3251 Foundations of Data Science

Data Science Fundamentals Certificate

Module 3

DISTRIBUTION OF RANDOM VARIABLES

Course Roadmap

Module / Week	Title
1	Introduction to Statistics for Data Science
2	Probability
3	Distribution of Random Variables
4	Inference
5	Model Building
6	Linear Regression
7	Multiple Linear Regression
8	Logistic Regression
9	Introduction to Bayesian Inference
10	Multi-level Models
11	Markov Chain Monte Carlo
12	Presentations
13	Final Exam

Module 3: Learning Objectives

- Review of Probability and Random Variable
- Introduction to Distributions

Key Topic Overview – Distribution of Random Variables

- Continuous Distributions
- Discrete Distributions
- Measures of Distribution

Required/Recommended Readings

Required:

1) OpenIntro, 3rd edition, by David Diez, Christopher Barr, Mine Centinkaya-Rundel, Copyright © 2015 OpenIntro.org

Recommended:

1) Think Stats: Probability & Statistics for Programmes Version 1.6.0, by Allen Downey, Copyright © 2011 Green Tea Press

2) Using R for introductory statistics, John Versani, Chapman and Hall/CRC Press

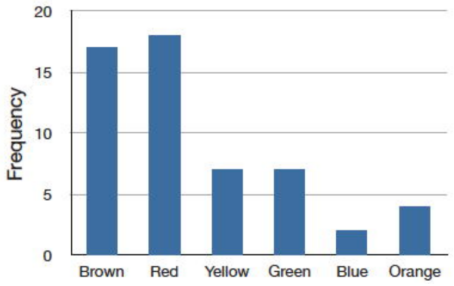
Section sub-header

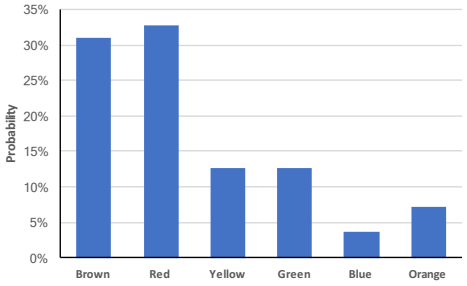
INTRODUCTION

Frequency Table

- Let's suppose that we buy a bag of M&M and count the number each colour
- This table is called a frequency table and it describes the distribution of M&M color frequencies.
- This kind of distribution is called a frequency distribution

Color	Frequency
Brown	17
Red	18
Yellow	7
Green	7
Blue	2
Orange	4





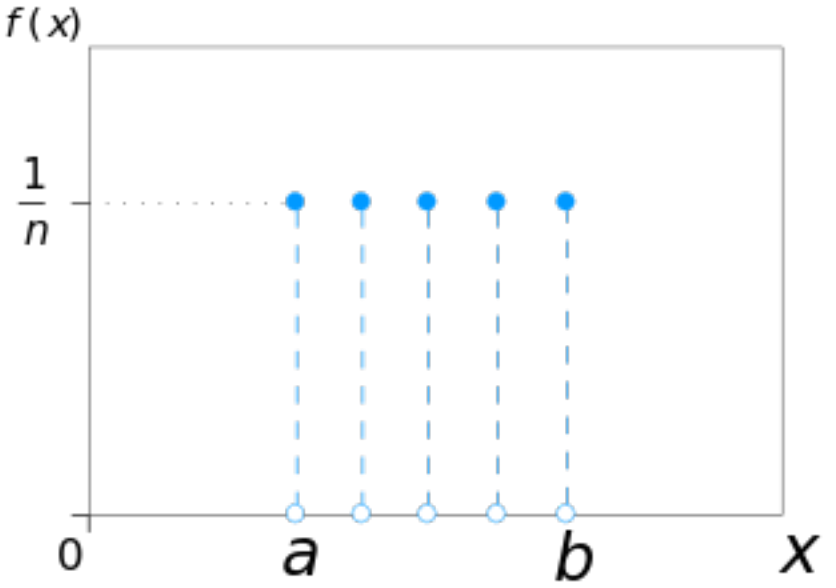
- If we report proportions rather than frequencies we obtain a probability distribution (probability of picking one specific colour when we pick one M&M) – Brown = 0.30

http://onlinestatbook.com/2/introduction/distributions.html

Probability Distribution

- Description of how likely a random variable or set of random variable is to take on each of the possible states.

- Discrete Variables:
  - Probability Mass Function (PMF)
  - Maps from a state of a random variable to the probability of that random variable taking that stated
  - Domain of P must be the set of all possible states of x
  - For all x, 0 <= P(x) <= 1
  - Sum[P(x)] = 1



https://en.wikipedia.org/wiki/Discrete_uniform_distribution

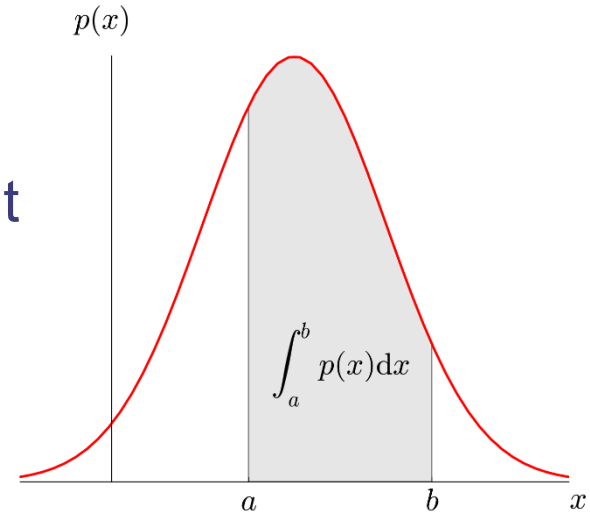
Probability Distribution – cont.

- For a discrete random variable, a probability distribution contains the probability of each possible outcome
- For a continuous random variable, the probability of any one outcome is zero (if you specify it to enough decimal places)
- The approach is to create a grouped frequency distribution. In a grouped frequency distribution, scores falling within various ranges are tabulated - histogram
- A probability density function is a formula that can be used to compute probabilities of a range of outcomes for a continuous (as the area below the curve)

http://onlinestatbook.com/2/glossary/probability_density.html

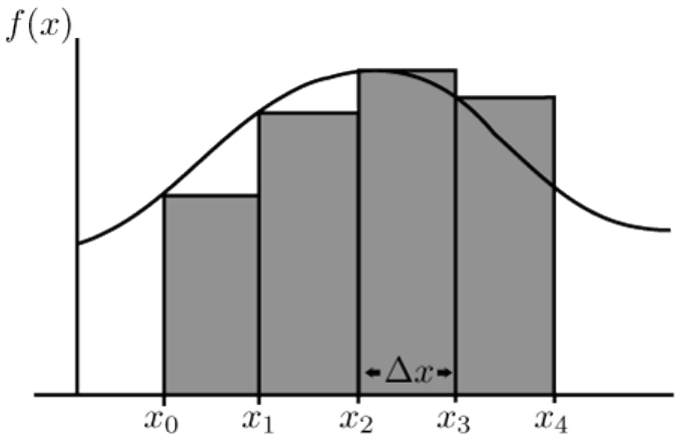
Probability Distribution – cont.

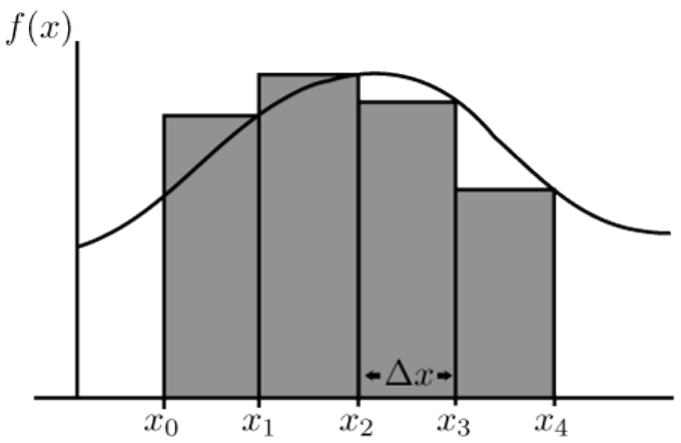
- Continuous Random Variables:
  - Probability Density Function (PDF)
  - Domain of P must be the set of all possible states of x
  - For all x, p(x) >0 , note no requirement for p(x) <=1
  - Integral[p(x)] = 1
- p(x) is the probability of landing inside an infinitesimal region with volume dx
- What's the probability of x being equal to a specific value?
- What's the P(a <= x <= a)



http://work.thaslwanter.at/Stats/html/statsDistributions.html

Meaning of $\int$





$$\lim_{n\to\infty} \sum_{k=1}^{n} f(x_k^*)\Delta x = \int_a^b f(x)\,dx$$

https://www.quora.com/How-do-use-the-limit-of-the-sequence-with-Riemann-Sums

http://teachtogether.chedk12.com/teaching_guides/view/73

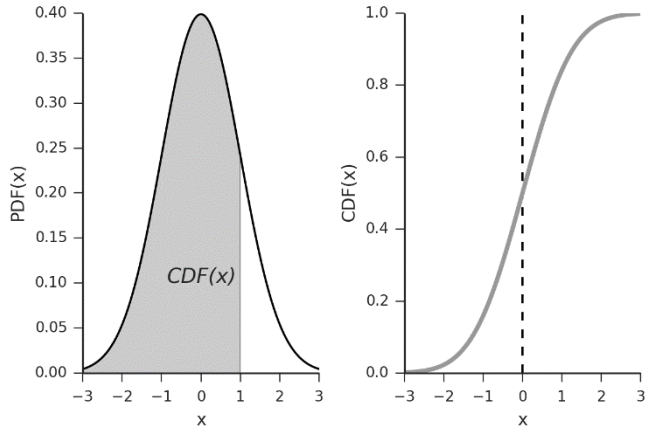
Cumulative Probability Distribution

- Probability that X will take a value equal or less than x
- In the case of a discrete distribution

$$F(x) = \mathrm{P}(X \leq x) = \sum_{x_i \leq x} \mathrm{P}(X = x_i) = \sum_{x_i \leq x} p(x_i)$$

- In the case of continuous distribution, corresponds to the area under the pdf

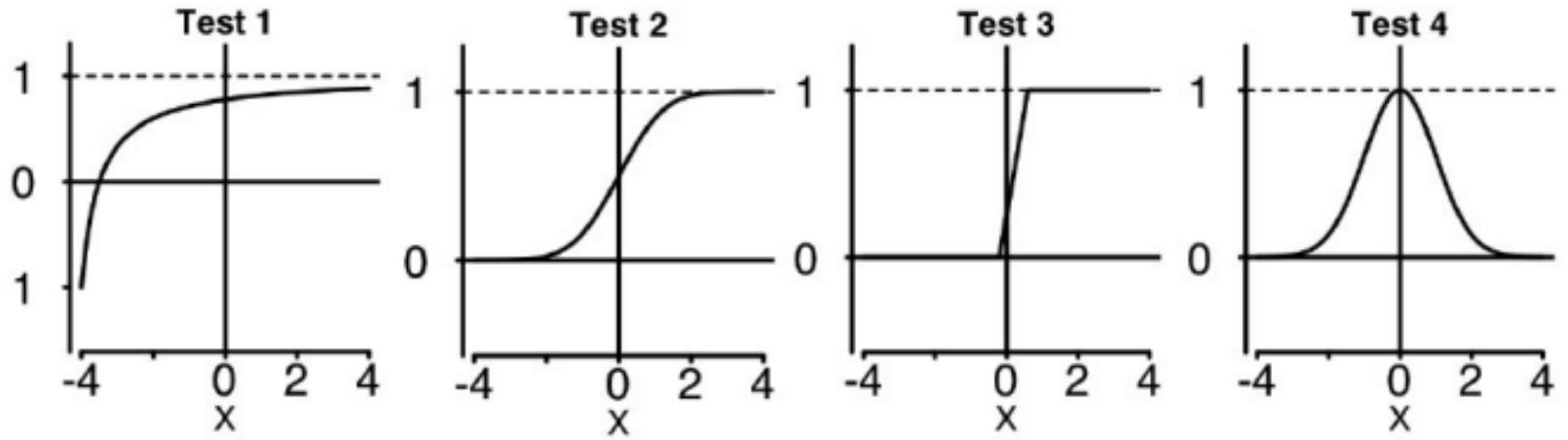
$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt.$$



http://work.thaslwanter.at/Stats/html/statsDistributions.html

Valid cdf?

- Which of the following is a valid cdf?



https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014

Question

- Suppose:
  - X has range [0, 2]
  - pdf $f(x) = cx$
  - a) What is the value of c
  - b) Compute the cdf $F(x)$
  - c) Compute $P(1 \leq X \leq 2)$

Section sub-header

MEASURES OF DISTRIBUTION

Expectation

- Mean or expected value is the probability weighted average of all values

$$E(X) = \sum_{n=1}^{\infty} x_n p_n$$

where X is a discrete random variable taking value x_n with probably p_n

- For continuous variables, it is computed with an integral. Where p(x) is the pdf

$$E(X) = \int_{-\infty}^{\infty} x p(x) dx$$

Variance

- The variance measures how much the values of a function of a random variable vary as we sample different values of x
- Variance is the expectation of the squared deviation of a random variable from its mean

$$Var(X) = \sum_{n=1}^{\infty} p_n (x_n - \mu)^2$$

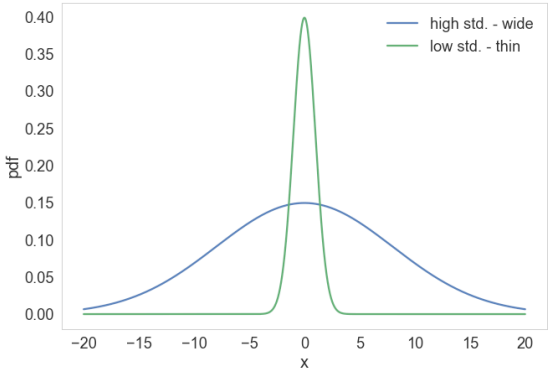
- where X is a discrete random variable taking value x_n with probably p_n and mean μ

$$Var(X) = \int_{-\infty}^{\infty} \left(x - \mu^2\right) f(x)\, dx$$

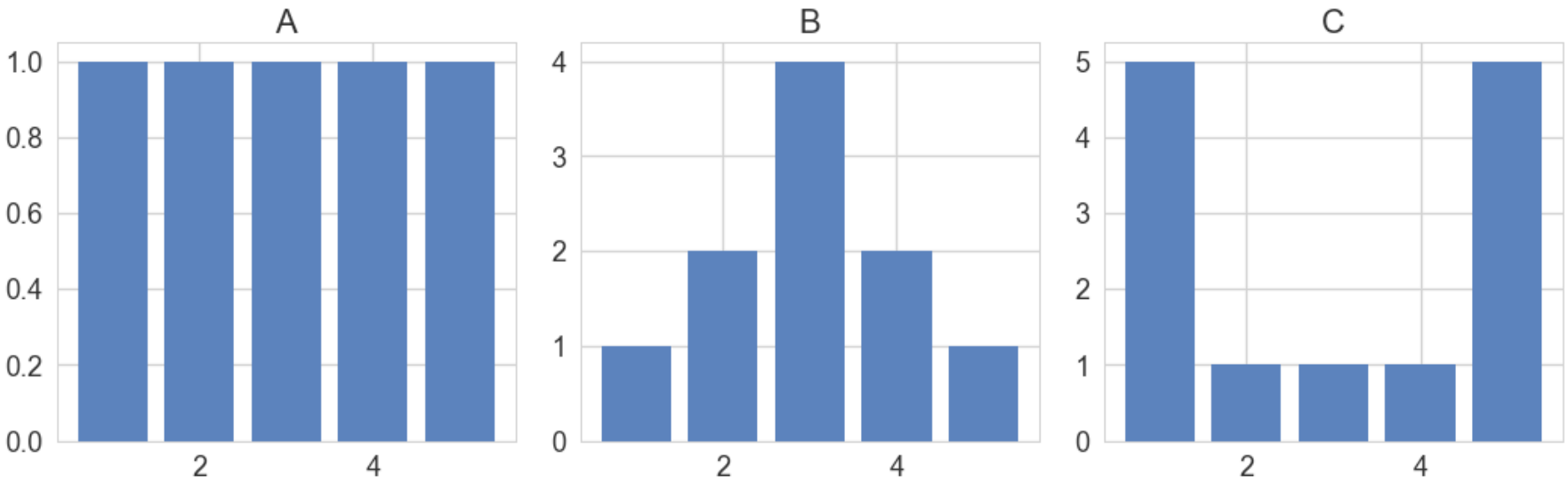
- where X is a continuous random variable having a probability density function f(x) and mean μ
- Measures the spread of the random variable from the mean

Standard Deviation

- Standard deviation is a measure that is used to quantify the amount of variation of a random variable.
- Calculated as the square root of variance
- A low standard deviation indicates that the values of a random variable are close to the expected value, while a high standard deviation indicates that the values are spread out over a wider range of values.



Standard Deviation?



- Sort A, B, and C by their standard deviation. From largest to smallest

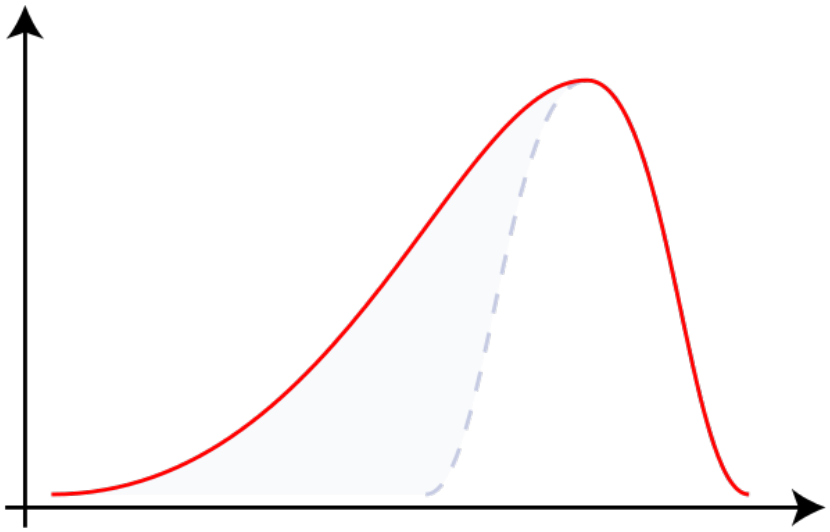
Skewness

- Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean.

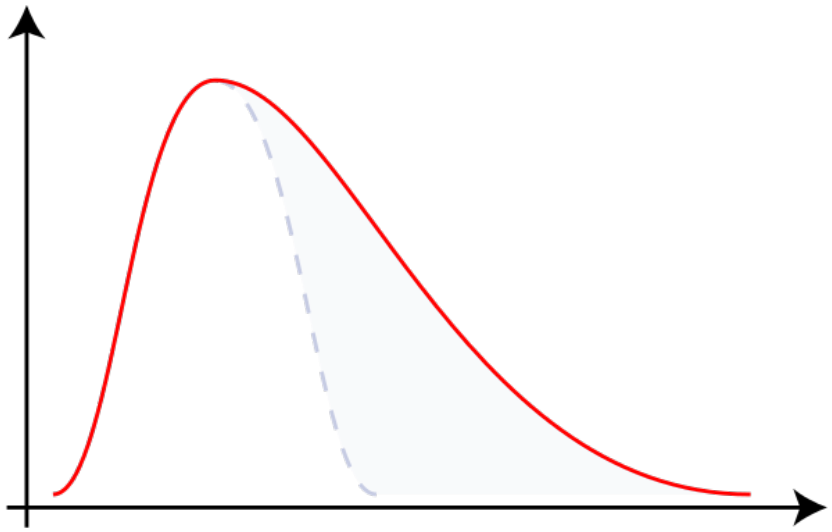
$$\gamma_1 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu^3}{\sigma^3}$$

- The normal distribution has a skewness of 0; therefore the skewness will tell us whether a random variable is concentrated to the left (+ve skew) or right (-ve skew) of the mean.

Skewness



Negative Skew



Positive Skew

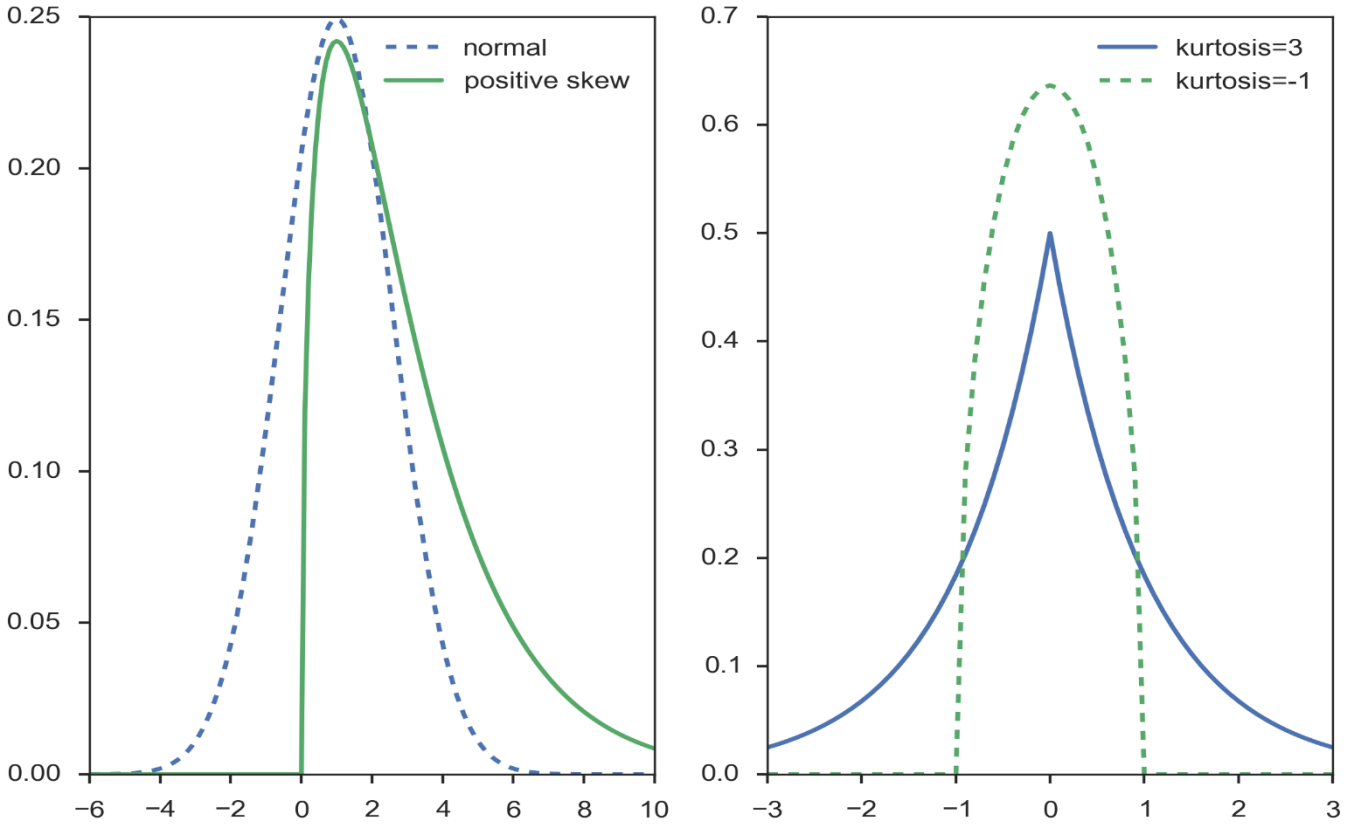
Kurtosis

- Kurtosis is a measure of the "tailedness" of the probability distribution of a real-valued random variable.

$$Kurt(X) = \frac{E\left[(X-\mu)^4\right]}{\left(E\left[(X-\mu)^2\right]\right)^2} = \frac{\mu^4}{\sigma^4}$$

- The normal distribution has a kurtosis of 3; for distributions with kurtosis less than 3, there are fewer outliers than the normal distribution and vice versa.

Kurtosis



http://work.thaslwanter.at/Stats/html/statsDistributions.html

Section sub-header

DISTRIBUTION OF RANDOM VARIABLES

Frequently Encountered Probability Distributions

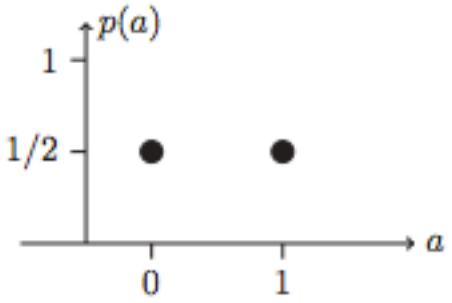
- Discrete
  - Bernoulli Distribution
  - Binomial Distribution
  - Poisson Distribution
- Continuous
  - Uniform Distribution
  - Normal Distribution

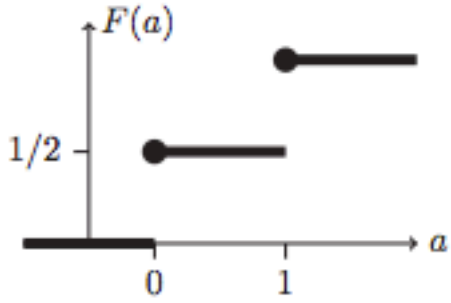
Bernoulli Distribution

- The Bernoulli distribution models one trial in an experiment that can result in either success or failure
  - Success happens with probability *p* , while failure happens with probability *1-p*
- The general terminology is to say X is 1 on success and 0 on failure, with success and failure defined by the context
- Many decisions can be modeled as a binary choice, such as whether to vote for or against a proposal. If *p* is the proportion of the voting population that favors the proposal, than the vote of a random individual is modeled by a Bernoulli(p)

Bernoulli Distribution

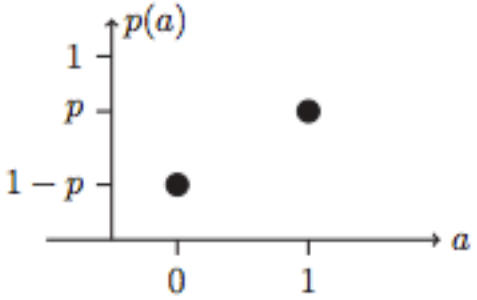
value a:	0	1
pmf $p(a)$:	1/2	1/2
cdf $F(a)$:	1/2	1

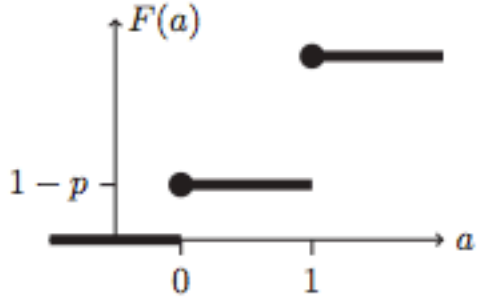




Table, pmf and cmf for the Bernoulli(1/2) distribution

values a:	0	1
pmf $p(a)$:	1-p	p
cdf $F(a)$:	1-p	1





Table, pmf and cmf for the Bernoulli(p) distribution

https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014

Bernoulli Distribution (Con't)

- The Bernoulli distribution with parameter p is characterized as

$$p(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \\ 0 & \text{if } x \neq 0, 1 \end{cases}$$

- Expected value $E(X) = p$
- Variance $Var(X) = p(1-p)$

Binomial Distribution

- Suppose an experiment having two possible outcomes: either success or failure, is repeated several times and the repetitions are independent of each other
- The total number of experiments where the outcome turns out to be a success is a random variable whose distribution is called binomial distribution with parameter number of repetitions n and probability of success p
- The binomial distribution is characterized as

$$p(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} \text{ if } x \in \{0,1,\ldots,n\} \\ 0 \quad \text{if x} \notin \{0,1,\ldots,n\} \end{cases}$$

- A binomial distribution can be seen as a sum of mutually independent Bernoulli random variables
- Expected value $E(X) = np$
- Variance $Var(X) = np(1-p)$

Binomial Example

- What is the probability of 3 or more heads in 5 tosses of a fair coin?
  - The binomial coefficients associated with n = 5 are

$$\binom{5}{0} = 1, \quad \binom{5}{1} = \frac{5!}{1!\,4!} = \frac{5\cdot 4\cdot 3\cdot 2\cdot 1}{4\cdot 3\cdot 2\cdot 1} = 5, \quad \binom{5}{2} = \frac{5!}{2!\,3!} = \frac{5\cdot 4\cdot 3\cdot 2\cdot 1}{2\cdot 1\cdot 3\cdot 2\cdot 1} = \frac{5\cdot 4}{2} = 10,$$

X ~ binomial(5, p)

values a:	0	1	2	3	4	5
pmf $p(a)$:	$(1-p)^5$	$5p(1-p)^4$	$10p^2(1-p)^3$	$10p^3(1-p)^2$	$5p^4(1-p)$	p^5

We were told $p = 1/2$ so

$$P(X \geq 3) = 10\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 + 5\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^1 + \left(\frac{1}{2}\right)^5 = \frac{16}{32} = \frac{1}{2}.$$

Binomial Example

- X ~ Binomial(5,p) n=5, k=2
- Computer p(2)
- List (2H, 3T) has 10 entries:
  - HHTTT, HTHTT, HTTHT, HTTTH, THHTT, THTHT, THTTH, TTHHT, TTHTH,TTTHH
  - Each with the same probability: $p^2 (1-p)^3$
  - So p(2) = 10 * $p^2 (1-p)^3$
- Shortcut:
  - Count the number of time we can choose 2 out of 5 things. (n choose x)
  - Multiply by the probability of occurrence

Geometric Distributions

- The geometric distribution models the number of tails before the first heads. (Bernoulli trials)
- A random variable X has a geometric distribution with parameter p if it takes the values 0, 1, 2, 3, . . . and its pmf is given by $p(k) = P(X = k) = (1 - p)^k p$
  - Mean: p / (1-p)
  - Variance: (1-p) / p^2

Geometric Distributions

- Suppose that the inhabitants of an island plan their families by having babies until the first girl is born. Assume the probability of having a girl with each pregnancy is 0.5 independent of other pregnancies. What is the probability that a family has k boys?
  - we can think of boys as tails and girls as heads. Then the number of boys in a family is the number of tails before the first heads.
  - Let X be the number of boys in a (randomly-chosen) family.
  - If X = k the sequence of children in the family from oldest to youngest is BBB ...BG with the first k children being boys.
  - The probability of this sequence is $(1/2)^k (1/2)$.
  - So X follows a geometric(1/2) distribution.

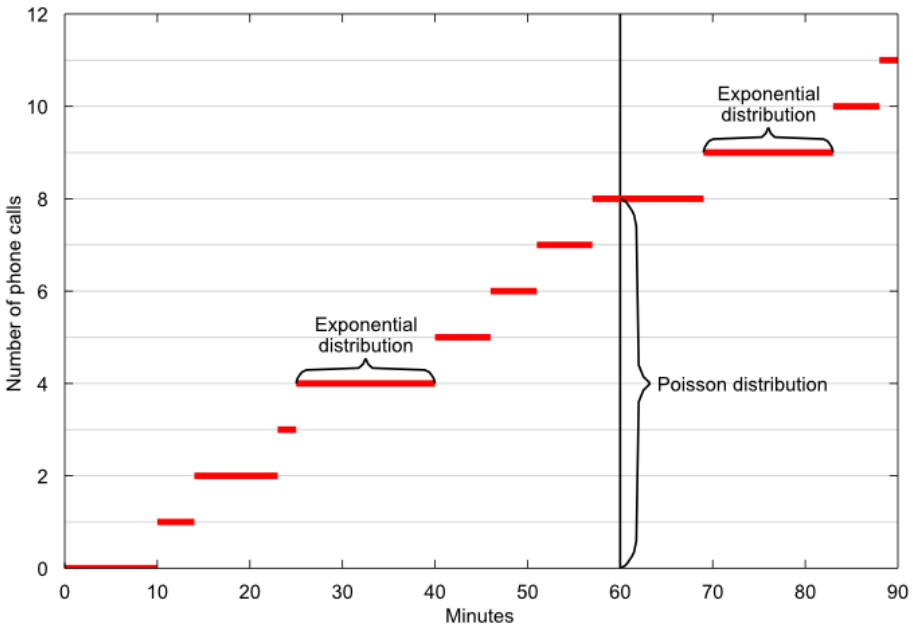
Negative Binomial Distribution

- The geometric distribution is a special case of the negative binomial distribution
- The negative binomial counts the number of failures until a fixed number of successes
- The binomial counts the successes, the negative binomial counts the number of failures until we get an x number of successes

$$f(x|r, p) = \begin{cases} \binom{r+x-1}{x} p^r (1-p)^x & \text{fox } x = 1,2,3, \ldots \\ 0 & otherwise \end{cases}$$

Poisson Distribution

- Counting Process
- Consider the number of phone calls received by a call center.
- If the time elapsed between two successive phone calls has an exponential distribution and it is independent of the time of arrival of the previous calls, then the total number of calls received in one hour has a Poisson distribution



Poisson Distribution (Con't)

- Independent increments:
  - the numbers of events occurred in disjoint time intervals are independent.
- Stationary increments:
  - the distribution of the number of events occurred in a time interval only depends on the length of the interval and does not depend on the position.

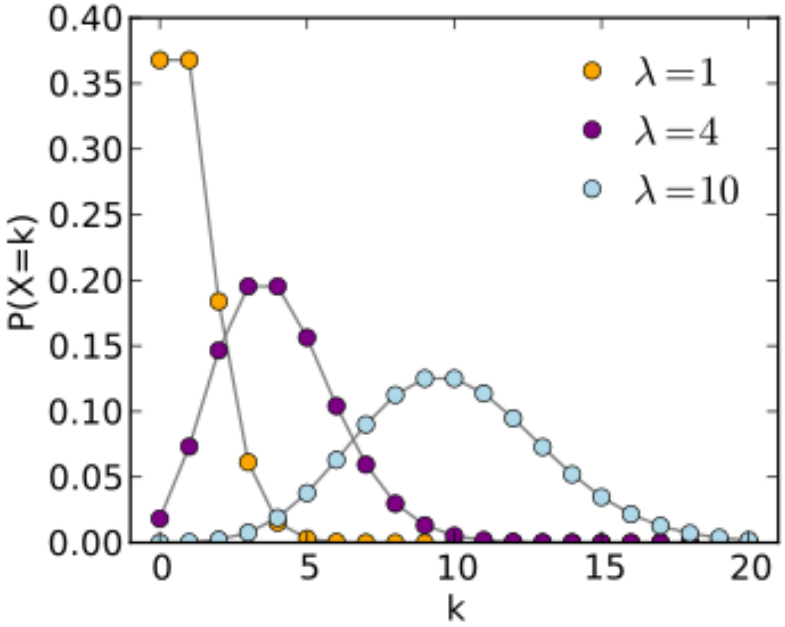
Poisson Distribution (Con't)

- PDF:

$$p(x) = \begin{cases} \exp(-\lambda)\frac{1}{x!}\lambda^x & \text{if } x \in Z_+ \\ 0 & \text{if x} \notin Z_+ \end{cases}$$

where Z is the set of integers

- Expected value $E(X) = \lambda$
- Variance $Var(X) = \lambda$



https://en.wikipedia.org/wiki/File:Poisson_pmf.svg

Example

- If three persons, on an average, come to ABC company for job interview, then find the probability that less than three people have come for interview on a given day?
  - $P(x<3; \lambda=3) = P(x=0; \lambda=3)+P(x=1; \lambda=3)+P(x=2; \lambda=3)$
  - $P(x=0; \lambda=3) = e^{-3}3^0 / 0! = 0.04978706837$
  - $P(x=1; \lambda=3) = e^{-3}3^1 / 1! = 0.1493612051$
  - $P(x=2; \lambda=3) = e^{-3}3^2 / 2! = 0.22404180766$

  $P(x<3; \lambda=3) = 0.42319008113$

Poisson Distribution

- In our example 3 persons show up per day for an interview, what is the probability that a person will arrive at any given hour
- The probability of a person arriving at any day is

$$p = \frac{3}{3600\ x\ 24} = 0.000035$$

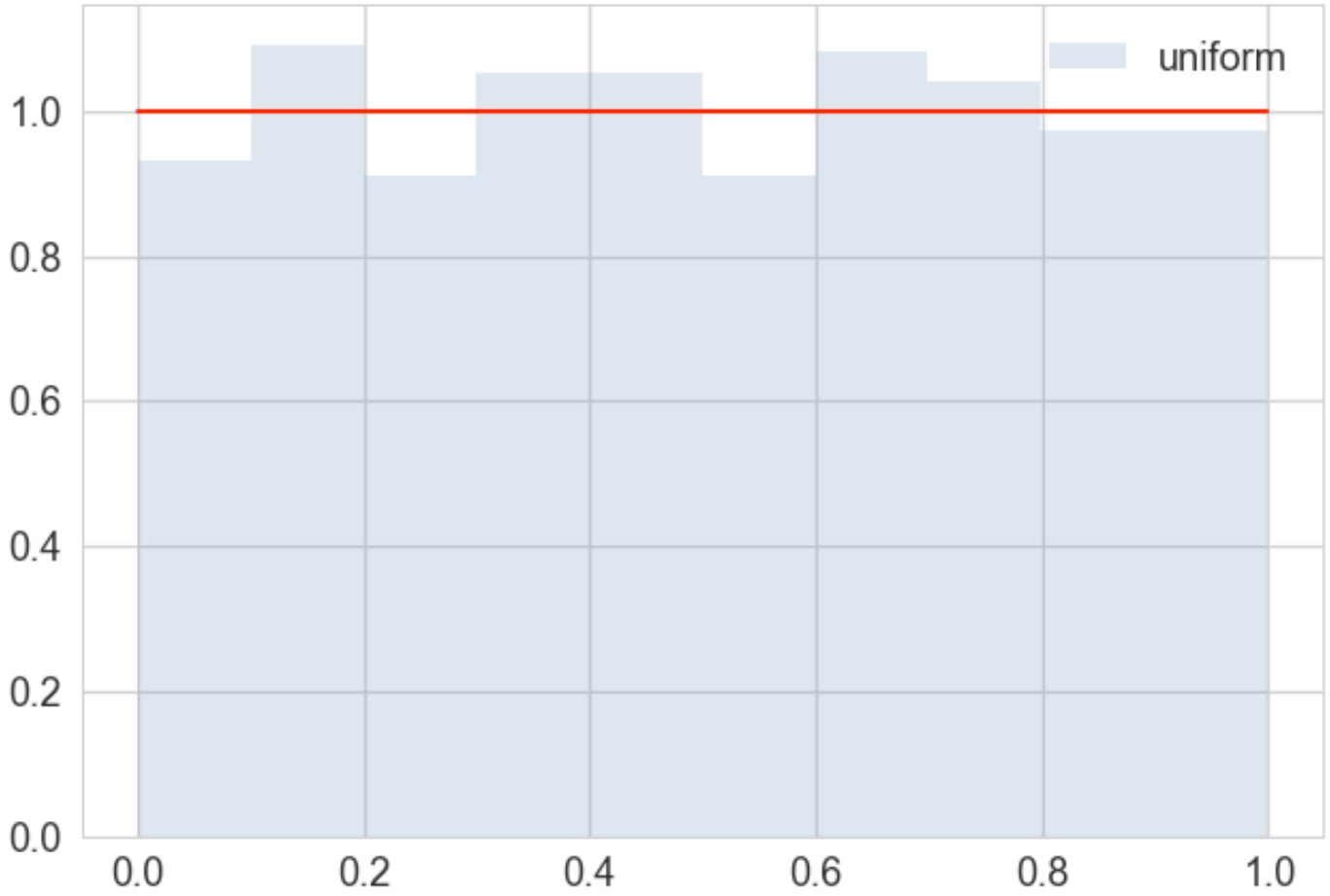
- Applying the binomial distribution, $n = 3600\ x\ 24, p = 0.000035$ the calculation is cumbersome….
- Using calculus, it can be approximated by the Poisson distribution (λ in our example is 5)

$$f(x|\lambda) = \begin{cases} \dfrac{e^{-\lambda x}\lambda^{x}}{x!} & \text{for x = 0, 1, 2, . . . ,} \\ 0 & \text{Otherwise} \end{cases}$$

Uniform Distribution

- Models situations where all outcomes are equally likely
- Consider the probability distribution when rolling a fair dice
  - Each side has an equal probability of turning up
  - Because there are six sides to each dice, there are six possible outcomes, with each outcome having a probability of 1/6

Uniform distribution



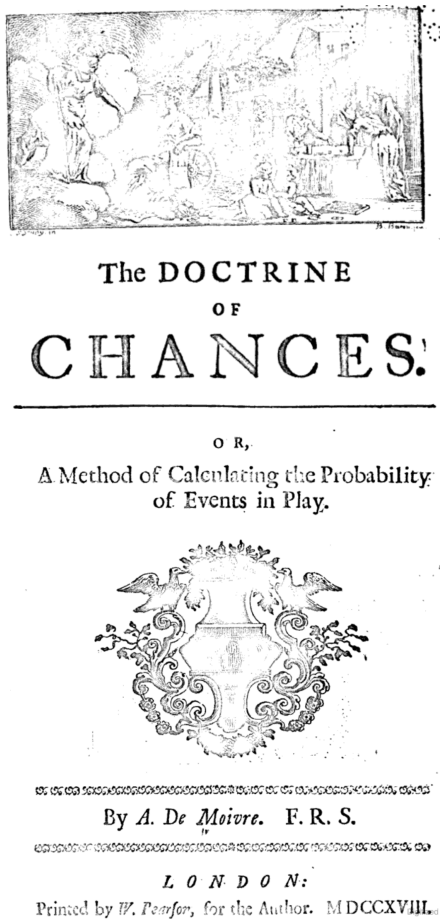
Uniform Distribution (Con't)

- The uniform distribution is characterized as having the same probability density for all values in its support.

$$f(x) = \begin{cases} \dfrac{1}{u-l} & \text{if } x \in [l,u] \\ 0 & \text{if } x \notin [l,u] \end{cases}$$

- A random variable with a uniform distribution is called a uniform random variable
- Expected value $E(X) = \dfrac{u+l}{2}$
- Variance $Var(X) = \dfrac{(u-l)^2}{12}$

Normal Distribution



The DOCTRINE
OF
CHANCES:

OR,
A Method of Calculating the Probability
of Events in Play.

By *A. De Moivre.* F. R. S.

LONDON:
Printed by *W. Pearson*, for the Author. MDCCXVIII.

- De Moivre used the normal distribution to approximate the binomial distribution for a big number of trials
- He was a 'consultant' to gamblers
- He based his work on Newton's

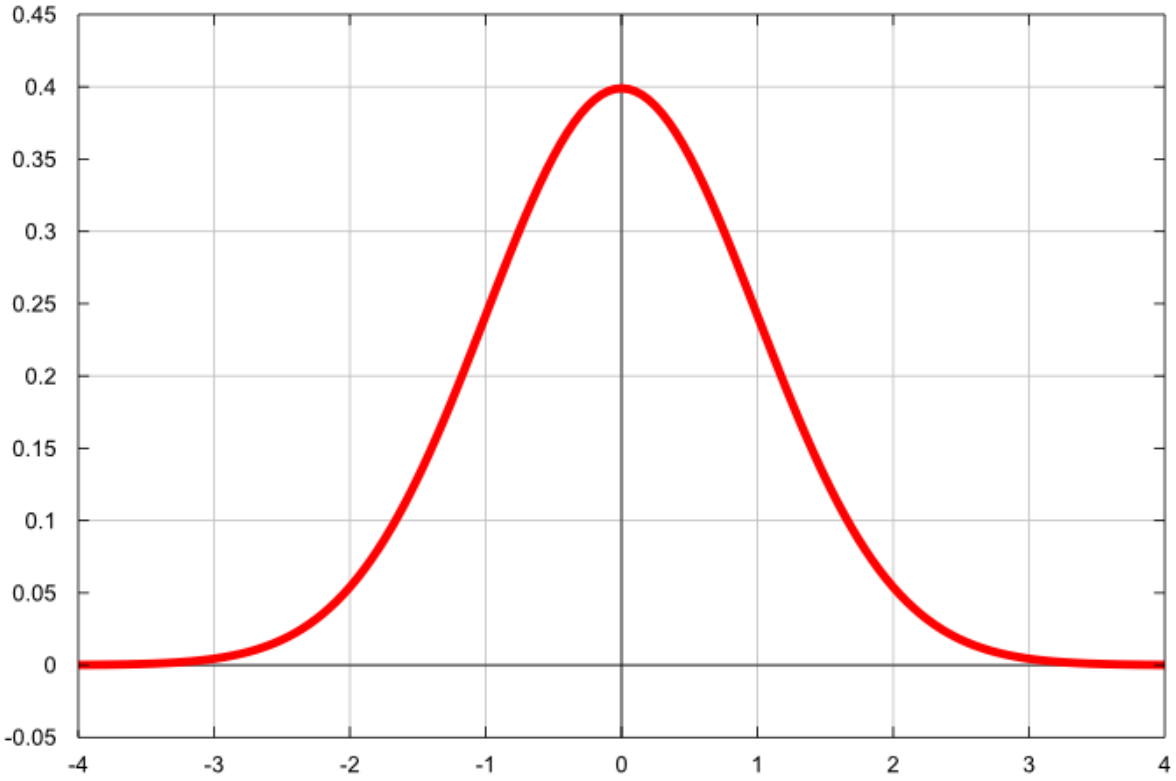
https://books.google.ca/books?id=3EPac6QpbuMC&pg=PA1&source=gbs_toc_r&cad=3#v=onepage&q&f=false

Normal Distribution

- It is also known as the Gaussian distribution
- Bell shaped curved. Although other bell shapes exist (such as the Cauchy, Student's t, and logistic distributions)
- Important in the context of the central limit: the sum of many random variables will have an approximately normal distribution. More specifically, where X1, …, Xn are independent and identically distributed random variables

Standard Normal Distribution

$$X \sim N(0,1)$$



Normal Distribution (Con't)

- The normal distribution is characterized as

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) \text{ for } x \in \Re$$

- Expected value $E(X) = 0$
- Variance $Var(X) = 1$
- A more generalized version $X \sim N(\mu, \sigma^2)$

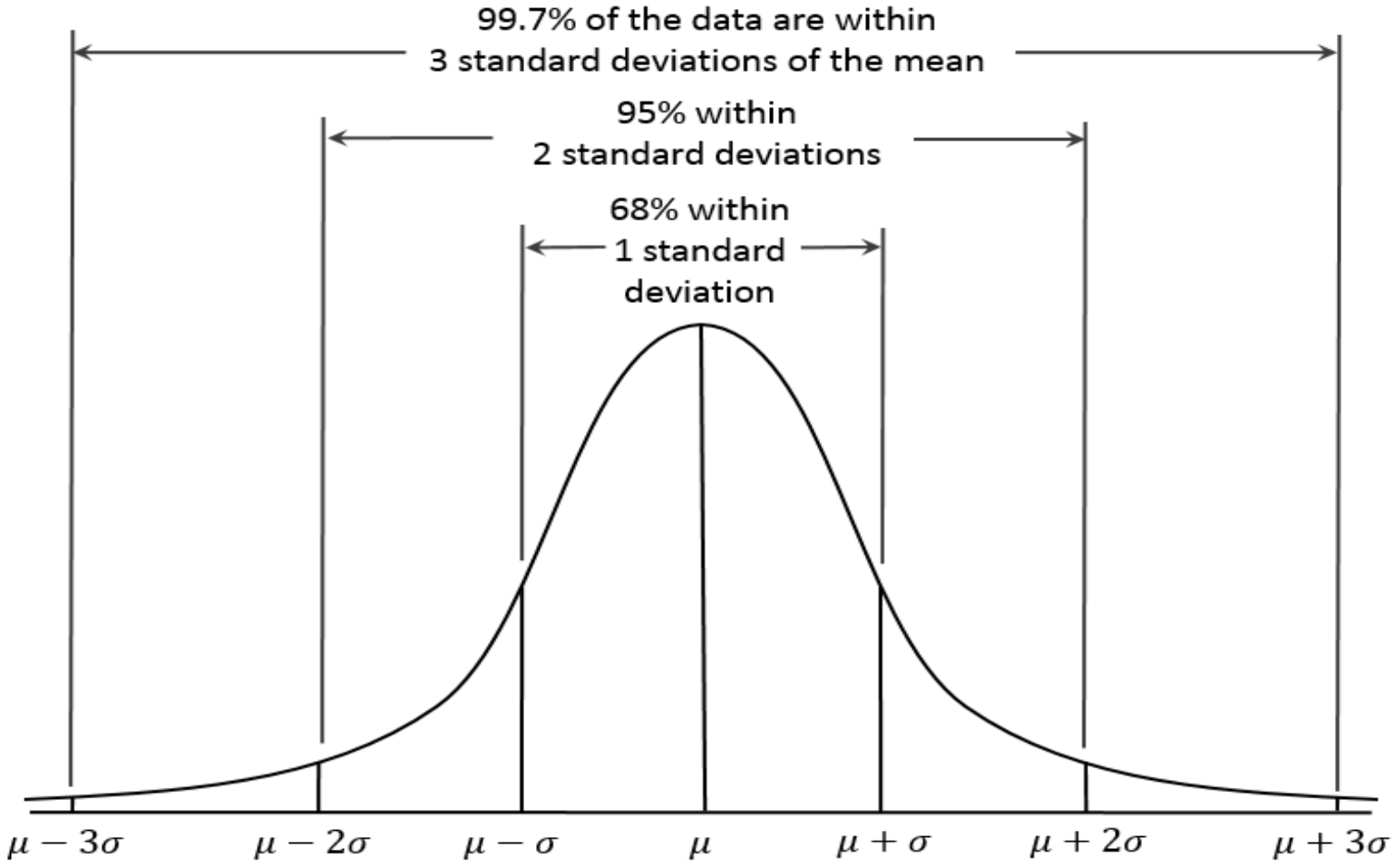
$$f(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \text{ for } x \in \Re$$

Normal Distribution

- There are many examples of normal distribution in real life
  - Heights of adults
  - Certain physiological measurements, such as blood pressure of adult humans.
  - IQ scores
  - Measurement errors in physical experiments are often modeled by a normal distribution.
  - Standardized test results
  - In finance, in particular the Black–Scholes model, changes in the logarithm of exchange rates, price indices, and stock market indices are assumed normal

https://en.wikipedia.org/wiki/Normal_distribution

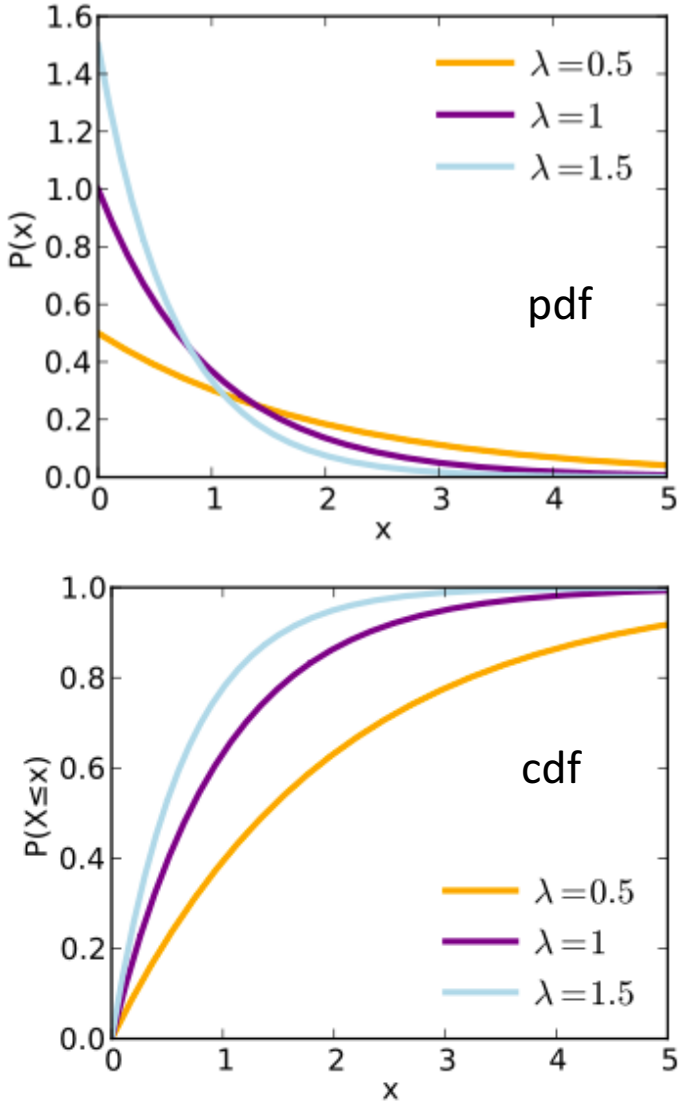
Measures of Distribution (Con't)



https://en.wikipedia.org/wiki/File:Empirical_Rule.PNG

Exponential distribution

- Models the waiting time for a continuous process to change state.
  - Parameter: λ
  - Range: [0, ∞)
  - Density: $f(x) = \lambda e^{-\lambda x}$ for $0 \leq x$
  - pdf $F(x) = 1 - e^{-\lambda x}$
  - Mean = $1/\lambda$
  - Variance = $1/(\lambda^2)$



Exponential distribution Memorylessness

- When T is interpreted as the waiting time for an event to occur relative to some initial time:
  - if T is conditioned on a failure to observe the event over some initial period of time s, the distribution of the remaining waiting time is the same as the original unconditional distribution.
  - Example: Suppose that the amount of time one spends in a bank is exponentially distributed with mean 10 minutes, $\lambda = 1/10$. What is the probability that a customer will spend more than 15 minutes in the bank? What is the probability that a customer will spend more than 15 minutes in the bank given that he is still in the bank after 10 minutes?
    - $P(X > 15) = \exp(-15\lambda) = \exp(-3/2) = 0.22$
    - $P(X > 15|X > 10) = P(X > 5) = \exp(-1/2) = 0.604$

Summary of Standard Distributions

- **Uniform**: Several outcomes are all equally likely, e.g. which number comes up on a thrown die
- **Normal** (aka. Gaussian): The probability density falls off as the square of the distance from the mean, e.g. heights of women
- **Bernoulli**: Takes value 1 when an experiment succeeds and 0 otherwise, e.g. tossing a coin
- **Binomial**: A series of independent Bernoulli trials, e.g. tossing a coin 20 times to how many heads appear
- **Poisson**: Rates of sparse events in space or time, e.g. lightening strikes, machine failures

Additional Probability Distributions

- Continuous
  - Chi-square Distribution
  - Beta and Gamma Distribution

Questions?

Additional Resources

- Interactive probability distributions
- https://ocw.mit.edu/ans7870/18/18.05/s14/applets/probDistrib.html
- http://www.bigdataexaminer.com/2015/12/14/how-to-implement-these-5-powerful-probability-distributions-in-python/
- https://www.johndcook.com/blog/distributions_scipy/

Next Class

- Hypothesis testing
- The t-distribution
- Paired data
- Difference between two means

Additional Slides

Chi-square Distribution

- The chi-square distribution is characterized as

$$f(x)=\begin{cases} cx^{\frac{n}{2}-1}\exp\left(-\frac{1}{2}x\right) & \text{if } x\in[0,\infty) \\ 0 & \text{if } x\notin[0,\infty) \end{cases} \quad \text{with n} \in \aleph$$

- n degrees of freedom $X \sim \chi^2(n)$
- $c=\dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}$
- Expected value $E(X)=n$
- Variance $Var(X)=2n$

Chi-square Distribution (con't)

- The chi-square random variable is constructed from n independent standard normal random variables
  - The square of a standard normal random variable is a chi-square random variable with one degree of freedom

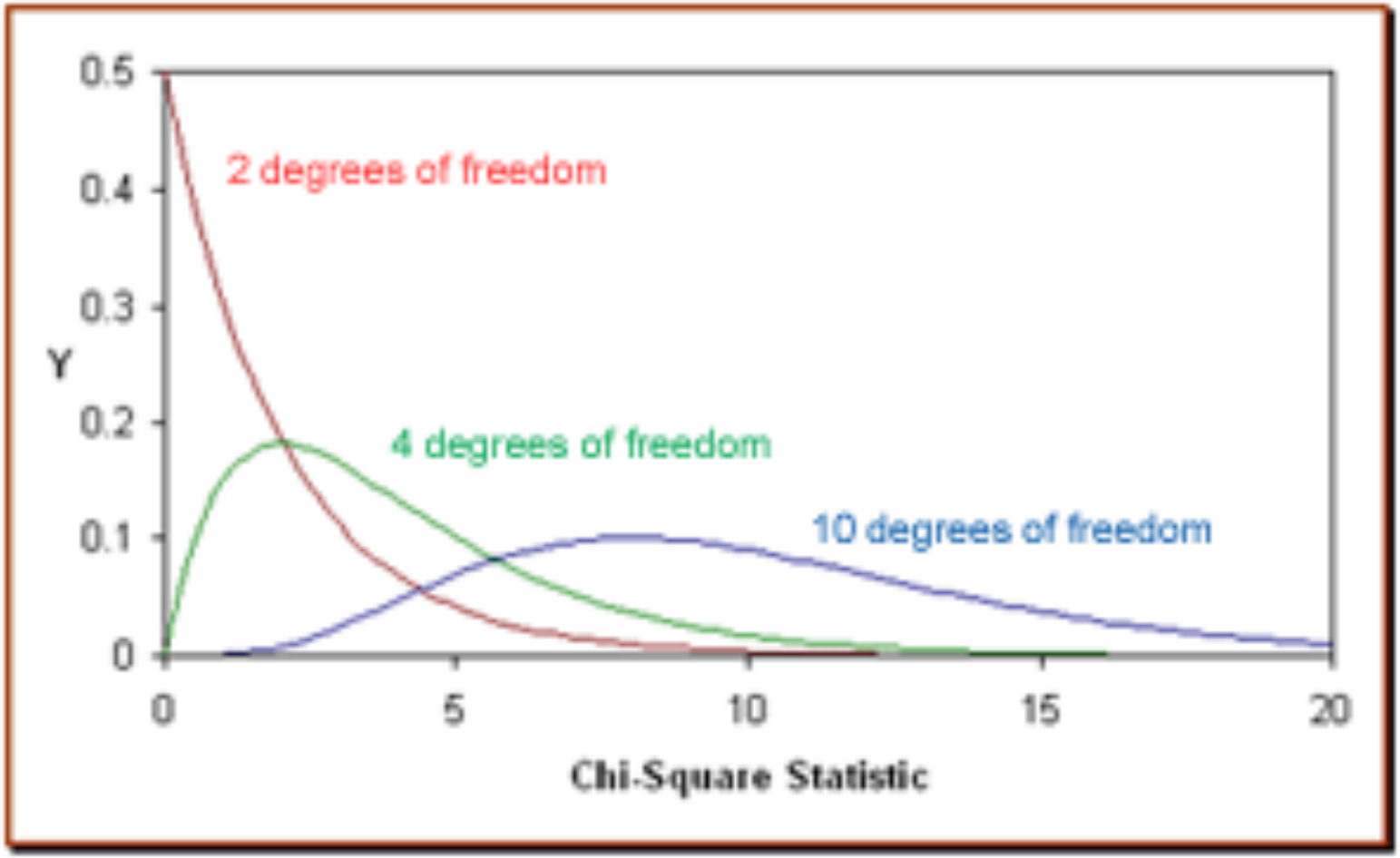
$$Y = Z^2 \sim \chi^2(1) \text{ where } Z \sim N(0,1)$$

  - The sum of squares of n independent standard normal random variables is chi-square distributed with n degrees of freedom

$$Y = X_1^2 + X_2^2 + \ldots + X_n^2 \sim \chi^2(n)$$

  - Expected value $E(Y) = n$
  - Variance $Var(Y) = 2n$

Chi-square Distribution (con't)



Chi-square Distribution (con't)

- The chi-square distribution is very important because many test statistics are approximated as chi-square distributions.
  - Deviations of differences between theoretically expected and observed frequencies (one-way tables)
  - Relationship between categorical variables (contingency tables)

Chi-square Distribution Application

- The Chi Square distribution can be used to test whether observed data differ significantly from theoretical expectations (more on this on the module of inference)
- We roll a 36 times, the expected frequency of any number is 6

Expected frequency = (1/6)(36) = 6

- $\frac{(E-O)^2}{E}$ follows a X^2 distribution with k-1 degrees of freedom, where k is the number of categories

Outcome	E	O	$\frac{(E-O)^2}{E}$
1	6	8	0.667
2	6	5	0.167
3	6	9	1.500
4	6	2	2.667
5	6	7	0.167
6	6	5	0.167

$\chi_5^2 = 5.333$

Beta and Gamma Distributions

- The beta function B is defined as

$$\mathrm{B}(\mathrm{a},\mathrm{b}) = \int_0^1 u^{a-1}(1-u)^{b-1}\,du;\ \mathrm{a}>0,\ \mathrm{b}>0$$

- The beta function is related to the gamma function

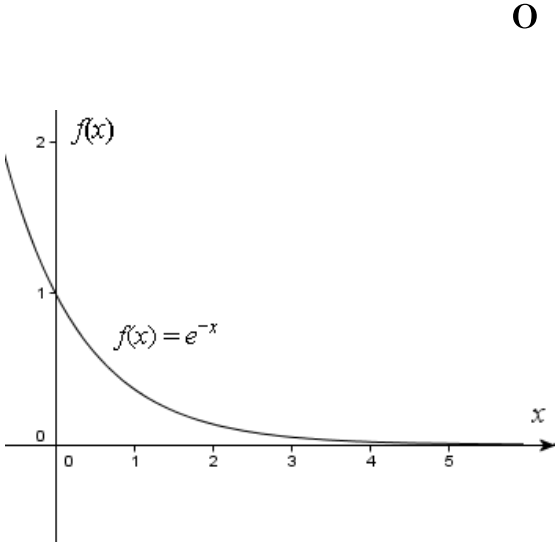
$$\mathrm{B}(\mathrm{a},\mathrm{b}) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

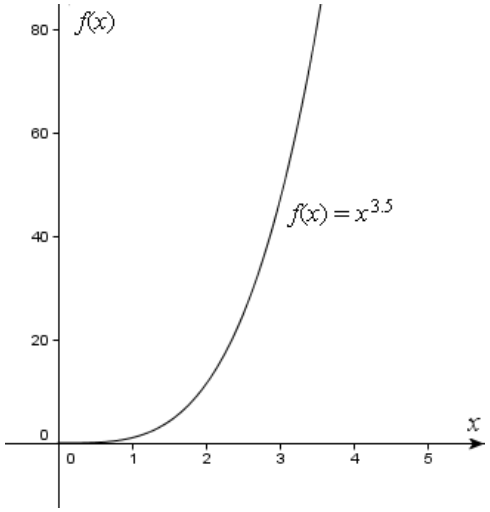
- The gamma function is defined for all complex numbers except the non-positive integers.

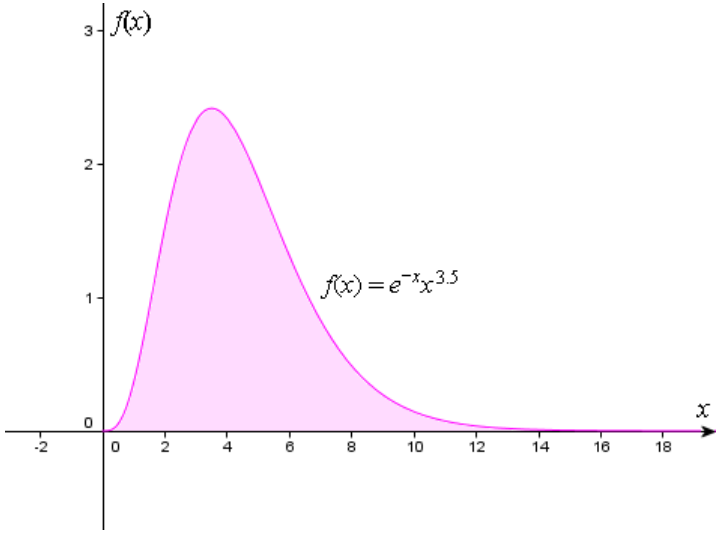
Beta and Gamma Distributions (Con't)

- Consider the following gamma function

$$\Gamma(n+1) = \int_0^\infty e^{-x} x^n dx \text{ if } n \text{ is a real number}$$



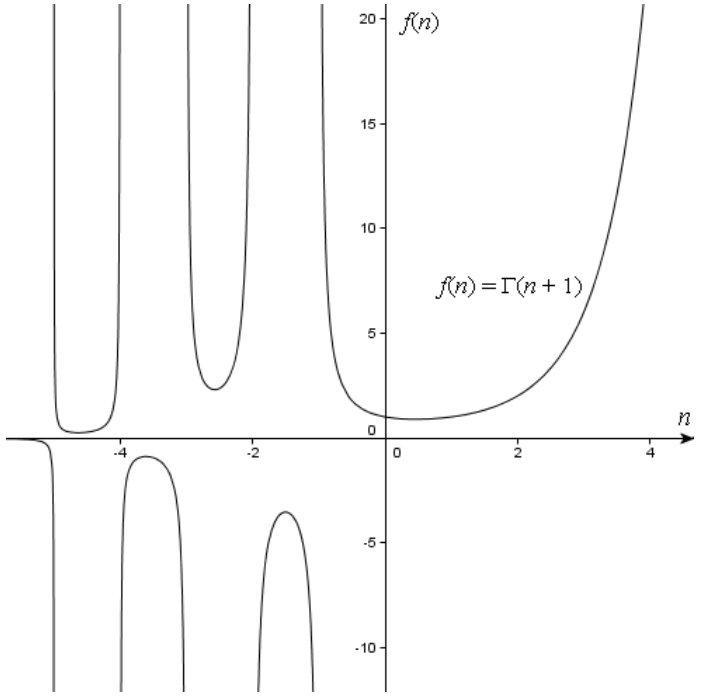




$$\Gamma(n+1) = n! \text{ if } n \text{ is a positive integer}$$

Beta and Gamma Distributions (Con't)

- Consider $f(n) = \Gamma(n+1)$



- There are "holes" corresponding to the negative integers

Beta and Gamma Distributions (Con't)

- The beta distribution is used to model a distribution *of probabilities (e.g.* probabilities of binomial outcomes) when we don't know the probabilities in advance, but which we have some reasonable guesses.
- The beta distribution with parameters $a, b \in (0, \infty)$ is the continuous distribution on (0,1) with probability density f given by

$$f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, \; 0 < x < 1$$