CAP4770/5771
Introduction to Data Science
Fall 2016

University of Florida, CISE Department
Prof. Daisy Zhe Wang

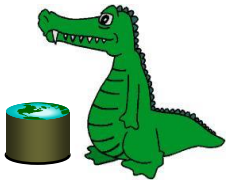
Based on notes from CS194 at UC Berkeley by Michael Franklin, John Canny, and Jeff Hammerbacher

Exploratory Data Analysis

Statistics of Data

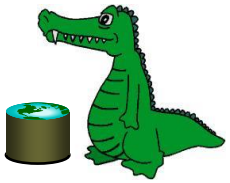
Hypothesis Testing

Data Visualization



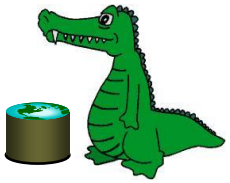
Logistics

- Readings:
  - Hypothesis testing DSS chapter 7 may need background from statistics and probability DSS chapters 5 & 6
- Lab 1 grades will be released this Friday
- Lecture covers core concepts and techniques
- Book chapters covers topics related to lectures
- Labs are hands-on session of the lectures
- Homework and quiz include lab (+ lecture)
- Midterm mainly covers lecture (+ book)



Review

- Data types and sources
- Data Models
- Data Preparation (Lab1)



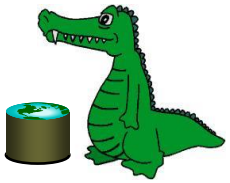
Outline

- Exploratory Data Analysis
  - Some basic statistics and distributions
    - Bernoulli, binomial, normal
  - Hypothesis Testing
    - A/B testing
  - Visualization: graphical representation of data
    - Chart Types

Statistical Notation

We'll use upper case symbols "X" to represent random variables, which you can think of as draws from the entire population.

Lower case symbols "x" represent particular samples of the population, and subscripted lower case symbols to represent instances of a sample: x_i

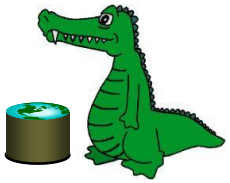


Random Variable

- A random variable is a variable whose realization is determined by chance according to a distribution
  - Rolling a die – what is the distribution? Biased vs. unbiased

- One can describe a random variable by its expected value. It is the sum of all possible realizations weighted by their probabilities

- The expected value is similar to an average, but with an important difference: the average is computed when you already have realizations, while the expected value is computed before you have realizations

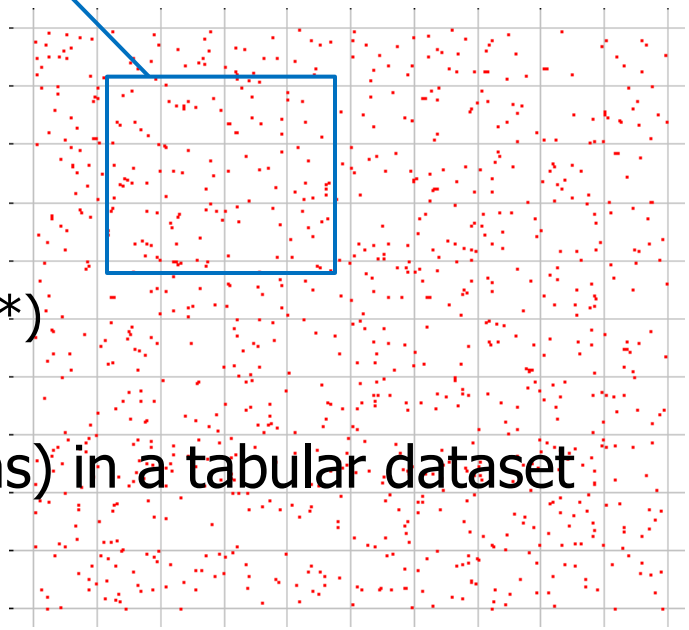


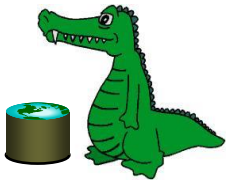
Measurement on Samples

- Many datasets are **samples** from an **infinite population**.
- **Measurement:** We often want to measure properties of data
- We are most interested in **measures on the population**, but we have access only to a **sample** of it.

A sample measurement is called a **"statistic".** Examples:

- Basic properties
  - Sample min, max, mean, range
  - Median, variance, std. deviation (outlier*)
  - Quartiles, modes, IQR
- Relationships: between fields (columns) in a tabular dataset
  - Co-variance, correlation





Mean, Variance, Standard Deviation

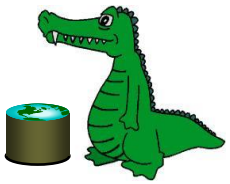
The mean is the arithmetic average of the observations.

The variance is a measure of spread of the individual observations from the sample mean:

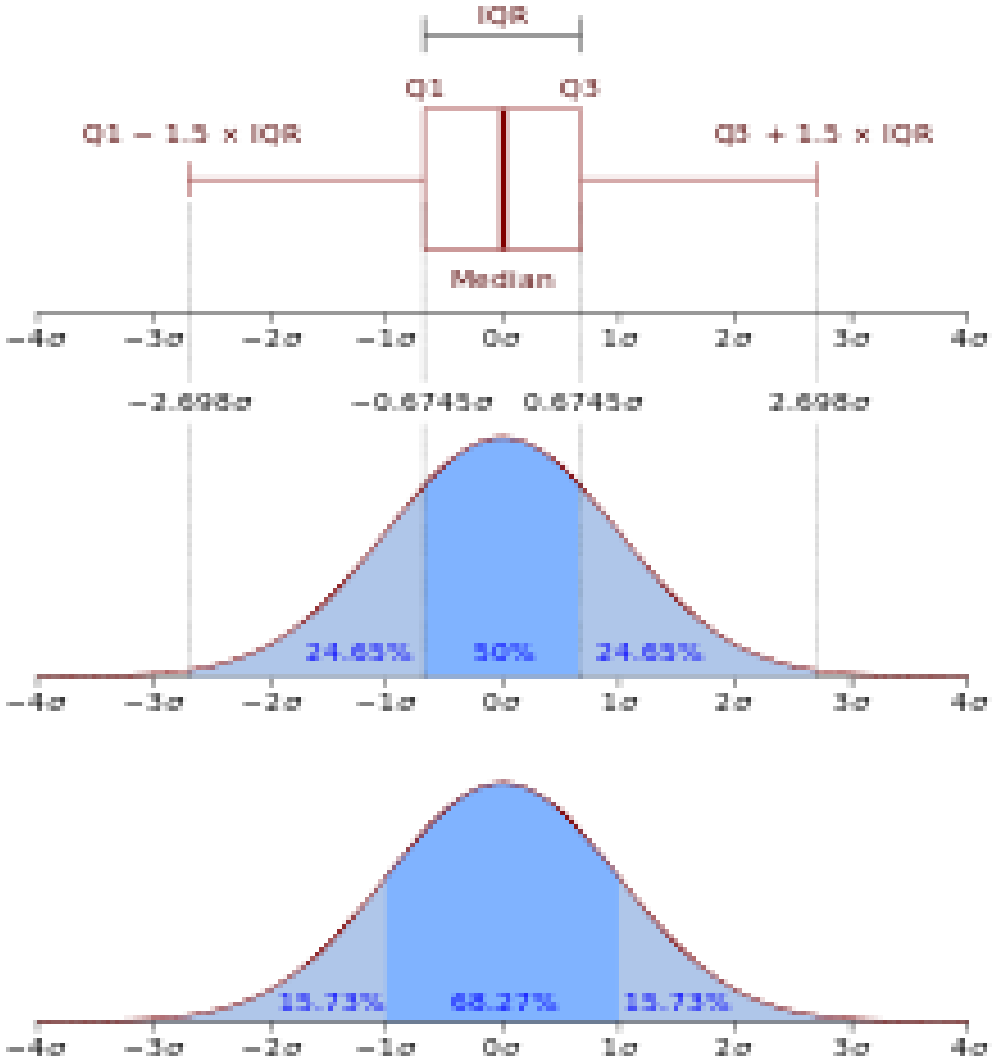
$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

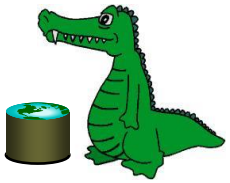
The standard deviation is the square root of variance.

The probability density function (pdf) covers an area representing the probability of realizations of the underlying values



Box-plot for 5 number summary





Covariance and correlation

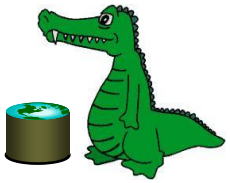
- Covariance: product of units, what is large CoVar?

$$CoVar(X,Y)$$

$$= 1/n \sum_{i=i}^{n} (x_i - E(X))(y_i - E(Y))$$

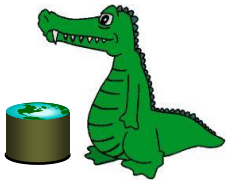
- Correlation: unitless, [-1,1]

$$Corr(X,Y) = \frac{CoVar(X,Y)}{Var(X) * Var(Y)}$$



Correlation Caveats

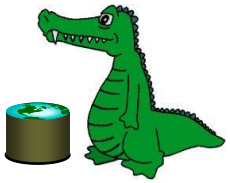
- Confounding Variables
  - Assumption of the dataset – "all else being equal"
  - A -> B -> C (Conditional Independence)
- Correlation does not always mean Causality
  - A -> B & A -> C (Conditional Independence)
  - A -> C & B -> C (Explain Away)



Probability Dependence, Independence and Bayes Rule

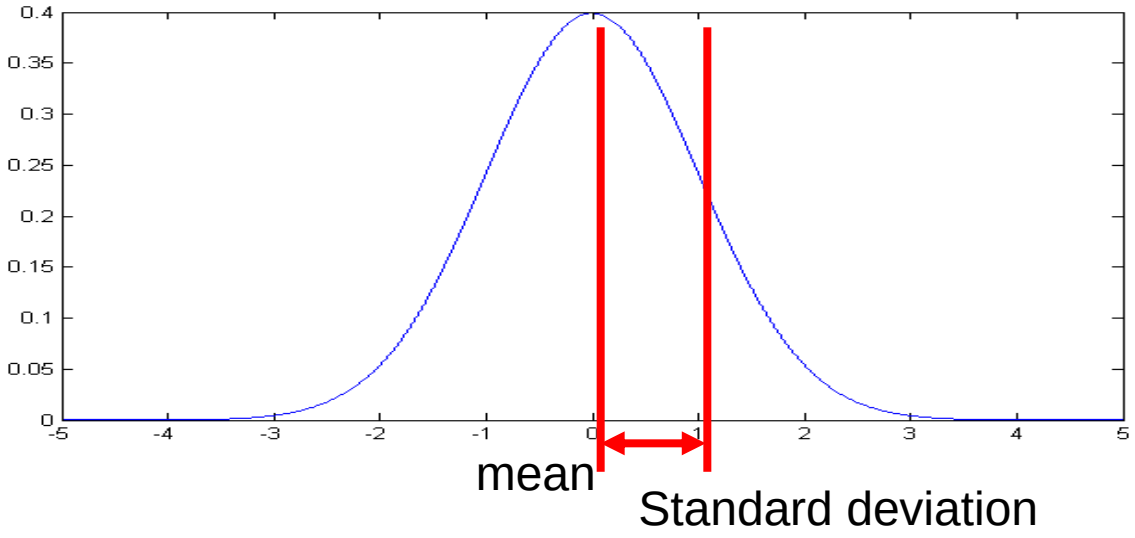
- P(E,F) = P(E|F)P(F)
- P(E,F) = P(E)P(F), if E ⫫ F
- P(A,B,C) = P(C|B)P(B|A)P(A), if ??
- P(A,B,C) = P(A)P(B|A)P(C|A), if ??
- P(A,B,C) = P(A)P(B)P(C|AB), if ??

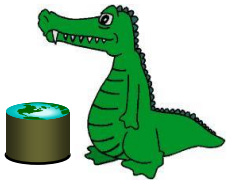
- Bayes Theorem: P(E|F) = P(F|E)P(E)/(P(F|E)P(E)+P(F|~E)P(~E))



Normal Distribution

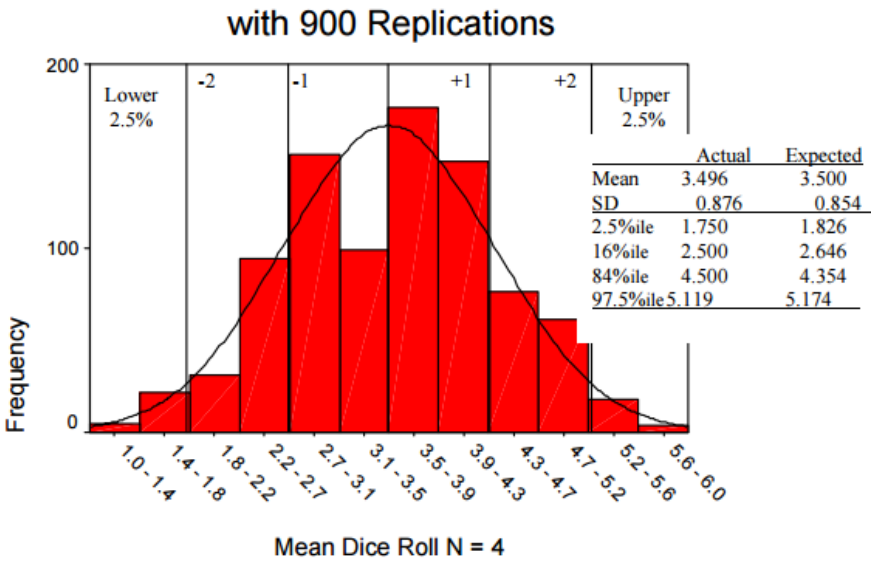
- Symmetric about the mean
- Mean, median and mode are the same
- Defined by mean (μ) and standard deviation (σ)
- Total area under the normal curve = 1 (pdf)
- Uniform Distribution (Continuous vs. Discrete Distribution)

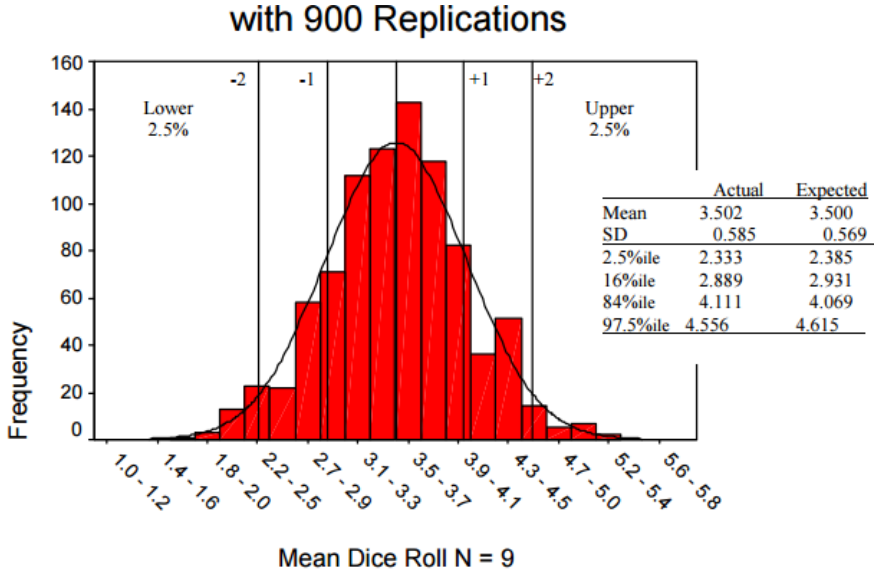




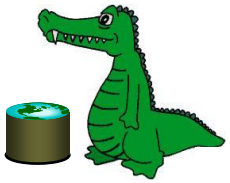
Central Limit Theorem

The distribution of the mean of a set of N identically-distributed random variables approaches a normal distribution as $N \rightarrow \infty$.

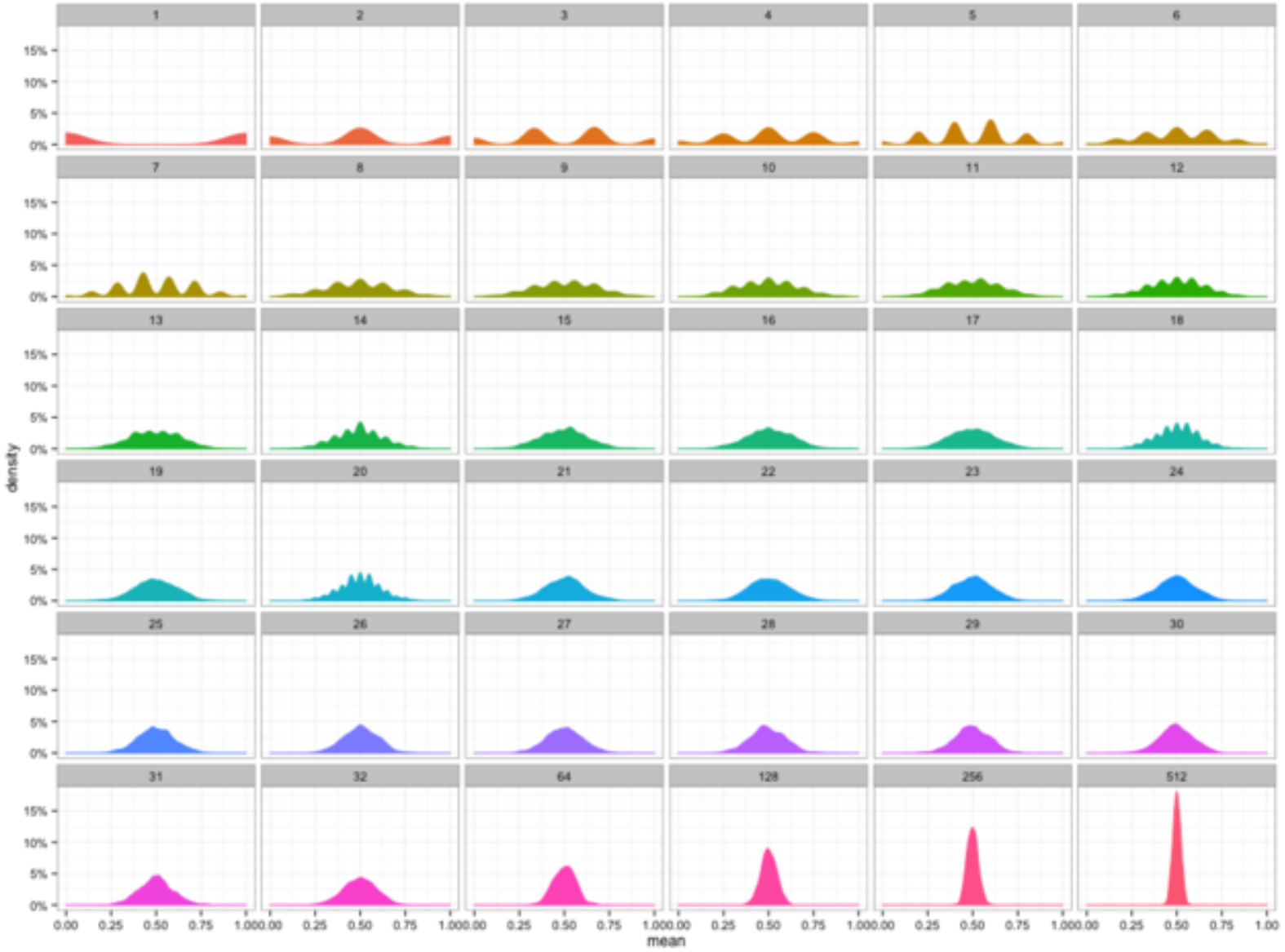




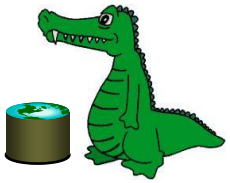
http://www.soph.uab.edu/ssg/files/People/MBeasley/Courses/CLT.pdf



Central Limit Theorem



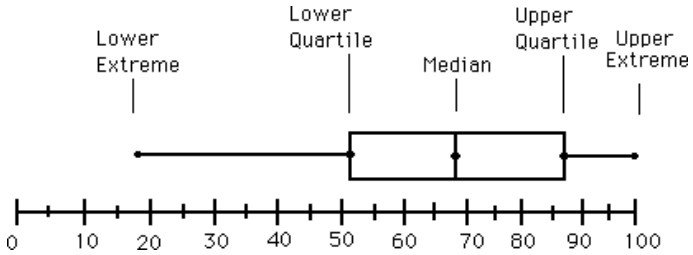
From the Wikipedia



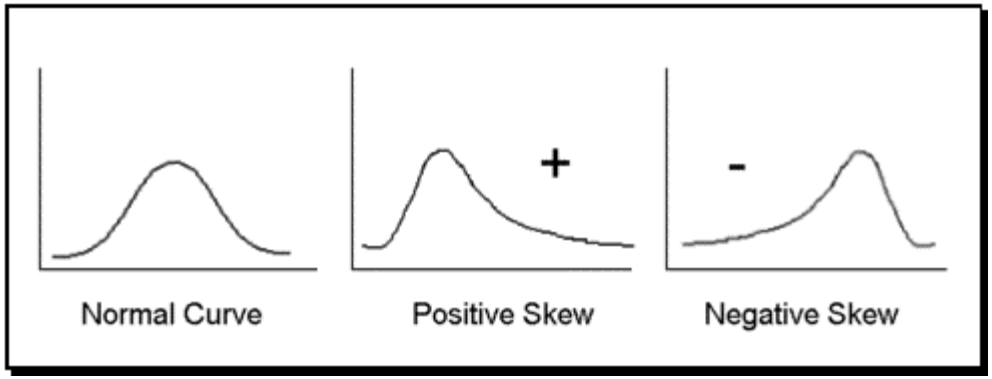
Normally distributed?

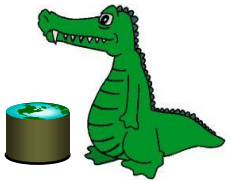
Many statistical tools, including mean and variance, t-test, ANOVA etc. assume **data** are **normally distributed**.

Very often this is not true. The box-and-whisker plot gives a good clue



Whenever its asymmetric, the data cannot be normal. The histogram gives even more information

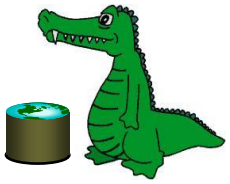




Normality Assumption and CLT

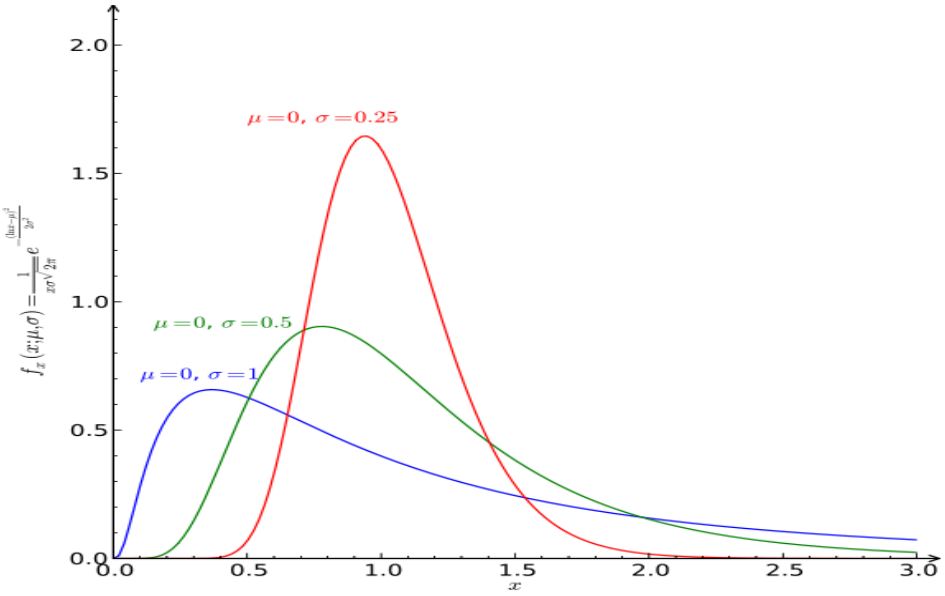
All parametric statistical tests (e.g. t-test and ANOVA) assume normally-distributed data, depend on sample mean and variance measures of the data.

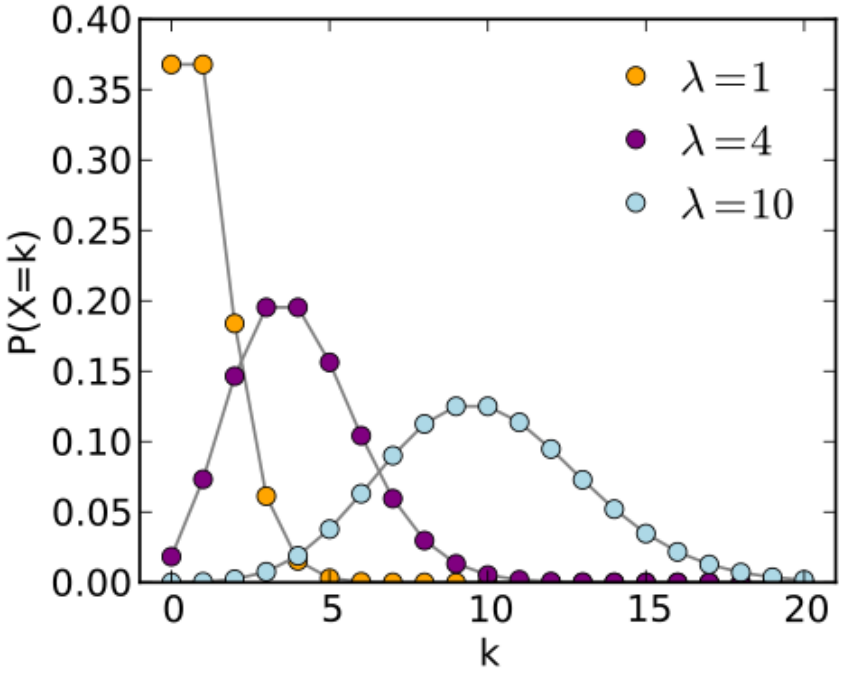
But, they work reasonably well for data that are not normally distributed as long as the samples are not too small (because of CLT).



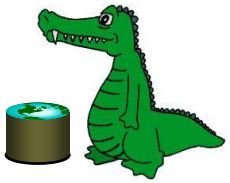
Correcting distributions

If X satisfies a log-normal distribution, $Y = \log(X)$ has a normal dist.





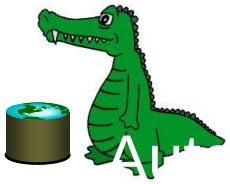
If X is a Poisson* with mean λ and sdev. $\sqrt{\lambda}$, then $\sqrt{X}$ is approximately normally distributed with sdev. 1 with $\lambda > 10$



Inference

By inference, we mean the research of the values of the parameters given some data

- **Estimation**: use the data to estimate the parameters
- **Hypothesis Testing**: guess a value for the parameters and ask the data whether this value is true



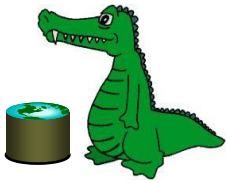
Rhine Paradox*

Joseph Rhine was a parapsychologist in the 1950's (founder of the *Journal of Parapsychology* and the *Parapsychological Society, an affiliate of the AAAS*).

He ran an experiment where subjects had to guess whether 10 hidden cards were red or blue.

He found that about 1 person in 1000 could guess the color of all 10 cards.

* Example from Jeff Ullman/Anand Rajaraman



Rhine Paradox

Q: Is this occurrence statistically significant?

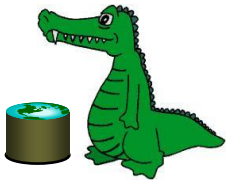
He called back the "psychic" subjects and had them do the same test again. They all failed.

He concluded that the act of telling psychics that they have psychic abilities causes them to lose it…(!)

Q: what's wrong with his conclusion?

Pitfalls: p-hacking (in book), data dredging (wikipedia)

- Low probability events would happen if huge # of hypothesis are test
- Incorrect testing procedure (hypothesis -> sample -> new hypothesis -> new sample)



Hypothesis Testing: Motivation

Suppose, for the past year, the mean of the monthly energy cost for families was $260 p.m.

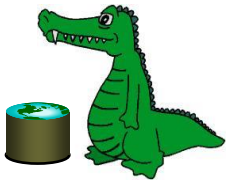
Determine whether the mean has changed, for the current year.

One solution: generate a **random sample** of 25 families and record energy costs for the current year.

Descriptive Statistics: Energy Cost

Variable	Total Count	Mean	SE Mean	StDev
Energy Cost	25	330.6	30.8	154.2

*http://blog.minitab.com

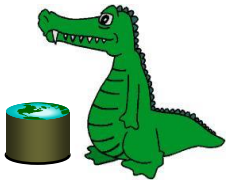


Hypothesis Testing: Motivation

Even though our *sample* mean is 330.6, the population mean could still be 260 due to **sampling error**.

Hypothesis testing can assess the likelihood of this possibility!

*http://blog.minitab.com

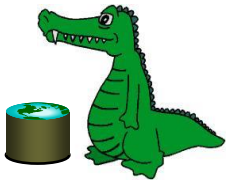


Null & Alternate Hypothesis

The **null hypothesis (H_0)**: The population mean (330.6) equals the hypothesized mean (260).

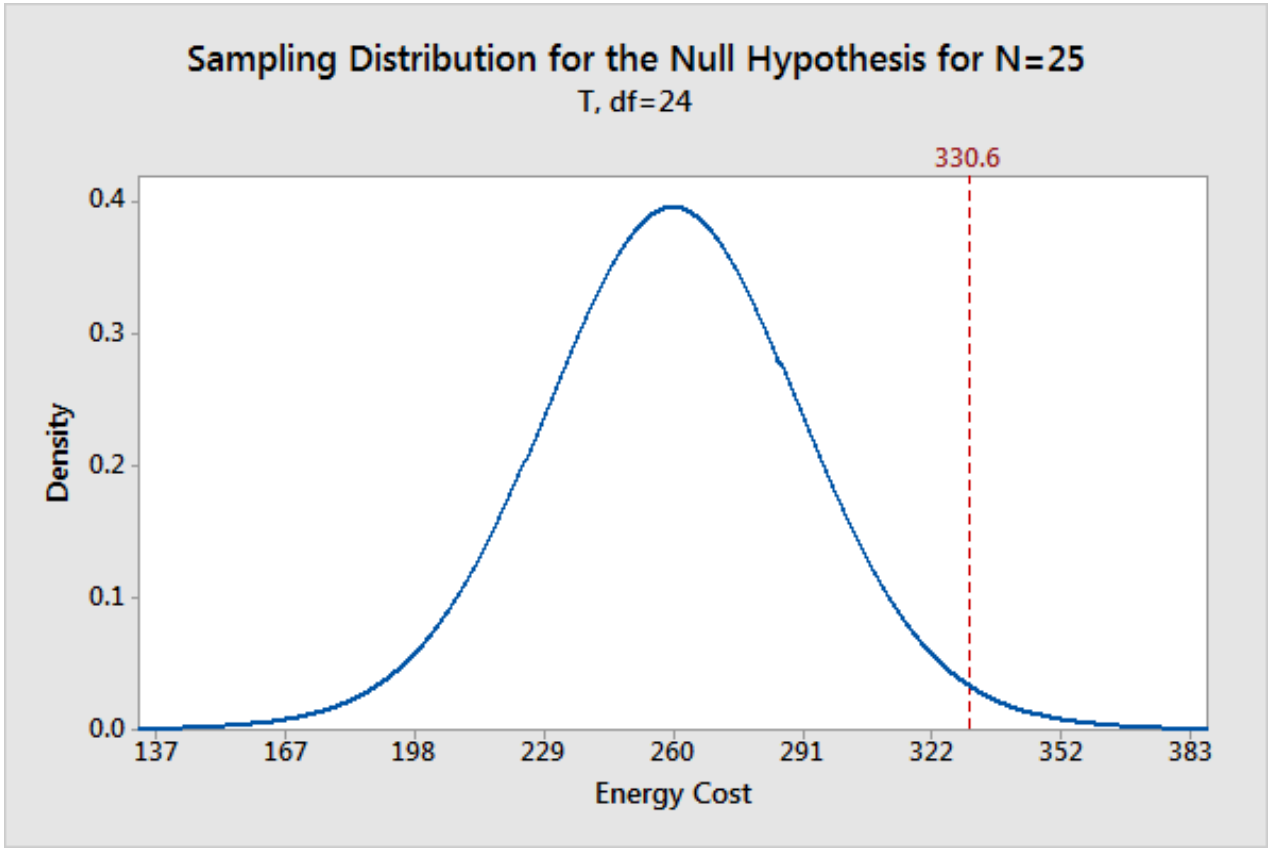
The **alternative hypothesis (H_a)**: The population mean (330.6) differs from the hypothesized mean (260).

A **sampling distribution** is the distribution of a **statistic**, such as the mean, that is obtained by repeatedly drawing a large number of samples from a specific population.

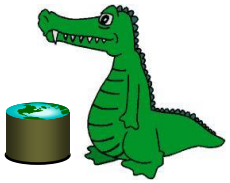


Sampling Distribution

Goal is to determine whether our sample mean (330.6) is significantly different from the null hypothesis mean (260)



*http://blog.minitab.com



Test Statistic

- We want to prove a hypothesis H_A, but its hard so we try to **disprove a null hypothesis H_0**.
- A **test statistic** is some measurement we can make on the data which is likely to be **big under H_A** but **small under H_0**.
- We chose a test statistic whose distribution we know if H_0 is true: e.g.
  - Two samples a and b, normally distributed, from A and B.
  - H_0 hypothesis that mean(A) = mean(B), test statistic is: s = mean(a) – mean(b).
  - s has mean zero and is normally distributed under H_0.
  - But it is "large" if the two means are different.