IE630: Simulation Modelling & Analysis
Input Data Analysis

Saurabh Jain

Assistant Professor

Department of Industrial Engineering and Operations Research

IIT Bombay

IIT BOMBAY

IEOR IIT BOMBAY

Quick Recap

Purpose & Overview

- Deterministic: Non-random, fixed values
  - Number of resources units
  - Entity arrival time, entity travel time
  - Processing time
- Random: Model as a distribution, ( ) or ( ) values from ( )?
  - Interarrival times, processing times
  - Distribution?
  - Parameters?
  - Affects on the simulation output?

Purpose & Overview

Manufacturing
Machine breakdown, Quality, processing times, arrival rates, worker availability and productivity

Healthcare
Patient arrival rate, service time for procedures, patient no show rates, treatment outcomes, equipment failures

Transportation
Arrival and Departure of vehicles, traffic congestions and delays, fuel consumption, loading/unloading times, weather conditions

Supply Chain & Retail
Customer arrivals, demand fluctuations, lead times, supplier reliability, stockout rates

Many more……

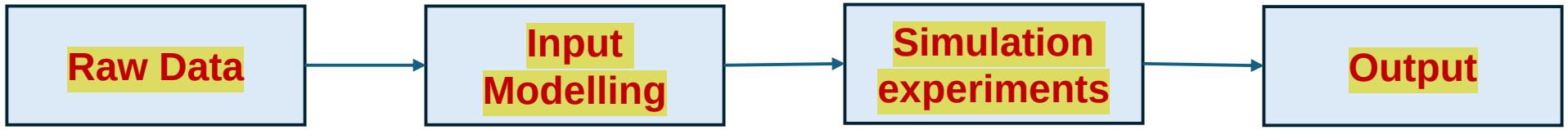
Purpose & Overview

- Key driving force for the simulation model
- Quality of output is no better than the quality of inputs
- Selection of probability distribution to model one or more sources of randomness
- **Focus:**
  - Data Collection
  - Identification of probability distribution
  - Parameter estimation of the distribution
  - Evaluation of the chosen distribution

Data Collection

Data Collection

- Tasks in studying/solving a real-world problem
  - GIGO – Garbage-In-Garbage-Out



- Even with valid model structure, simulation results can be misleading, if the input data is
  - ✔ inaccurately collected
  - ✔ inappropriately analyzed
  - ✔ not representative of the environment

Data Collection

- Data is an integral part of any model
- Matching between the quality of data fitting the requirements of the model
- Sensitivity of outputs to uncertainty in inputs
- Variability in data – Model validity
- Cost
- Access
- Methods
  - Sensors & Surveillance methods
  - Human centered data – Time & Motion study
  - Electronic records – Health records, tracking records
  - Survey, Polls, People interview

Data Collection

- Prepare in Advance:
  - Conduct practice or pre-observation sessions.
  - Watch for unusual circumstances that might skew data.
- Monitor Data During Collection:
  - Continuously check the adequacy and quality of data being gathered.
- Merge Similar Data Sets:
  - Combine data from successive time periods.
  - Integrate data from identical time frames on different days.
- Account for Data Censoring:
  - Identify instances where quantities are only partially observed.
  - Avoid omitting long process times or critical observations.

Data Collection

- Identify Variable Relationships:
  - Utilize scatter diagrams to assess potential correlations.
- Detect Autocorrelation:
  - Ensure that data points are not unduly influenced by preceding observations unless intended.
- Prioritize Input Data Collection:
  - Focus on collecting raw input data rather than performance metrics for accurate analysis.

Using Data: Alternatives and Issues

- Direct Use of Observed Data:
  - Utilize actual observed values (e.g., interarrival times, service durations, part types) to drive model inputs.
  - Ensures all values are valid and realistic.
  - **Limitations:**
    - Cannot extrapolate beyond the observed data range.
    - May lack sufficient data for extended or multiple simulation runs.
    - Computationally intensive due to reliance on disk file reads.
- Fitting Probability Distributions to Data:
  - Generate synthetic observations from a fitted distribution to drive model inputs.
  - **Advantages:**
    - Can extend beyond the range of observed data, offering more flexibility (which could be beneficial or detrimental).
  - **Challenges:**
    - Poor fit to the data may affect the model's validity.

Fitting Distributions: Some Important Issues

- ✔ No Exact Science: no single "correct" answer when selecting a distribution.
- ✔ Theoretical vs. Empirical Considerations: Weigh the benefits of using a theoretical probability distribution versus directly using observed data.
- ✔ Range of Distributions:
  - Unbounded (Infinite in Both Directions): Normal distribution.
  - Positive-Only Distributions: Exponential, gamma.
  - Bounded Distributions: Beta, uniform.
- ✔ Parameter Flexibility: Choose a distribution that allows easy adjustment of parameters to control mean and variance.
- ✔ Sensitivity Analysis: Assess how changes in distribution parameters impact the simulation model.
- ✔ Handling Outliers & Multimodal Data:
  - Consider splitting the dataset if multiple modes or extreme values exist.

Empirical Distribution

- Discrete Empirical Distribution

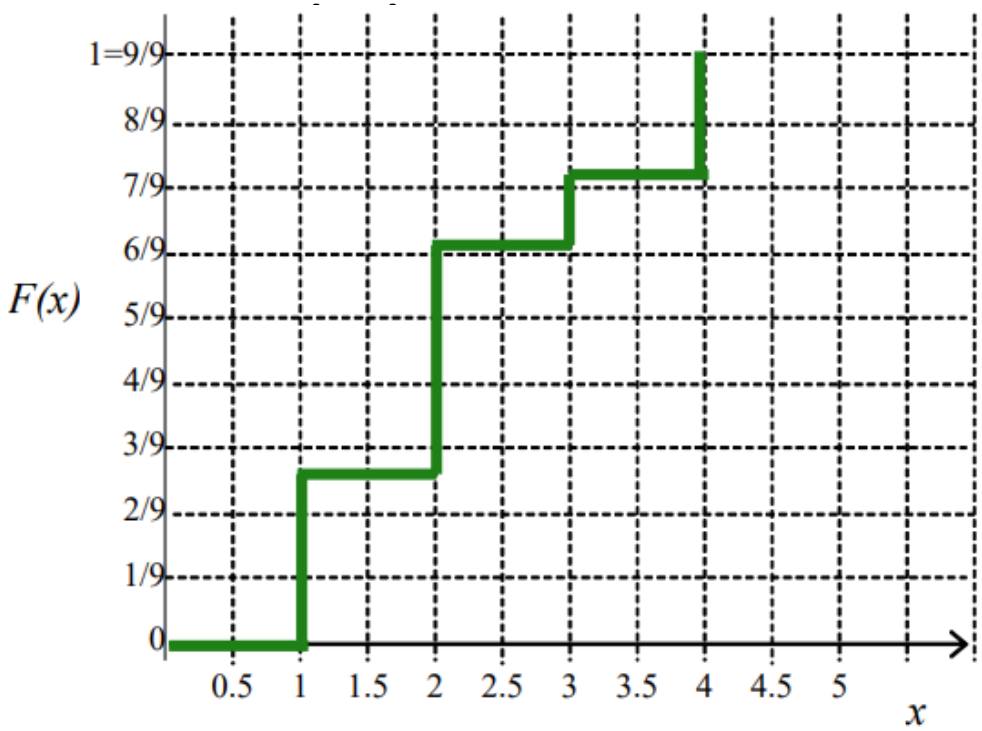
X_i = 1,2,2,1,2,4,3,4,1,2

- Continuous empirical distribution

- X_i = 1.7, 1.8, 1.9, 2.0, 2.1, 2.3, 2.5, 3.6, 4.1, 4.4

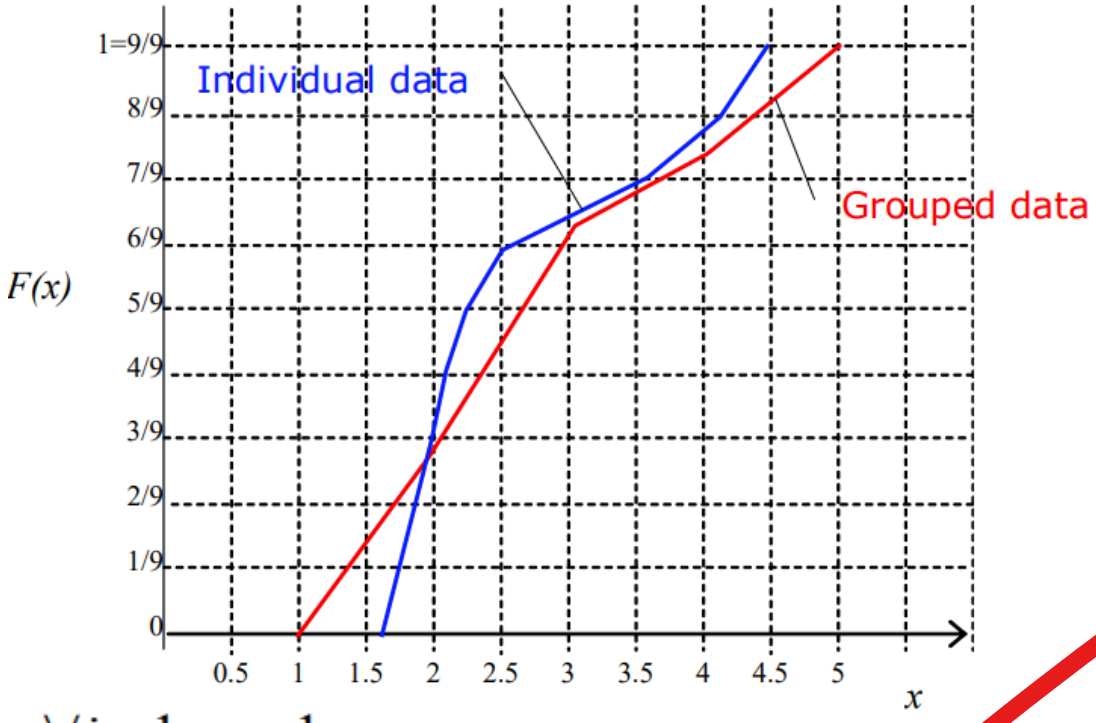
Empirical Distribution: Discrete

- Given X_1 , X_2 , …, X_n create a discrete cumulative distribution function
- Example: Number of samosa people a
  4, 1, 2
- Computation:



Empirical Distribution: Continuous

- Given X_1 , X_2 , …, X_n create a continuous piecewise linear distribution function
- Cumulative Distribution: Linearly interpolate between data points



$$F(x) = \begin{cases} 0 & , x < X_{(1)} \\ \dfrac{i-1}{n-1} + \dfrac{x - X_{(i)}}{(n-1)(X_{(i+1)} - X_{(i)})} & , X_{(i)} \le x < X_{(i+1)} \\ 1 & , X_{(n)} \le x \end{cases} \quad \forall i = 1 \ldots n-1$$

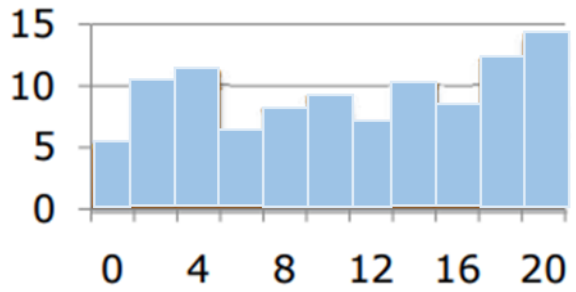
Identifying the Distribution: Histogram

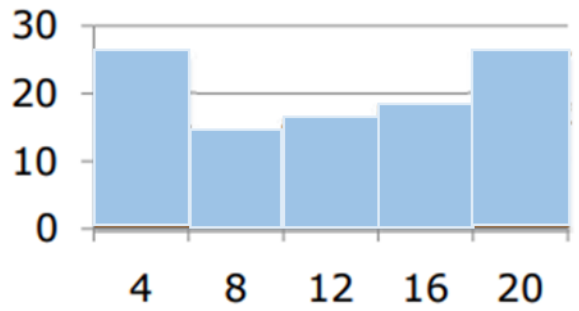
- Why Use a Histogram?
  - A histogram is a powerful tool for visualizing the shape of a dataset. It helps identify patterns, such as skewness, symmetry, or multimodal distributions.
- Determining the Number of Class Intervals
  - The choice of class intervals (bins) significantly affects how the data is represented. The number of intervals should be carefully chosen based on:
    - Sample size – A larger dataset allows for more intervals without losing clarity.
    - Data dispersion – A dataset with greater variability may require more intervals to capture meaningful differences.
  - A useful rule of thumb is to set the number of intervals to approximately the square root of the sample size.
- Continuous vs. Discrete Data
  - For continuous data, the histogram provides an approximation of the probability density function (PDF), illustrating how values are distributed.
  - For discrete data, the histogram corresponds to the probability mass function (PMF), where each bar represents the probability of a specific value occurring

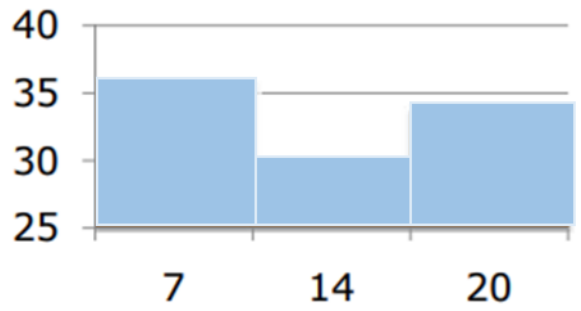
Identifying the Distribution: Histogram

- Dealing with Small Datasets
  - When working with a limited number of observations, the histogram may appear irregular or "jagged." To improve readability:
  - Combine adjacent intervals to create a smoother and more interpretable distribution.
  - Ensure the histogram still accurately reflects the underlying data trends.

Same data with different interval sizes

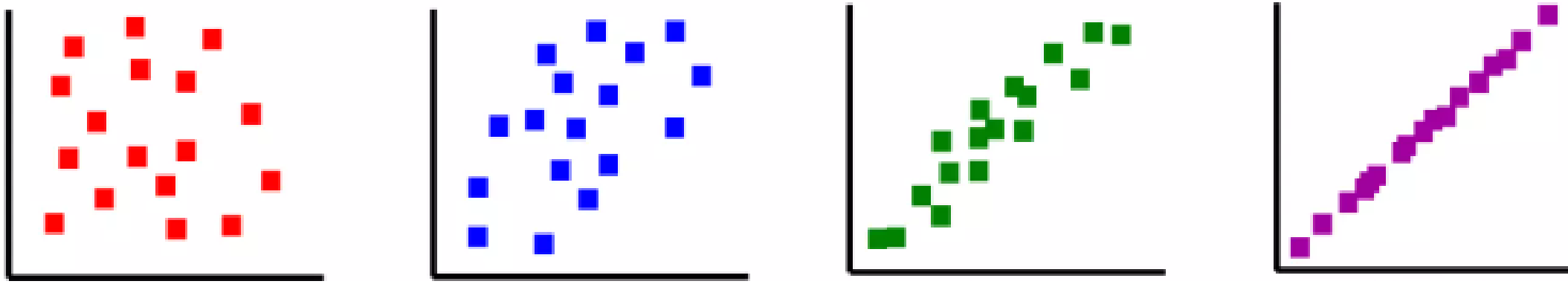






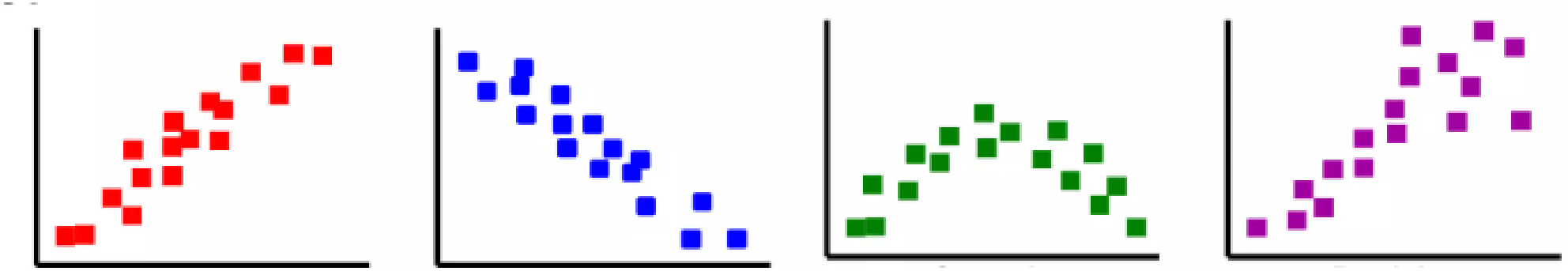
Identifying the Distribution: Scatter Plot

- A scatter diagram is a quality tool used to illustrate the relationship between paired data.
  - To create a scatter diagram:
    - Plot Random Variables X and Y on the x-axis and y-axis respectively
  - This visual representation helps identify ( ) or ( ) between the variables.



Identifying the Distribution: Scatter Plot

- A linear relationship between two variables can be described using the following key elements:
- Correlation: Measures how closely the data points align to a straight line (strength and direction of the relationship).
- Slope: Indicates the steepness or rate of change in the data (how much Y changes for a unit change in X).
- Direction: Specifies whether the relationship is positive (upward slope) or negative (downward slope).



Identifying the Distribution: Family of Distribution

- Binomial: Number of successes in *n* trials.
- Gamma: Time to complete task, e.g. customer service or machine repair

  gamma(α, β) $X \geq 0$
- Poisson: Independent events in fixed time/space.
- Normal: Sum of multiple processes.

  N(μ, σ^2)
- Lognormal: Product of multiple processes.

  LN(μ, σ2)
- Exponential: Time between independent, memoryless events.

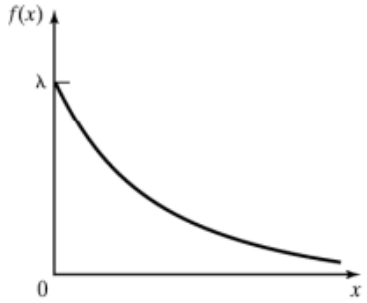
  expo(λ) $X \geq 0$
- Weibull: Time to component failure.

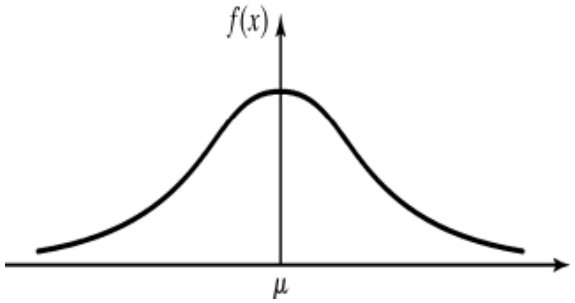
  weibull(α,β) $X \geq 0$
- Uniform: Models complete uncertainty (all outcomes equally likely).
- Triangular: Known minimum, most likely, and maximum values.

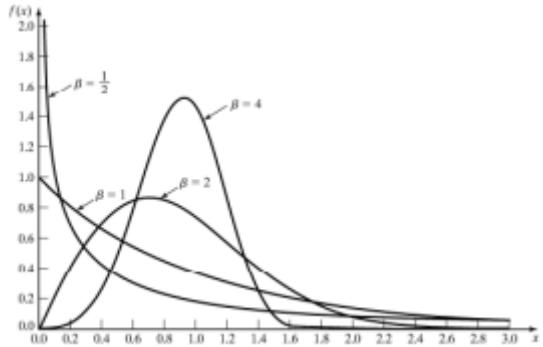
  tria(a, b, c)
- Empirical: Resamples actual collected data.

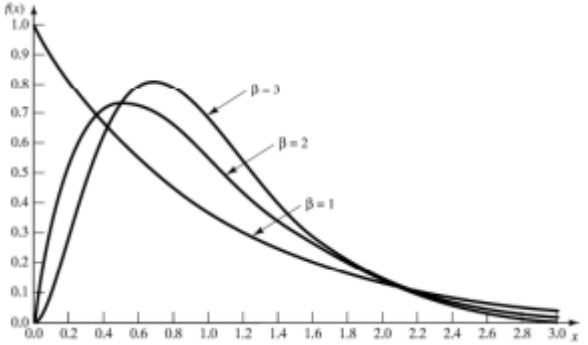
Identifying the Distribution: Family of Distribution

- The appropriate family of distributions is chosen based on the following factors:
- Context of the Input Variable: Understanding the nature of the variable being modeled (e.g., discrete or continuous).
- Shape of the Histogram: Analyzing the data's histogram to identify patterns or trends in its distribution.









Identifying the Distribution: Family of Distribution

- Nature of the Process: Is it inherently discrete or continuous?

- Boundaries and Range:
  - Does the data have bounds?
  - Is it restricted to positive values, negative values, or a specific interval (e.g., [-a, b])?

- True vs. Approximation: There is no single "true" distribution for any stochastic process.

- Objective: Aim to achieve a robust approximation that faithfully represents the process's characteristics.

Identifying the Distribution: Q-Q Plots

- Graphical tool to tests the conformity between ( ) and the ( )
- Steps -
- Arrange the data $x_1, x_2, \ldots, x_n$ in increasing order:

$$x[1] \leq x[2] \leq \ldots \leq x[n]$$

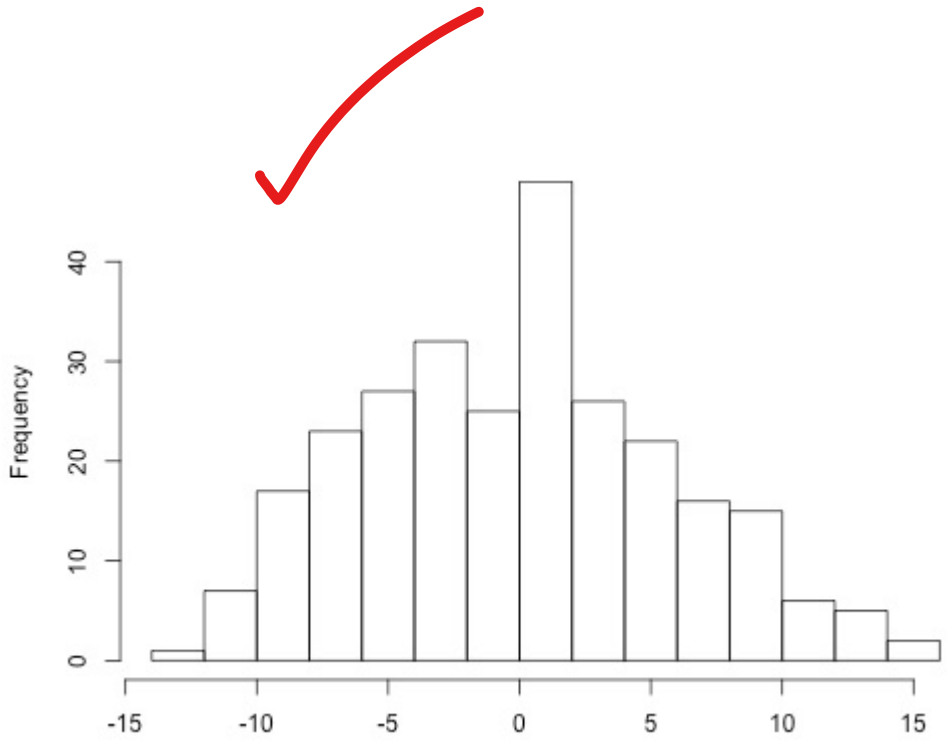
- Associate to each data point x[i] the i/(n+1)-**quantile** q_i of the standard normal distribution

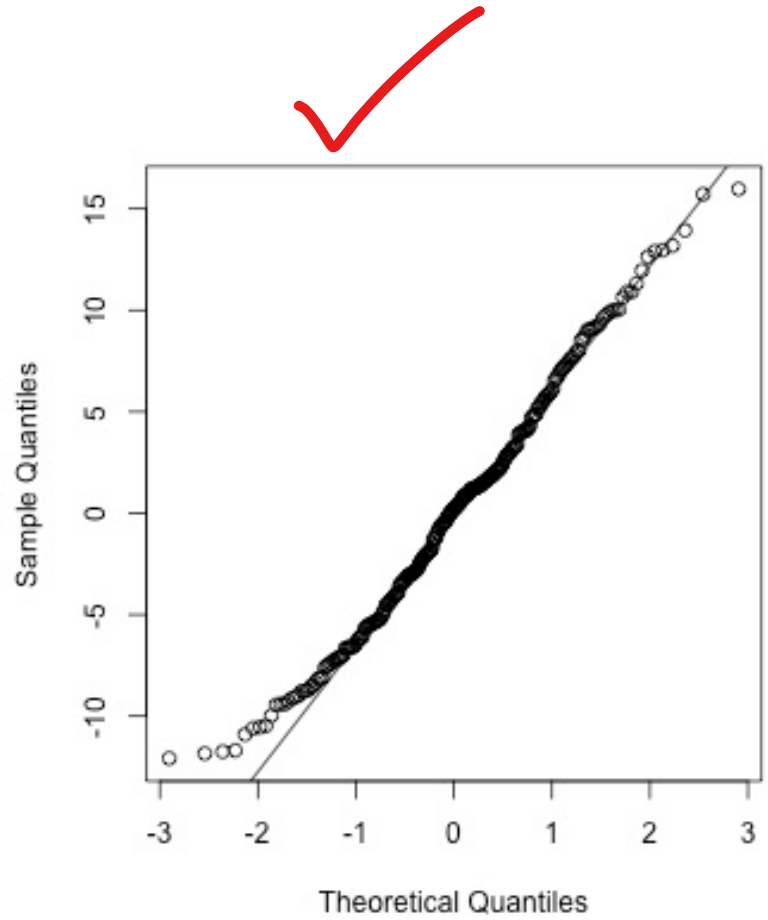
- Perfect Linearity Is Rare: Observed values rarely fall exactly on a straight line.
- Dependency of Ordered Values: Since values are ranked, they are not independent, so points are less likely to scatter widely around the line.

- Homogeneity Check: A Q-Q plot can test whether data come from distributions with the same variance.
- Comparing Two Samples: It evaluates whether a single distribution can represent two datasets.

Identifying the Distribution: Q-Q Plots

- The histogram is roughly bell-shaped, so it is an indication ?
- The pattern of the normal probability plot is straight





Example (TTF)

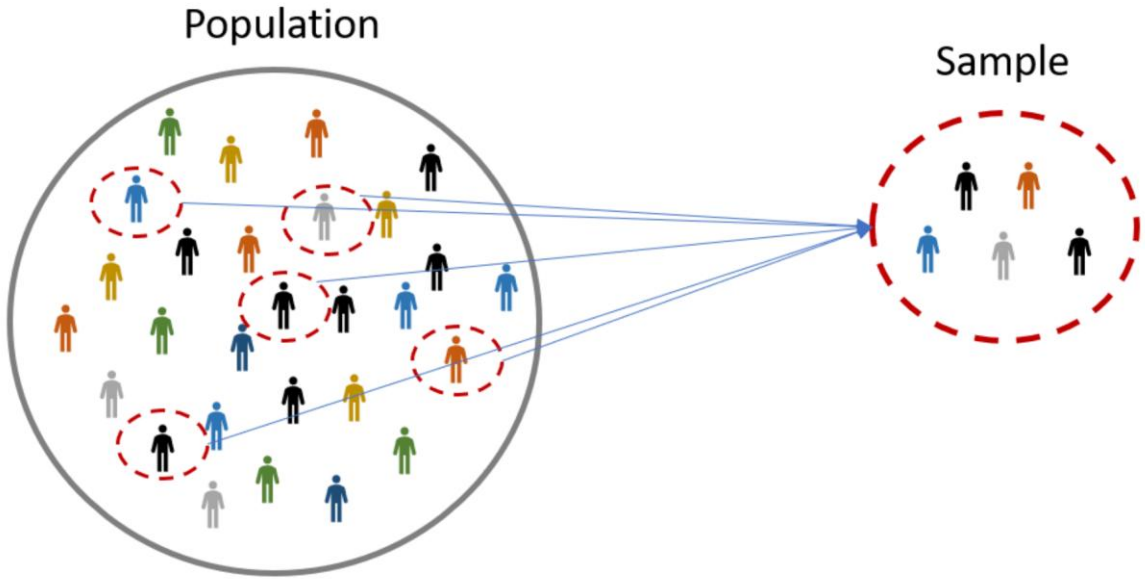
2.801	1.783	2.752	0.019	0.112
54.454	31.6	16.715	39.605	0.394
2.421	0.083	0.855	3.065	6.89
4.891	6.435	0.536	0.273	0.494
28.876	1.043	2.925	0.916	4.147
0.056	5.314	5.914	0.012	131.541
1.084	4.554	0.128	22.2	6.154
16.334	8.185	0.798	1.961	72.654
2.862	21.782	1.753	0.002	0.567
0.914	0.199	13.075	2.124	6.367

Example (TTF)

Statistical Inference

- How to use observed data (sample) to make inferences about the unknown population parameters?

- Let X_1 , X_2 , …, X_n be random sample from population distribution F_θ where θ is vector of unknown parameters
  - Estimate θ
  - Test if F_θ is a ‘good’ fit



Sample Statistics

- Next step after selecting a family of distributions
- If observations in a sample of size n are $X_1, X_2, \ldots, X_n$ (discrete or continuous), the sample mean, and sample variance are:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad S^2 = \frac{\left(\sum_{i=1}^{n} X_i^2\right) - n\bar{X}^2}{n-1}$$

- If the data are discrete and have been grouped in a frequency distribution:

$$\bar{X} = \frac{\sum_{j=1}^{n} f_j X_j}{n} \qquad S^2 = \frac{\left(\sum_{j=1}^{n} f_j X_j^2\right) - n\bar{X}^2}{n-1}$$

- Where f_j is the observed frequency of value X_j

Sample Statistics

- When raw data are unavailable (data are grouped into class intervals), the approximate sample mean and variance are:

$$\bar{X} = \frac{\sum_{j=1}^{c} f_j m_j}{n} \qquad S^2 = \frac{\left(\sum_{j=1}^{n} f_j m_j^2\right) - n\bar{X}^2}{n-1}$$

  - f_j is the observed frequency in the j-th class interval
  - m_j is the midpoint of the j-th interval
  - c is the number of class intervals
- The parameter is an unknown constant, but an estimator is the statistic

Parameter Estimation

- Likelihood function

Let $X_1, X_2, ..., X_n$ be a random sample from a population with pmf/pdf $f(x|\theta_1, ..., \theta_k)$, where $\theta_1, ..., \theta_k$ are unknown parameters. Another way to estimate these parameters is finding the values of $\hat{\theta}_1, ..., \hat{\theta}_k$ that maximize the likelihood that the observed sample is generated by $f(x|\hat{\theta}_1, ..., \hat{\theta}_k)$.

The joint pmf/pdf of the random sample, $f(x_1, x_2, ..., x_n|\theta_1, ..., \theta_k)$, is called the likelihood function, and it is denoted by $L(\theta|\mathbf{x})$.

$$L(\theta|\mathbf{x}) = L(\theta_1, ..., \theta_k|x_1, ..., x_n) = \begin{cases} f(x_1, ..., x_n|\theta_1, ..., \theta_k) & \text{generally} \\ \prod_{i=1}^{n} f(x_i|\theta_1, ..., \theta_k) & \text{random sample} \end{cases}$$

Parameter Estimation

- MLE for θ

$$f(x;\theta) = \begin{cases} (\theta+1)x^{\theta}, & 0 \le x \le 1 \\ 0, & \text{otherwise.} \end{cases}$$

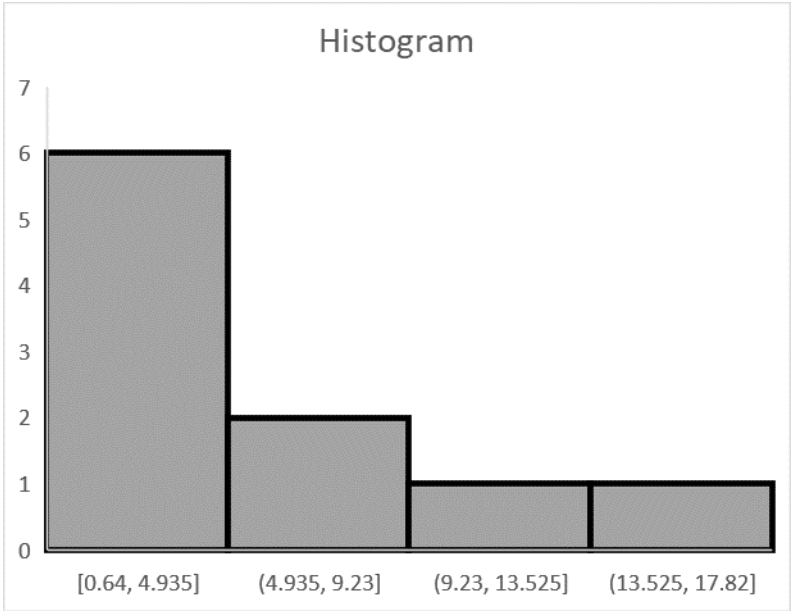
Parameter Estimation

- How to determine the unknown parameters of the distribution from the data given?
- Maximum Likelihood Estimation (MLE)
- Assuming the data are from a particular distribution, we need to estimate the unknown parameter by estimating the maximum likelihood of the parameter
- $f(x)=\lambda e^{-\lambda x}$
- Data – 3.11, 0.64, 2.55, 2.20, 5.44, 3.42,10.39, 8.93, 17.82, 1.30

Parameter Estimation

- $f(x)=\lambda e^{-\lambda x}$

Data – 3.11, 0.64, 2.55, 2.20, 5.44, 3.42,10.39, 8.93, 17.82, 1.30

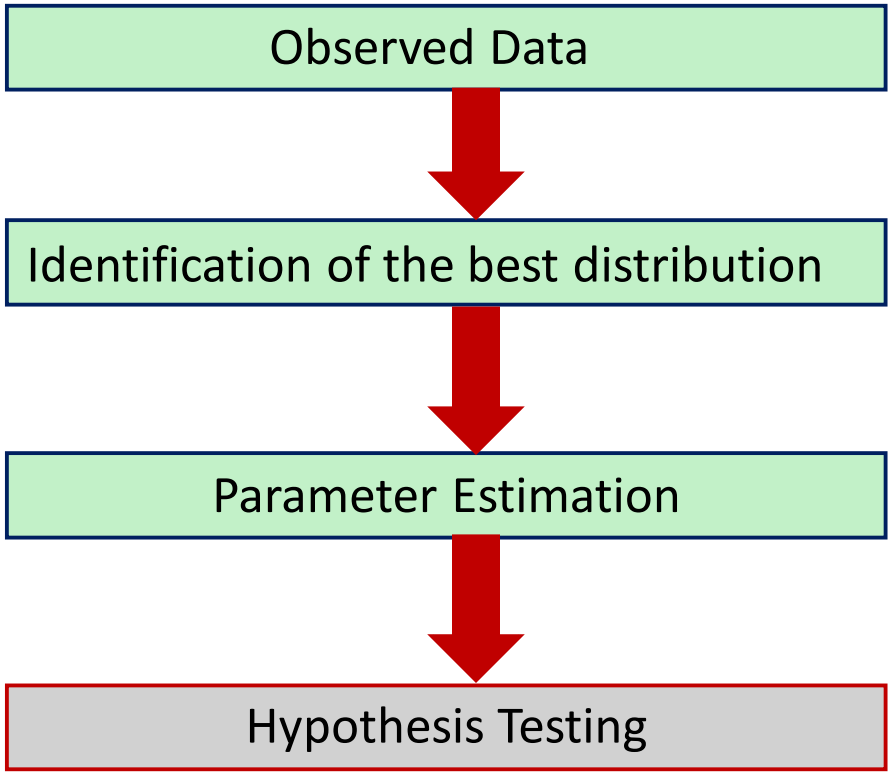


Parameter Estimation

Distribution	*Parameters*	*Estimators*
Exponential	λ	$\hat{\lambda} = \frac{1}{\bar{X}}$
Poisson	λ	$\hat{\lambda} = \bar{X}$
Normal	μ, σ^2	$\hat{\mu} = \bar{X}$ $\hat{\sigma}^2 = S^2 (unbiased)$
Lognormal	μ, σ^2	$\hat{\mu} = \bar{X}$ $\hat{\sigma}^2 = S^2$ (after taking ln of data)
Gamma	α, β	$\hat{\alpha}$ by using $T = \left[\ln \bar{X} - \frac{1}{n} \sum_{i=1}^{n} \ln X_i \right]^{-1}$ $\hat{\beta} = \frac{\bar{X}}{\hat{\alpha}}$

Goodness-of-Fit Testing

- Conduct hypothesis testing on input data distribution using
  - Kolmogorov-Smirnov test
  - Chi-square test
- No single correct distribution in a real application exists
  - If very little data are available, it is ( ) to reject any candidate distributions
  - If a lot of data are available, it is likely to ( ) all candidate distributions



Goodness-of-Fit Testing

- H_0: X_i's are i.i.d with fitted distribution & estimated parameters
- H_1: X_i's are not i.i.d and do not conform with fitted distribtions & estimated parameters

- Type I Error: α
  - Error of first kind, False positive
  - Reject H_0 when it is true
- Type II Error: β
  - Error of second kind, False negative
  - Retain H_0 when it is not true

Statistical Decisions	**H_0 is true**	**H_0 is False**
Accept H_0	Correct	Type II Error ( )
Reject H_0	Type I Error ( )	Correct

Chi-Square Test

- Intuition:

Histogram of the observed data ⟺ Shape of the candidate density/mass function

- Valid for large sample sizes when parameters are estimated by maximum-likelihood
- Can be applied to discrete as well as continuous distributions

Steps:

1. Arrange n observations in a set of k intervals
2. General guidelines number of class intervals (k):

20	Too less to perform	100	10-20
50	5-10	>100	$\sqrt{n}$ to $n/5$

Chi-Square Test

- Test Statistic:
-

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Expected Frequency

$E_i = n \times p_i$

where p_i is the theoretical prob. of the i-th interval.

O_i : Observed frequency in the i-th class

- Test Conditions:

$$\chi_0^2 \leq \chi_{k-1,1-\alpha}^2$$ Unable to reject Null Hypothesis

$$\chi_0^2 > \chi_{k-1,1-\alpha}^2$$ Reject Null Hypothesis

Example 1 (TTF)

2.801	1.783	2.752	0.019	0.112
54.454	31.6	16.715	39.605	0.394
2.421	0.083	0.855	3.065	6.89
4.891	6.435	0.536	0.273	0.494
28.876	1.043	2.925	0.916	4.147
0.056	5.314	5.914	0.012	131.541
1.084	4.554	0.128	22.2	6.154
16.334	8.185	0.798	1.961	72.654
2.862	21.782	1.753	0.002	0.567
0.914	0.199	13.075	2.124	6.367

Example (TTF)

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Example 2 (Arrivals)

Data - 1, 0, 1, 0, 0, 2, 1, 0, 0, 3, 1, 1, 2, 1, 2, 0, 1, 2, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 3, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0,1, 0, 0, 0, 1, 2, 3, 2, 0, 0, 0, 3, 1, 0, 0, 2, 0, 0, 1, 2

Number of Defects	Observed Frequency
0	32
1	15
2	9
3	4

Example (Arrivals)

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Kolmogorov-Smirnov Test (K-S Test)

- Intuition: Q-Q Plot
- Recall :

- Statistics:

- Use-cases:
  - Sample size is small
  - No estimation of the parameters

Kolmogorov-Smirnov Test (K-S Test)

- Test fitted cdf with empirical cdf
  - Step 1: Rank the data smallest to largest
  - Step 2: Compute D^+ and D^-
  - Step 3: $D = \text{Max}\{D^+, D^-\}$
  - Step 4: Test Statistics

Case	*Adjusted Test Statistics*	*1-α*				
		0.850	**0.900**	**0.950**	**0.975**	**0.990**
All parameters known	$\left(\sqrt{n}+0.12+\frac{0.11}{\sqrt{n}}\right)D$	1.138	1.224	1.358	1.480	1.628
Normal $(\hat{\mu},\hat{\sigma}^2)$	$\left(\sqrt{n}-0.01+\frac{0.85}{\sqrt{n}}\right)D$	0.775	0.819	0.895	0.955	1.035
Expo$(\hat{\lambda})$	$\left(\sqrt{n}+0.26+\frac{0.5}{\sqrt{n}}\right)\left(D-\frac{0.2}{n}\right)$	0.926	0.990	1.094	1.190	1.308