Data Science

Sample Statistics

October 1st, 2014

Jarret Petrillo, jarret.petrillo@gmail.com

How many different blueprints do we need?

Midtown Post Office

The Palace

Washington Square Monument

Residential home

1342 N. Leavitt Price History

Price History

Date	Event	Price		$/sqft	Source
09/06/13	Sold	$901,500	+0.2%	$200	Public Record
06/25/13	Listing removed	$900,000		$200	@properties
04/30/13	Listed for sale	$900,000	+63.9%	$200	@properties
01/19/10	Sold	$549,000	-56.4%	$122	Public Record
06/02/09	Sold: Foreclosed to lender	--		$0	Public Record
12/07/04	Sold	$1,260,000	+86.7%	$280	Public Record
11/17/04	Sold	$675,000	-29.8%	$150	Public Record
10/16/03	Sold: Foreclosed to lender	--		$0	Public Record
06/20/00	Sold	$961,500	+381%	$213	Public Record
09/08/99	Sold	$200,000		$44	Public Record

Sample mean: reducing data to one dimension

```
sample mean = sum(x)/n
```

How much information is lost?

Deviation from the mean
(x - x_bar)

Variance is the sum of square deviations
sample variance = sum(sqr(x-x_bar))/(n-1)

Why squared? And other problems of variance

- *We're trying to capture in one dimension the measure of misfit*
- *Square deviations penalizes large deviations more than small ones*
- *Variance has no units*
- *How do we interpret a variance of 1,000?*

Standard deviation is the square root of variance

Adding more data...

- *Consider a dataset of two columns*
- *We take the sample means and variances for both columns*
- *What are we missing?*