Review of Probability

Mark Craven and David Page
Computer Sciences 760

www.biostat.wisc.edu/~craven/cs760/

Goals for the lecture

you should understand the following concepts

- definition of probability
- random variables
- joint distributions
- conditional distributions
- independence
- union rule
- Bayes theorem
- expected values
- multinomial distribution
- probability density function
- normal distribution

Definition of probability

- *frequentist* interpretation: the probability of an event from a random experiment is the proportion of the time events of same kind will occur in the long run, when the experiment is repeated

- examples
  - the probability my flight to Chicago will be on time
  - the probability this ticket will win the lottery
  - the probability it will rain tomorrow

- always a number in the interval [0,1]

  0 means "never occurs"

  1 means "always occurs"

Sample spaces

- *sample space*: a set of possible outcomes for some event

- examples
  - flight to Chicago: {on time, late}
  - lottery: {ticket 1 wins, ticket 2 wins,…,ticket *n* wins}
  - weather tomorrow:

    {rain, not rain} or

    {sun, rain, snow} or

    {sun, clouds, rain, snow, sleet} or…

Random variables

- *random variable*: a variable representing the outcome of an event

- example
  - X represents the outcome of my flight to Chicago
  - we write the probability of my flight being on time as $P(X = \text{on-time})$
  - or when it's clear which variable we're referring to, we may use the shorthand $P(\text{on-time})$

Notation

- uppercase letters and capitalized words denote random variables
- lowercase letters and uncapitalized words denote values
- we'll denote a particular value for a variable as follows

$$P(X = x) \qquad P(Fever = true)$$

- we'll also use the shorthand form

$$P(x) \quad \text{for} \quad P(X = x)$$

- for Boolean random variables, we'll use the shorthand

$$P(fever) \;\; \text{for} \;\; P(Fever = true)$$

$$P(\neg fever) \;\; \text{for} \;\; P(Fever = false)$$

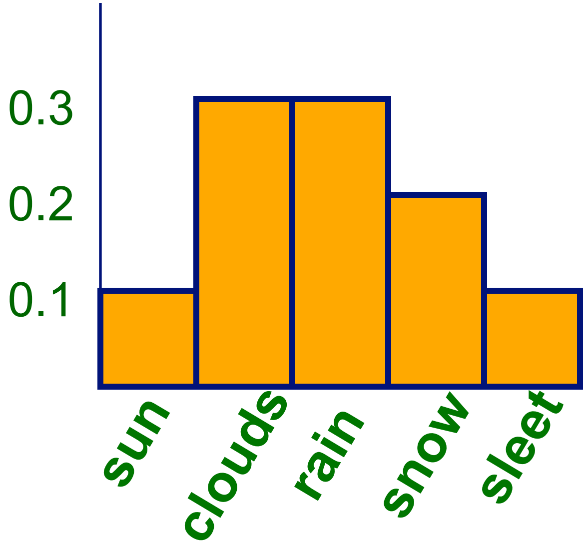
Probability distributions

- if X is a random variable, the function given by $P(X = x)$ for each x is the *probability distribution* of X

- requirements:

  $P(x) \geq 0$ for every x

  $$\sum_x P(x) = 1$$



Joint distributions

- *joint probability distribution*: the function given by $P(X = x, Y = y)$
- read "X equals x and Y equals y"
- example

x, y	$P(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

probability that it's sunny and my flight is on time

Marginal distributions

- the *marginal distribution* of X is defined by

$$P(x) = \sum_y P(x, y)$$

"the distribution of X ignoring other variables"

- this definition generalizes to more than two variables, e.g.

$$P(x) = \sum_y \sum_z P(x, y, z)$$

Marginal distribution example

joint distribution

x, y	$P(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

marginal distribution for X

x	$P(X = x)$
sun	0.3
rain	0.5
snow	0.2

Conditional distributions

- the *conditional distribution* of X given Y is defined as:

$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

"the distribution of X given that we know the value of Y"

- Rearranging yields *product rule*, that

$$P(X = x \mid Y = y)P(Y = y) = P(X = x, Y = y)$$

Chain Rule

Generalization of the product rule, derived by repeated application of product rule:

$$ChainRule : P(x_1,\ldots,x_n) =$$

$$P(x_n \mid x_{n-1},\ldots,x_1)P(x_{n-1} \mid x_{n-2},\ldots,x_1)\ldots P(x_2 \mid x_1)P(x_1)$$

$$= \prod P(x_i \mid x_{i-1},\ldots,x_1)$$

Conditional distribution example

joint distribution

x, y	$P(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

conditional distribution for X given Y=on-time

x	$P(X = x \mid Y = on\text{-}time)$
sun	0.20/0.45 = 0.444
rain	0.20/0.45 = 0.444
snow	0.05/0.45 = 0.111

Independence

- two random variables, X and Y, are *independent* if

$$P(x,y) = P(x) \times P(y) \quad \text{for all } x \text{ and } y$$

Conditional independence

- two random variables, X and Y, are *conditionally independent* given Z if

$$P(x, y \mid z) = P(x \mid z) \times P(y \mid z) \quad \text{for all } x, y \text{ and } z$$

Independence example #1

joint distribution

x, y	$P(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

marginal distributions

x	$P(X = x)$
sun	0.3
rain	0.5
snow	0.2

y	$P(Y = y)$
on-time	0.45
late	0.55

Are X and Y independent here? NO.

Independence example #2

joint distribution

x, y	$P(X = x, Y = y)$
sun, fly-United	0.27
rain, fly-United	0.45
snow, fly-United	0.18
sun, fly-Delta	0.03
rain, fly-Delta	0.05
snow, fly-Delta	0.02

marginal distributions

x	$P(X = x)$
sun	0.3
rain	0.5
snow	0.2

y	$P(Y = y)$
fly-United	0.9
fly-Delta	0.1

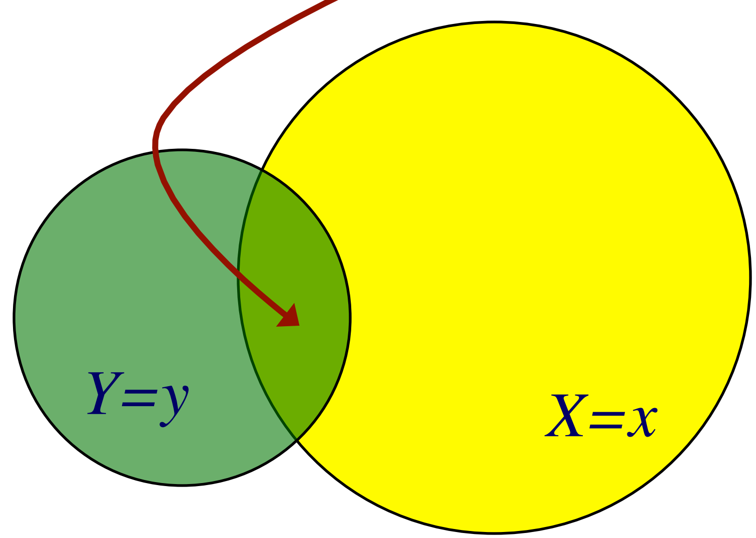
Are X and Y independent here? YES.

Probability of union of events

- the probability of the union of two events is given by:

$$P(x \vee y) = P(x) + P(y) - P(x,y)$$

this term needed to
avoid double counting



Bayes' Rule (or Theorem)

Recall product rule:

$$\mathrm{P}(a \wedge b) = \mathrm{P}(a|b)\,\mathrm{P}(b)$$

$$\mathrm{P}(a \wedge b) = \mathrm{P}(b|a)\,\mathrm{P}(a)$$

Equating right-hand sides and dividing by $\mathrm{P}(a)$:

$$\mathrm{P}(b|a) = \frac{\mathrm{P}(a|b)\,\mathrm{P}(b)}{\mathrm{P}(a)}$$

For multi-valued variables X and Y:

$$\mathbf{P}(Y|X) = \frac{\mathbf{P}(X|Y)\,\mathbf{P}(Y)}{\mathbf{P}(X)}$$

Why Use Bayes' Rule

- Causal knowledge such as P(*stiff neck*|*meningitis)* often is more reliable than diagnostic knowledge such as P*(meningitis*|*stiff neck*)*.*

- Bayes' Rule lets us use causal knowledge to make diagnostic inferences (derive diagnostic knowledge).

Normalization with Bayes' Rule

Might be easier to compute

$P(\mathit{stiff\ neck} \mid \mathit{meningitis})\ P(\mathit{meningitis})$ and

$P(\mathit{stiff\ neck} \mid \neg\, \mathit{meningitis})\ P(\neg\, \mathit{meningitis})$

than to directly estimate

$P(\mathit{stiff\ neck})$.

Bayes theorem with normalization

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{P(y)} = \frac{P(y \mid x)P(x)}{\sum_x P(y \mid x)P(x)}$$

- this theorem is extremely useful
- there are many cases when it is hard to estimate $P(x \mid \mathrm{y})$ directly, but it's not too hard to estimate $P(y \mid x)$ and $P(x)$

Bayes theorem example

- MDs usually aren't good at estimating $P(Disorder \mid Symptom)$
- they're usually better at estimating $P(Symptom \mid Disorder)$
- if we can estimate $P(Fever \mid Flu)$ and $P(Flu)$ we can use Bayes' Theorem to do diagnosis

$$P(flu \mid fever) = \frac{P(fever \mid flu)P(flu)}{P(fever \mid flu)P(flu) + P(fever \mid \neg flu)P(\neg flu)}$$

Another Example

- P(*stiff neck*|*meningitis*) = 0.5
- P(*meningitis*) = 1/50,000
- P(*stiff neck*) = 1/20
- Then P(*meningitis*|*stiff neck*) =

$$\frac{\mathrm{P}(\mathit{stiff\ neck}|\mathit{meningitis})\ \mathrm{P}(\mathit{meningitis})}{\mathrm{P}(\mathit{stiff\ neck})} =$$

$$\frac{(0.5)(1/50{,}000)}{1/20} = 0.0002$$

Expected values

- the *expected value* of a random variable that takes on numerical values is defined as:

$$E\left[X\right] = \sum_{x} x \times P(x)$$

this is the same thing as the *mean (*also written μ)

- we can also talk about the expected value of a function of a random variable

$$E\left[g(X)\right] = \sum_{x} g(x) \times P(x)$$

Variance

- $E[(X-\mu)^2]$

- $E[X^2] - (E[X])^2$

Expected value examples

$$E\left[Shoesize\right] =$$

$$5 \times P(Shoesize = 5) + \ldots + 14 \times P(Shoesize = 14)$$

- Suppose each lottery ticket costs \$1 and the winning ticket pays out \$100. The probability that a particular ticket is the winning ticket is 0.001.

$$E\left[gain(Lottery)\right] =$$

$$gain(\text{winning})P(\text{winning}) + gain(\text{losing})P(\text{losing}) =$$

$$(\$100 - \$1) \times 0.001 - \$1 \times 0.999 =$$

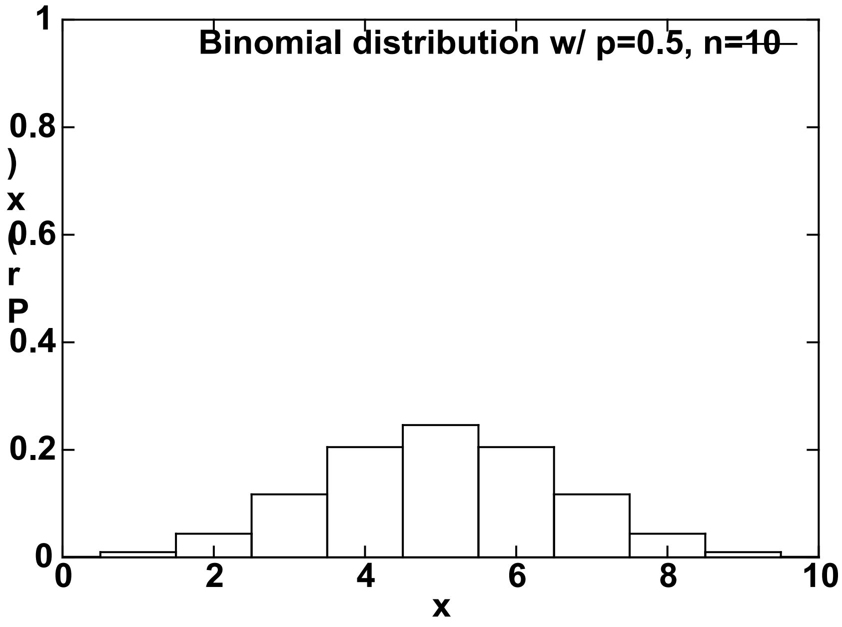
$$-\$0.90$$

The binomial distribution

- distribution over the number of successes in a fixed number n of independent trials (with same probability of success p in each)

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- e.g. the probability of x heads in n coin flips

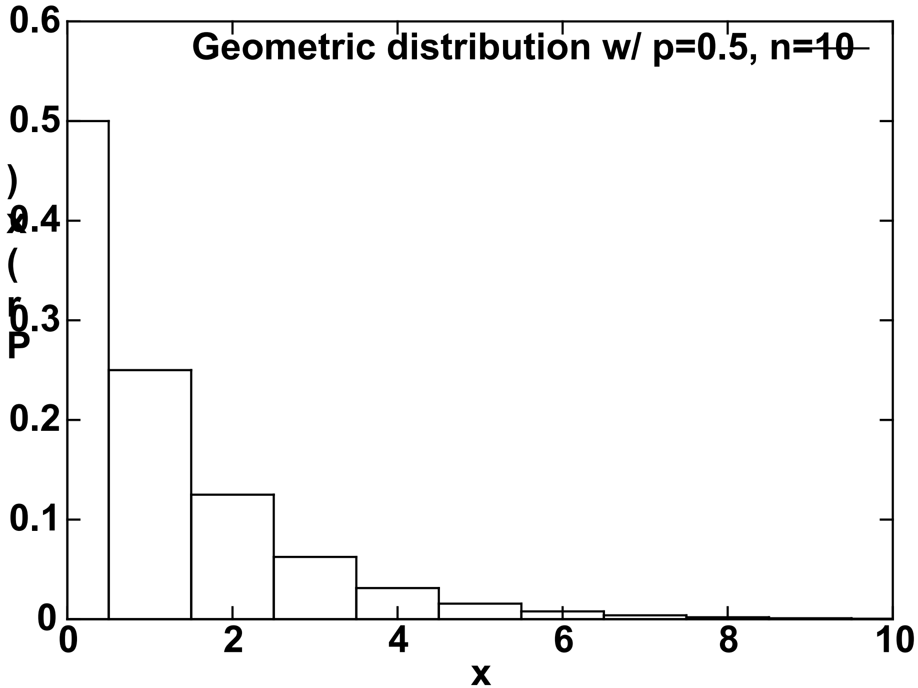


The geometric distribution

- distribution over the number of trials before the first failure (with same probability of success p in each)

$$P(x) = (1 - p)p^x$$

- e.g. the probability of x heads before the first tail



The multinomial distribution

- k possible outcomes on each trial
- probability p_i for outcome x_i in each trial
- distribution over the number of occurrences x_i for each outcome in a fixed number n of independent trials

vector of outcome occurrences

$$P(\mathbf{x}) = \frac{n!}{\prod_i (x_i!)} \prod_i p_i^{x_i}$$

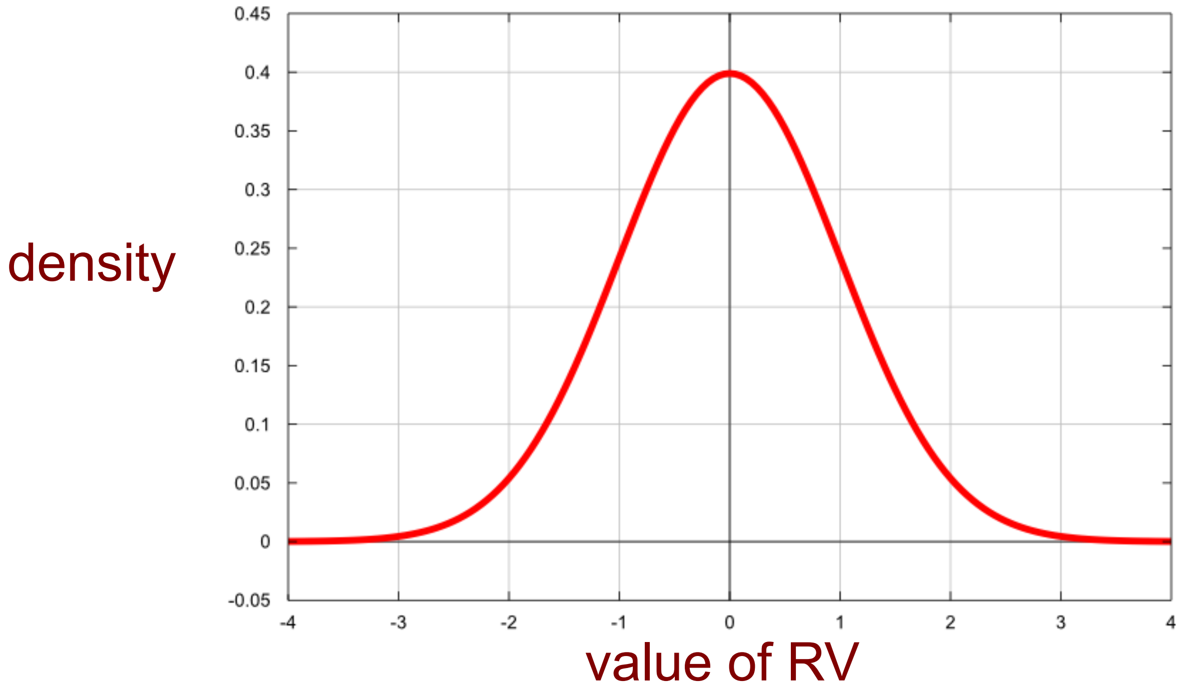
- e.g. with k=6 (a six-sided die) and n=30

$$P([7,3,0,8,10,2]) = \frac{30!}{7! \times 3! \times 0! \times 8! \times 10! \times 2!} \left(p_1^{\,7} p_2^{\,3} p_3^{\,0} p_4^{\,8} p_5^{\,10} p_6^{\,2} \right)$$

Continuous random variables

- up to now, we've considered only discrete random variables, but we can have RVs describing continuous variables too (weight, temperature, etc.)

- a continuous random variable is described by a *probability density function* (p.d.f.)



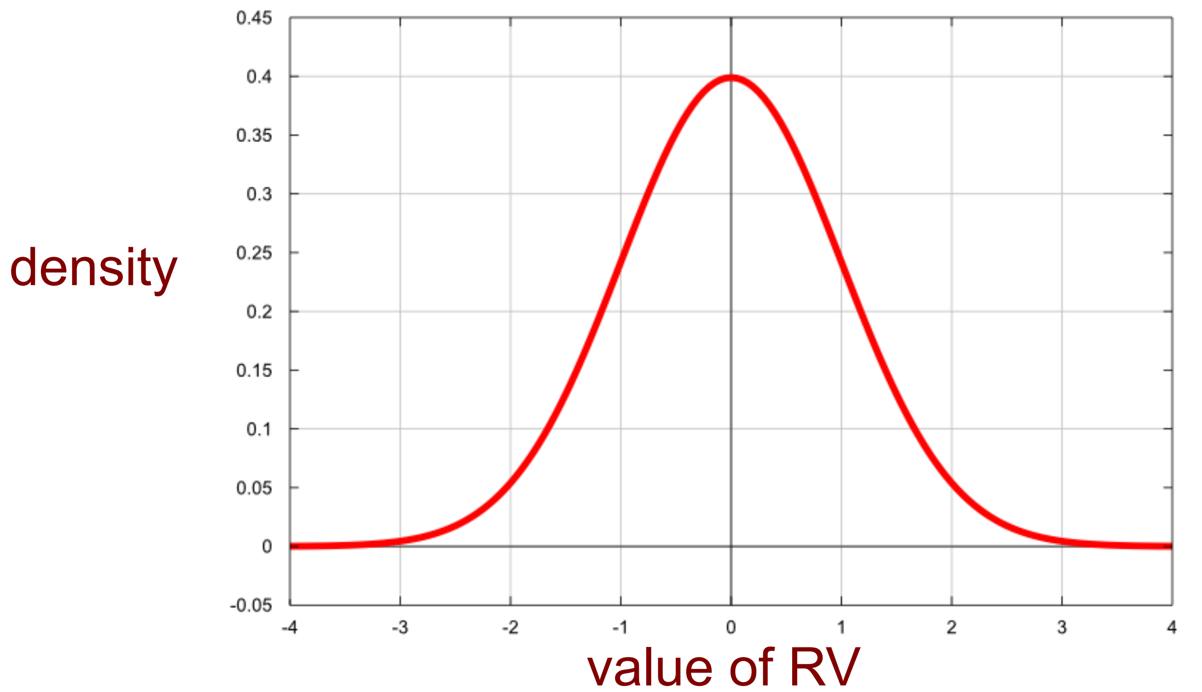
Probability density functions

- a continuous random variable is described by a *probability density function* $f(x)$

$$\forall x \;\; f(x) \geq 0$$

$$P\left[a \leq X \leq b\right] = \int_a^b f(x)\,dx$$

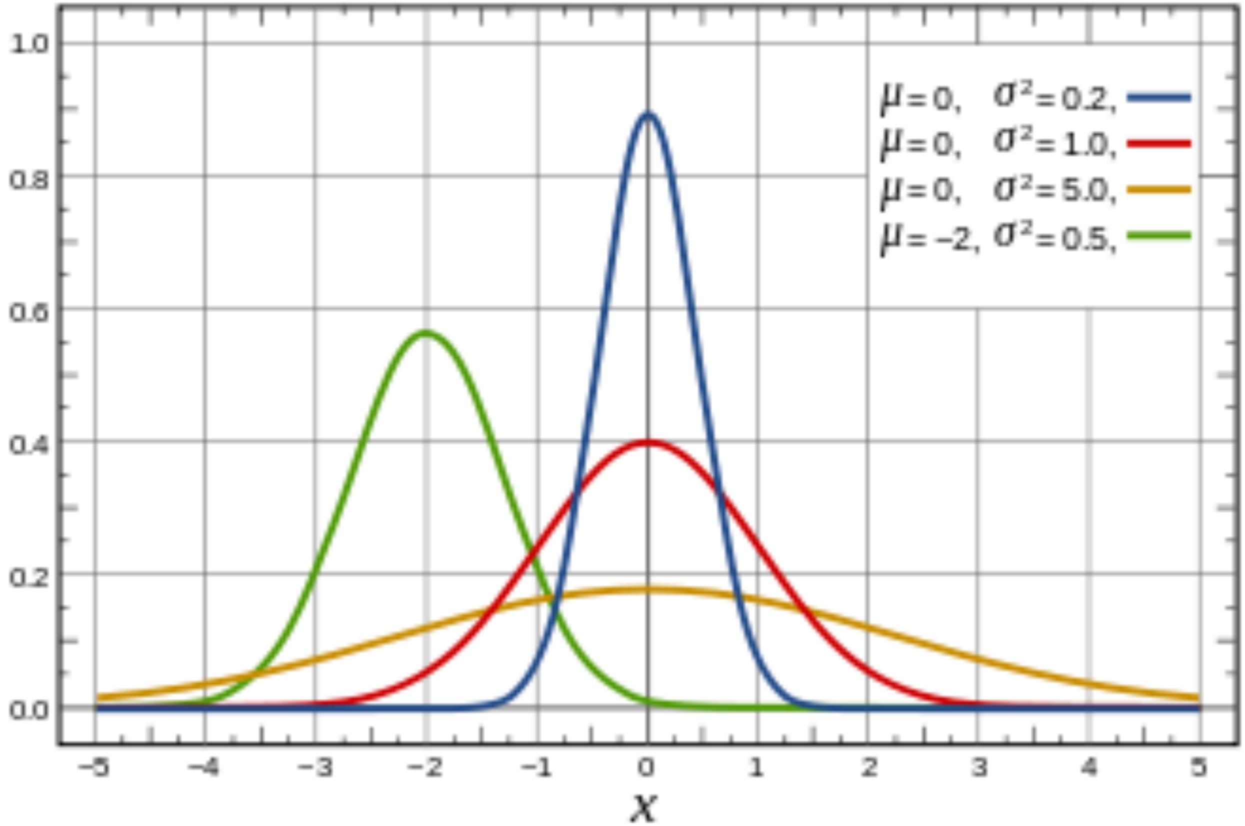
$$\int f(x)\,dx = 1$$



The normal (Gaussian) distribution

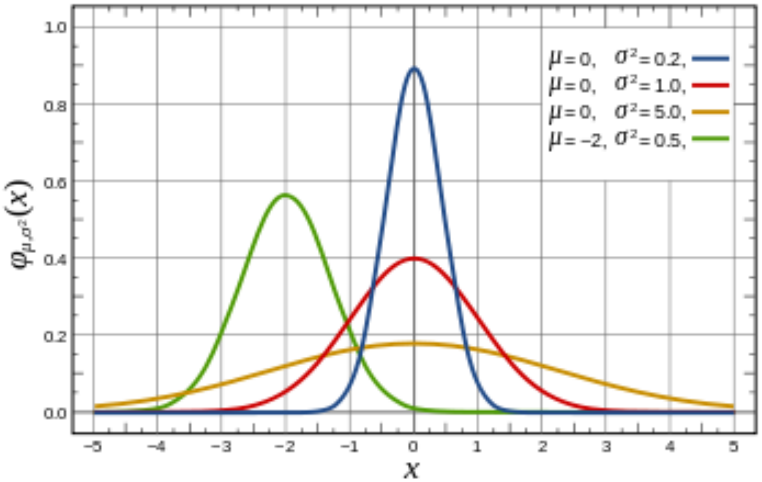
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$f(x)$

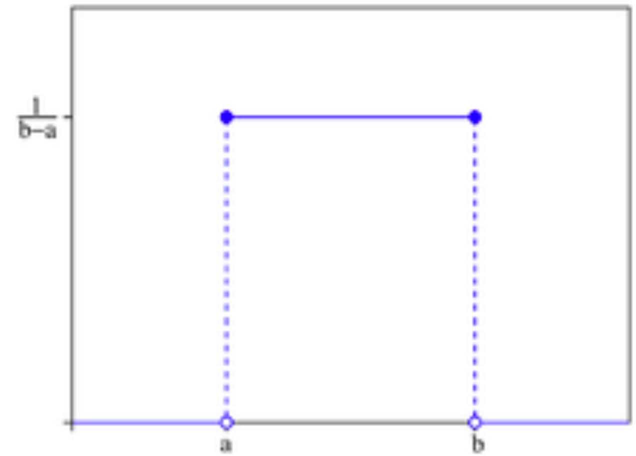


Some p.d.f.'s

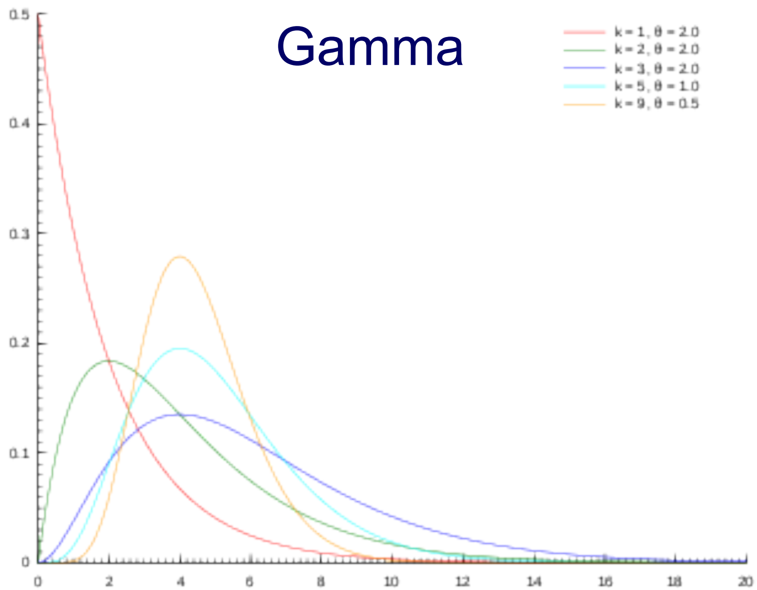
normal (Gaussian)



uniform



Gamma



student's *t*