Experimental Data Analysis

in ©MATLAB

Lecture 2:
Introduction to the statistics, probability distributions, and plotting statistical data

Jan Rusz
Czech Technical University in Prague
Motivation

Why to analyze data?

• We want to make good decisions
  – E.g. Patient has 39°C → He has got a fever!
    • Yes, we know the range of fever from medical books.
    • But, how did the authors of medical books know it?
• Good decisions are based on reality
  – Authors did measure temperature of many peoples, analyzed the data and found that temperatures higher than $\approx 38^\circ C$ are very rare and are related to unhealthy physical state
Motivation

Why statistical inference?

Blind faith vs. science

“There is a fundamental difference between religion, which is based on authority, and science, which is based on observation and reason. Science will win because it works.”


How confidently can we trust to our decision?

“Probability is common sense reduced to calculation“

Motivation

How to analyze the data?

– Direct approach
  • Look at the numbers:
    – 36.8°C, 36.7°C, 36.5°C, 36.6°C, 36.9°C, 36.7°C, ... it sucks!
  • Visualize
    – We all love pictures, but what if there is too many data?!
– Statistics
  • Empirical models
    – Summarize and describe data in convenient way
  • Statistical models
    – Fit the data into something more simple e.g. equation of probability distribution function
**Mean:** $\mu(x) = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

**Standard deviation:** $\sigma(x) = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$

**Median:** is the value separating the higher half of a data sample, from the lower half.

**Percentile:** $P$-th percentile of the $N$ ordered values (sorted from least to greatest)

$$n = \left\lceil \frac{P}{100} \times N \right\rceil$$
Mean:
Data set \(\{1, 2, 3, 4, 5\}\)
Mean = \(\frac{1 + 2 + 3 + 4 + 5}{5} = 3\)

Standard deviation:
Data set \(\{1, 2, 3, 4, 5\}\)
Mean = 3
\((1-3)^2 = 4\) \hspace{1cm} \((2-3)^2 = 1\) \hspace{1cm} \((3-3)^2 = 0\)
\((4-3)^2 = 1\) \hspace{1cm} \((5-3)^2 = 4\)
Variance (sum of the values/N-1) = \(\frac{4 + 1 + 0 + 1 + 4}{4} = 2.5\)
Standard deviation = \(\sqrt{2.5} \approx 1.58\)

Median:
Data set \(\{9, 5, 1, 4, 11, 2, 8\}\)
Sorted data set \(\{1, 2, 4, 5, 8, 9, 11\}\)
median = 5

Percentile: 25\(^{th}\) \(P = ?\)
data set \(\{10, 20, 30, 40, 50, 55, 60, 65, 70, 75\}\)
\(N = 10\)
\(n = \left\lfloor \frac{25}{100} \times 10 \right\rfloor = [2.5] = 3\)
data set \(\{10, 20, 30, 40, 50, 55, 60, 65, 70, 75\}\)
25\(^{th}\) \(P = 30\)
**Standard deviation:**
Data set {1, 2, 3, 4, 5}
Mean = 3
\[
\begin{align*}
(1-3)^2 &= 4 \\
(2-3)^2 &= 1 \\
(3-3)^2 &= 0 \\
(4-3)^2 &= 1 \\
(5-3)^2 &= 4
\end{align*}
\]
Variance (sum of the values/N-1) = \((4 + 1 + 0 + 1 + 4)/4 = 2.5\)
Standard deviation = \(\sqrt{2.5} \approx 1.58\)

**Median:**
Data set {9, 5, 1, 4, 11, 2, 8}
Sorted data set {1, 2, 4, 5, 8, 9, 11}
median = 5

**Median absolute deviation (MAD):**
\[
\text{MAD} = \text{median}(|X_i - \text{median}(X)|)
\]
\[
|X_i - \text{median}(X)| = \{4, 0, 4, 1, 6, 3, 3\}
\]
MAD = 3

Matlab example 1
Trimmed mean

Frequency

Value

Median
Trimmean (25%)
Mean

Matlab example 2
$p(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Gaussian probability distribution

- $\mu \pm 3\sigma$: 99.7% of data
- $\mu \pm 2\sigma$: 95% of data
- $\mu \pm \sigma$: 68% of data

Matlab example 3
Normal, Bell-shaped Curve

Percentage of cases in 8 portions of the curve:
- .13%
- 2.14%
- 13.59%
- 34.13%
- 34.13%
- 13.59%
- 2.14%
- .13%

Standard Deviations
-4σ -3σ -2σ -1σ 0 +1σ +2σ +3σ +4σ

Cumulative Percentages
- 0.1% 2.3% 15.9% 50% 84.1% 97.7% 99.9%

Percentiles
1 5 10 20 30 40 50 60 70 80 90 95 99

Z scores
-4.0 -3.0 -2.0 -1.0 0 +1.0 +2.0 +3.0 +4.0

T scores
20 30 40 50 60 70 80

Standard Nine (Stanines)
1 2 3 4 5 6 7 8 9

Percentage in Stanine
4% 7% 12% 17% 20% 17% 12% 7% 4%

http://mewannet.blogspot.cz/
Skewness & Kurtosis

Above: positively skewed, skewness > 0

Above: negatively skewed, skewness < 0

Below: kurtosis = 3

Below: kurtosis = 2

Below: kurtosis = 1
Standard Deviation (SD): \[ \sigma(x) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \]

Standard Error of the Mean (SEM): \[ SEM = \frac{\sigma(x)}{\sqrt{n}} \]

95% Confidence Interval (95% CI): \[ 95\% \, CI = SEM \times 1.96 \]
Histogram of non-normal data

Box plot

- Frequency
- Value

Interquartile range (IQR)

- Maximum Value in the Data: $Q_3 + 1.5 \times IQR$
- 75th Percentile ($Q_3$)
- Median ($Q_2$)
- 25th Percentile ($Q_1$)
- Minimum Value in the Data: $Q_1 - 1.5 \times IQR$

Matlab example 6
Empirical cumulative distribution function

75% of population had less than 33.6 kCZK income

20% of population had less than 20.7 kCZK income
Common probability distributions

- Uniform
- Bernoulli
- Hypergeometric
- Binomial
- Geometric
- Poisson
- Exponential
- Negative Binomial
- Log Normal
- Student’s t
- Normal (Gaussian)
- Chi-Squared
- Weibull
- Gamma
- Beta

https://blog.cloudera.com
Population

Sampling distribution of the mean

Random sample (n = 10), calculation of the mean
Bootstrapping data

Bootstrap distribution of the median

Value

Frequency

0 5 10 15 20

Value

0 5 10 15 20

Bootstrap sample, calculation of the median

Matlab example 7
Probability distribution

Probability distribution via kernel density estimation

Matlab example 8
You always **have to** report your descriptive statistics with mean & standard deviation considered as a minimum !!!

<table>
<thead>
<tr>
<th>Table 1. Clinical characteristics of PreHD subjects.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>n = 28 (14 men)</strong></td>
</tr>
<tr>
<td>Age (years)</td>
</tr>
<tr>
<td>UHDRS motor score</td>
</tr>
<tr>
<td>Cognitive score</td>
</tr>
<tr>
<td>Tapping</td>
</tr>
<tr>
<td>Pegboard</td>
</tr>
<tr>
<td>Disease burden score</td>
</tr>
<tr>
<td>Years to onset (years)</td>
</tr>
</tbody>
</table>