Descriptive Data Summarization

(Understanding Data)

- First: Some data preprocessing problems...
Missing Values

- The approach of the problem of missing values adopted in SQL is based on nulls and three-valued logic (3VL)
- null corresponds to UNK for unknown
- 3VL a mistake?

Boolean Operators

<table>
<thead>
<tr>
<th>AND</th>
<th>t</th>
<th>u</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>t</td>
<td>u</td>
<td>f</td>
</tr>
<tr>
<td>u</td>
<td>u</td>
<td>u</td>
<td>f</td>
</tr>
<tr>
<td>f</td>
<td>f</td>
<td>f</td>
<td>f</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OR</th>
<th>t</th>
<th>u</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>t</td>
<td>t</td>
<td>t</td>
</tr>
<tr>
<td>u</td>
<td>u</td>
<td>u</td>
<td>u</td>
</tr>
<tr>
<td>f</td>
<td>f</td>
<td>u</td>
<td>f</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NOT</th>
<th>t</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>f</td>
<td></td>
</tr>
<tr>
<td>u</td>
<td>u</td>
<td></td>
</tr>
<tr>
<td>f</td>
<td>t</td>
<td></td>
</tr>
</tbody>
</table>

- In scalar comparison in which either of the compared is UNK evaluates the unknown truth value
Another important Boolean operator is 

<table>
<thead>
<tr>
<th>MAYBE</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>f</td>
</tr>
<tr>
<td>u</td>
<td>t</td>
</tr>
<tr>
<td>f</td>
<td>f</td>
</tr>
</tbody>
</table>

Example

Consider the query “Get employees who may be- but are not definitely known to be- programmers born before January 18, 1971, with salary less then €40.000

EMP WHERE MAYBE ( JOB = 'PROGRAMMER' AND DOB < DATE ('1971-1-18') AND SALLARY < 40000 )
Without maybe we assume the existence of another operator called IS_UKN which takes a single scalar operand and returns true if operand evaluates UNK otherwise false.

EMP WHERE ( JOB = ‘PROGRAMMER’
    OR IS_UKN (JOB) )
   AND ( DOB < DATE (‘1971-1-18’)
   OR IS_UKN (DOB) )
   AND ( SALLARY < 40000
   OR IS_UKN (SALLARY) )
   AND NOT ( JOB = ‘PROGRAMMER’ AND
   DOB < DATE (‘1971-1-18’) AND
   SALLARY < 40000 )

Numeric expression

- WEIGHT * 454
  - If WEIGHT is UKN, then the result is also UKN
  - Any numeric expression is considered to evaluate UNK if any operands of that expression is itself UNK

- Anomalies
  - WEIGHT-WEIGHT=UNK (0)
  - WEIGHT/0=UNK (“zero divide”)
UNK is not $u$ (unk)

- UNK (the value-unknown null)
- $u$ (unk) (unknown truth value)
- ...are not the same thing
  - $u$ is a value, UNK not a value at all!

Suppose $X$ is BOOLEAN
- Has tree values: $t$ (true), $f$ (false), $u$ unk
- $X$ is unk, $X$ is known to be unk
- $X$ is UKN, $X$ is not known!

Some 3VL Consequences

- The comparison $x=x$ does not give true
  - In 3VL $x$ is not equal to itself it is happens to be UNK
- The Boolean expression $p \text{ OR NOT}(p)$ does not give necessarily true
  - unk OR NOT (unk) = unk
Example

- Get all suppliers in Porto and take the union with get all suppliers not in Porto
  - We do not get all suppliers!
- We need to add maybe in Porto
  - In 2 VL p OR NOT(p) corresponds to
  - p OR NOT(p) OR MAYBE(p) in 3VL
- While two cases my exhaust full range of possibilities in the real world, the database does not contain the real world - instead it contains only knowledge about real world

Some 3VL Consequences

- The expression r JOIN r does not necessarily give r
- A=B and B=C together does not imply A=C
- ....
- Many equivalences that are valid in 2VL break down in 3VL
- We will get wrong answers
Special Values

- Drop the idea of null and UNK, unk 3VL
- Use special values instead to represent missing information

- Special values are used in the real world
  - In the real world we might use the special value "?" to denote hours worked by a certain employee if actual value is unknown

Special Values

- General Idea:

- Use an appropriate special value, distinct from all regular values of the attribute in question, when no regular value can be used
- The special value must be of the applicable attribute is not just integers, but integers integers plus whatever the special value is
- Approach is not very elegant, but without 3VL problems, because it is in 2VL
3VL

- **NULL** corresponds to UNK for unknown
- u or unk also represented by **NULL**

**MySQL:**

- The **NULL** value can be surprising until you get used to it. Conceptually, **NULL** means a missing unknown value, and it is treated somewhat differently from other values. To test for **NULL**, you cannot use the arithmetic comparison operators such as =, <, or <>.

- 3VL a mistake?

---

- Visualizing one Variable
- Statistics for one Variable
- Joint Distributions
- Hypothesis testing
- Confidence Intervals
Visualizing one Variable

- A good place to start
  - Distribution of individual variables
- A common visualization is the frequency histogram
  - It plots the relative frequencies of values in the distribution

To construct a histogram

- Divide the range between the highest and lowest values in a distribution into several bins of equal size
- Toss each value in the appropriate bin of equal size
- The height of a rectangle in a frequency histogram represents the number of values in the corresponding bin
The choice of bin size affects the details we see in the frequency histogram.

- Changing the bin size to a lower number illuminates things that were previously not seen.

Bin size affects not only the detail one sees in the histogram but also one’s perception of the shape of distribution.
Statistics for one Variable

Sample Size
- The sample size denoted by $N$, is the number of data items in a sample

Mean
- The arithmetic mean is the average value, the sum of all values in the sample divided by the number of values

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Statistics for one Variable

Median
- If the values in the sample are sorted into a non decreasing order, the median is the value that splits the distribution in half
- $(1 1 1 2 3 4 5)$ the median is 2
- If $N$ is even, the sample has middle values, and the median can be found by interpolating between them or by selecting one of the arbitrary
Mode

- The mode is the most common value in the distribution
- \((1 \ 2 \ 2 \ 3 \ 4 \ 4 \ 4)\) the mode is \(4\)
- If the data are real numbers mode nearly no information
  - Low probability that two or more data will have exactly the same value
- Solution: map into discrete numbers, by rounding or sorting into bins for frequency histograms
- We often speak of a distribution having two or more modes
  - Distributions has two or more values that are common

The mean, median and mode are measures of location or central tendency in distribution

- They tell us where the distribution is more dense
- In a perfectly symmetric, unimodal distribution (one peak), the mean, median and mode are identical
- Most real distributions - collecting data - are neither symmetric nor unimodal, but rather, are skewed and bumpy
Skew

- In a skewed distribution the bulk of the data are at one end of the distribution
  - If the bulk of the distribution is on the right, so the tail is on the left, then the distribution is called left skewed or negatively skewed
  - If the bulk of the distribution is on the left, so the tail is on the right, then the distribution is called right skewed or positively skewed
- The median is to be robust, because its value is not distorted by outliers
  - Outliers: values that are very large or small and very uncommon

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data
Trimmed mean

- Another robust alternative to the mean is the trimmed mean
  - Lop off a fraction of the upper and lower ends of the distribution, and take the mean of the rest
    - 0, 0, 1, 2, 5, 8, 12, 17, 18, 18, 19, 19, 20, 26, 86, 116
  - Lop off two smallest and two largest values and take the mean of the rest
    - Trimmed mean is 13.75
    - The arithmetic mean 22.75

- Maximum, Minimum, Range

- Range is the difference between maximum and minimum
Interquartile Range
- Interquartile range is found by dividing a sorted distribution into four containing parts, each containing the same number
- Each part is called quartile
- The difference between the highest value in the third quartile and the lowest value in the second quartile is the interquartile range

Quartile example
- 1,1,2,3,3,5,5,5,5,6,6,100
- The quartiles are
  - (1 1 2),(3 3 5),(5 5 5), (6,6,100)
    - Interquartile range 5-3=2
    - Range 100-1=99
- Interquartile range is robust against outliers
Standard Deviation and Variance

- Square root of the variance, which is the sum of squared distances between each value and the mean divided by population size (finite population)

\[
\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N}(x_i - \bar{x})^2}
\]

- Example
  
  • 1, 2, 15 Mean=6
  
  \[
  \frac{(1-6)^2 + (2-6)^2 + (15-6)^2}{3} = 40.66 \quad \sigma = 6.37
  \]

Sample Standard Deviation and Sample Variance

- Square root of the variance, which is the sum of squared distances between each value and the mean divided by sample size

\[
s = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N}(x_i - \bar{x})^2}
\]

- Example
  
  • 1, 2, 15 Mean=6
  
  \[
  \frac{(1-6)^2 + (2-6)^2 + (15-6)^2}{3-1} = 61 \quad s = 7.81
  \]
Because they are averages, both the mean and the variance are sensitive to outliers
Big effects that can wreck our interpretation of data
For example:
- Presence of a single outlier in a distribution over 200 values can render some statistical comparisons insignificant

The Problem of Outliers
One cannot do much about outliers expect find them, and sometimes, remove them

Removing requires judgment and depend on one’s purpose
Joint Distributions

- Good idea, to see if some variable influence others

Scatter plot

- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane
Correlation Analysis

- Correlation coefficient (also called Pearson’s product moment coefficient)

\[ r_{XY} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sigma_X\sigma_Y} \]

where \( n \) is the number of tuples, \( \bar{x} \) and \( \bar{y} \) are the respective means of X and Y, \( \sigma \) and \( \sigma \) are the respective standard deviation of A and B, and \( \Sigma(XY) \) is the sum of the XY cross-product.

- If \( r_{XY} > 0 \), X and Y are positively correlated (X’s values increase as Y’s). The higher, the stronger correlation.
- \( r_{XY} = 0 \): independent; \( r_{XY} < 0 \): negatively correlated

- A is positive correlated
- B is negative correlated
- C is independent (or nonlinear)
Correlation Analysis

- $\chi^2$ (chi-square) test
  \[ \chi^2 = \sum \frac{(Observed - Expected)^2}{Expected} \]

- The larger the $\chi^2$ value, the more likely the variables are related
- The cells that contribute the most to the $\chi^2$ value are those whose actual count is very different from the expected count
- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

Chi-Square Calculation: An Example

<table>
<thead>
<tr>
<th></th>
<th>Play chess</th>
<th>Not play chess</th>
<th>Sum (row)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Like science fiction</td>
<td>250(90)</td>
<td>200(360)</td>
<td>450</td>
</tr>
<tr>
<td>Not like science fiction</td>
<td>50(210)</td>
<td>1000(840)</td>
<td>1050</td>
</tr>
<tr>
<td>Sum(col.)</td>
<td>300</td>
<td>1200</td>
<td>1500</td>
</tr>
</tbody>
</table>

- $\chi^2$ (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)
  \[ \chi^2 = \frac{(250 - 90)^2}{90} + \frac{(50 - 210)^2}{210} + \frac{(200 - 360)^2}{360} + \frac{(1000 - 840)^2}{840} = 507.93 \]
- It shows that like_science_fiction and play_chess are correlated in the group
Properties of Normal Distribution Curve

- The normal (distribution) curve
  - From $\mu - \sigma$ to $\mu + \sigma$: contains about 68% of the measurements ($\mu$: mean, $\sigma$: standard deviation)
  - From $\mu - 2\sigma$ to $\mu + 2\sigma$: contains about 95% of it
  - From $\mu - 3\sigma$ to $\mu + 3\sigma$: contains about 99.7% of it

Kinds of data analysis

- Exploratory (EDA) – looking for patterns in data
- Statistical inferences from sample data
  - Testing hypotheses
  - Estimating parameters
- Building mathematical models of datasets
- Machine learning, data mining…

- We will introduce hypothesis testing
The logic of hypothesis testing

- Example: toss a coin ten times, observe eight heads. Is the coin fair (i.e., what is its long run behavior?) and what is your residual uncertainty?
- You say, “If the coin were fair, then eight or more heads is pretty unlikely, so I think the coin isn’t fair.”
- Like proof by contradiction: Assert the opposite (the coin is fair) show that the sample result (≥ 8 heads) has low probability \( p \), reject the assertion, with residual uncertainty related to \( p \).
- Estimate \( p \) with a sampling distribution.

Probability of a sample result under a null hypothesis

- If the coin were fair (\( p = .5 \), the null hypothesis) what is the probability distribution of \( r \), the number of heads, obtained in \( N \) tosses of a fair coin? Get it analytically or estimate it by simulation (on a computer):
  - Loop \( K \) times
    - \( r := 0 \) // \( r \) is num.heads in \( N \) tosses
    - Loop \( N \) times // simulate the tosses
      - Generate a random \( 0 \leq x \leq 1.0 \)
      - If \( x \geq p \) increment \( r \) // \( p \) is the probability of a head
      - Push \( r \) onto sampling_distribution
  - Print sampling_distribution
Sampling distributions

This is the estimated sampling distribution of \( r \) under the null hypothesis that \( p = .5 \). The estimation is constructed by Monte Carlo sampling.

The logic of hypothesis testing

- Establish a null hypothesis: \( H_0: p = .5 \), the coin is fair
- Establish a statistic: \( r \), the number of heads in \( N \) tosses
- Figure out the sampling distribution of \( r \) given \( H_0 \)

The sampling distribution will tell you the probability \( p \) of a result at least as extreme as your sample result, \( r = 8 \)

If this probability is very low, reject \( H_0 \) the null hypothesis
A common statistical test: The Z test for different means

- A sample \( N = 25 \) computer science students has mean IQ \( m=135 \). Are they “smarter than average”?

- **Population** mean is 100 with standard deviation 15
  - The null hypothesis, \( H_0 \), is that the CS students are “average”, i.e., the mean IQ of the population of CS students is 100.
  - What is the probability \( p \) of drawing the sample if \( H_0 \) were true? If \( p \) small, then \( H_0 \) probably false.

- Find the sampling distribution of the mean of a sample of size 25, from population with mean 100

---

Central Limit Theorem:

The sampling distribution of the mean is given by the Central Limit Theorem

The sampling distribution of the mean of samples of size \( N \) approaches a normal (Gaussian) distribution as \( N \) approaches infinity.

If the samples are drawn from a population with mean \( \mu \) and standard deviation \( \sigma \), then the mean of the sampling distribution is \( \mu \) and its standard deviation is \( \sigma_{\bar{x}} = \sigma / \sqrt{N} \) as \( N \) increases.

These statements hold irrespective of the shape of the original distribution.
The sampling distribution for the CS student example

- If sample of $N = 25$ students were drawn from a population with mean 100 and standard deviation 15 (the null hypothesis) then the sampling distribution of the mean would asymptotically be normal with mean 100 and standard deviation $15/\sqrt{25} = 3$

The mean of the CS students falls nearly 12 standard deviations away from the mean of the sampling distribution

Only ~1% of a normal distribution falls more than two standard deviations away from the mean

The probability that the students are "average" is roughly zero

---

The Z test

\[
Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} = \frac{135 - 100}{15/\sqrt{25}} = \frac{35}{3} = 11.67
\]
Reject the null hypothesis?

- Commonly we reject the H0 when the probability of obtaining a sample statistic (e.g., mean = 135) given the null hypothesis is low, say < .05.

- A test statistic value, e.g. Z = 11.67, recodes the sample statistic (mean = 135) to make it easy to find the probability of sample statistic given H0.

- We find the probabilities by looking them up in tables, or statistics packages provide them.
  - For example, Pr(Z ≥ 1.67) = .05; Pr(Z ≥ 1.96) = .01.
  - Pr(Z ≥ 11) is approximately zero, reject H0.
The t test

- Same logic as the Z test, but appropriate when **population standard deviation** is unknown, samples are small, etc.
- Sampling distribution is t, not normal, but approaches normal as samples size increases
- Test statistic has very similar form but probabilities of the test statistic are obtained by consulting tables of the t distribution, not the normal

Suppose N = 5 students have mean IQ = 135, std = 27

Estimate the standard deviation of sampling distribution using the sample standard deviation

\[
t = \frac{\bar{x} - \mu}{s} = \frac{135 - 100}{27} = \frac{35}{12.1} = 2.89
\]
Summary of hypothesis testing

- H0 negates what you want to demonstrate; find probability $p$ of sample statistic under H0 by comparing test statistic to sampling distribution; if probability is low, reject H0 with residual uncertainty proportional to $p$.
- Example: Want to demonstrate that CS graduate students are smarter than average. H0 is that they are average. $t = 2.89$, $p \leq .022$

- Have we proved CS students are smarter? NO!
- We have only shown that mean = 135 is unlikely if they aren’t. We never prove what we want to demonstrate, we only reject H0, with residual uncertainty.
- And failing to reject H0 does not prove H0, either!
Confidence Intervals

- Just looking at a figure representing the mean values, we can not see if the differences are significant

Confidence Intervals ($\sigma$ known)

- Standard error from the sample standard deviation
  \[ \sigma_{\bar{x}} = \frac{\sigma_{Population}}{\sqrt{N}} \]

- 95 Percent confidence interval for normal distribution is about the mean
  \[ \bar{x} \pm 1.96 \cdot \sigma_{\bar{x}} \]
Confidence interval when (σ unknown)

- Standard error from the sample standard deviation
  \[ \hat{\sigma}_x = \frac{s}{\sqrt{N}} \]

- 95 Percent confidence interval for t distribution (t_{0.025} from a table) is
  \[ \bar{x} \pm t_{0.025} \cdot \hat{\sigma}_x \]

Previous Example:

Measuring the Dispersion of Data

- Quartiles, outliers and boxplots

  - **Quartiles**: Q₁ (25th percentile), Q₃ (75th percentile)
  - **Inter-quartile range**: IQR = Q₃ – Q₁
  - **Five number summary**: min, Q₁, M, Q₃, max
  - **Boxplot**: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually
  - **Outlier**: usually, a value higher/lower than 1.5 x IQR
Boxplot Analysis

- **Five-number summary** of a distribution:
  Minimum, Q1, M, Q3, Maximum

- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IRQ
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extend to Minimum and Maximum

Visualization of Data Dispersion: Boxplot Analysis
- Visualizing one Variable
- Statistics for one Variable
- Joint Distributions
- Hypothesis testing
- Confidence Intervals

Next:
- Noise
- Integration
- Data redundancy
- Feature selection