IERG 3050 Week5
Selecting Input Probability Distributions (Part II)

Bolei Zhou
Department of Information Engineering
The Chinese University of Hong Kong
Announcement

• Hand in Homework 1 by end of today!

• Homework 2 will be out by the end of tomorrow. No need to hand in, but important for your final exam and a good preparation for Quiz 1 on Oct. 16.

• Quiz 1 on Oct. 16:
  • 5-min discussion: on version 1
  • 15-min problem solving on your own: on version 2
Outline

Approach for specifying a probability distribution from collected data
Step 1: Selecting family of distributions.
Step 2: Estimating model parameters.
Step 3: Determining how representative the fitted distribution is.

Reading: Chapter 6 of the textbook
Maximum-Likelihood Estimation (MLE)

Suppose we have the following distribution of the Final Exam results for Class A and Class D:

Suppose that we randomly pick three students from the same class. Their scores are

\[ x_1 = 97, \quad x_2 = 99, \quad x_3 = 97. \]

Which class (Class A or Class D) do you think they come from?
Maximum-Likelihood Estimation (MLE)

Idea for MLE:

• Let the family of distributions under consideration be $f_X(x|\theta)$, where $\theta$ collects the parameters.

• Assume that we have observed the i.i.d. samples $x_1, x_2, \ldots, x_n$.

• The MLE chooses the distribution $f_X(x|\theta)$ for which the likelihood of observing $x_1, x_2, \ldots, x_n$ is the largest.
Maximum-Likelihood Estimation (MLE)

Idea for MLE:

• Let the family of distributions under consideration be \( f_X(x|\theta) \), where \( \theta \) collects the parameters.

• Assume that we have observed the i.i.d. samples \( x_1, x_2, \ldots, x_n \).

• The MLE chooses the distribution \( f_X(x|\theta) \) for which the likelihood of observing \( x_1, x_2, \ldots, x_n \) is the largest.

Mathematically:

⇒ MLE chooses the parameters \( \theta \) that maximize

\[
L(\theta) \triangleq f_X(x_1|\theta) \cdot f_X(x_2|\theta) \cdot \cdots \cdot f_X(x_n|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta),
\]
i.e.,

\[
\hat{\theta}_{\text{MLE}} \triangleq \arg \max_{\theta} L(\theta).
\]
Maximum-Likelihood Estimation (MLE)

Note:

The product expression

\[ L(\theta) \triangleq f_X(x_1|\theta) \cdot f_X(x_2|\theta) \cdots f_X(x_n|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta), \]

is a consequence of the independence of the data samples.
Maximum-Likelihood Estimation (MLE)

Because the $\log(\cdot)$-function is a strictly monotonically increasing function, finding the $\theta$ that maximizes the likelihood function

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i|\theta)$$

is equivalent to finding the $\beta$ that maximizes the log-likelihood function

$$\log(L(\theta)) = \sum_{i=1}^{n} \log(f_X(x_i|\theta))$$

Mathematically:

$$\hat{\theta}_{\text{MLE}} \triangleq \arg \max_{\theta} L(\theta)$$

is equivalent to

$$\hat{\theta}_{\text{MLE}} \triangleq \arg \max_{\theta} \log(L(\theta)).$$
Step 3: Determining how representative the fitted distribution is
Determining How Representative the Fitted Distribution is

Suppose that

- we have hypothesized a family of probability distributions (Step 1)
- and estimated the corresponding parameters (Step 2).

**Step 3** is to see if the fitted distribution agrees with the actually observed data.

**Heuristic methods:**

- Histogram over plot
- Frequency comparison
- Quantile-to-quantile plot (Q-Q plot)
- Probability-to-probability plot (P-P plot)
- ...  

**Formal statistical hypothesis tests:**

- Chi-square goodness-of-fit test \( (\chi^2 \text{ goodness-of-fit test}) \)
- ...
Heuristic methods
Heuristic method: Histogram over Plot

Let $h(x)$ be a “normalized” histogram based on the observed data.

Important assumptions:
- Bins of histogram have equal width.
- Histogram is “normalized” in the sense that the bars are vertically scaled such that the total area (i.e., total green area in the plot shown below) equals 1.

Let $\hat{f}(x)$ be the fitted PDF.
Plot $h(x)$ over $\hat{f}(x)$. Look for similarities.

Example: observed histogram vs. fitted exponential.
Heuristic method: Frequency Comparison

Frequency comparison for data from **continuous** distribution:

- Define histogram intervals interval \((b_{j-1}, b_j]\), \(j = 1, 2, \ldots, k\), each of width \(\Delta b\).
- Let \(h_j\) be the observed proportion of data in \(j\)-th interval.
- Let \(\hat{r}_j\) be the expected proportion of data in \(j\)-th interval if the fitted distribution is correct.
- Plot \(h_j\) and \(\hat{r}_j\) together. **Look for similarities.**
Chi-Square goodness-of-fit test
Chi-square Goodness-of-fit Test

Let \( x_1, x_2, \ldots, x_n \) be i.i.d. data samples (observed data).

We perform a formal statistical hypothesis test on the null hypothesis.

Here, our null hypothesis \( H_0 \) is:

The distribution that generated the i.i.d. data samples and the fitted distribution are the same.

Mathematically:

\[ H_0: \text{ the } x_i \text{'s are i.i.d. data samples generated from } \hat{f}(x), \text{ where } \hat{f}(x) \text{ is the fitted PDF.} \]

Caution:

- "Failure to reject \( H_0 \)" should not be interpreted as "accepting \( H_0 \) as being true."
- \( H_0 \) is virtually never exactly true!
- We are looking for an "adequate" fit of the distribution.
Chi-square Goodness-of-fit Test

Let \( x_1, x_2, \ldots, x_n \) be i.i.d. data samples (observed data).
Let \( \hat{f}(x) \) be the fitted PDF (continuous case).
Let \( \hat{p}(x) \) be the fitted PMF (discrete case).

1. Divide the range of data into \( k \) disjoint intervals, not necessarily of equal width:

\[
(b_0, b_1], \ (b_1, b_2], \ldots, (b_{k-1}, b_k]
\]

Note: \( b_0 \) could be \(-\infty\) and \( b_k \) could be \(+\infty\).

2. Compare the actual amount of observed data in each interval with what the fitted distribution predicts.

\[\Rightarrow\] Let \( N_j \) be the number of observed data points in the \( j \)-th interval.

\[\Rightarrow\] Let \( \hat{p}_j \) be the expected proportion of data samples in the \( j \)-th interval if the fitted distribution were true:

\[
\hat{p}_j = \begin{cases} 
\int_{b_{j-1}}^{b_j} \hat{f}(x) \, dx & \text{continuous case} \\
\sum_{b_{j-1} < x \leq b_j} \hat{p}(x) & \text{discrete case}
\end{cases}
\]
Chi-square Goodness-of-fit Test

Let $N_j$ be the number of observed data points in the $j$-th interval.
Let $\hat{p}_j$ be the expected proportion of data samples in the $j$-th interval if the fitted distribution were true:

⇒ If fitted distribution is correct, we expect that $N_j \approx n \cdot \hat{p}_j$
Chi-square Goodness-of-fit Test

Define test statistic

\[ D \triangleq \sum_{j=1}^{k} \frac{(N_j - n\hat{p}_j)^2}{n\hat{p}_j}. \]

If the data fit the distribution well (i.e., \( H_0 \) is true),

then \( D \) should be small.

If \( D \) is large,

then the distribution is not a good distribution and
we shall reject the hypothesis \( H_0 \).

But how large is large?

Under \( H_0 \): fitted distribution is correct and
\( D \) has (approximately) a chi-square distribution with \( k-1 \) degrees of freedom.
Chi-square Distribution

Let $X_i$, $i = 1, \ldots, r$, be i.i.d. normal r.v.s with mean 0 and variance 1.

Define $Y = X_1^2 + X_2^2 + \ldots + X_k^2$

$\Rightarrow$ $Y$ is distributed according to the chi-square ($\chi^2$) distribution with $k$ degrees of freedom.
Chi-square Goodness-of-fit Test

For large $n$:

If $D > \chi^2_{k-1, 1-\alpha}$ then we reject $H_0$ at level $\alpha$.

What does this mean?

This means that if the hypothesis is true, then the probability of having $D > \chi^2_{k-1, 1-\alpha}$ is not more than $100\alpha \%$.
**Table of $\chi^2_{\nu, r}$ values**

<table>
<thead>
<tr>
<th>$\nu \setminus r$</th>
<th>0.90</th>
<th>0.95</th>
<th>0.975</th>
<th>0.99</th>
<th>0.995</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.706</td>
<td>3.841</td>
<td>5.024</td>
<td>6.635</td>
<td>7.879</td>
</tr>
<tr>
<td>2</td>
<td>4.605</td>
<td>5.991</td>
<td>7.378</td>
<td>9.210</td>
<td>10.597</td>
</tr>
<tr>
<td>7</td>
<td>12.017</td>
<td>14.067</td>
<td>16.013</td>
<td>18.475</td>
<td>20.278</td>
</tr>
<tr>
<td>8</td>
<td>13.362</td>
<td>15.507</td>
<td>17.535</td>
<td>20.000</td>
<td>21.955</td>
</tr>
<tr>
<td>10</td>
<td>15.987</td>
<td>18.307</td>
<td>20.483</td>
<td>23.209</td>
<td>25.188</td>
</tr>
<tr>
<td>13</td>
<td>19.812</td>
<td>22.362</td>
<td>24.736</td>
<td>27.688</td>
<td>29.819</td>
</tr>
<tr>
<td>15</td>
<td>22.307</td>
<td>24.996</td>
<td>27.488</td>
<td>30.578</td>
<td>32.801</td>
</tr>
<tr>
<td>16</td>
<td>23.542</td>
<td>26.296</td>
<td>28.845</td>
<td>32.000</td>
<td>34.267</td>
</tr>
<tr>
<td>17</td>
<td>24.760</td>
<td>27.587</td>
<td>30.191</td>
<td>33.409</td>
<td>35.718</td>
</tr>
<tr>
<td>18</td>
<td>25.989</td>
<td>28.869</td>
<td>31.526</td>
<td>34.805</td>
<td>37.156</td>
</tr>
<tr>
<td>19</td>
<td>27.204</td>
<td>30.144</td>
<td>32.852</td>
<td>36.191</td>
<td>38.582</td>
</tr>
<tr>
<td>20</td>
<td>28.412</td>
<td>31.410</td>
<td>34.170</td>
<td>37.566</td>
<td>39.997</td>
</tr>
<tr>
<td>21</td>
<td>29.615</td>
<td>32.671</td>
<td>35.479</td>
<td>38.932</td>
<td>41.401</td>
</tr>
<tr>
<td>22</td>
<td>30.813</td>
<td>33.924</td>
<td>36.781</td>
<td>40.289</td>
<td>42.796</td>
</tr>
<tr>
<td>23</td>
<td>32.007</td>
<td>35.172</td>
<td>38.076</td>
<td>41.638</td>
<td>44.181</td>
</tr>
<tr>
<td>24</td>
<td>33.196</td>
<td>36.415</td>
<td>39.364</td>
<td>42.980</td>
<td>45.559</td>
</tr>
<tr>
<td>25</td>
<td>34.382</td>
<td>37.652</td>
<td>40.646</td>
<td>44.314</td>
<td>46.928</td>
</tr>
<tr>
<td>26</td>
<td>35.563</td>
<td>38.885</td>
<td>41.923</td>
<td>45.642</td>
<td>48.290</td>
</tr>
<tr>
<td>27</td>
<td>36.741</td>
<td>40.113</td>
<td>43.195</td>
<td>46.963</td>
<td>49.645</td>
</tr>
<tr>
<td>28</td>
<td>37.916</td>
<td>41.337</td>
<td>44.461</td>
<td>48.278</td>
<td>50.993</td>
</tr>
<tr>
<td>29</td>
<td>39.087</td>
<td>42.557</td>
<td>45.722</td>
<td>49.588</td>
<td>52.336</td>
</tr>
<tr>
<td>30</td>
<td>40.256</td>
<td>43.773</td>
<td>46.970</td>
<td>50.892</td>
<td>53.672</td>
</tr>
<tr>
<td>40</td>
<td>51.805</td>
<td>55.758</td>
<td>59.342</td>
<td>63.691</td>
<td>66.766</td>
</tr>
<tr>
<td>50</td>
<td>63.167</td>
<td>67.505</td>
<td>71.420</td>
<td>76.154</td>
<td>79.490</td>
</tr>
<tr>
<td>60</td>
<td>74.397</td>
<td>79.682</td>
<td>83.298</td>
<td>88.379</td>
<td>91.952</td>
</tr>
<tr>
<td>70</td>
<td>85.527</td>
<td>90.531</td>
<td>95.023</td>
<td>100.425</td>
<td>104.215</td>
</tr>
<tr>
<td>80</td>
<td>96.578</td>
<td>101.879</td>
<td>106.629</td>
<td>112.329</td>
<td>116.321</td>
</tr>
<tr>
<td>90</td>
<td>107.565</td>
<td>113.145</td>
<td>118.136</td>
<td>124.116</td>
<td>128.209</td>
</tr>
<tr>
<td>100</td>
<td>118.498</td>
<td>124.542</td>
<td>129.561</td>
<td>135.807</td>
<td>140.169</td>
</tr>
</tbody>
</table>
Chi-square Goodness-of-fit Test

Example: we want to test the uniformity of a random number generator, i.e., $H_0: x_1, x_2, \ldots, x_n$ are generated i.i.d. according to $U(0, 1)$.

Based on $n = 500$ observed samples, the histogram is:

![Histogram](image)

The expected frequency of each bin is $n \cdot 0.2 = 100$.

Therefore,

$$D = \frac{(115-100)^2}{100} + \frac{(120-100)^2}{100} + \frac{(90-100)^2}{100} + \frac{(105-100)^2}{100} + \frac{(70-100)^2}{100} = 16.5.$$ 

Since $\chi^2_{5-1, 0.95} = 9.488$ and $D > \chi^2_{5-1, 0.95}$,

the hypothesis $H_0$ is rejected at the 5% level.
## Error Type and Hypothesis Testing

<table>
<thead>
<tr>
<th>Table of error types</th>
<th>Null hypothesis ($H_0$) is</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True</td>
</tr>
<tr>
<td><strong>Decision about null hypothesis ($H_0$)</strong></td>
<td>Correct inference (true negative) (probability = $1 - \alpha$)</td>
</tr>
<tr>
<td>Fail to reject</td>
<td></td>
</tr>
<tr>
<td>Reject</td>
<td>Type I error (false positive) (probability = $\alpha$)</td>
</tr>
</tbody>
</table>

**Note:** “positivity” / “negativity” is based on the following:

- rejecting the null hypothesis $H_0$ is considered to be “positive”;
- failing to reject the null hypothesis $H_0$ is considered to be “negative.”
Error Type and Hypothesis Testing

- Null hypothesis: phenomenon being studied produces no effect or makes no difference
- Alternative hypothesis is opposed to the null hypothesis
- **Type I error**: the rejection of a true null hypothesis
- A test that shows a patient to have a disease when in fact the patient does not
- **Type II error**: the failure to reject a false null hypothesis
- When comparing two means, concluding the means were different when in reality they were not different is a type I error; concluding the means were not different when in reality they were different is a type II error.
Probability Plots
Probability Plots

Similar to the heuristic approaches of plotting the histograms as discussed before, probability plots are another approach to graphically compare the empirical distribution based on the sampled data and the fitted distribution. (The latter will be referred to as the fitted model.)

For a continuous distribution, we can determine how representative the fitted distribution is by comparing the CDFs of the observed data and the fitted model.

⇒ Quantile-quantile (Q-Q) plot.
⇒ Probability-probability (P-P) plot.
Probability Plots

Let \( x_1, x_2, \ldots, x_n \) be i.i.d. data samples (observed data).

Let \( x_{(1)}, x_{(2)}, \ldots, x_{(n)} \) be the increasingly sorted samples.

In the following, we will use an empirical data-based CDF \( F^d(x) \) which satisfies

\[
F^d(x_{(i)}) = \frac{i - \frac{1}{2}}{n} = \frac{2i - 1}{2n}, \quad i = 1, 2, \ldots, n.
\]

Such a CDF can, for example, be obtained as follows (with arbitrary \( \epsilon > 0 \)):

\[
F^d(x) = \int_{-\infty}^{x} f^d(x') \, dx' \quad \text{with} \quad f^d(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi\epsilon^2}} \cdot \exp\left(\frac{(x - x_{(i)})^2}{2\epsilon^2}\right).
\]
Quantile-to-quantile plot
Quantile-to-quantile Plot (Q-Q Plot)

Review: How do we define quantile?

Suppose $F(x)$ is the CDF of some continuous random variable.

For $0 \leq q \leq 1$, the $q$-quantile of $F(x)$ is defined to be the number $x_q$ such that

$$F(x_q) = q.$$
Quantile-to-quantile Plot (Q-Q Plot)

Review: How do we define quantile?

Side remark: note that the above definition of a $q$-quantile is different from the definition of a $q$-quantile in wikipedia.

Namely, for integers $q$ larger than 1, wikipedia defines:

<table>
<thead>
<tr>
<th>2-quantile</th>
<th>$F^{-1}(1/2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-quantiles</td>
<td>$F^{-1}(1/3), F^{-1}(2/3)$,</td>
</tr>
<tr>
<td>4-quantiles</td>
<td>$F^{-1}(1/4), F^{-1}(2/4), F^{-1}(3/4)$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Quantile-to-quantile Plot (Q-Q Plot)

For all \( q \) in the range \( 0 \leq q \leq 1 \),

plot the value of \( x^m_q \) (model) vs. the value of \( x^d_q \) (data).

Simplified version:

For several \( q \)'s in the range \( 0 \leq q \leq 1 \),

plot the value of \( x^m_q \) (model) vs. the value of \( x^d_q \) (data)

and connect the points with straight lines.
Quantile-to-quantile Plot (Q-Q Plot)

Special case of the above procedure:

For \( q = \frac{1-\frac{1}{2}}{n}, \frac{2-\frac{1}{2}}{n}, \ldots, \frac{n-3}{2}, \frac{n-1}{2}, \)

plot the value of \( x^m_q \) (model) vs. the value of \( x^d_q \) (data).

Note:

- Let \( x(1), x(2), \ldots, x(n) \) be the increasingly sorted samples.
- Define \( q_i \triangleq \frac{i-\frac{1}{2}}{n} \) for \( i = 1, \ldots, n \).
- Solving \( q_i = F^d(x^d_{(i)}) = \frac{i-\frac{1}{2}}{n} \), one obtains

\[
x^d_{(i)} = x(i).
\]
Q-Q Plot: An Example

Data: 10, 15, 16, 20 (interarrival times)

Fitted model: \( \text{expo}(\beta) \) with \( \beta = \beta_{\text{MLE}} = \frac{1}{4} \cdot (10 + 15 + 16 + 20) = 15.25 \).

Resulting \( q \)-quantiles:

<table>
<thead>
<tr>
<th>( q )</th>
<th>( x^d_q )</th>
<th>( x^m_q )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{1}{8} )</td>
<td>10</td>
<td>2.04</td>
</tr>
<tr>
<td>( \frac{3}{8} )</td>
<td>15</td>
<td>7.17</td>
</tr>
<tr>
<td>( \frac{5}{8} )</td>
<td>16</td>
<td>14.96</td>
</tr>
<tr>
<td>( \frac{7}{8} )</td>
<td>20</td>
<td>31.71</td>
</tr>
</tbody>
</table>

Computation for fitted model:

- The CDF of \( \text{expo}(\beta) \) is \( F^m(x) = 1 - e^{-x/\beta} \).
- Solving \( F^m(x^m_q) = q \), one obtains \( x^m_q = -\beta \cdot \log(1 - q) \).
Q-Q Plot: An Example

Note: The point \((0, 0)\) is included in the above curve because \(0\) is the 0-quantile for both data and model.
Issues with Q-Q Plots

For some model CDFs $F^m(x)$, there is no closed-form expression for their inverse (e.g., gamma and normal distributions).

Therefore it is difficult to find $x^m_q$ from $q$.

⇒ It might be better to use P-P plot (probability-probability plot).
Probability-to-probability plot
P-P Plot (probability-to-probability plot)

Let $x(1), x(2), \ldots, x(n)$ be the increasingly sorted samples.

For $i = 1, 2, \ldots, n$:

- let $q_i^d \triangleq F^d(x(i))$,
- let $q_i^m \triangleq F^m(x(i))$.

Plot the value of $q_i^m$ (model) vs. the value of $q_i^d$ (data).
P-P Plot (probability-to-probability plot)
P-P Plot: An Example

Data: 10, 15, 16, 20 (interarrival times)

Fitted model: \( \text{expo}(\beta) \) with \( \beta = \beta_{\text{MLE}} = \frac{1}{4} \cdot (10+15+16+20) = 15.25 \).

Resulting \( q \) values:

<table>
<thead>
<tr>
<th>( x(i) )</th>
<th>( q_i^d )</th>
<th>( q_i^m )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>( \frac{1}{8} )</td>
<td>0.48</td>
</tr>
<tr>
<td>15</td>
<td>( \frac{3}{8} )</td>
<td>0.63</td>
</tr>
<tr>
<td>16</td>
<td>( \frac{5}{8} )</td>
<td>0.65</td>
</tr>
<tr>
<td>20</td>
<td>( \frac{7}{8} )</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Computation for fitted model:

- The CDF of \( \text{expo}(\beta) \) is \( F^m(x) = 1 - e^{-x/\beta} \).
- From \( F^m(x(i)) = q_i^m \) one obtains \( q_i^m = 1 - e^{-x(i)/\beta} \).
P-P Plot: An Example

Note: The points $(0, 0)$ and $(1, 1)$ are included in the above curve because
- $F^d(x) = 0, F^m(x) = 0$ for $x = 0$,
- $F^d(x) = 1, F^m(x) = 1$ for $x = +\infty$. 
Comparison of Q-Q and P-P Plots
Thanks!

• Next week: Random Number Generation
• Required reading: Chapters 7 and 8

• Special thanks to Prof. Pascal Vontobel, Prof. Minghua Chen, Prof. Rosanna Chan, Prof. Jianwei Huang, Prof. Angela Zhang for improving slides over the years.
Preview on the course proposals

• [https://piazza.com/class/jzkcr8285wg6pl?cid=7](https://piazza.com/class/jzkcr8285wg6pl?cid=7)