ECE 295: Lecture 03 Histograms

Spring 2018

Prof Stanley Chan

School of Electrical and Computer Engineering
Purdue University

The Era of Big Data!

http://i1.wp.com/olap.com/wp-content/uploads/2013/11/bigstock-Big-data-concept-in-word-tag-c-49922318.jpg

Statistics

The science of making sense of data!

Why study statistics?

... Using fancy tools like neural nets, boosting, and support vector machines without understanding basic statistics is like doing brain surgery before knowing how to use a band-aid...

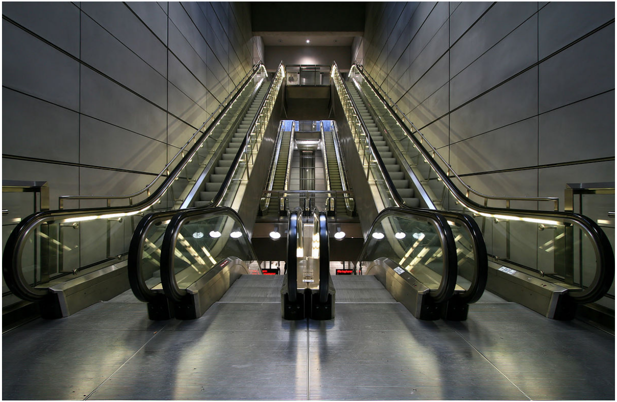
Larry Wasserman, "All of Statistics"

Today's Plan

Histogram!

Let's do a case study first ...

The Escalator Problem

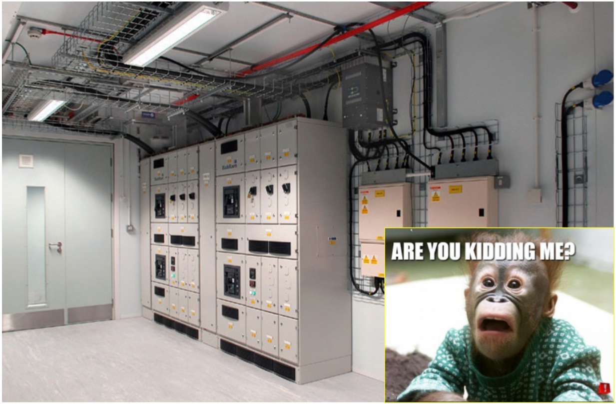


Energy efficient escalators:

- ON when there are pedestrians
- STAND-BY when there is no pedestrian for several seconds
- How much saving?

That's Easy!

- Go to the meter room, and
- Measure it!!!



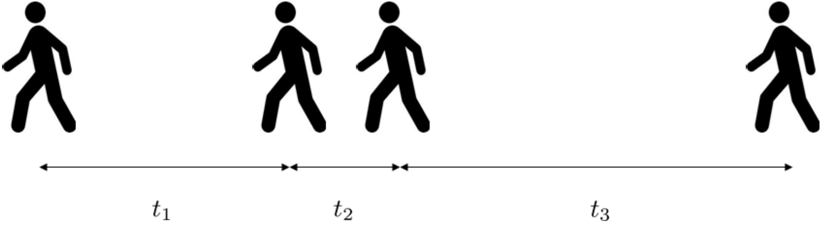
But what if you have not yet built the escalator?

Let's collect data

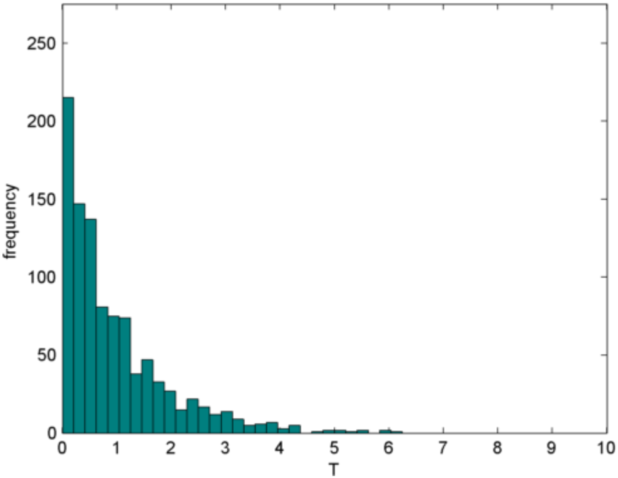
Inter-arrival Time

Let T be the inter-arrival time.

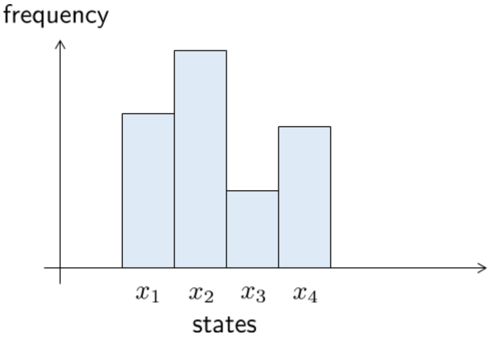
Possible values of T: Call them $t_1, t_2, t_3, \ldots,$



How does the histogram of T look like?



What can be told from a histogram?



- Set of all possible state: $x_1, x_2, \ldots, x_m$.
- Empirical **frequency** of each state: $\widehat{p}_1, \widehat{p}_2, \ldots, \widehat{p}_m$.

Important!

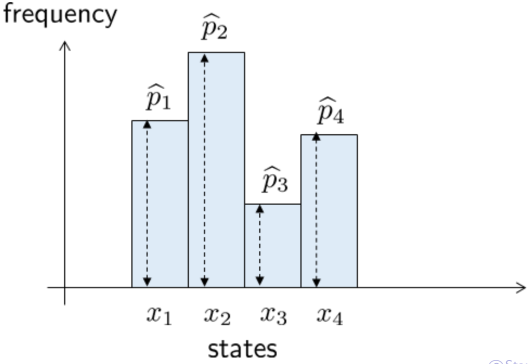
$$\widehat{p}_1 + \widehat{p}_2 + \ldots + \widehat{p}_m = 1.$$

What can be told from a histogram?

Sample Mean:

$$\overline{X} = \sum_{i=1}^{m} \widehat{p}_i x_i$$

- "Average" of computed from the histogram
- Could be different if you run another experiment

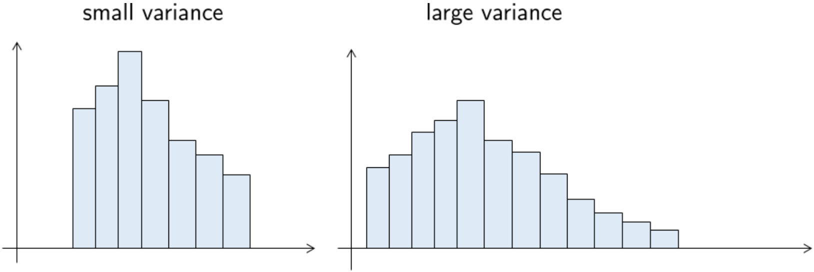


What can be told from a histogram?

Sample Variance:

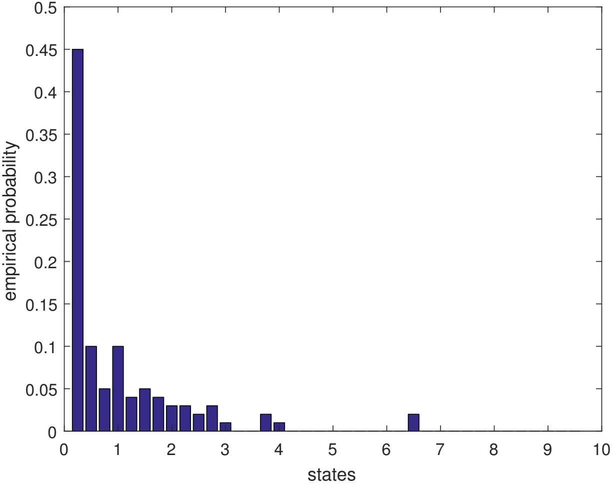
$$S^2 = \sum_{i=1}^{m} \widehat{p}_i (x_i - \overline{X})^2.$$

- Measures the deviation
- Large S^2 means that the histogram is wide-spread
- S is the sample standard deviation



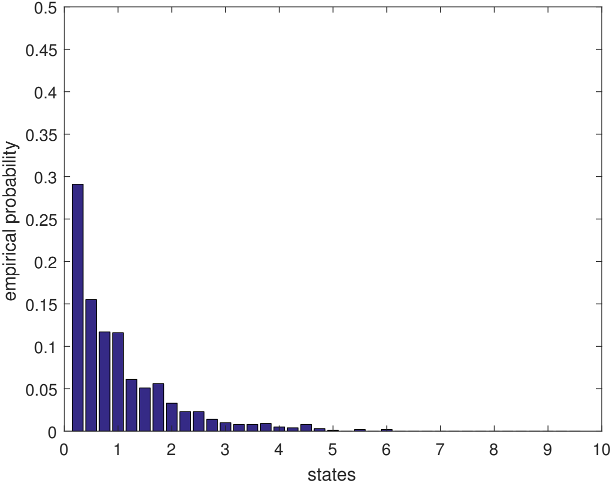
Histogram Grows

What if we have 100 measurements?



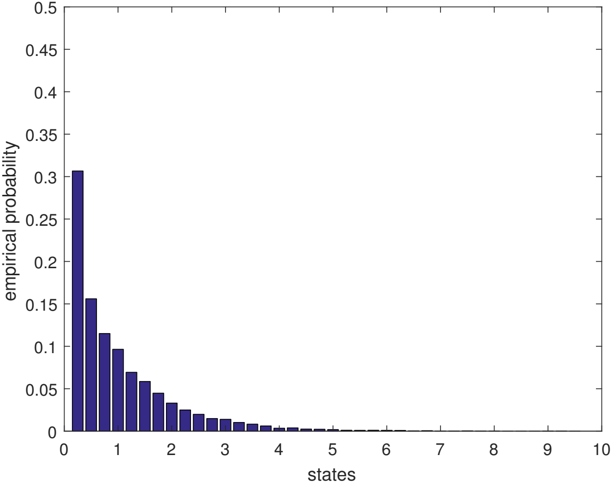
Histogram Grows

What if we have 1000 measurements?



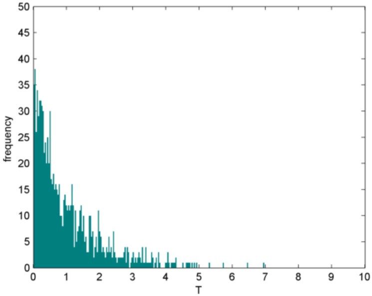
Histogram Grows

What if we have 10000 measurements?

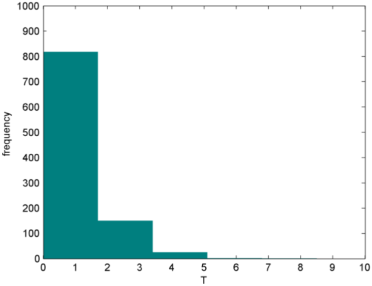


Bin-width of Histogram

Bad choice of bin-width:



200 bins



5 bins

- Too many bins: Not enough data!
- Too few bins: Not descriptive!

Optimal Bin-width

Here is a method to estimate the bin-width. The method is called **Cross-Validation**.

Notations

- n: number of data points
- m: number of bins
- h: bin-width: n/m. (Can round off to nearest integer.)
- $\widehat{p}_j$: frequency of the j-th bin.

Cross-validation Score:

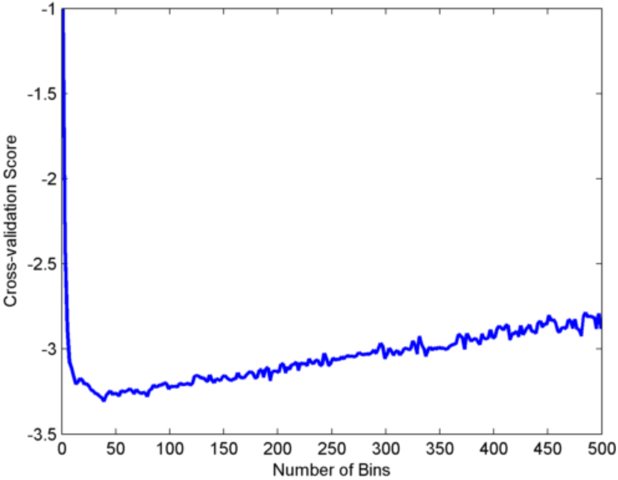
$$J(h) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)h}\left(\widehat{p}_1^2 + \widehat{p}_2^2 + \ldots + \widehat{p}_m^2\right).$$

Optimal Bin-width

Procedure:

- Pick the number of bins m.
- Since n is fixed, we can compute $h = n/m$.
- Build a histogram of m bins.
- The heights of the histogram bars are $\widehat{p_j}$.
- Calculate the Cross-Validation Score $J(h)$.
- If $J(h)$ is high, try another m until $J(h)$ is low enough.

Optimal Bin-width



Summary

Histogram:

- The most **basic** tool we use to analyze data.
- **Three components**: states, empirical probability, bin-width.
- Bin-width can be controlled by **Cross Validation**.
- **Sample Mean**: average of computed from the histogram.
- **Sample Variance**: deviation found of the states in the histogram.
- High-dimensional histograms.