


  
week4

Week 4

INTRODUCTION TO DATA SCIENCE IN PYTHON




### Distributions

- **Distribution:** Set of all possible random variables
- **Example:**
  - Flipping Coins for heads and tails
    - a binomial distribution (two possible outcomes)
    - discrete (categories of heads and tails, no real numbers)
    - evenly weighted (heads are just as likely as tails)
  - Tornado events in Ann Arbor
    - a binomial distribution
    - Discrete
    - evenly weighted (tornadoes are rare events)

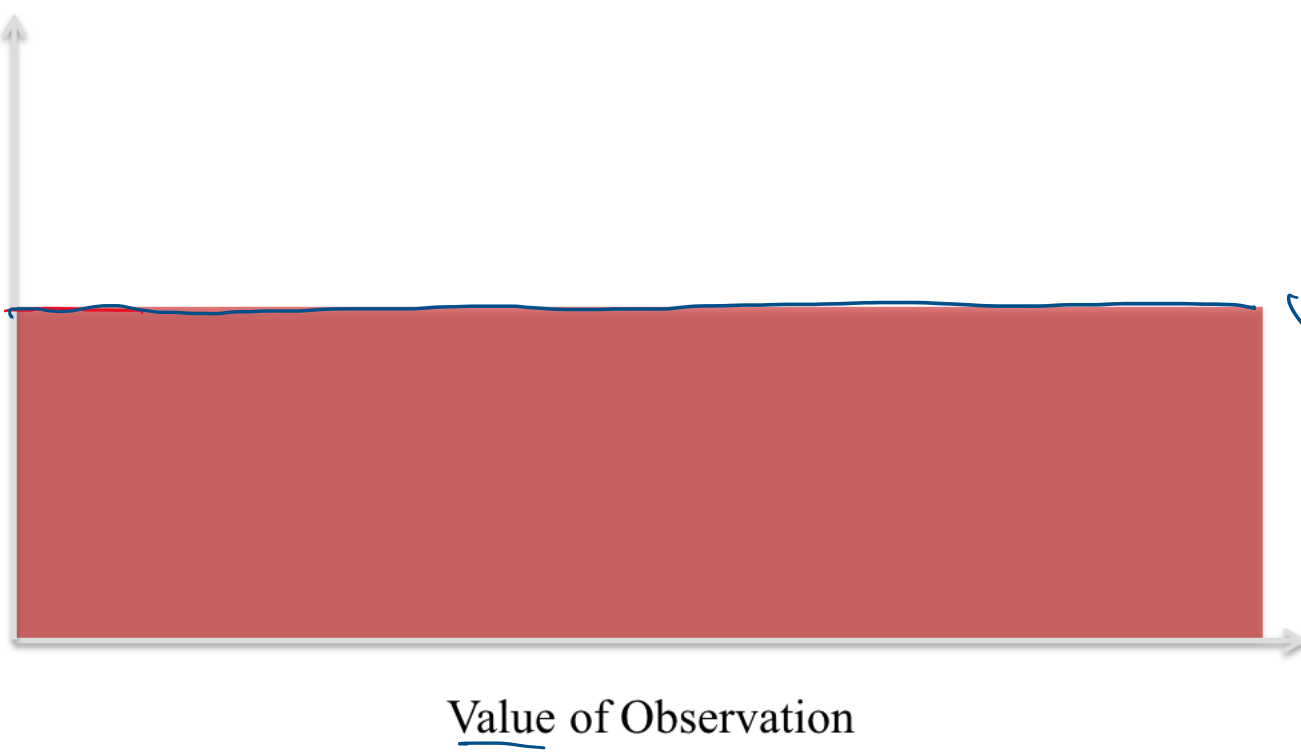
Binomial  $\rightarrow$  continues : value of the given observation & cate  $\rightarrow$  real number  
 $\hookrightarrow$  graph distribution

Week 4

INTRODUCTION TO DATA SCIENCE IN PYTHON



### Uniform Distribution (Continuous)




Probability Observation Occurs

Value of Observation

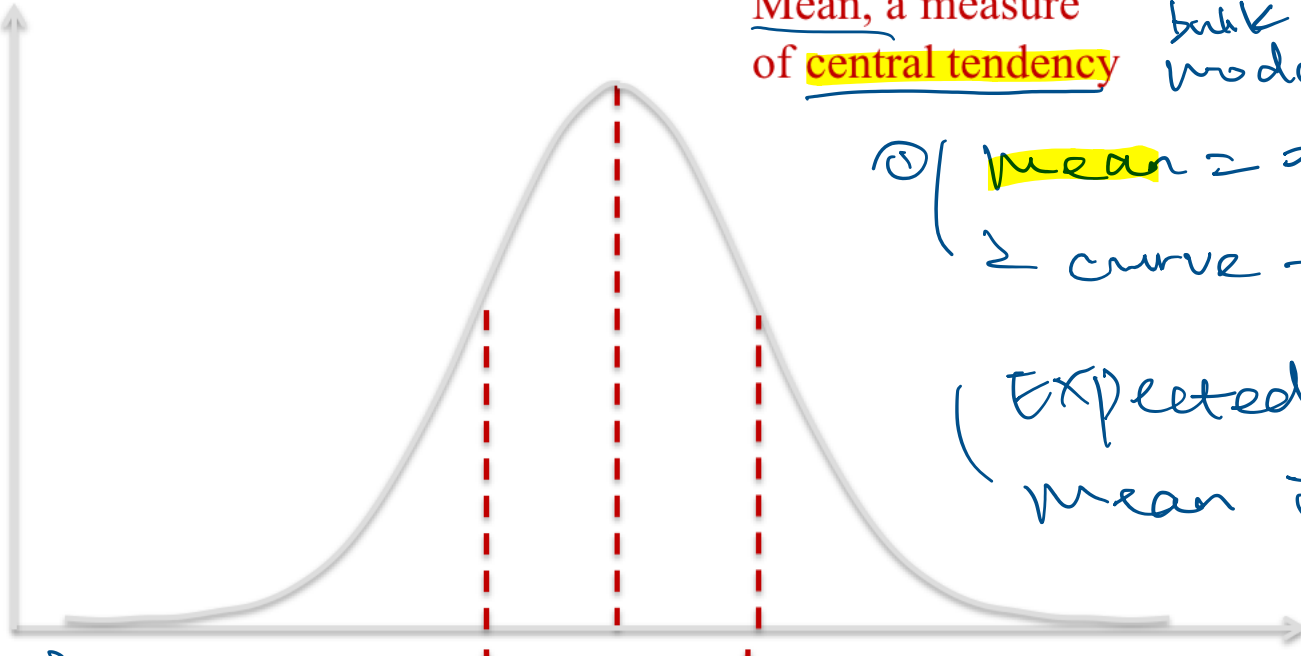
uniform distribution

Week 4

INTRODUCTION TO DATA SCIENCE IN PYTHON



### Normal (Gaussian) Distribution




Probability Observation Occurs

Value of Observation

Mean, a measure of central tendency  
mode, median, mean  
①  $\mu = 0$   
 $\hookrightarrow$  curve - symmetric  
Expected value = probability from distribution given a sufficient large sample set  
mean  $\bar{x} = \frac{\sum x}{n}$  sample from distribution samples taken  
Ex: roll a die 3 times 1 2 6  $\rightarrow \frac{1+2+6}{3} = 3$   
② variance of distribution  
how different each item from mean  
broadly values of samples spread out from mean

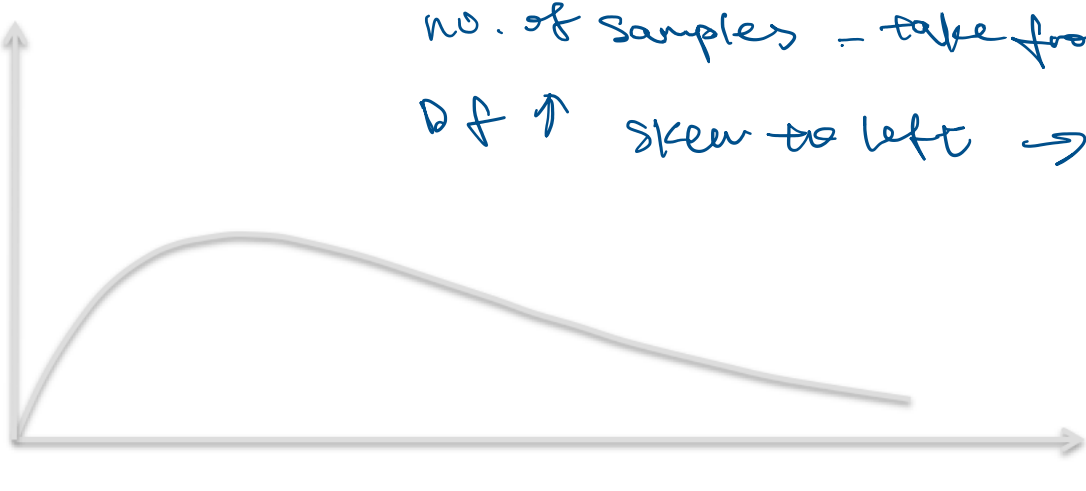
Week 4

INTRODUCTION TO DATA SCIENCE IN PYTHON



### Chi-Squared ( $\chi^2$ ) Distribution

- Left-skewed
- **Degrees of freedom = 4**  
only 1 parameter  
no. of samples - take from normal population  
Df  $\uparrow$  skew to left  $\rightarrow$  center




Probability Observation Occurs

Value of Observation

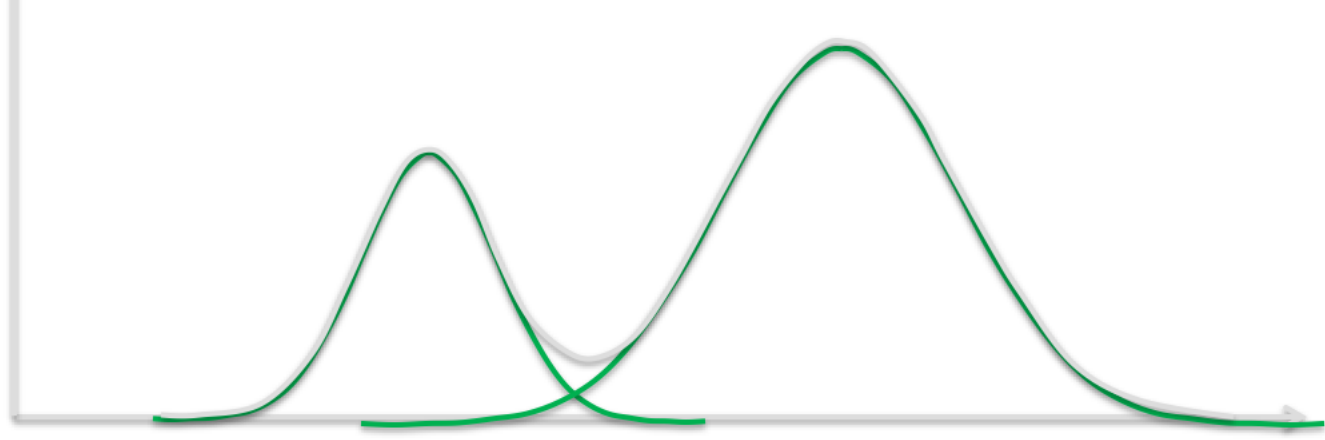
Week 4

INTRODUCTION TO DATA SCIENCE IN PYTHON



### Bimodal distributions

22 峰明  
a single high point / peak  
multiple peaks - Bimodal  
Model: 2 normal distribution with different parameters  
Gaussian Mixture Models - clustering data




Probability Observation Occurs

Value of Observation

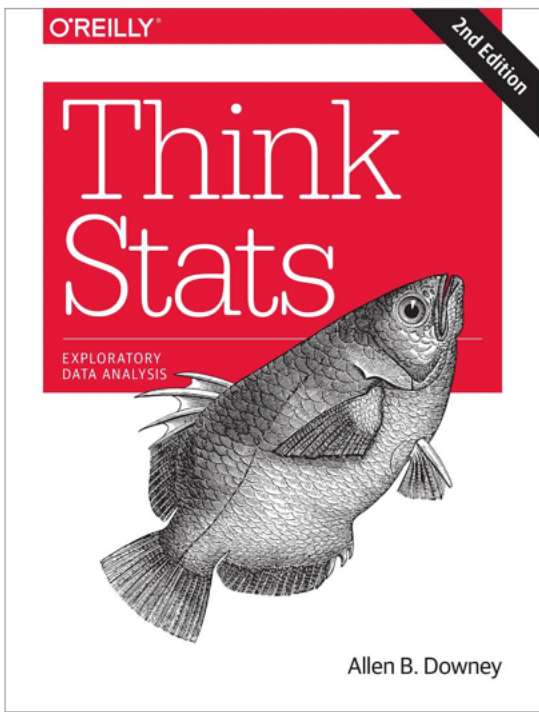
distribution  
shape - describe probability of a value being pulled  
When sample a population  
np.scipy - different distribution built-in  
 $\hookrightarrow$  sample from

Week 4

INTRODUCTION TO DATA SCIENCE IN PYTHON



### Think Stats




- Probability and Statistics for Programmers
  - Allen B. Downey
  - Available for free under CC license at:  
<http://greenteapress.com/thinkstats2/index.html>

A/B test

Week 4

INTRODUCTION TO DATA SCIENCE IN PYTHON




### Hypothesis Testing

- **Hypothesis:** A statement we can test
  - **Alternative hypothesis:** Our idea, e.g. there is a difference between groups
  - **Null hypothesis:** The alternative of our idea, e.g. there is no difference between groups
- **Critical Value alpha ( $\alpha$ )**  
Braining groups Null hypo  $\rightarrow$  no difference - reject null hypo  
Choose significance level accept alternative  
 $\hookrightarrow$  evidence against null hypo  
 $\hookrightarrow$  confident in alternative hypo  
– The threshold as to how much chance you are willing to accept  
– Typical values in social sciences are 0.1, 0.05, or 0.01  
interventions do - forced difference tolerance for prob 2% - 1%  
send email - low cost phone - high

Week 4

INTRODUCTION TO DATA SCIENCE IN PYTHON



### p-hacking

$\alpha = 0.05$  run more t-test  
expect positive result find positive result  
 $\hookrightarrow$  no. of test run  
 $\hookrightarrow$  to the chance  
5% risk  
• **P-hacking, or Dredging** spurious correlation  
– Doing many tests until you find one which is of statistical significance  
– At a confidence level of 0.05, we expect to find one positive result 1 time out of 20 tests  
– Remedies:

- Bonferroni correction 0.05 - 1 test  $\rightarrow$  3 test  $\frac{0.05}{3} = 0.01$
- Hold-out sets  $\frac{1}{n} \rightarrow$  cv.
- Investigation pre-registration  
outline what expect - why  
 $\hookrightarrow$  describe test - backup positive proof  
 $\hookrightarrow$  convince reviewer - experiment test fully a given hypo