# Statistics and Machine Learning 1

# Lecture 1C: Some Standard Distributions

Mark Muldoon
Department of Mathematics, Alan Turing Building
University of Manchester

Week 1

# Bernoulli

- The *Bernoulli* distribution represents the outcome of a single random 'trial' (e.g. quality control test), or 'toss of a (possibly biased) coin'.
- It is discrete with probability mass function (PMF)

$$\mathrm{P}(X=0) = p_0 = 1-q, \qquad \mathrm{P}(X=1) = p_1 = q, \tag{1}$$

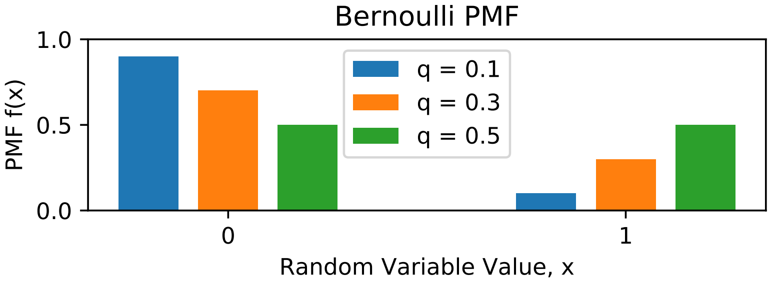
  for some $q \in [0,1]$.
- Its mean and variance are

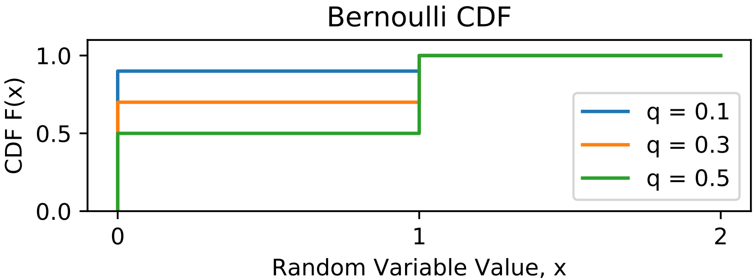
$$\begin{aligned} \mathrm{Mean}(X) &= \mathbb{E}[X] \\ &= 0 \times p_0 + 1 \times p_1 \\ &= q, \end{aligned} \tag{2}$$

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= 0^2 \times p_0 + 1^2 \times p_1 - q^2 \\ &= q(1-q), \end{aligned} \tag{3}$$

  respectively. Examples of probability mass functions and cumulative density functions (CDFs) follow.

# Bernoulli PMFs and CDFs



# Binomial

- The *binomial* distribution represents the outcome of $n$ independent 'trials'—formally a binomial random variable is the sum of $n$ Bernoulli random variables.
- Its probability mass function $\mathrm{Bin}(k \mid n, q)$ is the probability of $k$ positive outcomes out of $n$ given positive outcome probability $q$.
- It is discrete with probability mass function

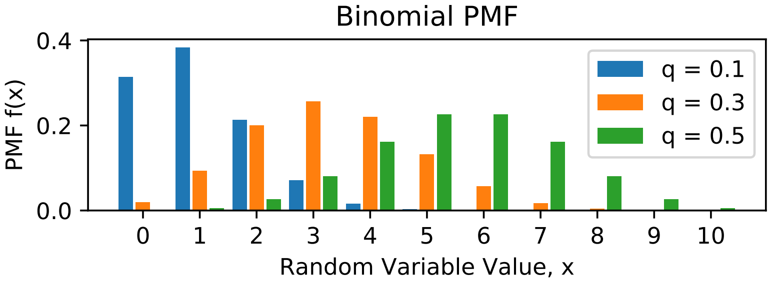
$$\mathrm{P}(X = k) = p_k = \mathrm{Bin}(k \mid n, q) = \binom{n}{k} q^k (1-q)^{n-k}, \tag{4}$$

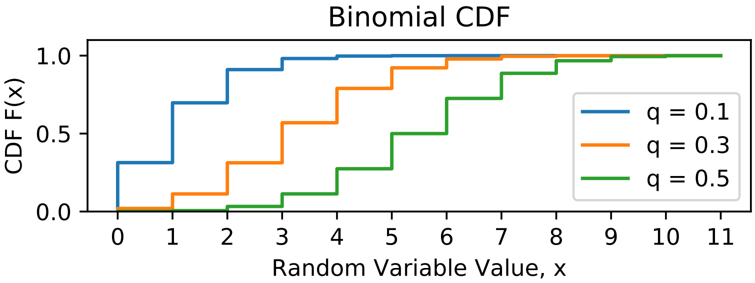
  for some integer $n$, for $0 \le k \le n$, for binomial coefficients $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, and $q \in [0, 1]$.
- The mean and variance are

$$\mathrm{Mean}(X) = nq, \qquad \mathrm{Var}(X) = nq(1-q), \tag{5}$$

  respectively—this is a special case of the general result that means and variances are *additive*.

# Binomial PMFs and CDFs



# Poisson

- The *Poisson* distribution represents the number of events that happen with a 'rate' $\lambda$ and can be thought of as the limit of the binomial as $n \to \infty$ at constant $nq$.
- It is discrete, with probability mass function

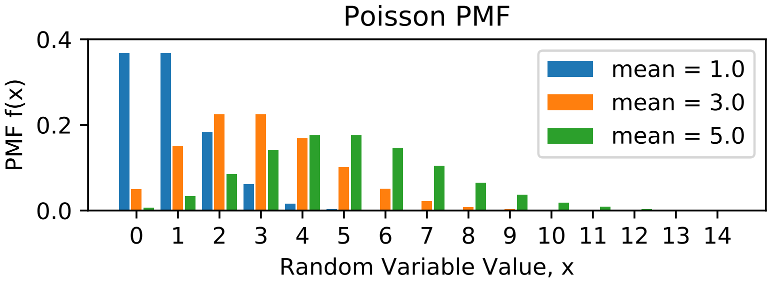
$$\mathrm{P}(X = k) = p_k = \mathrm{Po}(k|\lambda) = \frac{\lambda^k}{k!}\, e^{-\lambda}, \tag{6}$$

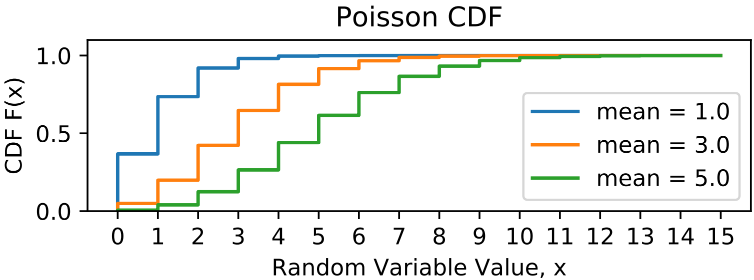
  for some $\lambda > 0$.
- The mean and variance are

$$\mathrm{Mean}(X) = \lambda, \qquad \mathrm{Var}(X) = \lambda. \tag{7}$$

- *Example:* For a disease that $0.003\%$ of the population suffer from, in a town of $100\,000$ the probability of seeing $k$ cases should be close to $\mathrm{Po}(k\,|\,3.0)$.

# Poisson PMFs and CMFs



# Beta

- The *Beta* distribution is most commonly used to represent uncertainty in probabilities or proportions.
- It is continuous with probability density function (PDF)

$$f(x) = \mathrm{Beta}(x \,|\, \alpha, \beta) = B_{\alpha,\beta}^{-1}\, x^{\alpha-1}(1-x)^{\beta-1}, \tag{8}$$

  for $x \in [0,1]$ and positive $\alpha, \beta$.
- The factor $B_{\alpha,\beta}^{-1}$ is a *normalizing constant*. It's chosen so that $\mathrm{P}(0 \leq x \leq 1) = 1$ and there is no simple expression for it. Instead, it's defined as an integral:

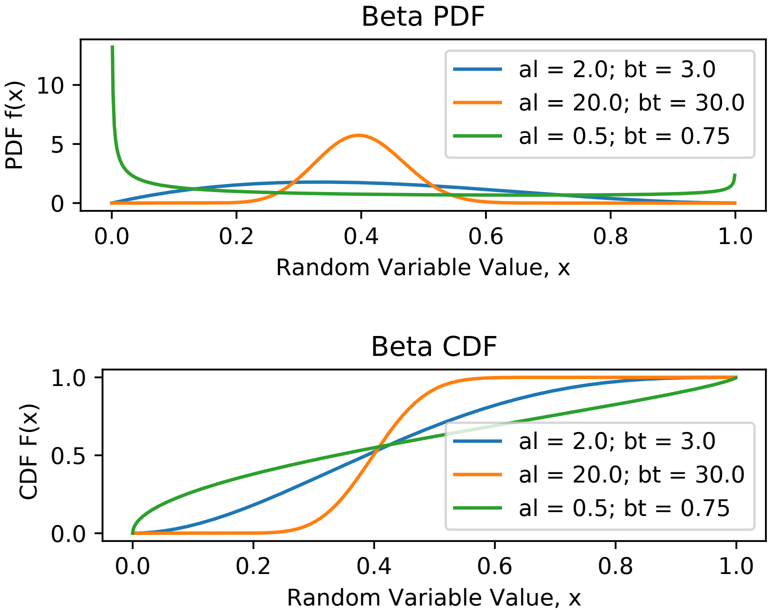
$$B_{\alpha,\beta} = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x.$$

- The mean of the Beta distribution is

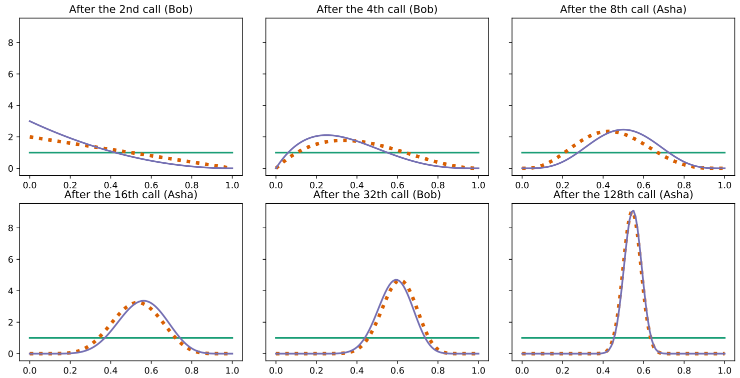
$$\mathrm{Mean}(X) = \frac{\alpha}{\alpha+\beta}, \qquad \mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(1+\alpha+\beta)} \tag{9}$$

  The variance decreases with increasing $\alpha$ and $\beta$ at constant mean.

# Beta PDFs and CDFs



# Application: estimating support from polling data




A simulated political poll from a race between Asha and Bob. Green curves show our initial knowledge about support for Asha while the dashed and solid curves show estimates informed by increasing numbers of voters' responses. All the curves here are Beta distributions.

# Gamma

- The *Gamma* distribution is often used to model times between events.
- It is continuous with probability density function

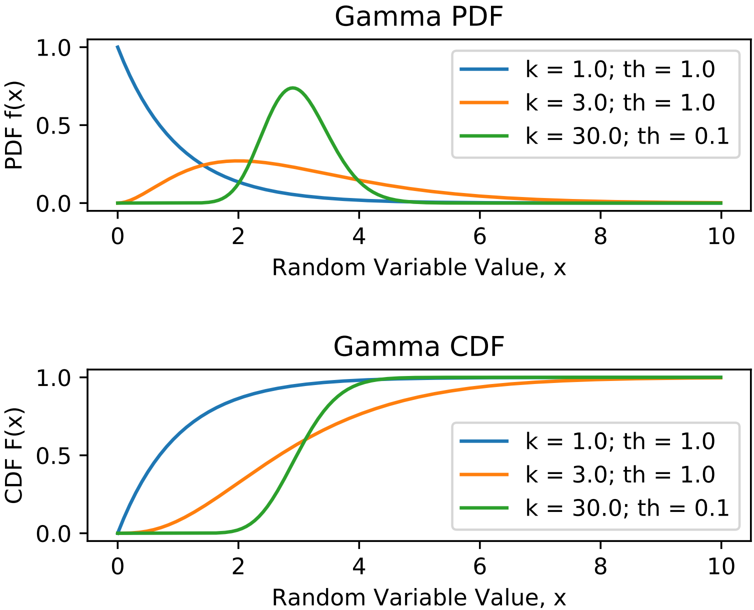
$$f(x) = \mathrm{Gamma}(x \,|\, \kappa, \theta) = \frac{1}{\Gamma(\kappa)\theta^{\kappa}}\, x^{\kappa-1} \mathrm{e}^{-x/\theta}, \tag{10}$$

  for positive $x$ and positive $\kappa, \theta$.
- The mean and variance are

$$\mathrm{Mean}(X) = \kappa\theta, \qquad \mathrm{Var}(X) = \kappa\theta^2. \tag{11}$$

- The special case $\kappa = 1$ is called the *exponential* distribution and it is the distribution of times between events for a *memoryless* process.

# Gamma PDFs and CDFs



# Normal

- The *Normal* (also called the *Gaussian*) distribution is continuous, with probability density function

$$f(x) = \mathcal{N}(x \,|\, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{12}$$

  for positive $\sigma$ but any $x$ and $\mu$.
- The mean and variance are

$$\mathrm{Mean}(X) = \mu, \qquad \mathrm{Var}(X) = \sigma^2. \tag{13}$$

- It appears throughout science, mainly due to the celebrated *Central Limit Theorem*, which says that averages tend to become normally distributed as the sample grows large.

# Arithmetic with normally-distributed variables

Suppose we have two random variables, $X_1$ and $X_2$ that are independent and are both normally distributed with means $\mu_1$ and $\mu_2$ and variances $\sigma_1^2$ and $\sigma_2^2$, respectively.

- A random variable $W$ defined by $W = X_1 + X_2$ will also be normally distributed, with mean and variance given by

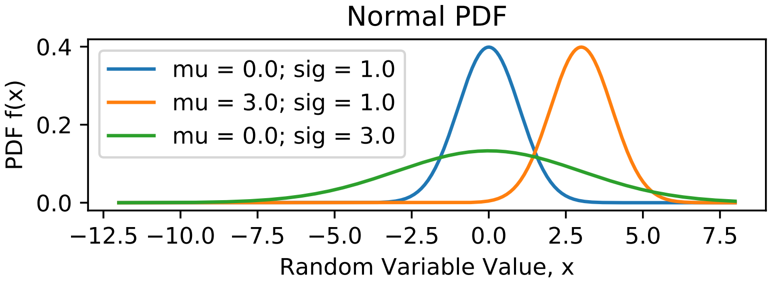
$$\mu_W = \mu_1 + \mu_2 \qquad \text{and} \qquad \sigma_W^2 = \sigma_1^2 + \sigma_2^2.$$

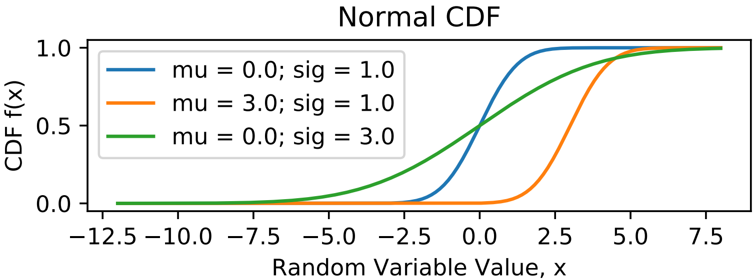
- A random variable $Y$ defined by $Y = aX_1 + b$ has will also be normally distributed, with mean and variance given by

$$\mu_Y = a\mu_1 + b \qquad \text{and} \qquad \sigma_Y^2 = a^2\sigma_1^2.$$

- A random variables defined by, for example, $X_1X_2$ or $X_2^3$ are *not* normally-distributed.

# Normal PDFs and CDFs



# Central Limits

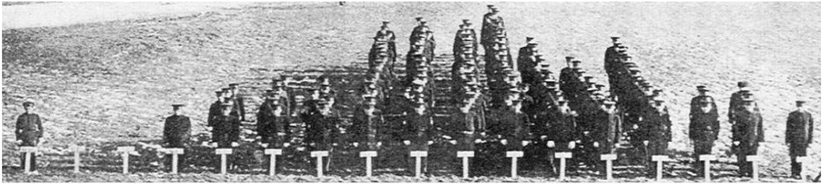
- Suppose we have a set of independent random variables $X_i$ for $i = 1, \ldots, n$ with

$$\mathrm{Mean}(X_i) = \mu, \qquad \mathrm{Var}(X_i) = V, \qquad \text{for all } i. \tag{14}$$

- Then as $n$ becomes large, the sum

$$S_n = \sum_{i=1}^{n} X_i \to \mathcal{N}(n\mu, nV) \tag{15}$$

  tends to become normally distributed.
- This (and generalisations) means that we see normal distributions everywhere, e.g. from Connecticut Agricultural College in 1914:



4:10 4:11 5:0 5:1 5:2 5:3 5:4 5:5 5:6 5:7 5:8 5:9 5:10 5:11 6:0 6:1 6:2

*Research Article*

## Common DNA Variants Accurately Rank an Individual of Extreme Height

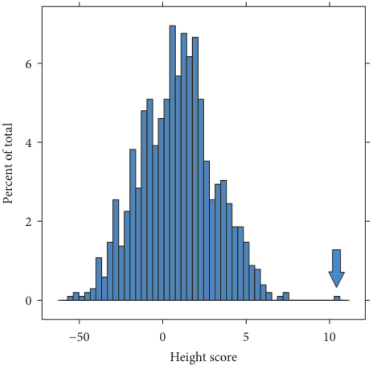



FIGURE 2: Height score distribution calculated using the 2910 SNPs. Mr. Bradley's height score (10.32, indicated by the arrow) ranked highest when compared to the 1020 individuals from ADNI and Cache County, while the next highest was 7.43. The mean height score within the ADNI and Cache County data was 0.98 with a standard deviation of 2.22, making Mr. Bradley's height score 4.2 standard deviations above the mean.

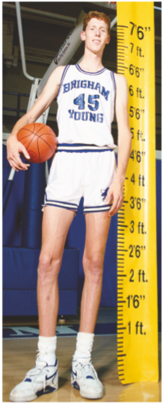
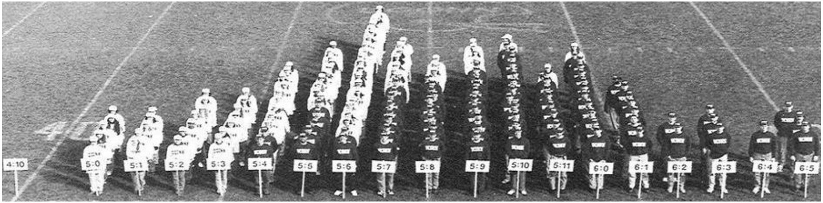



FIGURE 1: Shawn Bradley is 2.29 m (7′ 6″) tall with no known medical conditions. Mr. Bradley played basketball for Brigham Young University from 1990 to 1991. He played in the National Basketball Association from 1993–2005. Photo courtesy of BYU photography.


Hindawi
International Journal of Genomics
Volume 2018, Article ID 5121540, 7 pages
https://doi.org/10.1155/2018/5121540

# Absence of Central Limits

- Sometimes central limits do not hold.
- One example is when there are two underlying populations, for example Connecticut Agricultural College in 1996:



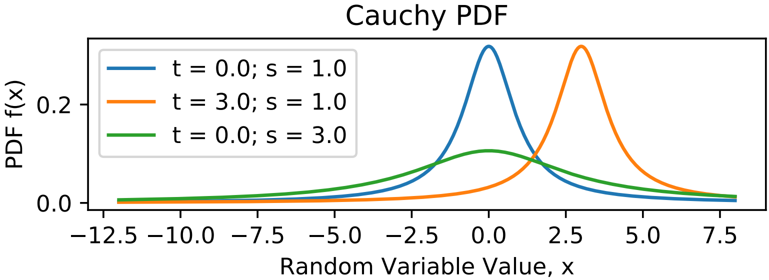

- Another case is where the moments are not defined / infinite.

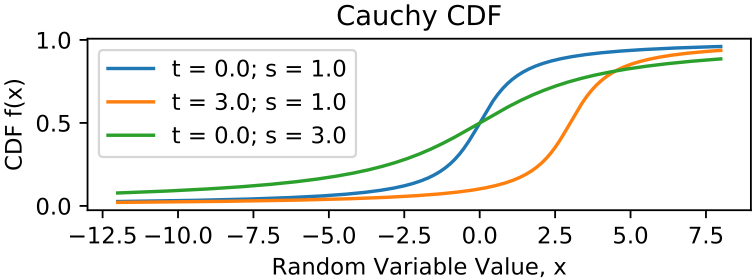
# Cauchy

- The *Cauchy* distribution has probability density function

$$f(x) = \frac{1}{\pi s \left(1 + ((x - t)/s)^2\right)} \tag{16}$$

  where $s$, which is positive, and $t$, which can be any number, are parameters.
- It has “heavy tails”, which means that large values are so common that the Cauchy distribution lacks a well-defined mean and variance!
- But the parameter $t$ gives the location of the mode and median, which are well-defined.
- The parameter $s$ determines the ‘width’ of the distribution as measured using e.g. the distances between percentiles, which are also well defined.


Cauchy PDF
t = 0.0; s = 1.0
t = 3.0; s = 1.0
t = 0.0; s = 3.0
PDF f(x)
0.2
0.0
−12.5 −10.0 −7.5 −5.0 −2.5 0.0 2.5 5.0 7.5
Random Variable Value, x


Cauchy CDF
t = 0.0; s = 1.0
t = 3.0; s = 1.0
t = 0.0; s = 3.0
CDF f(x)
1.0
0.5
0.0
−12.5 −10.0 −7.5 −5.0 −2.5 0.0 2.5 5.0 7.5
Random Variable Value, x