Foundations of Analytics
Probability

Today

- Introduction: Probability, Statistics, Variance
- Terminology
  - Random Experiments, Sample Spaces, Events
  - Interpretations and Axioms of Probability
  - Addition Rules
  - Conditional Probability
  - *Multiplication and Total Probability Rules*
  - *Independence*
- *Bayes Theorem*
- *Naive Bayes as a Classifier*

Probability vs. Statistics

Different subjects: both about random processes

Probability

- Logically self-contained
- A few rules for computing probabilities
- One correct answer

Statistics

- Messier and more of an art
- Get experimental data and try to draw probabilistic conclusions
- No single correct answer

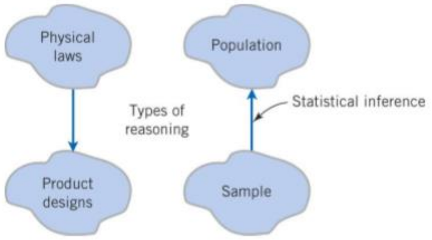
Probability vs. Statistics

Probability example

You have a fair coin (equal probability of heads or tails). You will toss it 100 times. What is the probability of 60 or more heads? There is only one answer (about 0.028444) and we will learn how to compute it.

Statistics example

You have a coin of unknown provenance. To investigate whether it is fair you toss it 100 times and count the number of heads. Let’s say you count 60 heads. Your job as a statistician is to draw a conclusion (inference) from this data.



*Statistical Inference is a type of (inductive) reasoning

Variability

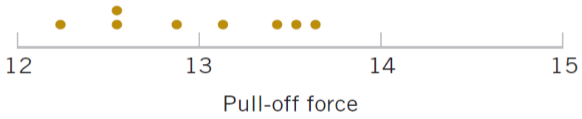
- Experiments & processes are not deterministic.
- Statistical techniques are useful to describe and understand variability.
- By variability, we mean successive observations of a system or phenomenon do *not* produce exactly the same result.
- Statistics gives us a framework for describing this variability and for learning about potential sources of variability.

An Example of Variability

Monthly credit hard expenditures of eight customers are reported (in K TL):
12.6, 12.9, 13.4, 12.3, 13.6, 13.5, 12.6, 13.1.

The **dot diagram** is a very useful plot for displaying a small body of data - say up to about 20 observations.

This plot allows us to see easily two features of the data; the **location,** or the middle, and the **scatter** or **variability**.



Random Variable

- Since credit card expenditures vary or exhibits variability, it is a **random variable**
- A random variable, X, can be modeled by:

$$X = \mu + \varepsilon$$

where μ is a constant and ε is a random disturbance.

An Experiment in Variation: An attempt to control variability

W. Edwards Deming, a famous industrial statistician & contributor to the Japanese quality revolution, conducted a illustrative experiment on process **over-control** or **tampering**.

Let’s look at his apparatus and experimental procedure.

Deming’s Experimental Set-up

Marbles were dropped through a funnel onto a target and
the location where the marble struck the target was recorded.

Variation was caused by several factors:
Marble placement in funnel & release dynamics, vibration, air currents, measurement errors.

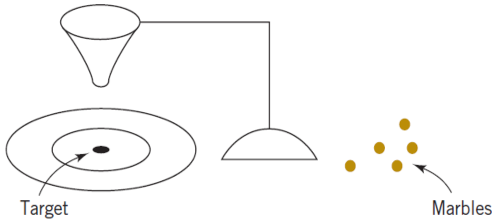
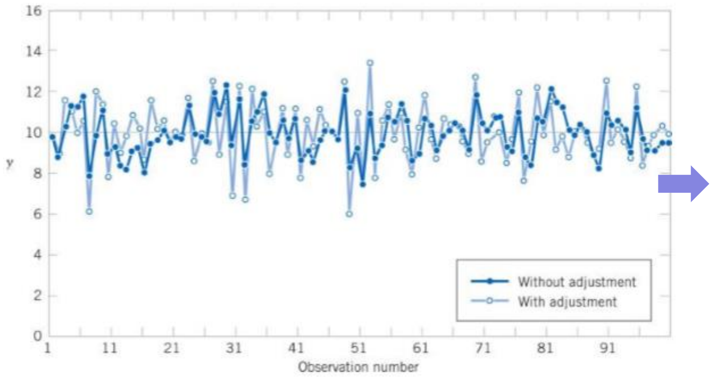


Figure: Deming’s Funnel experiment

Deming’s Experimental Procedure

- The funnel was aligned with the center of the target. Marbles were dropped. The distance from the strike point to the target center was measured and recorded.
- Strategy 1: The funnel was not moved. Then the process was repeated.
- Strategy 2: The funnel was moved an equal distance in the opposite direction to compensate for the error. He continued to make this type of adjustment after each marble was dropped. Then the process was repeated.
- Which strategy do you think yields smaller variation?

Deming’s Experimental Procedure



Adjustment Increases Variability

- Distance from the target for strategy 2 was approximately twice as large than for strategy 1.
- The deviations from the target is increased due to the adjustments to the funnel.
- This experiment explains that the adjustments to a process based on random disturbances can actually increase the variation of the process. This is referred to as **overcontrol** or **tampering.**

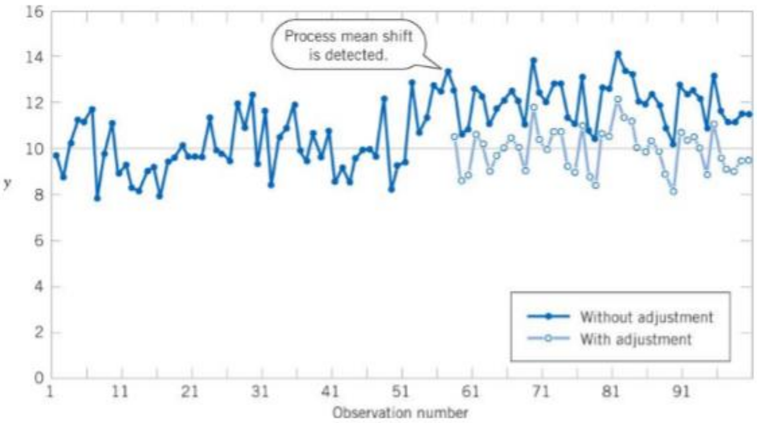
Conclusions from the Deming Experiment

The lesson of the Deming experiment is that a process should not be adjusted in response to random variation, but only when a clear shift in the process value becomes apparent.

Then a process adjustment should be made to return the process outputs to their normal values.

To identify when the shift occurs, a control chart may be used. Output values, plotted over time along with the outer limits of normal variation, pinpoint when the process leaves normal values and should be adjusted.

Detecting and Correcting the Process



Process mean shift is detected at observation #57, and an adjustment (a decrease of two units) reduces the deviations from target.

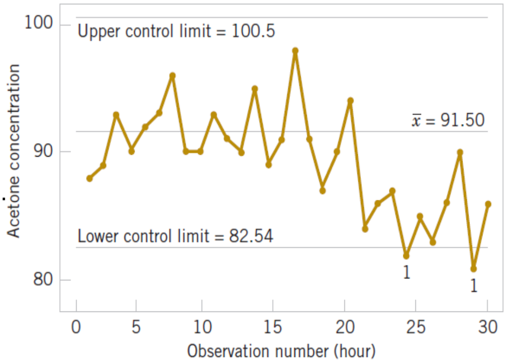
How Is the Change Detected?

- A control chart is used. Its characteristics are:
  - – Time-oriented horizontal axis, e.g., hours.
  - – Variable-of-interest vertical axis, e.g., % drop rate.
- Long-term average is plotted as the center-line.
- Long-term usual variability is plotted as an upper and lower control limit around the long-term average.
- A sample of size *n* is taken and the averages are plotted over time. If the plotted points are between the control limits, then the process is normal; if not, it needs to be adjusted.

How Is the Change Detected Graphically?

An example from a production process is selected to depict both UCL & LCL.

A control chart for the chemical process concentration data. Process steps out at hour 24 & 29. Shut down & adjust process.



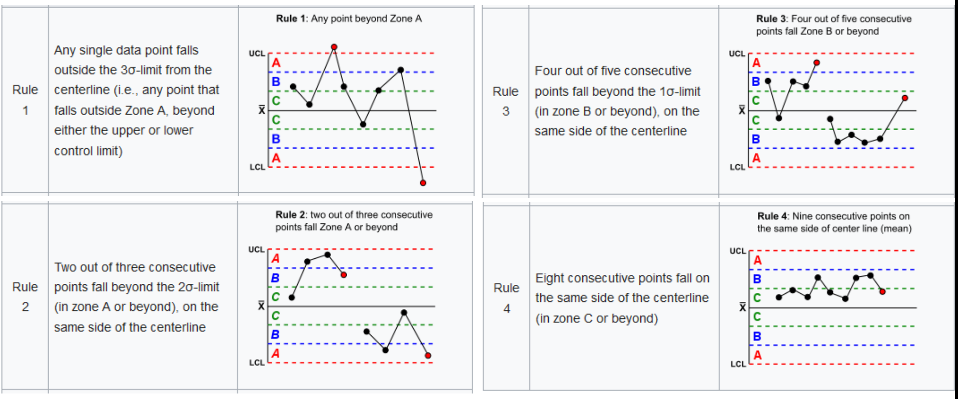
The center line on the control chart is just the average of the concentration measurements for the first 20 samples

$$\overline{X} = 91.5g/l$$

when the process is stable. The upper control limit and the lower control limit are located 3 standard deviations of the concentration values above and below the center line.

Anomaly Detection

- An application: Western Electronic Rules



Rule 1	Any single data point falls outside the 3σ-limit from the centerline (i.e., any point that falls outside Zone A, beyond either the upper or lower control limit)	**Rule 1**: Any point beyond Zone A	Rule 3	Four out of five consecutive points fall beyond the 1σ-limit (in zone B or beyond), on the same side of the centerline	**Rule 3**: Four out of five consecutive points fall Zone B or beyond
Rule 2	Two out of three consecutive points fall beyond the 2σ-limit (in zone A or beyond), on the same side of the centerline	**Rule 2**: two out of three consecutive points fall Zone A or beyond	Rule 4	Eight consecutive points fall on the same side of the centerline (in zone C or beyond)	**Rule 4**: Nine consecutive points on the same side of center line (mean)

Mechanistic and Empirical Models

A **mechanistic model** is built from our underlying knowledge of the basic physical mechanism that relates several variables.

Example: Ohm's Law

$$\text{Current} = V/R$$

$$I = V/R + \varepsilon$$

where ε is a term added to the model to account for the fact that the observed values of current flow do not perfectly conform to the mechanistic model.

- The form of the function is known.

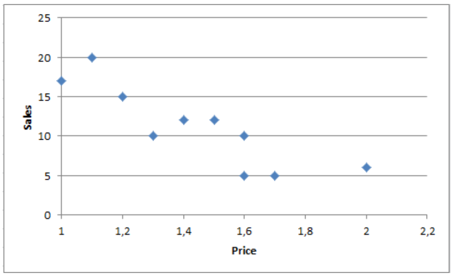
Mechanistic and Empirical Models

An **empirical model** is built from our engineering and scientific knowledge of the phenomenon, but is not directly developed from our theoretical or first-principles understanding of the underlying mechanism.

The form of the function is not known *a priori*.

An Example of an Empirical Model: Price vs. Demand

	sales	price
week	(thousands of gallons)	per gallon
1	10	1,3
2	6	2
3	5	1,7
4	12	1,5
5	10	1,6
6	15	1,2
7	5	1,6
8	12	1,4
9	17	1
10	20	1,1



An Example of an Empirical Model:

What would be your selection for the «empirical model» that represents the relation between price and sales?

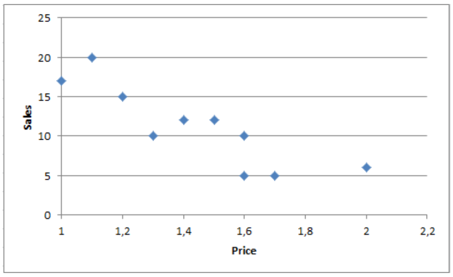
$Sales = \beta_0 + \beta_1 (price)$

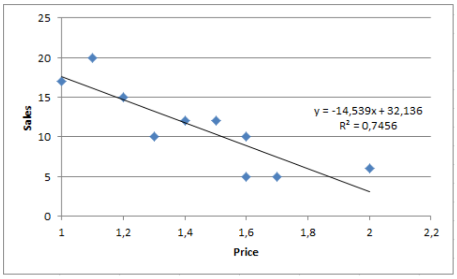
In general, this type of empirical model is called a **regression model**.

The **estimated** regression relationship is given in the next slide

An Example of an Empirical Model: Price vs. Demand

week	sales (thousands of gallons)	price per gallon
1	10	1,3
2	6	2
3	5	1,7
4	12	1,5
5	10	1,6
6	15	1,2
7	5	1,6
8	12	1,4
9	17	1
10	20	1,1


Sales
Price


Sales
y = -14,539x + 32,136
R² = 0,7456
Price