Statistics and Data Analysis

Lecture by Dr. Zohar Yakhini
Typeset by Steven Karas

2016-11-22

1 Data Visualization

The goal of visualization is to make large datasets coherent, and support visual comparison of the data. Each of these has visual examples in the lecture slides.

1.1 Histograms

Split data into bins, draw on two axes. Labeling, axis range is important to mark. Bin height/length is also important to mark if bins are categorical. Note that histograms specifically represent the distribution, so they **must** sum to 1. Bin size is extremely important, because overly summarized data hides complexity, and overly binned data hides patterns.

1.2 Density Plots

Density plots smooth out the bins. Can be done by fitting a spline over the bins.

1.3 Pie chart

Can be useful for categorical bins, but horrible for distributions. Less than useful for comparing data, especially if it's in different pie charts.

1.4 Bar diagrams

Error bars.

1.5 CDF plot

Cumulative distribution function. Good for showing various relationships, and overall risk.

1.6 Scatter plot

Show relationships, non-linear patterns within the data. Good for sussing out correlation. 3 data axes can be represented using a 3D plot, or by including color/size as an additional axes.

Correlation is not Causation Can't believe I need to write this. Just because two random variables are dependent does not imply any causality between them.

1.6.1 Quantile-Quantile plot

Type of scatter plot. Loop over data, plot marks w.r.t. the quantile of each distribution. I did not understand what he just said. His chart has confusing axes, and we've been doing this for too long.

Moth longevity Step appearance of data implies methodology issues/sampling bias.

1.6.2 Monotonicity

A plot of the points (x_i, y_i) for $i = 1, ..., n$ is called monotone if $x_i <= x_j$ implies that $y_i <= y_j$. Not all scatter plots are monotone, but QQ plots are.

1.7 Maximal Information Coefficient

from slide 30,31

1.8 Heatmaps

1.9 Box plots

Graph box around 25% to 75% quantiles, mark median, circle mean, with "whiskers". The whiskers can represent many things, such as the min/max, the 10/90% quantiles, or something like $3Q + 1.5IQR$. Outliers can be marked directly, but aren't always.

1.10 Violin Plots

Similar to box plot, but plot the distribution instead of the box. Quartiles are marked with dashed lines. See example on slide 37. Two categories can be grafted as two halves of a "violin", for easy comparison.

2 Continuous Probability Distributions

A continuous random variable can assume any value in an interval in a collection of intervals (The reals is one such collection, and most common). It's impossible to say that the probability of a random variable gets a particular value. Instead, we refer to the probability of a random variable lying within a given interval.

2.1 Uniform Distribution

A random variable is uniformly distributed if the probability is proportional to the length of the interval.

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{else} \end{cases}$$

$$E(X) = \frac{a+b}{2}$$

$$Var(X) = \frac{(b-a)^2}{12}$$

$$E(X) = \sum_x xP(x) = \int_{-\infty}^{\infty} xf(x)dx$$

Example: HTTP requests

2.2 Cumulative Distribution Function

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

Empirical The empirical CDF is a step function.

NEXT WEEK central limit theorem