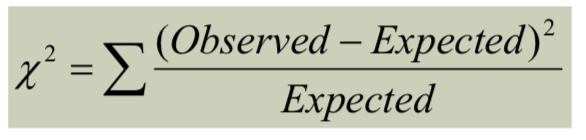


**Deterministic model**: all data is known before. Once you start the system, you know exactly what is going to happen.

**Probabilistic model**: Element of chance is involved. know the likelihood that something will happen, but don’t know when it will happen

|  |  |
| --- | --- |
| **Correlation:** Measure of strength of linearity between 2 variables. How strongly related 2 variables are. (between -1 and +1)    If , then a relationship exists | **Covariance** Measure of strength of correlation. Indicates extent to which two variables change in tandem. (any value)  Sample: Population: |

Chi-square Test: if chi-squared>critical chi-squared reject null hypothesis

**Binary Variables:** 2 states (0 or 1), 2 types are Symmetric and Asymmetric

|  |  |
| --- | --- |
| **Symmetric:** both equally important  Distance Measure:    q= Both A&B, r=A not B, s=B not A, t=neither | **Asymmetric:** not equally important  Similarity Measure(Jaccard):    Distance/Dissimilarity measure: |

**Nominal Data:** unordered data. categories Ex: profession, eye color

**Ordinal Data: (STAGES)** Discrete or continuous, order is important Ex: doneness of meat(rare, medium, well done)

Can map range onto [0,1] so range of 4 becomes (0, 0.33, 0.66, 1) Ex: Class (freshman, sophomore, junior, senior)

d(0, 3)=1 and d(2,3)=0.33

**Interval:** discrete units, not for comparison, nonlinear Ex: years of surgeries, time left for work, income

**Ratio data:** can be compared, linear Ex: height of 2 year old, commute time

**Dimensionality Reduction:** remove unimportant variables(redundant, irrelevant, similar to another)

**Numerosity Reduction(Data Reduction):** regression and log-linear models. Remove actual data points, view overview/graphs

**Multi-collinearity:** add more IVs creates relationships between IVs

Refers to relation of the dimensions.

**Overfitting:** add more independent variables to regression procedure making it worse(IV too similar may overfit)

Refers to number of dimensions. Too many predictors, can detect noise(not supposed to)

You have multicollinearity with overfitting, but you cannot have overfitting without multicollinearity.

**Equidepth Binning:** Divide range into N intervals, approx same number of samples. Good data scaling.categorical attributes tricky

**Equiwidth Binning:** divide range into N intervals, exact same number of samples. Outliers dominate, skewed data not handled well

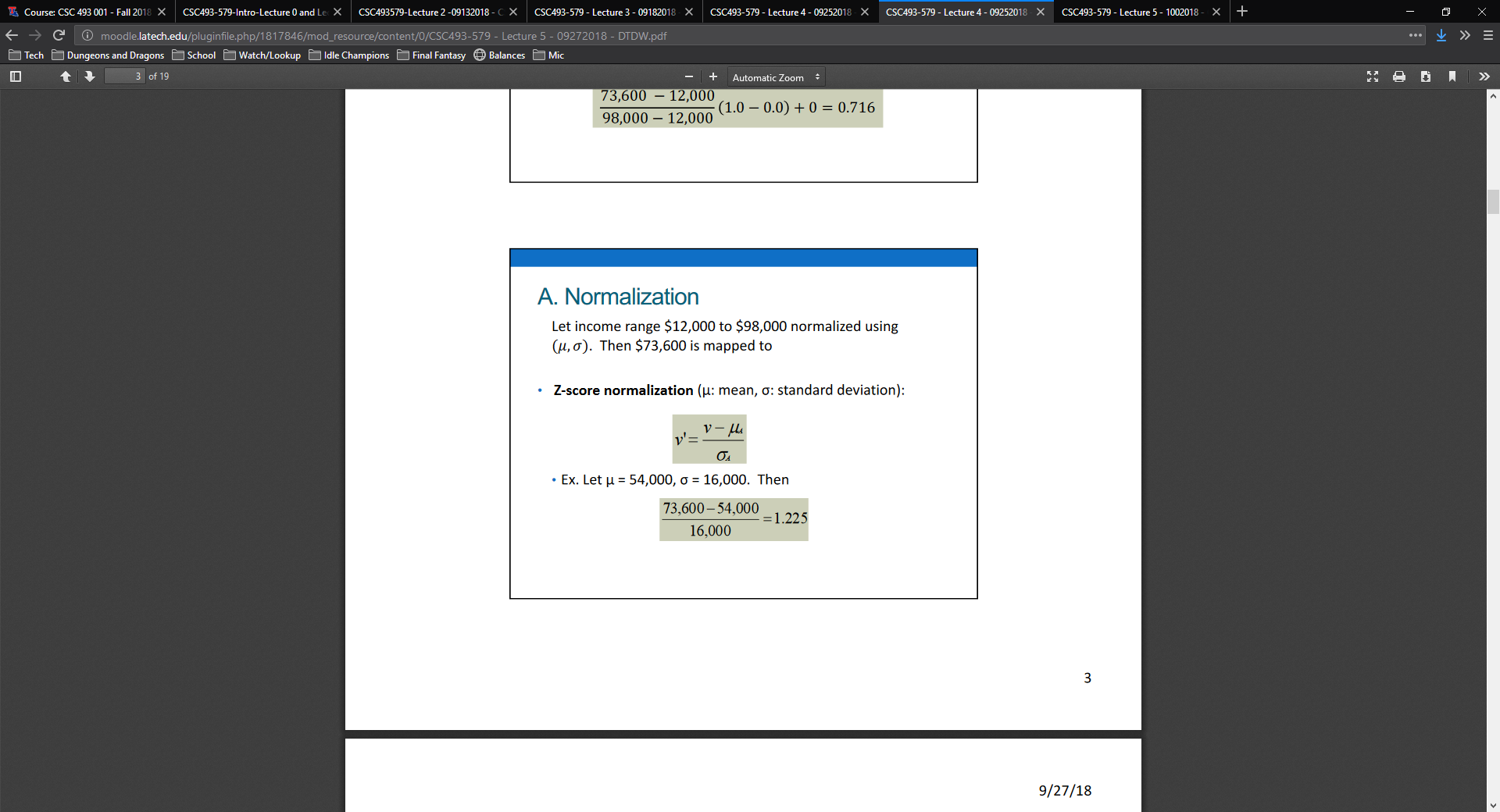
Width=(High-Low)/N

**Measure:** raw value measured.  **Metric:**calculation between 2 or more measures, usually percent, ratio, fraction...

**Similarity/Dissimilarity:** (Correlation) numerical measure of how alike/different 2 objects are(usually 0 to 1, w/ 1 as identical)

**Skewed Distribution:** to right = +, to left = -

**Min-Max Normalization:**

**Z-Score Normalization:** 

**Measures of Data Quality:** Accurate, Complete, Consistent, Timely, Believable, Interpretable

**Data Prep:** Cleaning, integration, reduction, transformation, discretization

