Lecture Notes 4
Modeling Input Distributions

- Selecting a Distribution
  - No Data Available vs. Data Available
- Estimating Parameters of Distribution
  - Maximum Likelihood Estimation
  - Method of Moments
- Testing Goodness of Fit of Selected Distribution
  - Chi-Squared Test
  - Kolmogorav-Smirnov Test
  - ARENA's Input Analyzer
- Be sure to check course webpage for reading assignment and HW
- Download and install Arena

Introduction

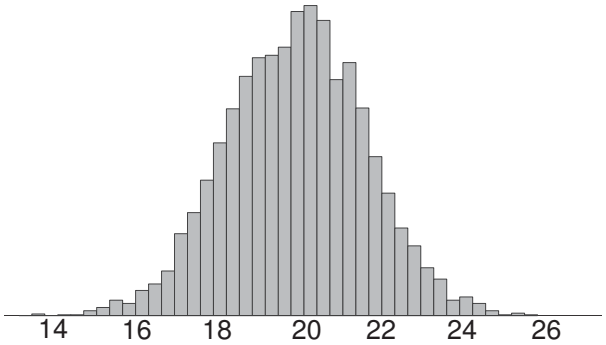
- Developing good intuition for choosing distributions takes some time and experience.
- But a few simple principles are useful in many settings.
- Important to see beyond the mathematical notation and to understand that *every distribution tells a story.*
- The goal is to find the distribution that tells the story you want.

Selecting a Distribution: No Data Available

- In many settings, data needed to carry out simulation study is not available.
- Various factors may account for the absence of data.
- *Data nonexistent.*
  - Suppose you are planning to open a McDonald's restaurant in a foreign country that has no McDonald's yet.
  - Want to know how many cashier lines to operate.
  - You might want to simulate the process of arrivals and demands.
  - But where can you get data?
  - Measuring arrivals and demands at some existing McDonald's is unlikely to provide the information you want.

- *Cost.*
  - There are many kinds of costs associated with sampling.
  - Time required to collect data might render the study obsolete.
  - Physical process of measurement may itself be expensive.
- *Legal barriers.*
  - If you wanted to simulate a competitor's operations, limited (or no) access to task times, demand data, etc.
  - Even in your own factory, union agreements might preclude measuring task times involving workers.

- In the absence of data, try to combine opinions of those involved in operation with informed modeling judgement.
- Usually easier to elicit a subjective range of values than a mean or variance.
- Under an assumption that the range corresponds to plus-or-minus three standard deviations, it can be used to estimate a variance.
- Supplement a likely range of values with considerations of shape, tails, etc., to arrive at a meaningful model.



- Suppose we are given only the histogram, not individual values.
- Based on shape of histogram, we might choose a normal distribution to fit to the data.
- Mean μ appears to be around 20.

- Rough calculation to estimate standard deviation σ:
  - Most of the data lies between 14 to 26.
  - For a normal distribution, 0.997 probability of lying within 3 standard deviations of mean.
  - So $20 \pm 3\sigma$ should roughly give the range of 14 to 26
    $\Longrightarrow \; 6\sigma = 12$
    $\Longrightarrow \; \sigma = 2$

Selecting a Distribution: Data Available

A reasonable procedure:

Step 1. Select a theoretical family of distributions

- Make selection based on shape of histogram of data (and possibly other considerations).

Step 2. Using data, estimate parameters of family (e.g., mean, variance, etc.)

- Once we have chosen values for the parameters, this completely determines the specific member of the family.

Step 3. Evaluate fit of specific distribution to data using goodness-of-fit test

Step 1. Select theoretical family of distributions

- Data to be modeled by a distribution should always be summarized in a histogram.
- With even modest amount of data, histogram reveals important properties of underlying distribution, particularly shape.
- Select theoretical distn (e.g., normal or exponential) based primarily on histogram, but also on physical grounds
  - support
  - tails
  - shape
  - possibly failure rate.

Step 2. From data, estimate parameters of family

- Most theoretical distributions require parameter values for their complete specification.
  - Normal distn has 2 parameters: μ and σ
  - Exponential distn has 1 parameter: rate λ ($=1/\text{mean}$)
  - Gamma distn has 2 parameters: α and β
- After choosing family of distributions, need to specify values for its parameters.
- We consider two methods for estimating parameters of distribution
  - Maximum likelihood estimation (MLE)
  - Method of moments

Step 2. Estimating Parameters of Distributions. Approach 1: Maximum Likelihood Estimation (MLE)

- **Example**: Suppose we have collected 5 interarrival times: 3, 1, 4, 3, 8.
- Suppose data IID and we decided to use exponential distn to describe data. Why is this reasonable?
- Recall density of exponential with rate λ:

$$f_\lambda(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

- Need to estimate value of λ.
- **Main Idea of MLE**:
  - Choose the value of λ under which observed sample is most likely to occur.
  - Given observation x, *likelihood* of parameter λ is $f_\lambda(x)$.

- For our example, first compute likelihood of parameter λ for each observation:
  - Likelihood for first observation (3): $\lambda e^{-3\lambda}$
  - Likelihood for second observation (1): $\lambda e^{-1\lambda}$
  - Likelihood for third observation (4): $\lambda e^{-4\lambda}$
  - Likelihood for fourth observation (3): $\lambda e^{-3\lambda}$
  - Likelihood for fifth observation (8): $\lambda e^{-8\lambda}$
- Since data assumed independent, likelihood $L(\lambda)$ for whole sample is product of individual likelihoods:

$$\begin{aligned} L(\lambda) &= \lambda e^{-3\lambda} \lambda e^{-1\lambda} \lambda e^{-4\lambda} \lambda e^{-3\lambda} \lambda e^{-8\lambda} \\ &= \lambda^5 e^{-19\lambda} \end{aligned}$$

- The estimator $\widehat{\lambda}$ of λ is maximizer of $L(\lambda)$, or equivalently, of the **log likelihood** $\tilde{L}(\lambda) = \ln L(\lambda)$.

- In this example,
$$\tilde{L}(\lambda) \;=\; 5\,\ln(\lambda) - 19\lambda$$
- Find $\lambda > 0$ that maximizes $\tilde{L}(\lambda)$.
  - Solving $\frac{d}{d\lambda}\tilde{L}(\lambda) = 0$ yields $\hat{\lambda} = 5/19$.
  - $\frac{d^2}{d\lambda^2}\tilde{L}(\hat{\lambda}) < 0$, so $\hat{\lambda} = 5/19$ is a local maximum.
  - $\tilde{L}(0) = -\infty$ at boundary $\lambda = 0$.
- Is $\hat{\lambda} = 5/19$ a reasonable estimate of λ?
  - $19/5$ is sample mean of data $(3, 1, 4, 3, 8)$.
  - Sample mean should be close to true mean.
  - But for exponential distribution,
$$\lambda = (\text{true mean})^{-1}$$
  - So reasonable to estimate λ by
$$\hat{\lambda} = (\text{sample mean})^{-1}$$

- In general, given sample $X_1, \ldots, X_n \sim$ IID exponential(λ) with unknown rate $\lambda > 0$, likelihood of λ is
$$L_n(\lambda) \;=\; \prod_{i=1}^{n} \lambda e^{-\lambda X_i} \;=\; \lambda^n \exp\left(-\sum_{i=1}^{n} \lambda X_i\right)$$
- So log likelihood $\tilde{L}_n(\lambda) = \ln L_n(\lambda)$ is just
$$\tilde{L}_n(\lambda) = n\,\ln(\lambda) - \sum_{i=1}^{n} \lambda X_i$$
- Solving for $\frac{d}{d\lambda}\tilde{L}_n(\lambda) = 0$ yields global maximizer of $\tilde{L}_n(\lambda)$ as
$$\hat{\lambda} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^{-1}$$
(also check $\frac{d^2}{d\lambda^2}\tilde{L}_n(\lambda) < 0$ and boundary $\lambda = 0$).
- So $\hat{\lambda}$ is the MLE.

General Formulation of MLE

- Given sample $X_1, \ldots, X_n$ of IID observations from distn F
- F has k unknown parameters $\theta_1, \ldots, \theta_k$; so write $F_{\theta_1,\ldots,\theta_k}$
- Likelihood $L_n(\theta_1, \ldots, \theta_k)$ of parameters $\theta_1, \ldots, \theta_k$ for sample $X_1, \ldots, X_n$:
  - If $F_{\theta_1,\ldots,\theta_k}$ is continuous distn with density $f_{\theta_1,\ldots,\theta_k}$,
$$L_n(\theta_1, \ldots, \theta_k) = \prod_{i=1}^{n} f_{\theta_1,\ldots,\theta_k}(X_i)$$
  - If $F_{\theta_1,\ldots,\theta_k}$ is discrete distn with pmf $p_{\theta_1,\ldots,\theta_k}$,
$$L_n(\theta_1, \ldots, \theta_k) = \prod_{i=1}^{n} p_{\theta_1,\ldots,\theta_k}(X_i)$$
- MLE is the value $(\hat{\theta}_1, \ldots, \hat{\theta}_k)$ of $(\theta_1, \ldots, \theta_k)$ that maximizes $L_n(\theta_1, \ldots, \theta_k)$ [or equivalently $\ln L_n(\theta_1, \ldots, \theta_k)$].
  - Often use calculus to find $(\hat{\theta}_1, \ldots, \hat{\theta}_k)$ but also check boundary values.

Example:

- Suppose $X_1, \ldots, X_n$ are IID $\mathcal{N}(\mu, \sigma^2)$ with unknown μ, σ^2.
- Recall density of $\mathcal{N}(\mu, \sigma^2)$ is
$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$
- Then
$$L_n(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{-\frac{(X_i-\mu)^2}{2\sigma^2}\right\}$$
- Can show the pair $(\hat{\mu}, \hat{\sigma}^2)$ maximizing $L_n(\mu, \sigma^2)$ is
$$\hat{\mu} \;=\; \bar{X}_n \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i,$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2 \;\neq\; S_n^2.$$

Example:

- Suppose $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are IID multivariate normal $\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - $\mathbf{X}_k = (X_{k,1}, X_{k,2}, \ldots, X_{k,d})^\top \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - Covariance matrix $\boldsymbol{\Sigma} = (\Sigma_{i,j} : i, j = 1, 2, \ldots, d)$, with $\Sigma_{i,j} = \mathcal{E}[(X_{k,i} - \mu_i)(X_{k,j} - \mu_j)]$
- Can show that MLE of unknown $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are

$$\widehat{\boldsymbol{\mu}} = \bar{\mathbf{X}}_n = \frac{1}{n} \sum_{k=1}^{n} \mathbf{X}_k,$$

$$\widehat{\Sigma}_{i,j} = \frac{1}{n} \sum_{k=1}^{n} \left(X_{k,i} - \bar{X}_{n,i}\right)\left(X_{k,j} - \bar{X}_{n,j}\right),$$

  where $\bar{X}_{n,i} = (1/n) \sum_{k=1}^{n} X_{k,i}$
- Also, MLE of correlation coefficient $\rho_{i,j}$ of $X_{k,i}$ and $X_{k,j}$ is

$$\widehat{\rho}_{i,j} = \frac{\widehat{\Sigma}_{i,j}}{\sqrt{\widehat{\Sigma}_{i,i}\,\widehat{\Sigma}_{j,j}}}$$

Example:

- Suppose $X_1, \ldots, X_n$ are IID gamma(β, α) with β, α unknown.
- Recall density of gamma(β, α) is

$$f_{\beta,\alpha}(x) = \frac{\beta^{-\alpha}\, x^{\alpha-1}\, e^{-x/\beta}}{\Gamma(\alpha)}, \quad x \geq 0$$

- Thus,

$$L_n(\beta, \alpha) = \prod_{i=1}^{n} \frac{\beta^{-\alpha}\, X_i^{\alpha-1}\, e^{-X_i/\beta}}{\Gamma(\alpha)}$$

- No closed-form solution for maximizer $(\widehat{\beta}, \widehat{\alpha})$ of $L_n(\beta, \alpha)$.
  - MLE must be computed numerically.

Remarks on MLE

- When parametric family has nice mathematical properties, MLE can be computed in closed-form.
- However, one must typically be prepared for having to compute MLEs by numerically optimizing likelihood function.
- Given potential difficulty of this task, why use MLEs?
- Answer: Vast statistical literature yields meta-principle:
  - For large samples (i.e., n large), MLEs squeeze all available statistical information about the parameters out of sample.
  - Mathematically, MLEs (asymptotically) have smallest variance of any possible parameter estimator.
  - In other words, MLEs maximize "statistical efficiency".

More Remarks on MLE

- Suppose $\widehat{\theta}$ is MLE of parameter θ of distribution F.
- Consider $\tau = g(\theta)$ for some function g.
- Then MLE of τ is $\widehat{\tau} = g(\widehat{\theta})$.
- Thus, suppose we have simulation model with input distributions $F_1, F_2, \ldots, F_m$ having parameters $\theta_1, \theta_2, \ldots, \theta_m$.
  - Each θ_i can be a vector of parameters.
- Can think of performance measure α of system as

$$\alpha = g(\theta_1, \theta_2, \ldots, \theta_m).$$

- Hence, if $\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_m$ are MLEs of $\theta_1, \theta_2, \ldots, \theta_m$, then MLE of α is

$$\widehat{\alpha} = g(\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_m).$$

Step 2. Estimating Parameters of Distributions. Approach 2: Method of Moments

- Alternative to MLE is the **method of moments** (MoM)
- For some families of distributions,
  - MLEs may be analytically intractable (and hence require numerical optimization).
  - Method of moments may still have closed-form expressions for estimators.
- Idea:
  - Equate true (population) moments to sample moments
  - Solve for unknown parameters from resulting equations
  - Specifically, if k parameters to estimate, equate true jth moment to sample jth moment, for $j = 1, 2, \ldots, k$.

- Suppose
  - Sample $X_1, \ldots, X_n$ of IID observations from some distn F
  - F has k unknown parameters $\theta_1, \ldots, \theta_k$.
- True jth moment of X is $\mu_j \equiv \mathcal{E}[X^j]$, where $X \sim F$.
  - Each μ_j is fcn of the k parameters; i.e., $\mu_j = \mu_j(\theta_1, \ldots, \theta_k)$
- Sample jth moment of $X_1, \ldots, X_n$: $M_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$
- Solve for $(\theta_1, \ldots, \theta_k)$ in k simultaneous eqns:

$$\begin{aligned} \mu_1(\theta_1, \ldots, \theta_k) &= M_1, \\ \mu_2(\theta_1, \ldots, \theta_k) &= M_2, \\ &\vdots \\ \mu_k(\theta_1, \ldots, \theta_k) &= M_k. \end{aligned}$$

- Solution $(\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ is method-of-moments estimator of $(\theta_1, \ldots, \theta_k)$.

Example:

- Suppose $X_1, \ldots, X_n \sim$ exponential with unknown rate λ.
- So number of unknown parameters is $k = 1$.
- Recall the true (population) mean is $\mu_1 = 1/\lambda$.
- Compute the sample mean of $X_1, \ldots, X_n$:

$$\bar{X}_n = M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Equate true and sample first moments, and solve for λ:

$$\mu_1 = \frac{1}{\lambda} = \bar{X}_n \quad \Rightarrow \quad \widehat{\lambda} = \frac{1}{\bar{X}_n}$$

- So method-of-moments estimator of λ is same as MLE of λ.
  - For other distributions, MoM and MLE estimators may differ.

Example:

- Suppose $X_1, \ldots, X_n$ are IID gamma(β, α).
  - Number of unknown parameters is $k = 2$
  - Mean is $\alpha\beta$, so $\mu_1 = \alpha\beta$
  - Variance is $\alpha\beta^2$, so $\mu_2 = \alpha\beta^2 + \alpha^2\beta^2$
- Compute sample first moment: $M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Compute sample second moment: $M_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$
- Set $\alpha\beta = M_1$ and $\alpha\beta^2 + \alpha^2\beta^2 = M_2$ and solve for α, β:

$$\widehat{\alpha} = \frac{M_1^2}{M_2 - M_1^2}, \qquad \widehat{\beta} = \frac{M_2 - M_1^2}{M_1}$$

- So for Gamma distribution,
  - Can solve analytically for method-of-moments estimator
  - We previously saw that we needed to use numerical methods for MLE.
- However, the MLE squeezes more information out of the sample than does the method of moments.
- In particular, the MLEs have smaller variance.

Step 3. Evaluate fit of specific distribution to data using goodness-of-fit test

- Once distribution and its parameters have been selected, evaluate fit between
  - chosen theoretical distribution, and
  - observed data.
- We restrict ourselves to
  - χ^2 test
  - Kolmogorov-Smirnov test
- Other methods exist (Anderson-Darling test, etc.)

Step 3. Goodness-of-Fit Test. Approach 1: χ^2 Test

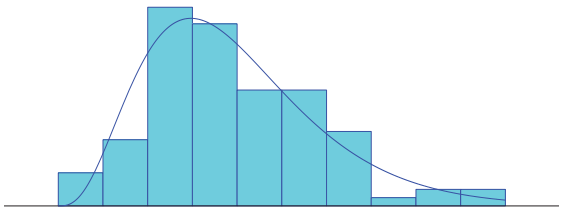
- When observed data is available, can use χ^2 test as method for evaluating how well a theoretical distribution fits observed data.
- Its relevance to simulation arises as follows:
  - Suppose you have collected data on, e.g., intearrival times, processing times, stock's daily returns.
  - You want to select a theoretical distribution that approximates your data and use that distribution in a simulation.
  - Which distribution should you use?
- Basic Idea of χ^2 test:
  - **Compare the density** of chosen distribution **to histogram** of the data.

The Statistical Framework of χ^2 Test

- As a statistical procedure, χ^2 test follows the general paradigm of hypothesis testing.
- Specify null and alternative hypotheses:

  H_0 : data come from selected distribution

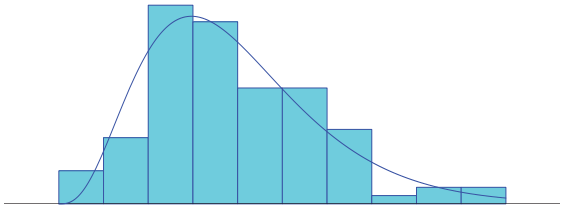
  H_A : data do not come from selected distribution
- **Basic Idea:** Under H_0, density of hypothesized distribution should look like histogram of data.



- Recall that H_0 gets benefit of the doubt.
  - As long as test statistic is not too extreme, do not reject H_0.
- Hence, hypothesis testing makes sense only if we have good reason to believe H_0 in the first place.
  - **Simply picking a distribution arbitrarily and then applying a χ^2 test is a bad idea.**
- To execute the test, we need to select a significance level, α.
- Recall that α is the maximum probability of a Type I error
  - $\alpha =$ the maximum probability of incorrectly rejecting H_0.
- Failing to reject H_0 with a *small* α gives more credibility to H_0.
  - Typically choose $\alpha = 0.1$, 0.05, or 0.01

The Mechanics of χ^2 Test

1. Partition range of possible values into k bins.
   - We discuss how to select bins below.
2. For $i = 1, \ldots, k$, let
   - O_i be observed number of data points in bin i.
   - E_i be expected number of observations in bin i under hypothesized theoretical distribution F.



- Specifically, suppose
  - Bins $(a_1, a_2]$, $(a_2, a_3], \ldots,$ $(a_k, a_{k+1}]$
  - Sample size is n.
- Then probability of falling in bin i is

$$\mathcal{P}(X \in (a_i, a_{i+1}]) = F(a_{i+1}) - F(a_i)$$

- Thus, if n observations, expected number falling in bin i is

$$E_i = n \cdot (F(a_{i+1}) - F(a_i)).$$

  - e.g., in exponential case $F(x) = 1 - \exp(-x/\beta)$, so $E_i = n \cdot ([1 - e^{a_{i+1}/\beta}] - [1 - e^{a_i/\beta}]) = n \cdot (e^{a_i/\beta} - e^{a_{i+1}/\beta})$.
- If theoretical distribution is discrete, and if the bins exactly correspond to the possible values, then $E_i = np_i$
  - p_i is the probability of the i-th value in distn's support.

3. Compute *test statistic*

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

4. Test statistic is compared with critical value $\chi^2_{\alpha,\nu}$ from χ^2 table, where ν is the number of degrees of freedom given by

$$\nu = \text{number of bins} - (1 + \text{number of estimated parameters})$$

   - Exponential has one parameter (the mean)
   - Normal has two parameters (mean and variance)

Alternatively, compute p-value of test statistic

- Reject H_0 iff p-value $<$ level α.

Selecting Bins for χ^2 Test

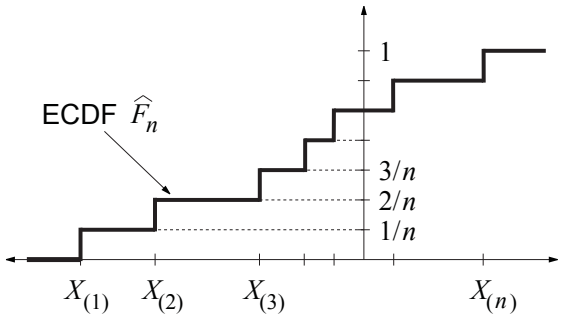
- A serious shortcoming of χ^2 test is the lack of a systematic way of selecting bins.
- There are, however, some rules-of-thumb that can be followed.
- Since histogram shape guides our subsequent analysis, proper graphing is important.
- Generally, bins for initial histogram should be equally spaced.
- Some suggest choosing number of bins roughly equal to square-root of sample size; i.e., $k = n^{1/2}$
- Also generally suggested that number of expected observations in every bin be greater than or equal to 5.

- Once a theoretical distribution has been selected, it's a good idea to tailor bins to hypothesized fit.
  - If theoretical dist discrete, each possible value should be a bin.
  - If it is continuous, then good idea to make bins *equiprobable*.
    - ▷ i.e., select bin endpoints so that
    $$F(a_2) - F(a_1) = \cdots = F(a_{k+1}) - F(a_k) = 1/k.$$
    - ▷ This makes expected number of observations per bin equal for all bins.
    - ▷ Generally not easy unless table of values available.
    - ▷ In case of exponential, we can achieve this by setting $a_1 = 0$, and solving
    $$F(a_{i+1}) - F(a_i) = 1/k$$
    to get $a_{i+1} = -\beta \ln(1 - (i/k))$.

Step 3. Goodness-of-Fit Test. Approach 2: Kolmogorov-Smirnov (K-S) Test

- Hypothesis $H_0 : X_1, X_2, \ldots, X_n \sim F$, IID
- Under H_0, CDF $F(x) = \mathcal{P}(X_i \le x)$
- Define **empirical CDF**
$$\begin{aligned} \widehat{F}_n(x) &= (\text{fraction of observations} \le x) \\ &= \frac{1}{n} \sum_{i=1}^{n} I\{X_i \le x\} \end{aligned}$$
  - $I\{A\}$ is the indicator function of the event $\{A\}$; $I\{A\} = 1$ if $\{A\}$ is true, and $I\{A\} = 0$ if $\{A\}$ is false.
- **Basic Idea:** Under H_0, likely that $F \approx \widehat{F}_n$

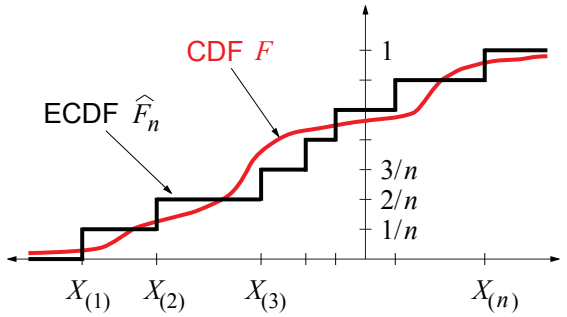
- Recall: empirical CDF
$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I\{X_i \le x\}$$
- Define **order statistics** $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ of observations $X_1, \ldots, X_n$
  - $X_{(i)}$ is ith smallest value from $X_1, X_2, \ldots, X_n$.



- Under null hypothesis $H_0 : X_1, \ldots, X_n \sim F$, IID
  - Hypothesized CDF F likely to be "close" to empirical CDF $\hat{F}_n$
  - Likely that $F \approx \hat{F}_n$; i.e., error

$$D_n = \max_{x \in \mathbf{R}} | \hat{F}_n(x) - F(x) | \text{ likely to be "small"}$$

  - Glivenko-Cantelli Thm: $D_n \to 0$ as $n \to \infty$ with prob 1.



- Error D_n measures max vertical distance between $\hat{F}_n$ and F:

$$D_n = \max_{x \in \mathbf{R}} | \hat{F}_n(x) - F(x) |$$

- Can compute D_n as follows:
  - Define

$$D_n^+ = \max_{1 \le i \le n} \left\{ \frac{i}{n} - F(X_{(i)}) \right\},$$
$$D_n^- = \max_{1 \le i \le n} \left\{ F(X_{(i)}) - \frac{i-1}{n} \right\}$$

  - Then $D_n = \max\{D_n^+, D_n^-\}$
- Natural test : Reject hypothesis H_0 if D_n is "large"
  - What does "large" mean?

Case	Test Statistic	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
All pars known	$\left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}\right) D_n$	1.224	1.358	1.628
Normal $\mathcal{N}(\bar{x}, s_n^2)$	$\left(\sqrt{n} - 0.01 + \frac{0.85}{\sqrt{n}}\right) D_n$	0.819	0.895	1.035
Exponential $\exp(1/\bar{x})$	$\left(\sqrt{n} + 0.26 + \frac{0.5}{\sqrt{n}}\right) \left(D_n - \frac{0.2}{n}\right)$	0.990	1.094	1.308
Weibull (α, β)	$\left(\sqrt{n}\right) D_n$	0.803	0.874	1.007

- **Kolmogorov-Smirnov test** :
  - Choose level $\alpha = 0.01, 0.05$, or 0.1.
  - Calculate the value of the error D_n corresponding to the data
  - **Reject** H_0 iff test statistic is greater than critical value in table above (or p-value $<$ level α)

Kolmogorov-Smirnov Test: Example

- Unsorted sample: $X_1, X_2, \ldots, X_{10}$

0.360 0.549 0.262 0.597 0.049
0.571 0.701 0.962 0.751 0.740

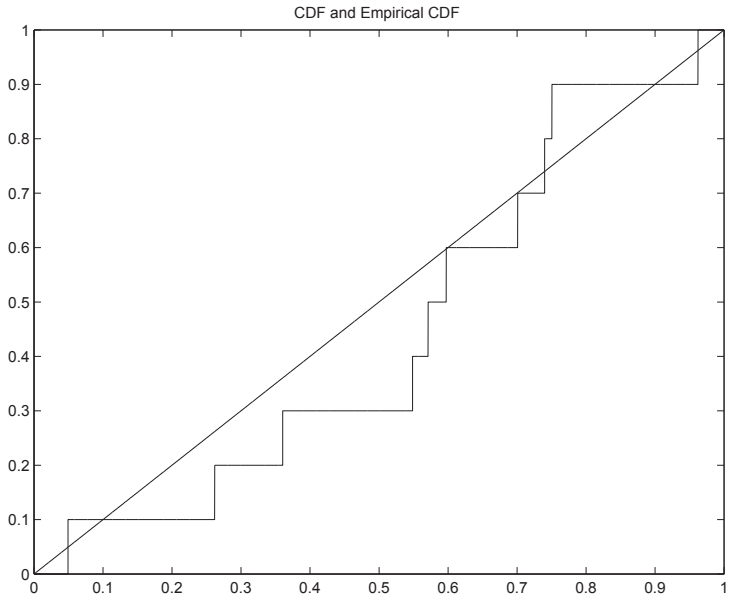
$n = 10$ and set level $\alpha = 0.05$

- Data hypothesized to be unif[0, 1], which has CDF

$$F(x) = x \quad \text{for} \quad x \in [0, 1]$$

- Construct empirical CDF
$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n I\{X_i \le x\} =$ (fraction of X_i's $\le x$)
- Sorted sample: $X_{(1)}, X_{(2)}, \ldots, X_{(10)}$

0.049 0.262 0.360 0.549 0.571
0.597 0.701 0.740 0.751 0.962



- Compute the error $D_n = \max_{x \in [0,1]} \mid \hat{F}_{10}(x) - F(x) \mid$

  $D_n = 0.2485$ and is achieved at $x = 0.549$

- Compare test statistic to critical value for $\alpha = 0.05$ from table:

$$\left(\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}}\right) D_n \;=\; 0.8243 \;<\; 1.358$$

- Reject or don't reject H_0?
  - Test statistic $0.8243 < 1.358$
  - Therefore, **don't reject**.

Comparison of K-S and χ^2 tests

K-S test

- Compares hypothesized CDF F to empirical CDF $\hat{F}_n$ from data.
- Value of test statistic depends on each observation X_i

χ^2 test

- Compares hypothesized density function f to histogram of data
  - Histogram is like an empirical density function.
- Value of test statistic depends on how bins are selected.
- Test statistic *does not* differentiate between observations in bin.
  - χ^2 test is less powerful than K-S test
- χ^2 test requires larger sample sizes than K-S test

ARENA Input Analyzer

- Arena's built-in tool to fit distributions to data.
- We now look at some examples using this tool.
- Table 1 contains 25 points generated from exponential distribution with mean 1.
- Figure 1 displays a histogram of the data.
- Suppose we didn't know data came from exponential distribution.
- We might still hypothesize an exponential fit, i.e.,

  H_0 : data came from exponential distribution,

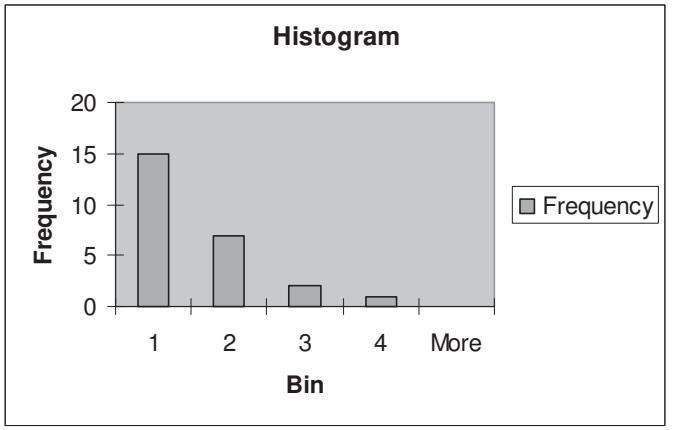
  H_A : data did not come from exponential distribution.

Example 1

0.770	0.238	1.948	0.478	0.882
1.326	1.024	1.306	1.773	2.026
0.012	0.429	2.411	1.747	0.076
0.110	0.290	3.248	0.348	1.405
0.348	0.127	0.427	0.940	0.055

- Table 1: 25 random observations generated from an exponential distribution with mean 1

Figure 1: Histogram of Data from Table 1



ARENA's Input Analyzer

1. Create text file
   - Individual data values separated by one or more "white space characters" (blank spaces, tabs, or linefeeds).
   - Give file any name with a .dst extension.

2. Start the Input Analyzer from the Start Menu in Windows (Programs → Rockwell Software → Arena → Input Analyzer).
   - Alternatively, you can launch the Input Analyzer from within Arena.

3. Go to File menu, and choose New.
   - Alternatively, you can click on the New button on the tool bar.

4. Go to File menu, choose *Data File* → *Use Existing*.
   - Alternatively, click on *Use Existing Data File* button on tool bar.
   - A window then appears, and choose the data (.dst) file that you created in step 1.

5. Go to Fit menu, and choose dist to which you want to fit data.
   - Can change number of intervals used in displayed histogram
     - Go to *Options* menu
     - Choose *Parameters* $\rightarrow$ *Histogram*.
     - Specify number of intervals in the *Intervals* section.
       - ▷ This will change the displayed histogram
       - ▷ But not exactly clear how this changes number of bins used in the χ^2 test.
       - ▷ Increasing *Intervals* seems to increase number of bins for χ^2 test, but the way this occurs is not clear.

6. You then will get a window that shows the results of the fit(s).

Bad Idea: choose *Fit All* in Step 5.

- Fit all of Arena's built-in distributions to data
- Rank them in terms of the squared error.
  - Go to *Window* menu and choose *Fit All Summary*.
- The problem with *Fit All* is that hypothesis testing should only be done when you have an a priori belief that H_0 might be true.
  - With *Fit All*, you are indiscriminately testing hypotheses that you don't *a priori* believe are true.

- First, suppose that we fit exponential distribution to data in Table 1.

```
Distribution Summary

Distribution: Exponential
Expression: EXPO(0.943)
Square Error: 0.005978

Chi Square Test
  Number of intervals = 2
  Degrees of freedom  = 0
  Test Statistic      = 0.535
  Corresponding p-value < 0.005
```



```
Kolmogorov-Smirnov Test
  Test Statistic = 0.11
  Corresponding p-value > 0.15

Data Summary

Number of Data Points = 25
Min Data Value        = 0.0121
Max Data Value        = 3.25
Sample Mean           = 0.943
Sample Std Dev        = 0.863

Histogram Summary

Histogram Range     = 0 to 3.58
Number of Intervals = 5
```

- Input Analyzer fit an exponential distribution with mean 0.943
  - p-value for χ^2 test is less than 0.005
  - p-value is probability of getting a data set that's more inconsistent with the fitted distribution than the data set you actually got, if the fitted distribution is truly "the truth"
  - Reject H_0 iff p-value $< \alpha$.
  - This indicates that we should reject null hypothesis.
  - However, note that *Number of intervals* is 2, which leads to the degrees of freedom being 0, which doesn't make any sense.

- Change number of intervals to 6 using *Options* $\rightarrow$ *Parameters* $\rightarrow$ *Histogram*
  - Then χ^2 test uses 3 intervals and 1 degree of freedom
  - p-value is now 0.129
  - So we would not reject (accept) the H_0 at level $\alpha = 0.05$
  - Arena does not specify if its χ^2 test uses equiprobable or equal-length bins.

- If we choose *Fit All* option in Step 5, we get following results (go to *Window* $\rightarrow$ *Fit All Summary*):

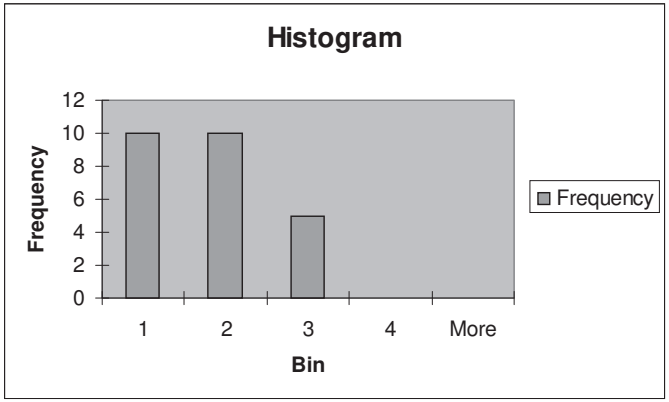
Function	Sq Error
Beta	0.00471
Erlang	0.00598
Exponential	0.00598
Gamma	0.00611
Weibull	0.0066
Lognormal	0.0317
Triangular	0.0501
Normal	0.0531
Uniform	0.138

- Beta distribution has finite support (i.e., a beta random variable can only take on values within a finite interval)
- Exponential distribution has infinite support.
- Thus, using beta distribution in model rather than an exponential distribution can seriously affect outcome of simulation.
- **Lesson**: Indiscriminately trying to fit a distribution to data may lead to misleading conclusions.

Example 2

0.789	0.664	1.386	2.256	0.243
0.454	0.458	2.296	0.414	2.799
2.094	1.070	2.015	1.095	1.564
0.880	0.216	1.878	1.986	0.483
0.257	1.597	1.661	1.129	1.507

- Table 2: Another 25 random observations from an exponential distribution with mean 1.



- Results from *Fit All* applied to data in Table 2:

Function	Sq Error
Beta	0.00926
Triangular	0.017
Erlang	0.0241
Weibull	0.0269
Gamma	0.0284
Lognormal	0.034
Uniform	0.0352
Normal	0.0354
Exponential	0.036

- Beta distribution ranked first, with p-value of 0.564 from χ^2 test.
- Exponential distribution is ranked last!

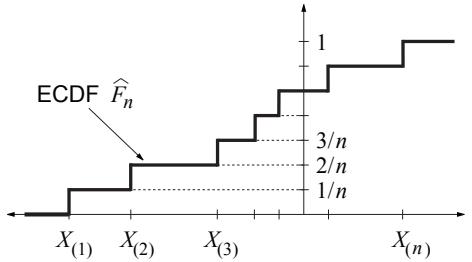
Conclusions

- What can we make of this?
- Failing to reject (accepting) H_0 should not be taken as confirmation that data indeed came from hypothesized dist.
- Especially true when amount of data is small.
- χ^2 test is most useful in knocking out an extremely bad fit.
- In general, a subjective assessment of shape of histogram combined with understanding of data source is more powerful than any statistical procedure.

- Arena's Input Analyzer can fit a distribution with specific parameters (e.g., an exponential with mean 1) to the data.
  - *Options* menu, selecting *Parameters* $\rightarrow$ *Distribution*
- Many other commercial software packages will automatically fit standard probability distributions to empirical data
  - "Best Fit" by Palisade Corp.
  - "UniFit II" by Averill M. Law & Associates.

Alternative approach for input modeling

- Rather than fit hypothesized distribution F to data $X_1, \ldots, X_n$, can instead directly use empirical CDF $\widehat{F}_n$ in simulation model.



  - $\widehat{F}_n$ assigns probability $1/n$ to each original observation X_i
- **Drawback:** For $\widehat{F}_n$, not possible to get observation $X \neq X_i$
- **Solution:** Use "smoothed" version of empirical CDF, e.g., via linear interpolation