Probability and Statistics

Willard Williamson

Adjunct Professor

iSchool, Syracuse University

linkedin.com/in/ncc-1701-d

Credits

- Thanks to the openintro.org project for providing lecture slides
- Thanks to St. Francis Xavier University for lecture slides on the uniform distribution: http://people.stfx.ca/jquinn/STAT%20231/Class%20Power%20points/Lecture%209%20Uniform%20and%20normal%20distributions.ppt
- A nice review on joint, marginal, and conditional probability: https://sites.nicholas.duke.edu/statsreview/jmc/

Objectives

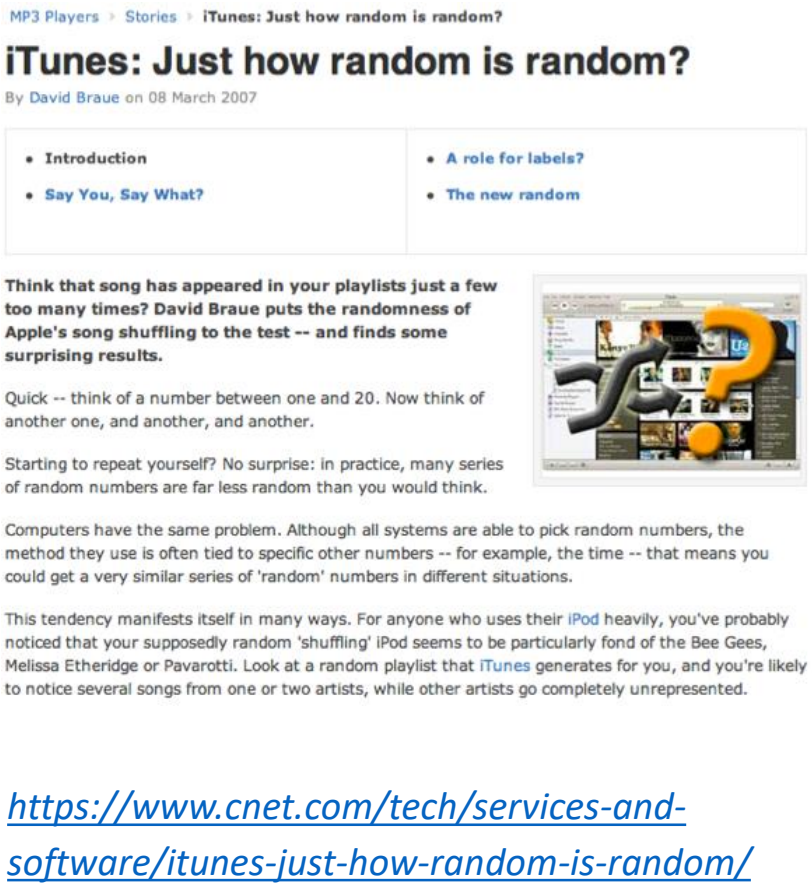
- Review of common probability concepts
- Not an exhaustive survey
- Highly recommended to take a probability and statistics class
- Probability and statistics is very important to data scientists

What is Probability

- Probability is a measure of uncertainty
- We can try to compute the probability of almost any event
- An event is the basic element to which probability can be applied
- Events:
  - An observation or the result of an experiment like heads vs. tails
  - A description of a potential outcome

Random processes

- A *random process* is a situation in which we know what outcomes could happen, but we don't know which particular outcome will happen.
- Examples: coin tosses, die rolls, iTunes shuffle, whether the stock market goes up or down tomorrow, etc.
- It can be helpful to model a process as random even if it is not truly random.



MP3 Players › Stories › iTunes: Just how random is random?

iTunes: Just how random is random?

By David Braue on 08 March 2007

- Introduction
- Say You, Say What?
- A role for labels?
- The new random

Think that song has appeared in your playlists just a few too many times? David Braue puts the randomness of Apple's song shuffling to the test -- and finds some surprising results.

Quick -- think of a number between one and 20. Now think of another one, and another, and another.

Starting to repeat yourself? No surprise: in practice, many series of random numbers are far less random than you would think.

Computers have the same problem. Although all systems are able to pick random numbers, the method they use is often tied to specific other numbers -- for example, the time -- that means you could get a very similar series of 'random' numbers in different situations.

This tendency manifests itself in many ways. For anyone who uses their iPod heavily, you've probably noticed that your supposedly random 'shuffling' iPod seems to be particularly fond of the Bee Gees, Melissa Etheridge or Pavarotti. Look at a random playlist that iTunes generates for you, and you're likely to notice several songs from one or two artists, while other artists go completely unrepresented.

https://www.cnet.com/tech/services-and-software/itunes-just-how-random-is-random/

Probability

There are several possible interpretations of probability but they (almost) completely agree on the mathematical rules probability must follow.

- P(A) = Probability of event A
- $0 \leq P(A) \leq 1$

Two Views on Statistical Learning

- Bayesian statistics
- Frequentist Statistics

Bayesian

- A **Bayesian** interprets probability as a subjective degree of belief
- Provides an equation to manipulate conditional probabilities
- Uses *prior* knowledge / belief about the phenomenon and then combines observed evidence with the prior knowledge / belief using the equation below
- $P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{likelihood * prior}{evidence}$
  - P(A): The prior knowledge or belief
  - P(B): The posterior knowledge or evidence gained from additional observations
  - P(A|B): The posterior degree of belief after having observed B
  - P(B|A): Likelihood: The likelihood of A conditioned on B
  - The quotient $\frac{P(B|A)P(A)}{P(B)}$ represents the support B provides for A
  - Example: A = temperature at noon tomorrow. B = temperature at noon today.

Frequentist

- **Frequentist statistics**:
  - The probability of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.
  - Makes an estimate of a parameter based on sample data.
  - It assumes a procedure where samples from the same population will be observed an infinite number of times.
  - It uses sampling with replacement
  - The samples build the distribution
  - We will use the frequentist interpretation in IST-718

Disjoint and non-disjoint outcomes

Disjoint (mutually exclusive) outcomes: Cannot happen at the same time.

- The outcome of a single coin toss cannot be a head and a tail.
- A student both cannot fail and pass a class.
- A single card drawn from a deck cannot be an ace and a queen.

Disjoint and non-disjoint outcomes

Disjoint (mutually exclusive) outcomes: Cannot happen at the same time.

- The outcome of a single coin toss cannot be a head and a tail.
- A student both cannot fail and pass a class.
- A single card drawn from a deck cannot be an ace and a queen.

Non-disjoint outcomes: Can happen at the same time.

- A student can get an A in Stats and A in Econ in the same semester.

Probability Distributions

Probability distributions

A *probability distribution* lists all possible events and the probabilities with which they occur.

- The probability distribution for the gender of one kid:

Event	Male	Female
Probability	0.5	0.5

Probability distributions

A *probability distribution* lists all possible events and the probabilities with which they occur.

- The probability distribution for the gender of one kid:

Event	Male	Female
Probability	0.5	0.5

- Rules for probability distributions:

  1. The events listed must be disjoint
  2. Each probability must be between 0 and 1
  3. The probabilities must total 1

Random Variables

Random variables

A *random variable* is a numeric quantity whose value depends on the outcome of a random event

- We use a capital letter, like X, to denote a random variable
- The values of a random variable are denoted with a lowercase letter, in this case x
- For example, $P(X = x)$

There are two types of random variables:

- *Discrete random variables* often take only integer values
    - Example: Number of credit hours, Difference in number of credit hours this term vs last
- *Continuous random variables* take real (decimal) values
    - Example: Cost of books this term, Difference in cost of books this term vs last

Expectation

- We are often interested in the average outcome of a random variable.
- We call this the *expected value* (mean), and it is a weighted average of the possible outcomes

$$\mu = E(X) = \sum_{i=1}^{k} x_i P(X = x_i)$$

Expected value of a discrete random variable

In a game of cards you win $1 if you draw a heart, $5 if you draw an ace (including the ace of hearts), $10 if you draw the king of spades, and nothing for any other card you draw. Write the probability model for your winnings, and calculate your expected winning.

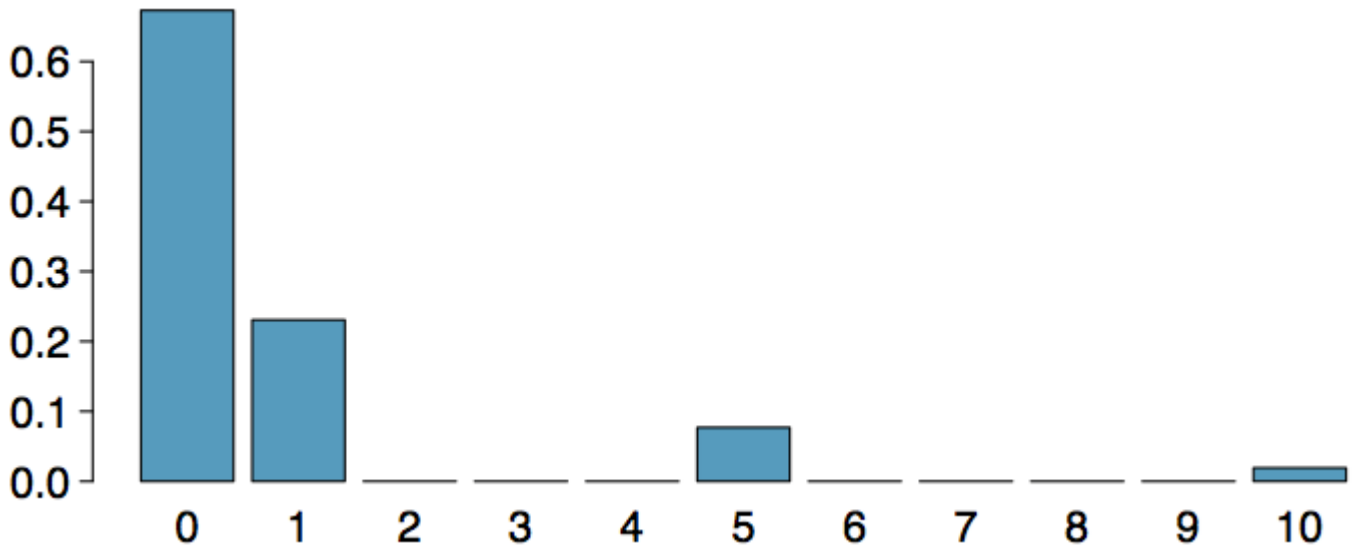
Expected value of a discrete random variable

In a game of cards you win $1 if you draw a heart, $5 if you draw an ace (including the ace of hearts), $10 if you draw the king of spades and nothing for any other card you draw. Write the probability model for your winnings, and calculate your expected winning.

Event	X	$P(X)$	$X\,P(X)$
Heart (not ace)	1	$\frac{12}{52}$	$\frac{12}{52}$
Ace	5	$\frac{4}{52}$	$\frac{20}{52}$
King of spades	10	$\frac{1}{52}$	$\frac{10}{52}$
All else	0	$\frac{35}{52}$	0
Total			$E(X) = \frac{42}{52} \approx 0.81$

Expected value of a discrete random variable (cont.)

Below is a visual representation of the probability distribution of winnings from this game:



Variance

Variance is roughly the average squared deviation from the mean.

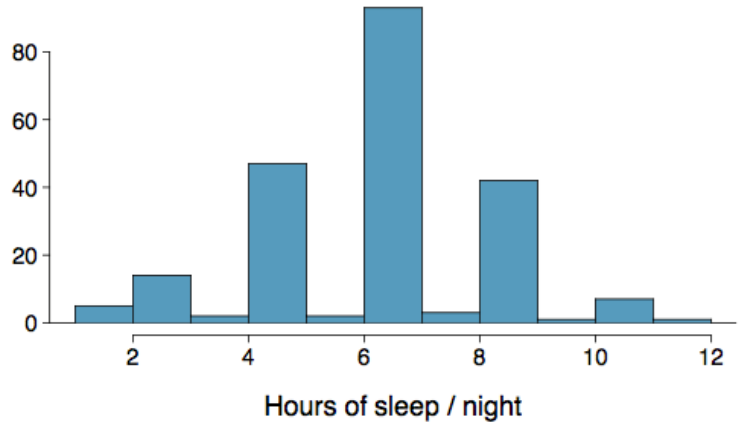
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Variance

Variance is roughly the average squared deviation from the mean.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- Calculate the variance for student sleep

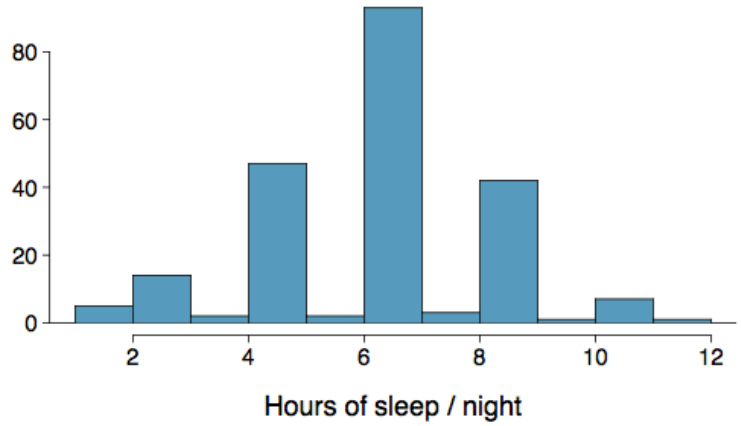


Variance

Variance is roughly the average squared deviation from the mean.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- Calculate the variance for student sleep
- The sample mean is $\bar{x} = 6.71$, and the sample size is n = 217.
- The variance of amount of sleep students get per night can be calculated as:



$$s^2 = \frac{(5-6.71)^2 + (9-6.71)^2 + \cdots + (7-6.71)^2}{217-1} = 4.11\ hours^2$$

Variance (cont.)

Why do we use the squared deviation in the calculation of variance?

Variance (cont.)

Why do we use the squared deviation in the calculation of variance?

- To get rid of negatives so that observations equally distant from the mean are weighed equally.
- To weigh larger deviations more heavily.

Standard Deviation

The *standard deviation* is the square root of the variance, and has the same units as the data.

$$s = \sqrt{s^2}$$

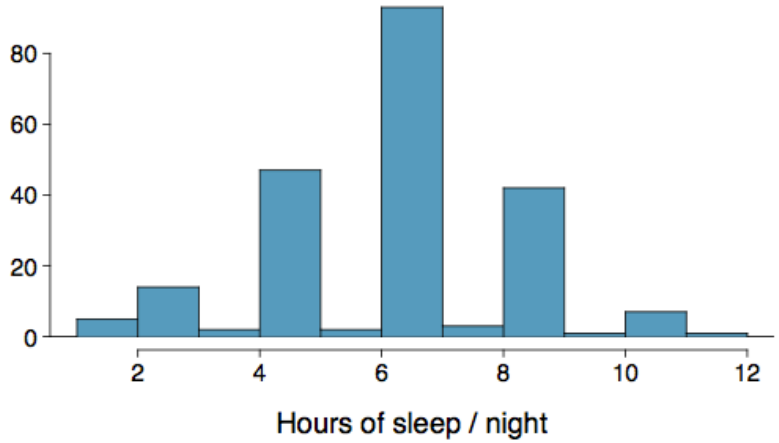
Standard Deviation

The *standard deviation* is the square root of the variance, and has the same units as the data.

$$s = \sqrt{s^2}$$

- The standard deviation of amount of sleep students get per night can be calculated as:

$$s = \sqrt{4.11} = 2.03 \; hours$$



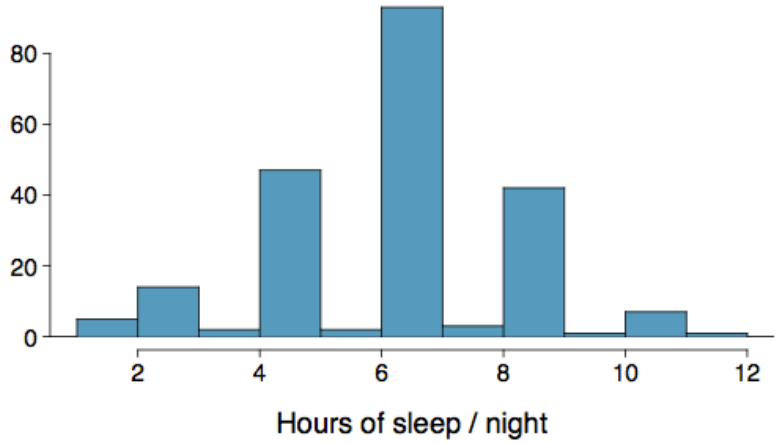
Standard Deviation

The *standard deviation* is the square root of the variance, and has the same units as the data.

$$s = \sqrt{s^2}$$

- The standard deviation of amount of sleep students get per night can be calculated as:

$$s = \sqrt{4.11} = 2.03 \; hours$$

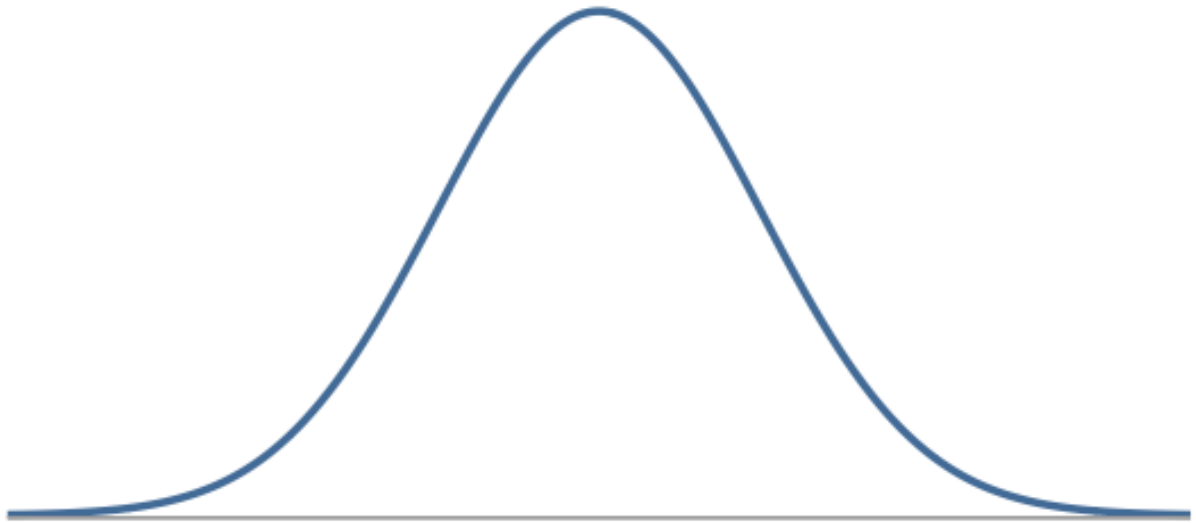


- We can see that all of the data are within 3 standard deviations of the mean.

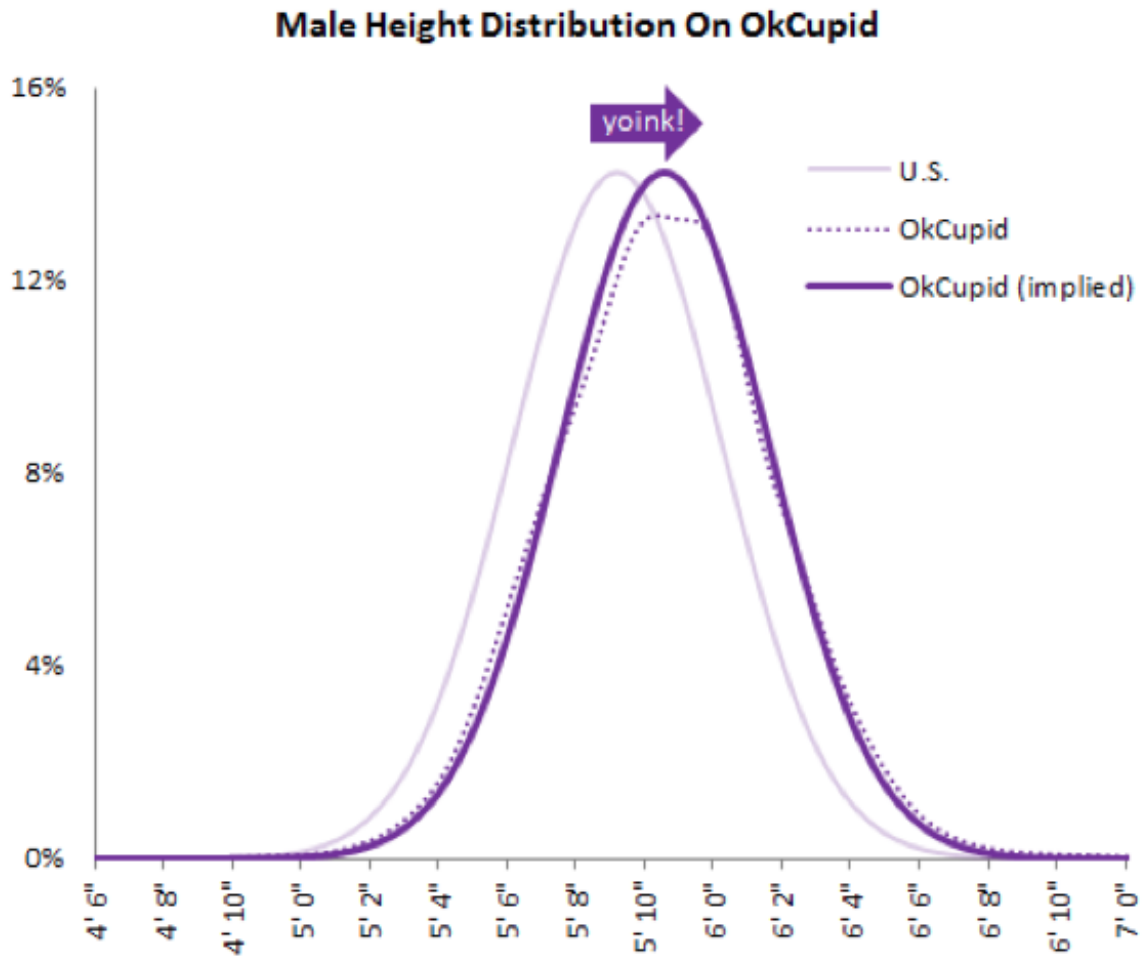
Normal distribution

Normal Distribution

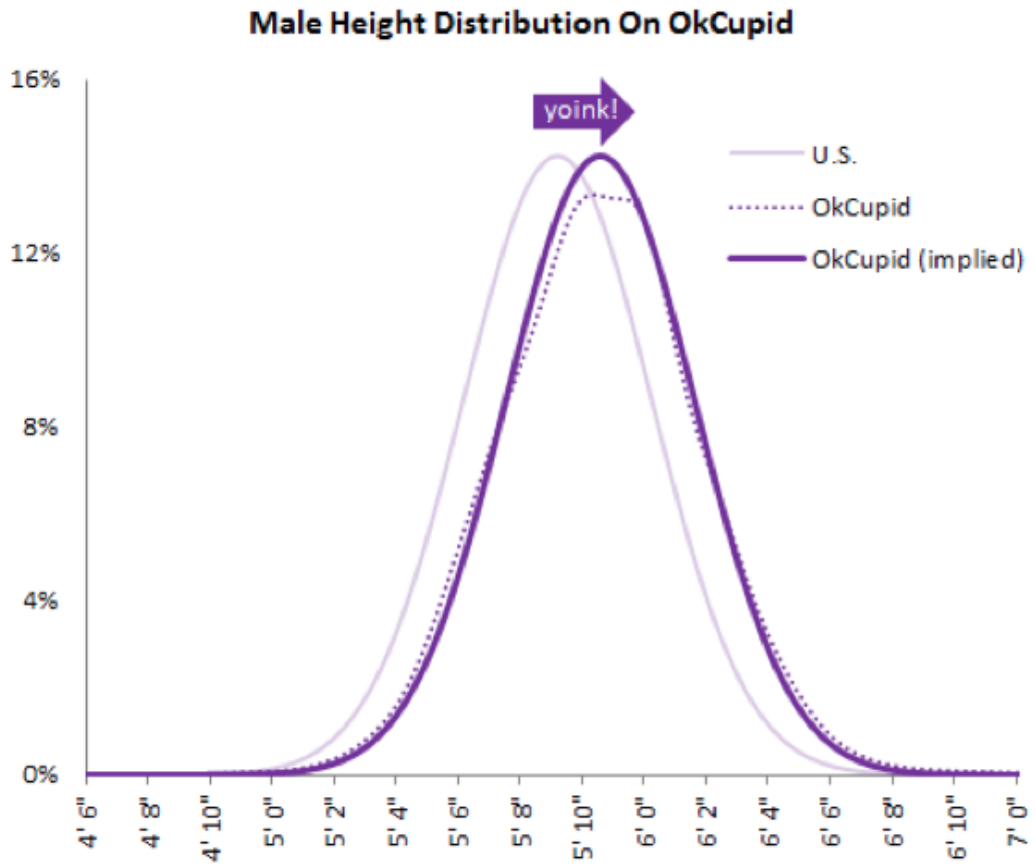
- Unimodal and symmetric, bell shaped curve
- Many variables are nearly normal, but none are exactly normal
- Denoted as $N(\mu, \sigma)$ → Normal with mean μ and standard deviation σ



Heights of males



Heights of males

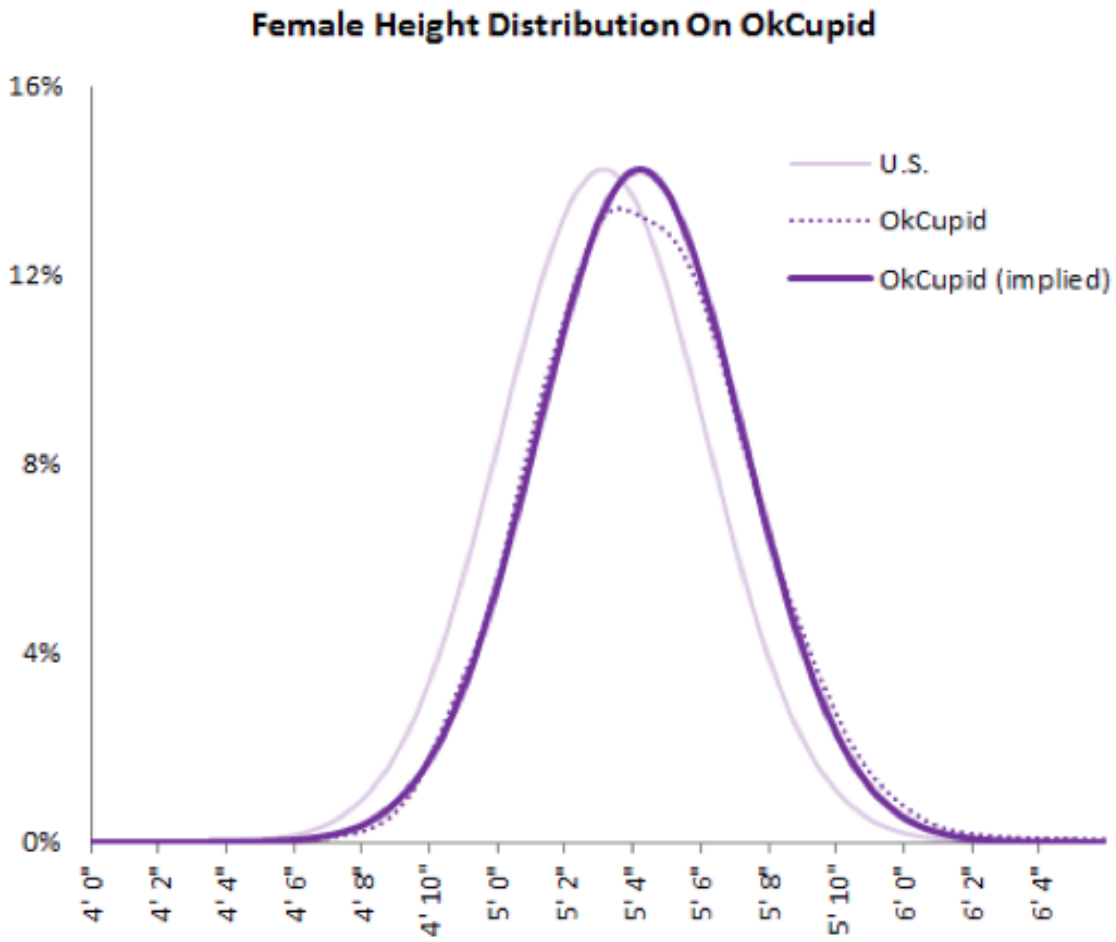


"The male heights on OkCupid very nearly follow the expected normal distribution -- except the whole thing is shifted to the right of where it should be. Almost universally guys like to add a couple inches."

"You can also see a more subtle vanity at work: starting at roughly 5' 8", the top of the dotted curve tilts even further rightward. This means that guys as they get closer to six feet round up a bit more than usual, stretching for that coveted psychological benchmark."

http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating

Heights of females



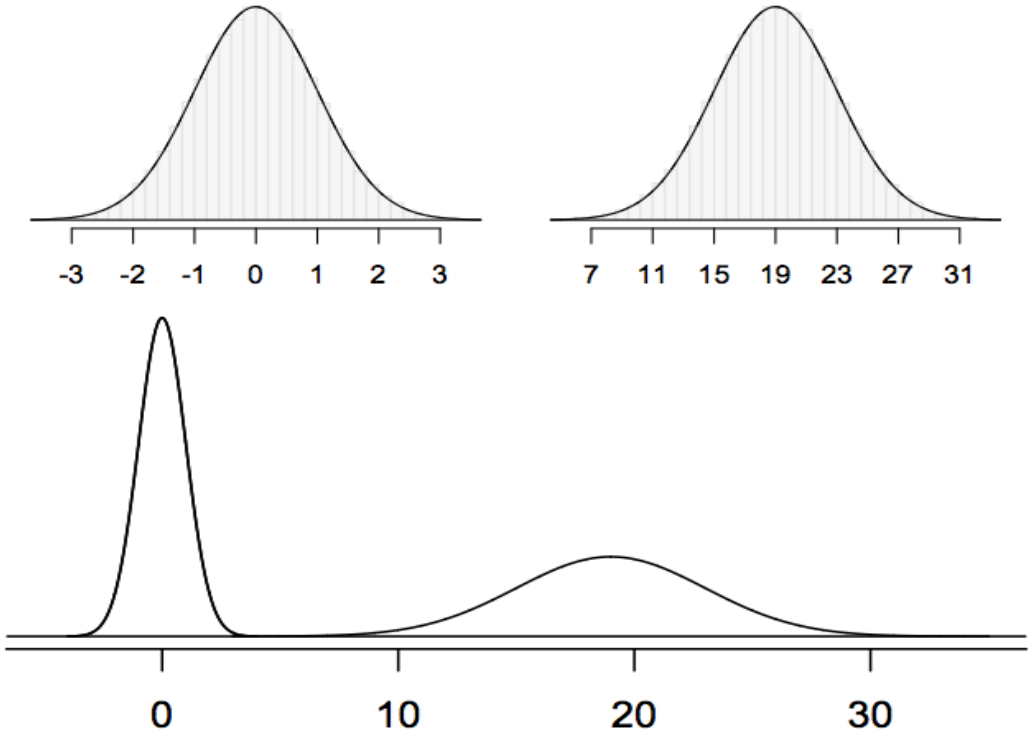
http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating

"When we looked into the data for women, we were surprised to see height exaggeration was just as widespread, though without the lurch towards a benchmark height."

Normal distributions with different parameters

μ: mean, σ: standard deviation

$N(\mu = 0, \sigma = 1)$ $N(\mu = 19, \sigma = 4)$

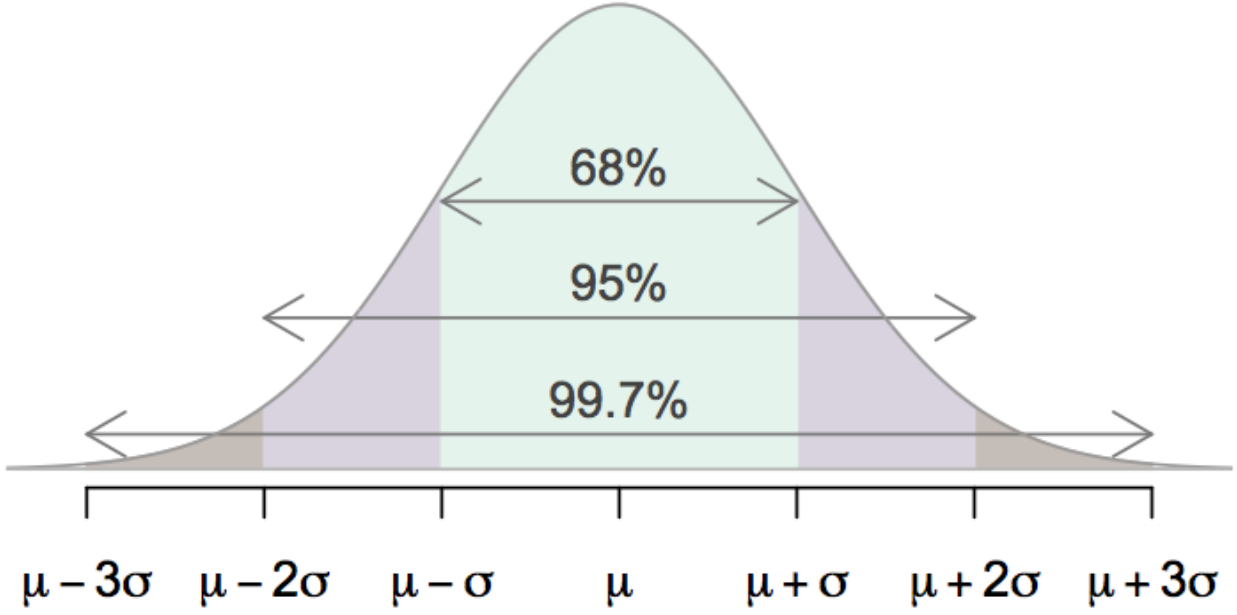


68-95-99.7 Rule

For nearly normally distributed data,

- about 68% falls within 1 SD of the mean,
- about 95% falls within 2 SD of the mean,
- about 99.7% falls within 3 SD of the mean.

It is possible for observations to fall 4, 5, or more standard deviations away from the mean, but these occurrences are very rare if the data are nearly normal.



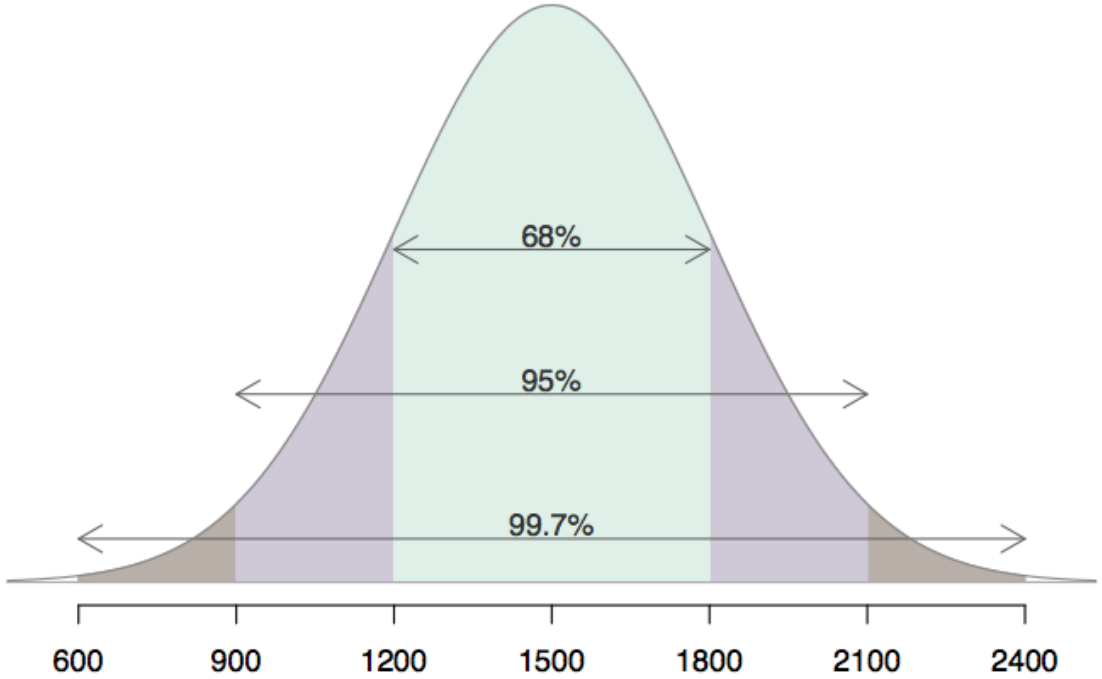
Describing variability using the 68-95-99.7 Rule

SAT scores are distributed nearly normally with mean 1500 and standard deviation 300.

Describing variability using the 68-95-99.7 Rule

SAT scores are distributed nearly normally with mean 1500 and standard deviation 300.

- ~68% of students score between 1200 and 1800 on the SAT.
- ~95% of students score between 900 and 2100 on the SAT.
- ~$99.7% of students score between 600 and 2400 on the SAT.



Uniform Distribution

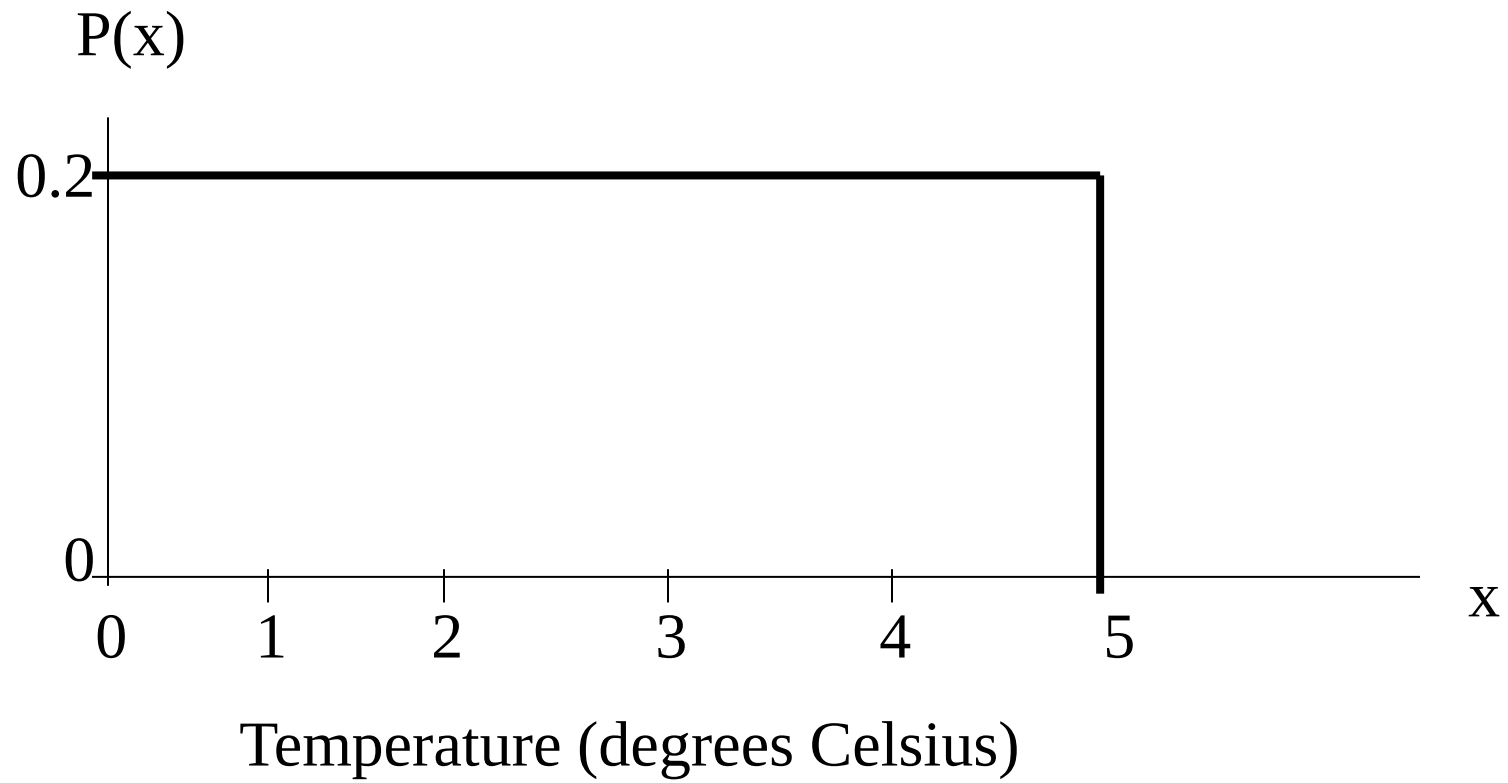
Uniform Distribution

A **Uniform Distribution** has equally likely values over the range of possible outcomes.

A graph of the uniform probability distribution is a rectangle with area equal to 1.

Example

The figure below depicts the probability distribution for temperatures in a manufacturing process. The temperatures are controlled so that they range between 0 and 5 degrees Celsius, and every possible temperature is equally likely.



General Uniform Distribution

A **Uniform Distribution** has equally likely values over the range of possible outcomes, say c to d (x axis ranges from c to d).

$$\text{Height of the density function}: f(x) = \frac{1}{d-c}$$

$$\text{Mean} = \mu = \frac{c+d}{2}$$

$$\text{Standard Deviation} = \sigma = \frac{d-c}{\sqrt{12}}$$

Sampling

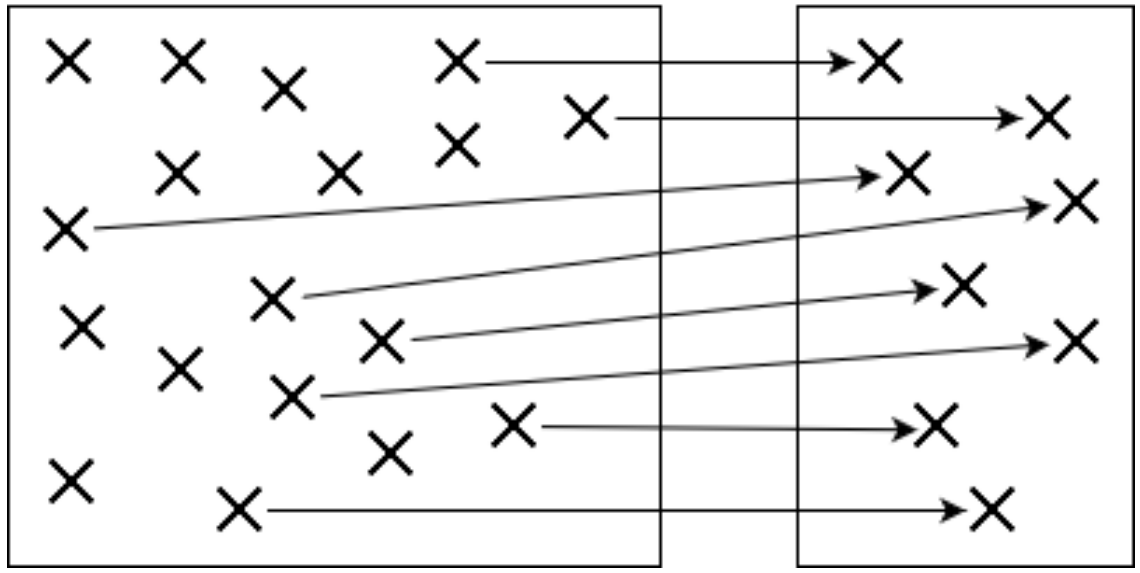
Population

- **Population**: A population is an entire collection of objects or individuals about which information is desired.
- Populations are often hard to measure exactly.
- If we want to measure the percentage of Trump voters in the United States, the population is all eligible voters in the United States.
- Some populations are known in their entirety.
- If we want to know the percentage of female employees at Lockheed in Liverpool, the population is all Lockheed Liverpool employees

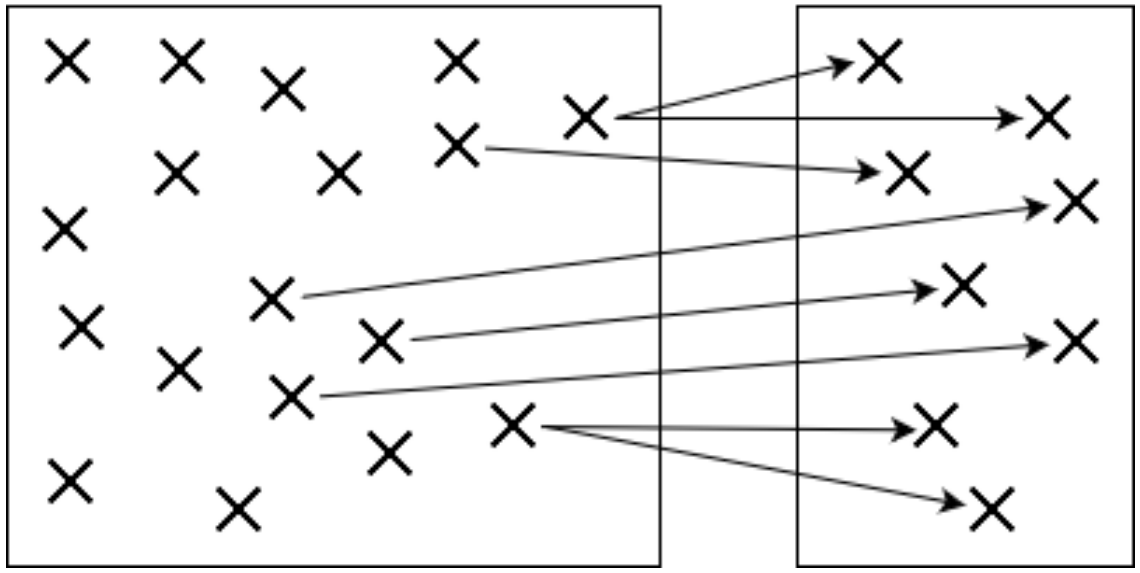
Sampling

- What do we do if the population is too large to work with: Like the United states population?
- **Sample**: A sample is a subset of a population.
  - Ex: A sample of 1000 prospective voters for the Trump vs. Clinton election
  - Samples can be drawn from a population with or without replacement.
  - Sampling with or without replacement depends on context

Sampling Without Replacement



Sampling With Replacement



Sampling Continued …

- The process of sampling is a natural process, we do it in our everyday lives …
- If you walked into an unfamiliar store and wanted to determine if the store was expensive or not, you would not check every price in the store; but rather, you would check the prices of a variety of items.
- If you are cooking a batch of spaghetti sauce and wanted to check the quality, you would taste a small sample instead of the entire batch.

Degrees of Freedom

- Degrees of freedom indicate the number of independent values that can vary when estimating a population parameter.
- Example:
  - Say you are trying to estimate a fixed population mean
  - You collect a sample containing 10 independent observations and calculate the mean
  - $\mu = \sum_{i=1}^{10} x_i$ / 10
  - Rearrange the above formula: $\mu * 10 = \sum_{i=1}^{10} x_i$ (or mean * 10 = sample sum)
  - Only 9 out of 10 values can vary; the 10th sample value must be fixed
  - When estimating population parameters (like mean), the degrees of freedom = sample size – number of parameter estimates (10 – 1 in this example)

Bessel's Correction

- Bessel was a French Mathematician
- Bessel's Correction uses the notion of degrees of freedom to improve the estimate of population variance when estimated from sample data.
- Math Fact: Sample variance is always less than or equal to population variance (assuming sampling with replacement).
- Bessel's correction: Subtracting 1 from the number of sample observations in the variance formula produces a better estimate of the population variance.
- See: https://en.wikipedia.org/wiki/Bessel%27s_correction

Covariance

- Covariance measures the linear relationship between 2 random variables.
- Positive covariance means as X increases, Y also increases
- Negative covariance means as X increases, Y decreases
- Does have units, ranges from + infinity to - infinity
- Not normalized so it’s hard to tell the degree of co-variation based on the resulting number.
- The sign is what matters when interpreting covariance

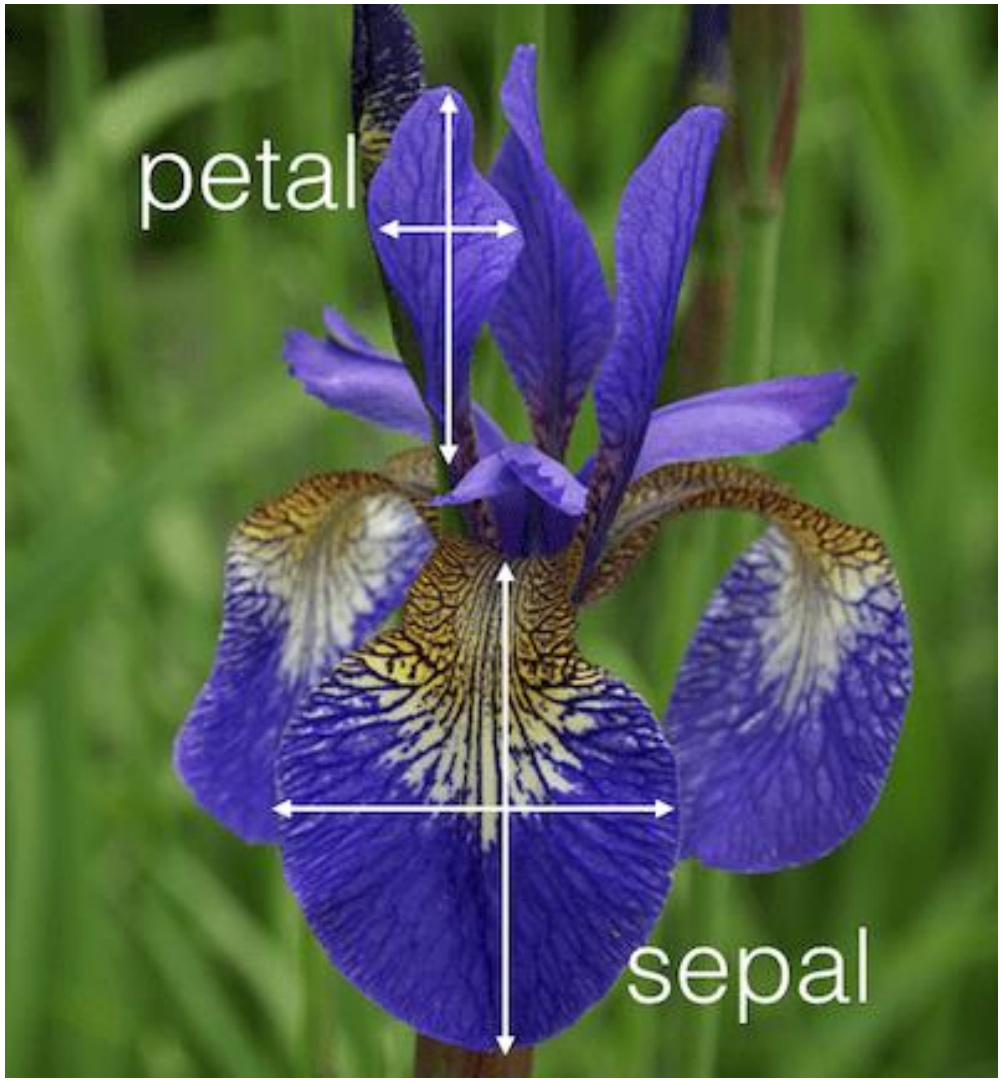
Covariance Continued …

- $COV(X,Y) = \frac{1}{n-ddof}(x_i - E(x))(y_i - E(y))$
- **numpy.cov(*m*, *y=None*, *rowvar=True*, *bias=False*, *ddof=None*, *fweights=None*, *aweights=None*)**
- ddof ==1, E(x) and E(y) are sample means
- ddof == 0, E(x) and E(y) are population means
- NumPy only has a function to calculate a covariance matrix, will not directly calculate covariance between 2 variables
- Note: Default is sample covariance (ddof == 1)
- Assumes that features in rows and observations in columns (rowvar == True)

Iris Data Covariance Matrix Example

- Iris Dataset comes from Statistician Ronald Fisher in 1936
- Covariance matrix for Iris data set.
- Cols: Sepal Length, Sepal Width, Petal Length and Petal Width
- Rows: Sepal Length, Sepal Width, Petal Length and Petal Width
- Diagonal represents variance. For example, matrix[0][0] is sepal length variance

	0	1	2	3
0	0.685694	-0.0392685	1.27368	0.516904
1	-0.0392685	0.188004	-0.321713	-0.117981
2	1.27368	-0.321713	3.11318	1.29639
3	0.516904	-0.117981	1.29639	0.582414



Correlation Coefficient

- Correlation measures the linear relationship between 2 random variables.
- Dimensionless (no units)
- Ranges from -1 to 1 where the absolute correlation suggests the degree of correlation.
- Size and sign matters
- +1 is perfect positive correlation, -1 is perfect negative correlation, 0 is completely uncorrelated.

Correlation Continued

- Correlation is essentially a normalized covariance.
- $r = \frac{COV(X,Y)}{std_x * std_y}$
- std_x and std_y are the standard deviations of X and Y respectively
- **numpy.corrcoef(*x, y=None, rowvar=True, bias=<no value>, ddof=<no value>*)**
- Assumes features are in rows and observations are in columns.
- r = correlation coefficient
- Note: ddof is ignored in this command

Iris Data Correlation Matrix Example

- Correlation matrix for Iris dataset
- Cols: Sepal Length, Sepal Width, Petal Length and Petal Width
- Rows: Sepal Length, Sepal Width, Petal Length and Petal Width
- Assumes that features in rows and observations in columns (rowvar == True)

	0	1	2	3
0	1	-0.109369	0.871754	0.817954
1	-0.109369	1	-0.420516	-0.356544
2	0.871754	-0.420516	1	0.962757
3	0.817954	-0.356544	0.962757	1

Correlation Rules of Thumb

Correlation Magnitude	Interpretation
0.00 – 0.20	Very Weak
0.20 – 0.40	Weak to Moderate
0.40 – 0.60	Medium to Substantial
0.60 – 0.80	Very Strong
0.80 – 1.00	Extremely Strong

Correlation / Covariance Summary

- Covariance provides the direction of a linear relationship between 2 variables
- Correlation provides the direction and strength of a linear relationship between 2 variables
- Covariance has units and ranges from negative infinity to positive infinity
- Correlation is unitless and ranges from -1 to +1
- Both can provide misleading results if provided with outliers or non linear data
- Both provide a measure of association, not causation

Central Limit Theorem (CLT)

- The Central Limit Theorem is possibly the most important theorem in all of Statistics
- CLT is certainly the most important theorem for statisticians and very important in general to data scientists
- The CLT states that if random samples of size n are repeatedly drawn (with replacement) from **any** population with mean μ and variance σ2 , then **when n is large**, the distribution of the sample means will follow the characteristics of a normal distribution

CLT Assumptions

- How large does the sample need to be?
  - A common rule of thumb is n >= 30 for ANY population distribution
  - Use n >= 15 if the population distribution is symmetric
- If the population follows a normal distribution, then n can be any size
- Samples must be independent (sample with replacement)
- Samples must be identically distributed
- As sample size increases, the resulting sample distribution of the mean will more closely approximate a normal distribution.

CLT Summary

- CLT is very powerful because it provides us with a tool to transform any population distribution into a normal distribution
- Once transformed into a normal distribution, operations can be performed using the well-known Gaussian distribution
- See CLT demo code in the numpy tutorial notebook.