Overview of Univariate Samples
Geog 210C
Introduction to Spatial Data Analysis

Chris Funk
Examples of Filtering

No Smoothing
Random Data?
Two Temperature Time-series
Temperature Time-series as Z-scores

Western Pac T-Z
Global GISS T-Z
Some Definitions (1)

- **Statistical population versus sample:**
  - **Population:** total set of elements_measurements that could be (hypothetically) observed in a study, e.g., all U.S. college students
  - **Sample:** subset of elements_measurements from population, e.g., college students in western U.S.

- **Population variables:**
  - Characteristics that describe a population, e.g., age or height of all college students in the U.S.

- **Population parameters versus sample statistics:**
  - **Parameters:** summary measures that describe a population variable, e.g., average age of college students in the U.S.
  - **Statistics:** summary measures that describe a sample variable, e.g., average age of college students in western U.S.
Some Definitions (2)

- **Statistical sampling:**
  - Procedure of getting a representative sample of a population, e.g., a random visit of all U.S. colleges
  - *Random sample* = sample in which every individual in population has same chance of being included

- **Descriptive statistics:**
  - Procedure of determining sample statistics, e.g., determination of the average student age of all randomly visited colleges

- **Statistical inference:**
  - Procedure of making statements regarding population parameters from sample statistics, e.g., average student age of all randomly visited colleges = average age of college students in the U.S.?

- **Statistical estimate:**
  - Best (educated) guess about the value of a population parameter

- **Hypothesis testing:**
  - Procedure of determining whether sample data support a hypothesis that specifies the value (or range of values) of a certain population parameter
Some Notation

- Sum of $n$ values (outcomes) from a variable $X$:
  \[
  \sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n
  \]

- Sum of linear combination of $n$ pairs of values (one set belongs to variable $X$, the other to variable $Y$):
  \[
  \sum_{i=1}^{n} (ax_i + by_i) = (ax_1 + by_1) + \ldots + (ax_n + by_n) = a\left(\sum_{i=1}^{n} x_i\right) + b\left(\sum_{i=1}^{n} y_i\right)
  \]
  $a$ and $b$ denote constants

- Sum of product of $n$ values of two variables $X$ and $Y$:
  \[
  \sum_{i=1}^{n} x_i y_i = (x_1 y_1) + \ldots +(x_n y_n) \neq (x_1 + \ldots + x_n)(y_1 + \ldots + y_n) = \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)
  \]

- Sum of a constant $k$:
  \[
  \sum_{i=1}^{n} k = k + \ldots + k = nk
  \]

- Sum of $n$ values of a variable $X$ each raised to a power $w$:
  \[
  \sum_{i=1}^{n} x_i^w = x_1^w + \ldots + x_n^w \neq (x_1 + \ldots + x_n)^w = \left(\sum_{i=1}^{n} x_i\right)^w
  \]

- Product of $n$ values of a variable $X$:
  \[
  \prod_{i=1}^{n} x_i = x_1 \cdot \ldots \cdot x_n
  \]
**Sample Histogram**

- **Setting:** Consider 10 hypothetical sample values:

  ![Sample Histogram Table]

- **Definition:** bar graph of # of sample values (counts) falling within a set of classes (bins)

  ![Counts of sample values within each class]

- **Estimated relative frequency table:**

  \[
  \hat{f}_k = \left( \frac{\text{# of data in } k\text{-th class}}{\text{total } \# \text{ of data}} \right)
  \]

<table>
<thead>
<tr>
<th>(x_k)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\hat{f}_k)</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
<td>0.4</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
</tr>
</tbody>
</table>

- **Histogram shape depends on number and width of bins:** use non-overlapping equal intervals with simple bounds
  - Rule of thumb for number of classes: \(5 \times \log_{10}(\# \text{of data})\)
  - For a **density histogram**, total area of bars = 1
Histogram Shape Characteristics

- **Peaked or not:**

- **Number of peaks:**

- **Symmetric or not:**
Definition: proportion of sample values less than, or equal to, any given datum value \( x_i \)

\[
\hat{p}_i = \hat{F}_X(x_i) = \frac{\text{# of data} \leq x_i}{\text{total # of data}}
\]

\approx \text{estimated probability that any sample chosen at random } \leq x_i

RANKED sample data and their estimated relative frequency:

<table>
<thead>
<tr>
<th>( x_k )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{f}_k )</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
<td>0.4</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.0</td>
</tr>
</tbody>
</table>

No bin width enters the construction of a CDF
Constructing a Sample CDF

Flowchart:
1. sort the $n$ sample data, $x$-values, in ascending order
2. construct a set of $n$ probability values, $p$-values, as:
   \[ \hat{p}_i = \hat{F}_X(x_i) = \frac{i}{n}, \quad i = 1, \ldots, n \]
   there are different ways to construct the $p$-values: the most widely used is actually
   \[ \hat{p}_i = \frac{i-0.5}{n}, \quad i = 1, \ldots, n \]
   this accommodates values below the lowest sample $x$-value and beyond the largest $x$-value
3. plot the sorted $x$-values against the corresponding $p$-values

Sample CDF = look up table of sorted $x$-values versus $p$-values

Note: $\hat{p}_{i+1} - \hat{p}_i = \frac{1}{n}, \quad 1 \leq i \leq n - 1$
Increasing the Resolution of a Sample CDF

**Objective:** construct a non-parametric (or is it multi-parametric) sample CDF (without fitting any parametric function, e.g., Gaussian) to that CDF

**Flowchart:**
1. choose a smallest possible x-value, $x_{\text{min}}$, and a largest possible x-value, $x_{\text{max}}$
2. associate probability $p_{\text{min}} = 0$ with $x_{\text{min}}$ and $p_{\text{max}} = 1$ with $x_{\text{max}}$
3. linearly interpolate between x- and p-value pairs to construct a piecewise linear sample CDF; there are variants on how to interpolate between such pairs of x- and p-values
**Quantiles**

**Definition:** sample value $x_p$ corresponding to specific cumulative relative frequency value $p$

**Famous quantiles:**
- min: $x_{0.0}$, lower quartile: $x_{0.25}$, median: $x_{0.5}$, upper quartile: $x_{0.75}$, max: $x_{1.0}$
- e.g., upper quartile is the number (in data units) with 75% of data being less than or equal to this value

**Percentiles:** $x_{0.01}, x_{0.02}, \ldots, x_{0.98}, x_{0.99}$

**Deciles:** $x_{0.1}, x_{0.2}, \ldots, x_{0.8}, x_{0.9}$

**Quantiles are not sensitive to extreme values (outliers)**

The graph above constitutes the sample quantile function of the inverse sample CDF

$$x_p = F_X^{-1}(p)$$
Measures of Central Tendency

- **Mid-range:**
  - arithmetic average of highest and lowest data: \( (x_{\text{max}} + x_{\text{min}})/2 \)

- **Mode:**
  - most frequently occurring value in data set

- **Median:**
  - datum value that divides data set into two halves;
    also defined as 50-th percentile: \( x_{0.5} \)

- **Mean:**
  - arithmetic average of data set
  - **sample mean:** \( \bar{x} \) or \( m_x = \frac{1}{n} \sum_{i=1}^{n} x_i \)
  - **population mean:** \( \mu_x = \frac{1}{N} \sum_{i=1}^{N} x_i \)

- Expressed in data units
- Also, \( m_x = \hat{\mu}_x \) the sample mean is an estimate of the population mean

- Most appropriate measure of central tendency depends on distribution shape
Measures of Dispersion (1)

- **Range:**
  - Difference between highest and lowest data: \( x_{\text{max}} - x_{\text{min}} \)

- **Interquartile range:**
  - Difference between upper and lower quartiles: \( x_{0.75} - x_{0.25} \)

- **Mean absolute deviation from mean:**
  - Average absolute difference between each datum and the mean:
    \[
    \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|
    \]

- **Median absolute deviation from median:**
  - Median absolute difference between each datum and the median:
    \( \text{median}|x_i - x_{0.5}| \)

- **Variance:**
  - Average squared difference between any datum and the mean
  - Sample variance:
    \[
    s_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - m_X)^2
    \]
  - Population variance:
    \[
    \sigma_X^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_X)^2
    \]

\[
\frac{s_X^2}{\sigma_X^2} = \hat{\sigma}_X^2 \quad \text{the sample variance is an estimate of the population variance}
\]

When will we care?
Measures of Dispersion (2)

- **Variance:**
  - Alternative definition: difference between average squared data and the mean squared
  - Sample variance
    \[ s^2_X = \frac{1}{n-1} \sum_{i=1}^{n} x_i^2 - \frac{n}{n-1} m_X^2 \]
  - Population Variance
    \[ \sigma^2_X = \frac{1}{N} \sum_{i=1}^{N} x_i^2 - \mu_X^2 \]

- **Coefficient of variation:**
  - Ratio of standard deviation and the mean
  - Sample coefficient
    \[ \frac{s_X}{m_X} = s_X = \sqrt{s^2_X} : \text{sample std deviation} \]
  - Population coefficient
    \[ \frac{\sigma_X}{\mu_X} = \sigma_X = \sqrt{\sigma^2_X} : \text{population std} \]

- **Choosing alternative measures of dispersion:**
  - Summary statistics involving squared values are sensitive to outliers
  - Summary statistic based on quantiles are robust to outliers
  - Coefficient of variation: useful for comparing spread of different data sets

Variance is expressed in data units SQUARED

The coefficient of variation is UNIT-LESS
Normalizing data to **zero mean** and **unit variance** (i.e. 1) allows more meaningful comparison of different data sets.

**Standardization procedure:**
1. Compute mean $\mu_x$ and standard deviation $\sigma_x$ of data set.
2. Subtract the mean from each datum: $x_i - \mu_x$.
3. Divide by the standard deviation: $z_i = (x_i - \mu_x) / \sigma_x$.

*NOTE* – Only applicable for **normal** data!

**Normalization procedure for non-normal data:**
1. Fit appropriate distribution.
2. Translate data into percentiles.
3. Translate percentiles into quantiles from a standard normal distribution ($\mu_x=0, \sigma_x=1$).

Normalized data are unit free; shape of distribution does not change (e.g., modes remain the same).
An Alternative Histogram Transformation (1)

- **Objective:** Transform original data set of $x$-values into a new data set of $z$-values with an arbitrary CDF

- **Flowchart:**
  1. construct piecewise linear CDF $F_X(x_p)$ of $x$-values and target CDF $F_Z(z_p)$ (with or without interpolation)
  2. find quantile $z_p$ of target CDF that corresponds to same quantile $x_p$ of sample CDF
  3. **Forward transformation:**

$$z_p = F_Z^{-1}(p) = F_Z^{-1}[\hat{F}_X^{-1}(x_p)]$$

![Diagram](image-url)
An Alternative Histogram Transformation (2)

- **Transformation characteristics:**
  - one-to-one (bijective)
  - non-linear and rank preserving; known as histogram equalization in digital image processing
  - can match any target CDF; that target CDF can be another sample CDF or a parametric CDF

- **Inverse transformation:**
  \[ x_n = F_X^{-1}(p) = F_X^{-1}[F_X(z_p)] \]

- Some CDF
- Std Gaussian CDF
Quantile-Quantile (Q-Q) Plots

- **Graph for comparing the shapes of two distributions**

- **Procedure:**
  1. rank both data sets from smallest to largest value
  2. compute quantiles of each data set
  3. cross-plot each quantile pair

- **Example:**

- **Interpretation**
  - straight plot aligned with 45° line implies two similar distribution shapes