta}\right)\left(\frac{1}{\beta} e^{-x_2/\beta}\right)\cdots\left(\frac{1}{\beta} e^{-x_S/\beta}\right) \\ &= \frac{1}{\beta^S} \exp\left(-\frac{1}{\beta}\sum_{\omega=1}^{S} x_\omega\right). \end{aligned}$$

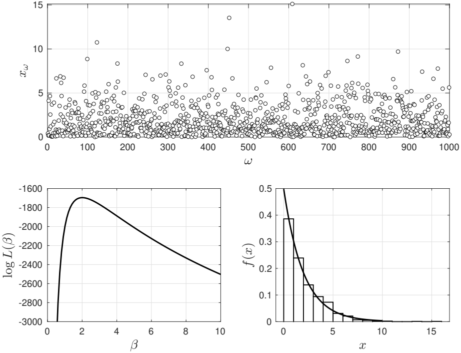
- We find β that maximizes log likelihood $\log L(\beta)$:

$$\max_{\beta}\ \log L(\beta) = -S\log\beta - \frac{1}{\beta}\sum_{\omega=1}^{S} x_\omega$$

- Value of β that maximizes log likelihood satisfies $\frac{\partial \log L(\beta)}{\partial \beta} = 0$.
- Best estimate is $\hat{\beta} = \dfrac{1}{S}\displaystyle\sum_{\omega=1}^{S} x_\omega$; *best* estimate $\hat{\beta}$ is empirical mean $\hat{\mathbb{E}}_X$.

Example: Estimation for Exponential using MLE `ch4_example_mle_exp.m`

- From data provided we find that $\hat{\beta} = \hat{\mathbb{E}}_X = 2$.
- Note that this is precisely location where $\log L(\beta)$ is maximized.
- Note that values of likelihood function $L(\theta)$ are extremely small.
- This is why using a log transformation of likelihood function is necessary.



Data Collection and Asymptotic Properties

- So far, we have assumed that we have data $x_\omega,\ \omega \in \mathcal{S}$.
- However, we have not discussed how this data is being *collected*.
- We also want to know how sample approximations and estimates $\hat{\theta}$ behave as we accumulate data.

We make following observations:

- Data $x_\omega \in \mathcal{S}$ is a set of observations of X collected from a 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 same underlying pdf $f_X(x)$ and if it is drawn *independently* from others; i.e., samples are independent and identically distributed (i.i.d).
- If sample x_ω is selected at random, the sample itself is an RV. Consequently, sometimes we denote a data sample sequence as a sequence of RVs $X_\omega \in \mathcal{S}$.

Data Collection and Asymptotic Properties

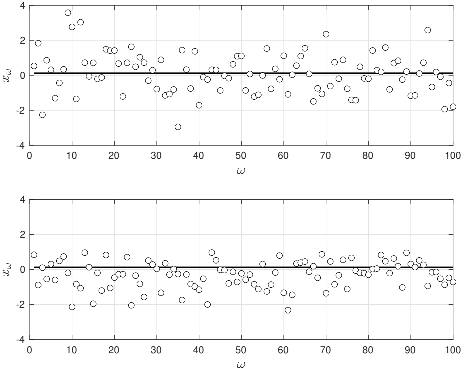
- Random sample X_ω has same probability $1/S$ of being selected and is *unbiased*.
- Lack of bias indicates that $\mathbb{E}[X_\omega] = \mathbb{E}[X]$ (drawing sample many times and averaging results gives expected value of underlying RV X).
- Random samples can be used to construct *approximation* techniques with powerful asymptotic properties.
- 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 difficult to select something randomly (due to inherent biases).

Example: Random and Non-Random Samples `ch4_random_sample.m`

- Random sequence X_ω, $\omega \in \mathcal{S}$ generated by drawing random samples from $X \sim \mathcal{N}(0,1)$. Samples distribute evenly around mean $\mathbb{E}[X] = 0$ and there is no bias.

- Nonrandom sequence X_ω, $\omega \in \mathcal{S}$ is generated by drawing a random sample from $\mathcal{N}(0,1)$ and discarding it if it is above a value of one.

- This introduces a discrimination mechanism that biases sample. Note samples do not distribute evenly around $\mathbb{E}[X] = 0$ and there is a bias.



Monte Carlo Approximations

- *Monte Carlo* (MC) is a computational technique that uses *random samples* to estimate properties of an RV and of its derived quantities (e.g., statistics).

- Consider a random sample $x_\omega \in \mathcal{S}$ of the multivariate X. MC uses the samples to compute a sample (empirical) approximation of an *expectation*:

$$\hat{\mathbb{E}}^{S}_{\phi(X)} = \frac{1}{S} \sum_{\omega \in \mathcal{S}} \phi(x_\omega) \approx \mathbb{E}[\phi(X)]$$

  where $\phi(X)$ is vector-valued function of X.

- We recall that the true (theoretical) expectation for a continuous RV is given by:

$$\mathbb{E}[\phi(X)] = \int_{x \in \mathcal{D}_X} \phi(x) f_X(x) dx,$$

  where $f_X(x)$ is the joint pdf of X.

- Computing expectation thus involves computing a multi-dimensional integral (operation cannot be carried out analytically).

- MC provides a numerical technique to approximate such an integral.

Monte Carlo Approximations

- Many quantities of interest can be expressed as expectation operations and thus can be computed via MC.
- Some examples include:
  - Expectation of X: $\hat{\mathbb{E}}_X^S := \dfrac{1}{S}\displaystyle\sum_{\omega\in\mathcal{S}} x_\omega \approx \mathbb{E}[X]$
  - Variance of X: $\hat{\mathbb{V}}_X^S := \dfrac{1}{S}\displaystyle\sum_{\omega\in\mathcal{S}} (x_\omega - \hat{\mathbb{E}}_X^S)^2 \approx \mathbb{V}[X]$
  - CDF of X: $\hat{F}_X^S(x) := \dfrac{1}{S}\displaystyle\sum_{\omega\in\mathcal{S}} \mathbf{1}[x_\omega \le x] \approx F_X(x)$
  - Expectation of function $\hat{\mathbb{E}}_\varphi^S := \dfrac{1}{S}\displaystyle\sum_{\omega\in\mathcal{S}} \varphi(x_\omega, u) \approx \mathbb{E}[\varphi(X, u)]$.
  - Probability of event: $\hat{\mathbb{P}}(X \in \mathcal{A}) := \frac{1}{S}\sum_{\omega\in\mathcal{S}} \mathbf{1}[x_\omega \in \mathcal{A}] \approx \mathbb{P}(X \in \mathcal{A})$.
- Note dependence of empirical approxs on number of samples S.
- When empirical approx use data, these are called data-driven approxs.
- Empirical approximations are also often called sample average approxs.

Monte Carlo Approximations

- Important questions related to MC approximations are:
  - Do the approximations become exact as we accumulate data?
  - How accurate are these approximations when we have a limited amount of data?
  - How much data do we need to obtain a desired accuracy?
  - How to approximation errors behave as we change our sample data?
- We will now explore tools to answer these questions; the use of these tools is enabled by the fact that samples are collected at random.
- As part of our discussion we will also see that RVs exhibit "universal" behavior as we accumulate data (i.e., as $S \to \infty$).

Law of Large Numbers

- Lets assess quality of empirical approximations as $S \to \infty$.
- Consider an i.i.d. random sequence $X_1, X_2, ..., X_S$ for univariate RV X. Since the samples are i.i.d, they have same underlying pdf with $\mathbb{E}[X]$.
- Since samples are random, they are unbiased $\mathbb{E}[X_1] = \mathbb{E}[X_2] = \cdots = \mathbb{E}[X_S] = \mathbb{E}[X]$.
- Consider now the MC approximation of $\mathbb{E}[X]$:

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

  Because the samples X_ω, $\omega \in \mathcal{S}$ are random, approx $\hat{\mathbb{E}}_X^S$ is also random (i.e., different samples give different approxs).

- There is thus inherent variability in approx $\hat{\mathbb{E}}_X^S$ (variability is a approx accuracy).
- Law of large numbers (LLN) indicates that:

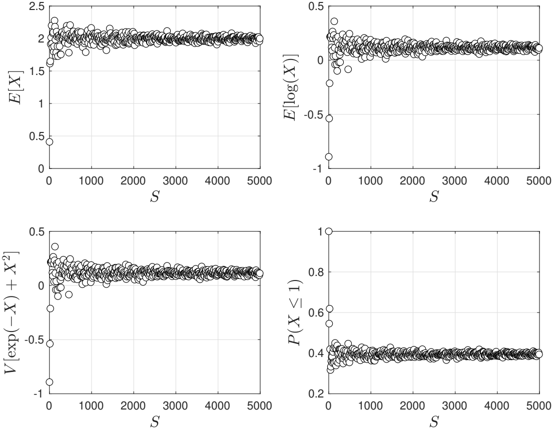
$$\lim_{S \to \infty} \hat{\mathbb{E}}_X^S = \mathbb{E}[X]$$

- LLN guarantees *stable long-run* behavior of sample averages.

Example: Monte Carlo Approximations `ch4_montecarlo_approx.m`

- We generate random samples $X_\omega,\ \omega \in \mathcal{S}$ from $X \sim \text{Weibull}(2,1)$.

- Compute MC approxs for $\mathbb{E}[X], \mathbb{E}[\log(X)]$ and $\mathbb{V}[\exp(X)+X^2]$, and $\mathbb{P}(X >\geq 1)$ for different S.



.

Central Limit Theorem

- Now turn our attention to assessing accuracy of sample approx $\hat{\mathbb{E}}_X^S$.
- Because MC approx is an RV, accuracy is measured in terms of its uncertainty.
- If we want uncertainty of an RV, we need to know its underlying pdf.
- Obtaining pdf of $\hat{\mathbb{E}}_X^S$ can achieved by using a powerful result in statistics known as the central limit theorem (CLT).
- Consider a sample sequence $X_1, X_2, ..., X_S$ that is i.i.d and that has *known* expected value $e = \mathbb{E}[X]$ and variance $v^2 = \mathbb{V}[X]$.
- CLT states that:

$$\lim_{S \to \infty} \hat{\mathbb{E}}_X^S \sim \mathcal{N}(e, v^2/S)$$

- CLT states that, *regardless* of the nature of X (e.g., Weibull, Poisson), sample approximation $\hat{\mathbb{E}}_X^S$ will *always* behave as a Gaussian RV as S increases.

Central Limit Theorem

- CLT states that expected value of $\hat{\mathbb{E}}_X^S$ is true value e. In other words, $\hat{\mathbb{E}}_X^S$ is an unbiased estimate of e.

- CLT states that variance of $\mathbb{V}[\hat{\mathbb{E}}_X^S]$ is given by v^2/S (std dev is $v/\sqrt{S}$) and that this shrinks with S. In other words, $\hat{\mathbb{E}}_X^S$ becomes more accurate as S increases.

- Since we know that $\hat{\mathbb{E}}_X^S \sim \mathcal{N}(e, v/\sqrt{S})$ then we have pdf for $\hat{\mathbb{E}}_X^S$ and thus we can compute quantities of interest for it.

- For example, we can compute $1-\alpha$ confidence intervals:

$$\mathbb{E}_X \in \left[\hat{\mathbb{E}}_X^S \pm z_{\alpha/2} v/\sqrt{S}\right].$$

- Note that this interval shrinks as S increases.

- CLT assumes that we know variance of X; as a result, we need to know these quantities to compute confidence interval.

- In practice, one often replaces this quantity with its MC approx (sample variance):

$$\hat{v}^2 = \frac{1}{S-1} \sum_{\omega \in \mathcal{S}} (x_\omega - \hat{\mathbb{E}}_X^S)^2.$$

Example: Verifying the Central Limit Theorem `ch4_clt_example.m`

- We i.i.d. samples $X_1, ..., , X_S$ from $X \sim \mathrm{Weibull}(2,1)$ and compute $\hat{\mathbb{E}}_X^S = \frac{1}{S}\sum_{\omega=1} X_\omega$.
- We repeat procedure for different sets of samples and visualize pdf of $\bar{X} = \hat{\mathbb{E}}_X^S$.

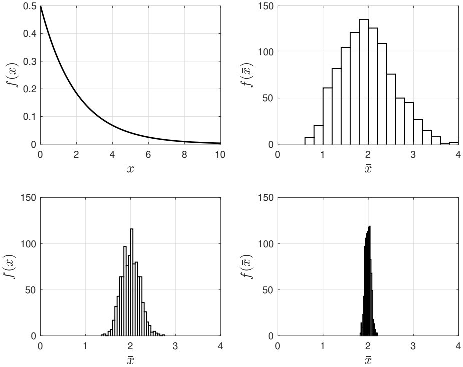


Figure: Pdf for $X \sim \mathrm{Weibull}(2,1)$ (top-left) and sample average $\bar{X}$ with $S = 10$ (top-right), $S = 100$ (bottom-left) and $S = 1000$ (bottom-right).

Extreme Value Theorem

- In CLT we study limiting behavior of *sample average* $\hat{\mathbb{E}}_X^S = \frac{1}{S}\sum_{\omega\in\mathcal{S}} X_\omega$.

- What if we are interested in a different statistic? For instance, the *sample max*:

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

- *Extreme value theorem* (EVT) characterizes pdf of X_{max}^S as $S \to \infty$.

- Consider, as before, an i.i.d. sequence $X_1, X_2, ..., X_S$ for RV X. The EVT states:

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

where $\mathrm{GEV}(a, b, c)$ is generalized extreme value (GEV) RV with params a, b, c.

Extreme Value Theorem

- GEV has cdf:

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

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

- GEV is reverse Weibull (for $c < 0$), Frechet (for $c > 0$), and Gumbel (for $c = 0$).

- Reverse Weibull is a variant of the two-parameter Weibull (explored before).

- GEV RV is used in failure analysis because sample max characterizes *extreme events*.

- This is relevant because such events often have small probabilities (are *rare* events).

- GEV thus gives a mechanism to estimate probabilities for rare events.

Example: Verifying Extreme Value Theorem `ch4_evt_example.m`

- Collect i.i.d. samples $X_1, ..., , X_S$ form $X \sim \mathcal{N}(2,1)$ and compute $X^S_{max} = \max\{X_1, \cdots, X_S\}$.
- Repeat procedure for different sets of samples and visualize pdf of X^S_{max}.
- We do this for different S to verify that pdf follows that of a GEV as we increase S.

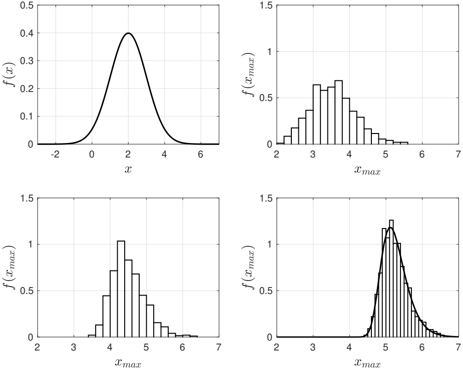


Figure: Pdf for $X \sim \mathcal{N}(2,1)$ (top-left) and sample max X^S_{max} with $S = 10$ (top-right), $S = 100$ (bottom-left) and $S = 1000$ (bottom-right).