CSE 519: Data Science

Steven Skiena
Stony Brook University

Lecture 2: Mathematical Preliminaries

Probability

Probability theory provides a formal framework for reasoning about the likelihood of events.

The probability *p(s)* of an outcome *s* satisfies:

- *0 <= p(s) <= 1*
- $\sum_{s \in S} p(s) = 1$

These basic properties are often violated in casual use of “probability” in data science.

Probability vs. Statistics

- Probability deals with predicting the likelihood of future events, while statistics analyzes the frequency of past events.
- Probability is theoretical branch of mathematics on the consequences of definitions, while statistics is applied mathematics trying to make sense of real-world observations.

Compound Events and Independence

Suppose half my students are female (event A)
Half my students are above median (event B).

What is the probability a student is both A & B?

Events A and B are independent iff

$$P(A \cap B) = P(A) \times P(B)$$

Independence (zero correlation) is good to simplify calculations but bad for prediction.

Conditional Probability

The conditional probability P(A|B) is defined:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

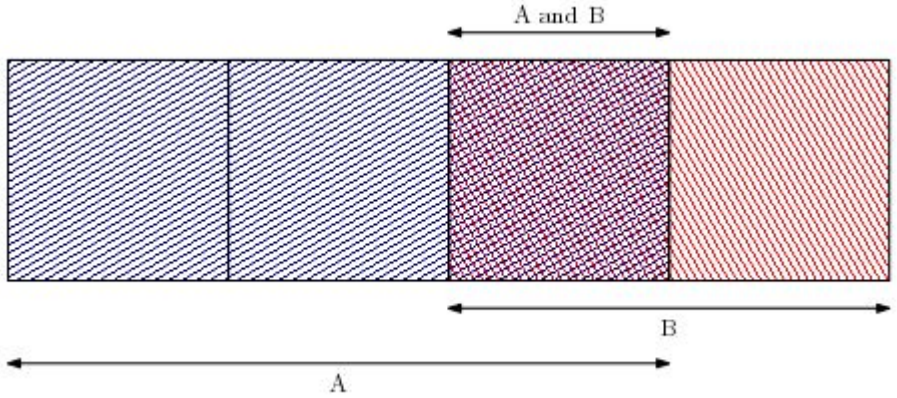
Conditional probabilities get interesting only when events are **not** independent, otherwise:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

Bayes Theorem

Bayes theorem is an important tool which reverses the direction of the dependences:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

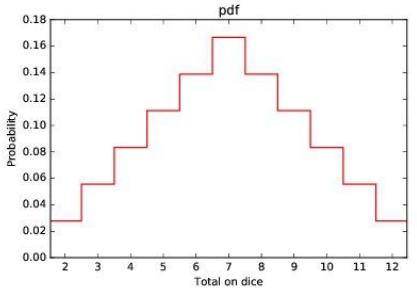


Distributions of Random Variables

Random variables are numerical functions where values come with probabilities.

Probability density functions (pdfs) represent RVs, essentially as histograms.

Here V is the sum of two dice.

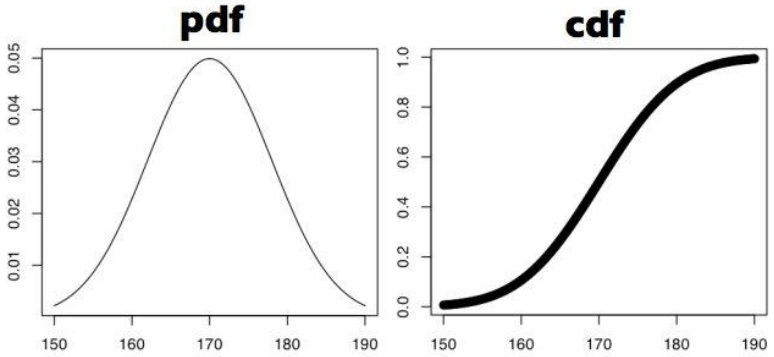


Probability/Cumulative Distributions

The cdf is the running sum of the pdf:

$$C(X \leq k) = \sum_{x \leq k} P(X = x)$$

The pdf and cdf contain exactly the same information, one being the integral / derivative of the other.



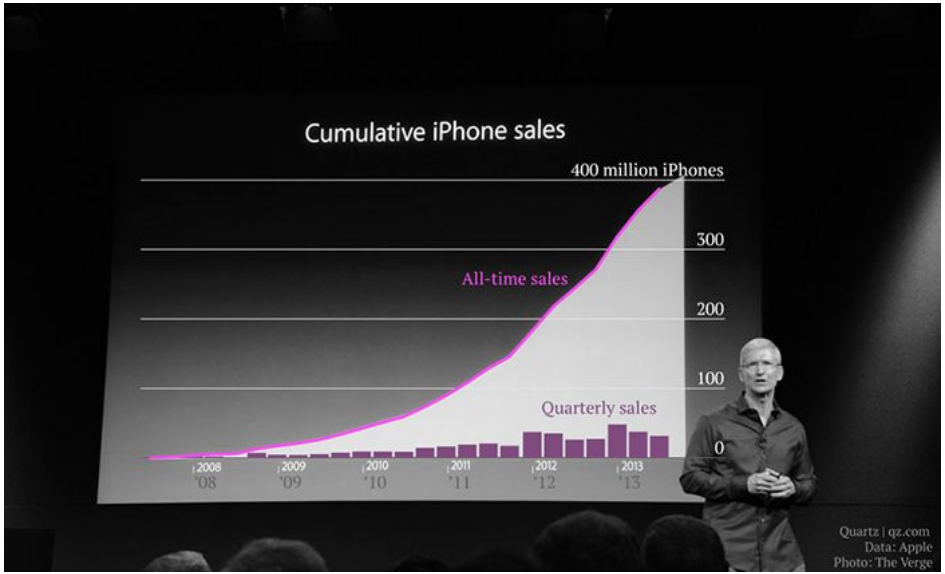
Visualizing Cumulative Distributions

Apple iPhone sales have been exploding, right?

How explosive is that growth, really?

Cumulative distributions present a misleading view of growth rate.

The incremental change is the derivative of this function, which is hard to visualize



Descriptive Statistics

Descriptive statistics provides ways to capture the properties of a given data set / sample.

- Central tendency measures describe the center around the data is distributed.
- Variation or variability measures describe data spread, i.e. how far the measurements lie from the center.

Centrality Measure: Mean

To calculated the mean, sum values and divide by number of observations: $\mu_X = \frac{1}{n}\sum_{i=1}^{n} x_i$

Mean is meaningful for symmetric distributions without outliers.

Other Centrality Measures

The median represents the middle value.

The geometric mean is the nth root of the product of n values: $\left(\prod_{i=1}^{n} a_i\right)^{1/n} = \sqrt[n]{a_1 a_2 \cdots a_n}.$

The geometric mean is always <= arithmetic mean, and more sensitive to values near zero.

Geometric means make sense with ratios:

½ and 2/1 *should* average to 1.

Which Measure is Best?

Mean is meaningful for symmetric distributions without outliers: e.g. height and weight.

Median is better for skewed distributions or data with outliers: e.g. wealth and income.

Bill Gates adds $250 to the mean per capita wealth but nothing to the median.

Aggregation as Data Reduction

Representing a group of elements by a new derived element, like mean, min, count, sum reduces a large dataset to a small summary statistic.

Such statistics can become features when taken over natural groups or clusters in the full data set.

Variance Metric: Standard Deviation

The variance is the square of the standard deviation sigma.

Do we divide by n or n-1?

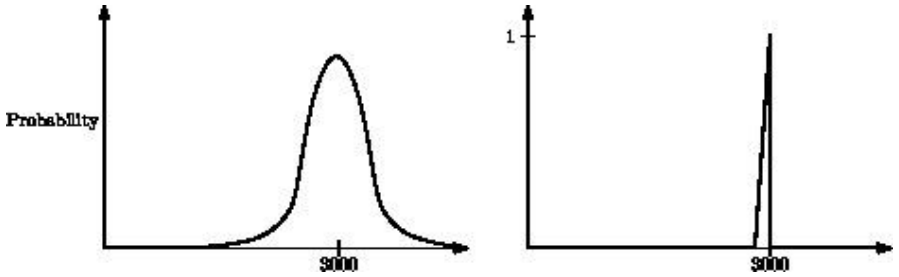
$$\hat{\sigma} = \sqrt{\frac{\sum_{i}^{n} (x_i - \bar{x})^2}{n-1}}$$

The population SD divides by n, the sample SD by n-1, but for large n, n ~ (n-1) so it doesn't really matter.

The Printer Cartridge Life Distribution

Distributions with the same mean can look very different.

But together, the mean and standard deviation fairly well characterize any distribution.



Parameterizing Distributions

Regardless of how data is distributed, at least $(1 - 1/k^2)$*th* of the points must lie within *k* sigma of the mean.

Thus at least 75% must lie within two sigma of the mean.

Even tighter bounds apply for normal distributions.

Interpreting Variance (Stock Market)

It is hard to measure “signal to noise” ratio, cause much of what you see is just variance.

Consider measuring the relative “skill” of different stock market investors.

Annual fluctuation in performance among funds is such that investor performance is random, meaning there is little real difference in skill.

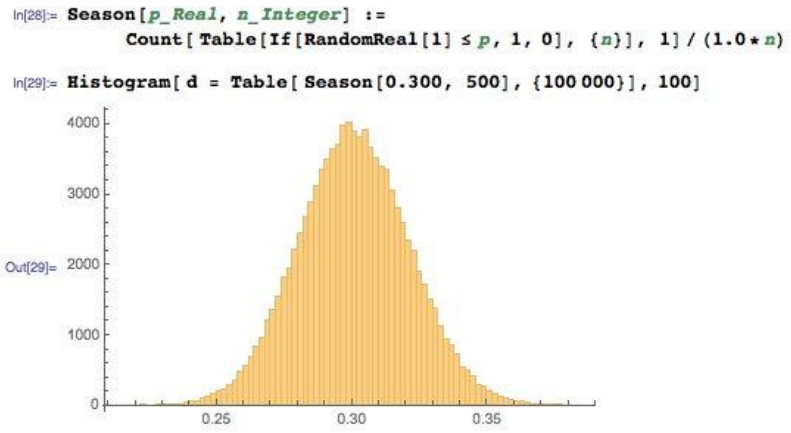
Interpreting Variance (Batting Avg)

In baseball, 0.300 hitters (30% success rate) represent consistency over 500 at-bats/season.

But simulations show a real 0.300 hitter has a 10% chance of hitting 0.275 or below.

They also have a 10% chance of hitting 0.325 or above.

Good or bad season, or lucky/lucky?



Interpreting Variance (Many Models)

We will typically develop several models for each challenge, from very simple to complex.

Some difference in performance will be explained by simple variance: which training/evaluation pairs were selected, how well parameters were optimized, etc.

Small performance win argue for simpler models.