STATISTICAL METHODS FOR THE PHYSICAL SCIENCES

Week 3: Random variables and probability distributions

Random variables: discrete

- Function that maps the sample space Ω on to real numbers which can further be mapped on to a probability.
- Simplest case to consider is a *discrete random variable.*
- E.g. tossing a coin: Ω = {*heads*, *tails*}. Define a variable X

$$X = \begin{cases} 0 & \text{if tails} \\ 1 & \text{if heads} \end{cases}$$

- Now we can write the probability that the variable *X* has a value *x*

$$p(x) = \Pr(X = x)$$

- So in this simple example: $p(0) = p(1) = 0.5$

- We can combine probabilities for discrete random variables, as for events. E.g. the probability that $X \leq x$ is:

$$F(x) = \Pr(X \leq x) = \sum_{x_i \leq x} \Pr(x_i)$$

This is the *cumulative distribution function* (cdf) for *X*

Random variables: continuous

- Here, we have a problem, since the probability of a specific value of infinite precision is zero (but is clearly not impossible!). So instead we must consider probability *ranges*. We define the cdf as:

$$F(x) = \Pr(X \leq x)$$

We can choose the limiting values of our distribution but note that it must satisfy:

$$\lim_{x \to -\infty} F(x) = 0 \quad \lim_{x \to +\infty} F(x) = 1$$

We can specify the probability that *X* falls in the range of values *a* to *b*:

$$\Pr(a \leq X \leq b) = F(b) - F(a)$$

- It is also useful to define the ***probability density function (pdf)***:

$$\frac{\Pr(x \leq X \leq x + \delta x)}{\delta x} = \frac{F(x + \delta x) - F(x)}{\delta x}$$

$$p(x) = \lim_{\delta x \to 0} \frac{\Pr(x \leq X \leq x + \delta x)}{\delta x} = \frac{\mathrm{d}F(x)}{\mathrm{d}x}$$

Probability density function properties

- It follows then that:

$$\Pr(X \leq x) = F(x) = \int_{-\infty}^{x} p(x')dx'$$

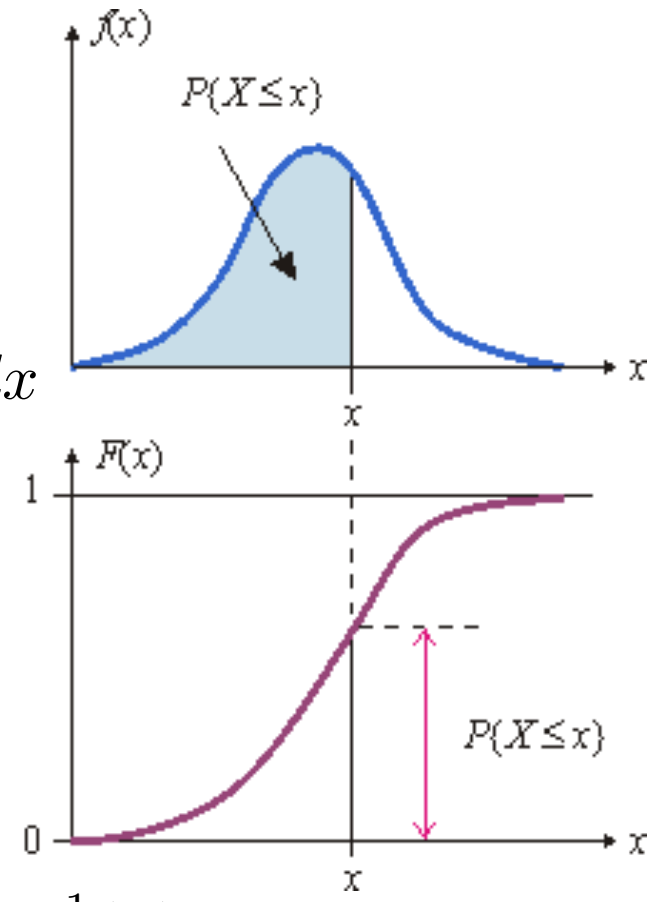
(where *x'* is a dummy variable)

- Also we have:

$$\Pr(a \leq X \leq b) = F(b) - F(a) = \int_{a}^{b} p(x)dx$$

which means that:

$$\int_{-\infty}^{+\infty} p(x)dx = 1$$



- We can also define quantiles α :

$$F(x_\alpha) = \int_{-\infty}^{x_\alpha} p(x)dx = \alpha \Longleftrightarrow x_\alpha = F^{-1}(\alpha)$$

(note that F^{-1} denotes the inverse function of *F*, not 1/*F*!)

Two random variables

- We can generalise to obtain the ***joint probability density function*** of two variables, *X* and *Y*:

$$p_{X,Y}(x,y) = \lim_{\delta x, \delta y \to 0} \frac{\Pr(x \leq X \leq x + \delta x \text{ and } y \leq Y \leq y + \delta y)}{\delta x \delta y}$$

- In general, the probability of *X* and *Y* having values in some region *R* is:

$$\Pr(X \, and \, Y \, in \, R) = \int \int_R p_{X,Y}(x,y) \mathrm{d}x \mathrm{d}y$$

- We can write the continuous form of ***conditional probability*** of *x*, given that *y* takes on a fixed value y_0:

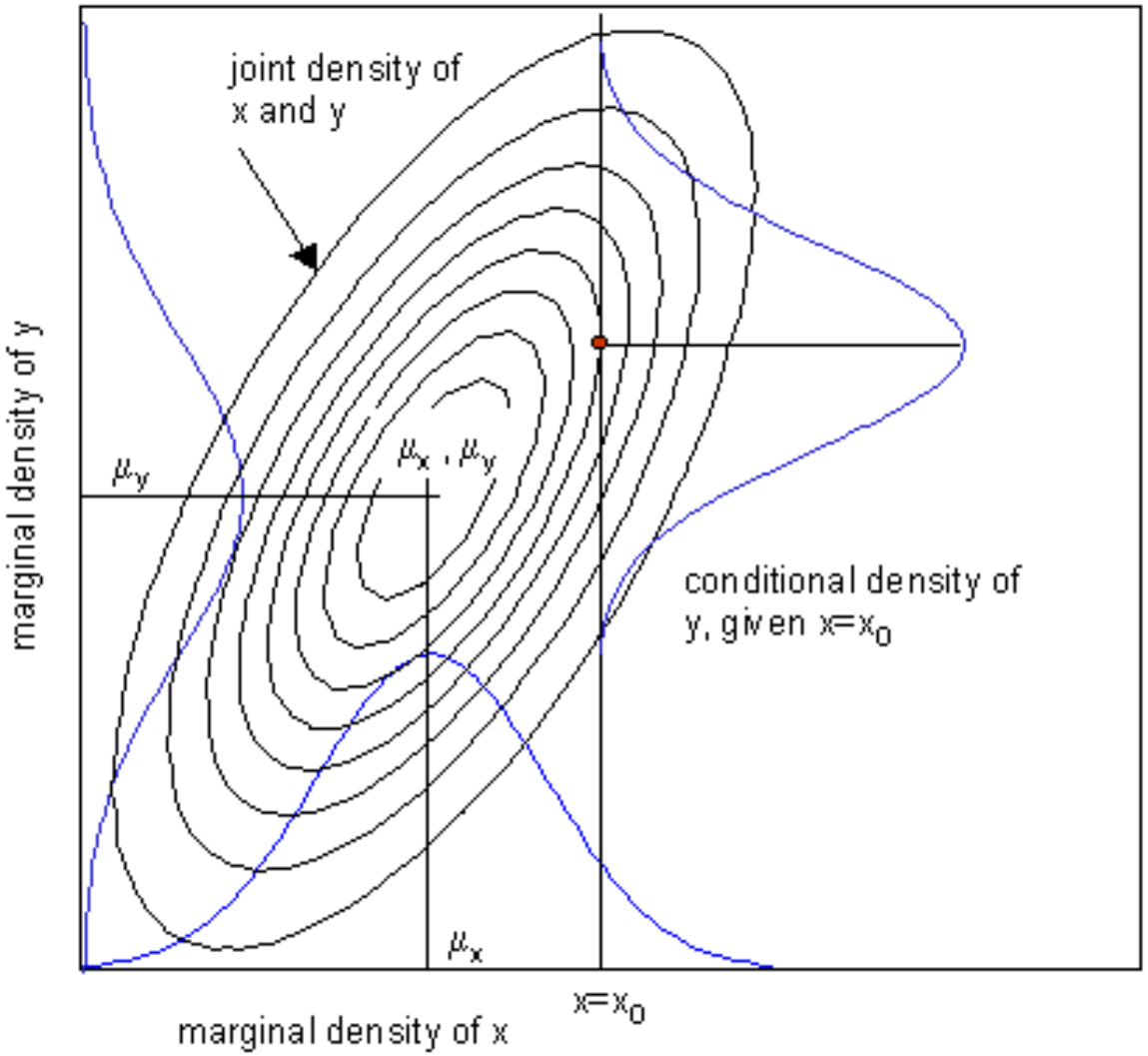
$$p(x|y_0) = \frac{p(x, Y = y_0)}{p(y_0)} = \frac{p(x, Y = y_0)}{\int p(x, Y = y_0) \mathrm{d}x}$$

- We can also write a continuous form of the ***law of total probability***:

$$p_X(x) = \int_{-\infty}^{+\infty} p_{X,Y}(x,y) \mathrm{d}y$$

This is the ***marginal pdf***, i.e. the pdf marginalised over conditional variable *y*

Two random variables: conditional and marginal distributions



Two random variables: multiplication rule and Bayes' theorem

- The ***multiplication rule*** can then be written as:

$$p(x, y) = p(y, x) = p(x|y)p(y) = p(y|x)p(x)$$

- With *X* and *Y* being *independent* if and only if:

$$p(x, y) = p(x)p(y)$$

- The marginal density of *X* can thus be rewritten :

$$p(x) = \int_{-\infty}^{\infty} p(x|y)p(y)\mathrm{d}y$$

- Which leads us to the continuous form of Bayes' theorem:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int_{-\infty}^{\infty} p(x|y)p(y)\mathrm{d}y}$$

Properties of random variables: expectation

- For a discrete random variable *X*, the expectation is equal to the sum of possible values weighted by their probabilities:

$$\mathrm{E}[X] = \sum_{i=1}^{n} x_i p(x_i)$$

This is closely related to the sample mean (consider a histogram of measured values of *X*)

- For the continuous case:

$$\mathrm{E}[X] = \int_{-\infty}^{+\infty} x p(x) \mathrm{d}x$$

Expectation of a continuous variable is usually called the mean of the distribution or population mean, μ

- Even more generally, we can obtain the expectation of some function of *X*, *f(X)*:

$$\mathrm{E}[f(X)] = \int_{-\infty}^{+\infty} f(x) p(x) \mathrm{d}x$$

- It follows that the expectation is a linear operator. So we can also consider the expectation of a scaled sum of variables:

$$\mathrm{E}[aX + bY] = a\mathrm{E}[X] + b\mathrm{E}[Y]$$

Properties of random variables: variance

- The variance of a discrete variable *X* with mean *μ* is:

$$\mathrm{V}[X] = \sigma_x^2 = \mathrm{E}[(X-\mu)^2] = \sum_{i=1}^{n}(x_i-\mu)^2 p(x_i)$$

- And in the continuous case:

$$\mathrm{V}[X] = \sigma_x^2 = \mathrm{E}[(X-\mu)^2] = \int_{-\infty}^{\infty}(x-\mu)^2 p(x)\mathrm{d}x$$

- We can rearrange things (similar to what we can do with the sample variance):

$$\begin{aligned}\mathrm{V}[X] &= \mathrm{E}[(X-\mu)^2] = \mathrm{E}[X^2 - 2X\mu + \mu^2] \\ &= \mathrm{E}[X^2] - \mathrm{E}[2X\mu] + \mathrm{E}[\mu^2] = \mathrm{E}[X^2] - 2\mu^2 + \mu^2 \\ &= \mathrm{E}[X^2] - \mu^2\end{aligned}$$

i.e. the variance is: ***expectation of squares – square of expectations***

- Thus for a function of *X*:

$$\mathrm{V}[f(X)] = \mathrm{E}[f(X)^2] - \mathrm{E}[f(X)]^2$$

Moments and skewness

- Much more generally, we can define the *n*th moment of a random variable *X* as:

$$\mu'_n = \mathrm{E}[X^n]$$

(e.g. the mean is the 1st moment, while the variance is the 2nd central moment)

- And the *n*th central moment as:

$$\mu_n = \mathrm{E}[(X - \mu)^n]$$

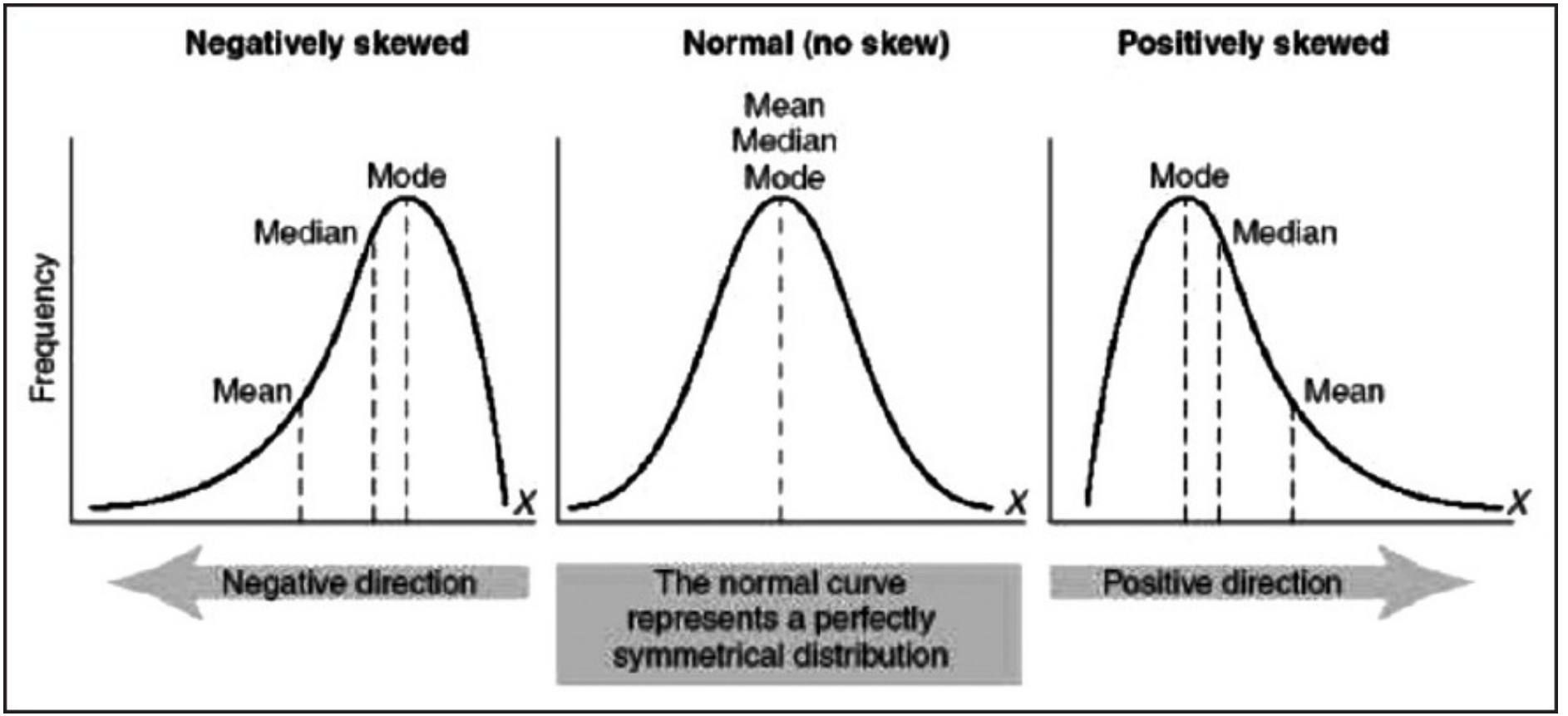
- We shall consider one more moment: the 3rd central moment, or skewness. Typically we normalise it by the standard deviation, so that the moment coefficient of skewness of *X* is:

$$\mathrm{Skew}[X] = \gamma_1 = \mathrm{E}\left[\left(\frac{X - \mu}{\sigma_x}\right)^3\right] = \frac{\mathrm{E}[(X - \mu)^3]}{(\mathrm{E}[(X - \mu)^2])^{3/2}}$$

- We can also define the sample skewness (i.e. for actual data):

$$b_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{3/2}}$$

What does skewness mean?



Properties of multivariate distributions

- We can expand our treatment of the properties of random variables to look at those drawn from multivariate distributions:

$$\mu_x = \mathrm{E}[X] = \int_{-\infty}^{+\infty} x p(x) \mathrm{d}x = \int_{-\infty}^{+\infty} x \int_{-\infty}^{+\infty} p(x, y) \mathrm{d}y \mathrm{d}x$$

- Where we have used our earlier result for the marginal distribution $p(x)$. We can generalise the approach to other quantities, e.g.:

$$\sigma_x^2 = \mathrm{V}[X] = \mathrm{E}[(X - \mu_x)^2] = \int_{-\infty}^{+\infty} (x - \mu_x)^2 \int_{-\infty}^{+\infty} p(x, y) \mathrm{d}y \mathrm{d}x$$

- We can also consider combinations of different functions of X and Y:

$$\mathrm{E}[af(X) + bg(Y) + c] = a\mathrm{E}[f(X)] + b\mathrm{E}[g(Y)] + c$$

Covariance

- We can also define the covariance of *X* and *Y*:

$$\mathrm{Cov}(X,Y) = \sigma_{xy} = \mathrm{E}[(X-\mu_x)(Y-\mu_y)] = \mathrm{E}[XY] - \mu_x\mu_y$$

- Note that the *Cauchy-Schwarz inequality* states that:

$$|\sigma_{xy}|^2 \le \sigma_x^2\sigma_y^2 \qquad \text{or} \qquad |\mathrm{Cov}(X,Y)|^2 \le \mathrm{V}[X]\mathrm{V}[Y]$$

- The correlation coefficient is:

$$\rho(X,Y) = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

By the Cauchy-Schwarz inequality, this must lie in the range: $-1 \le \rho(X,Y) \le 1$

- For independent variables *X* and *Y* we have:

$$\mathrm{E}[XY] = \mu_x\mu_y$$ i.e. covariance is zero

- For multivariate distributions, E.g. for *X*, *Y*, *Z* we can expand the covariance to measure a ***covariance matrix***.

$$\Sigma = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{pmatrix}$$

The *leading diagonal* contains the variances.
The matrix is symmetric about the leading diagonal

Linear functions of random variables

- Now we consider scaled linear combinations of random variables.

$$Y = a_1 X_1 + a_2 X_2 + \cdots a_n X_n = \sum_{i=1}^{n} a_i X_i$$

We simply state these results here (see Vaughan for deeper discussion/proof):

$$\mathrm{E}[Y] = \mathrm{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i \mathrm{E}[X_i] = \sum_{i=1}^{n} a_i \mu_i$$

$$\mathrm{V}[Y] = \sum_{i=1}^{n} a_i^2 \sigma_i^2 + \sum_{i=1}^{n} \sum_{j \neq i}^{n} a_i a_j \sigma_{ij}^2$$

If the X_i are mutually uncorrelated we have: $\mathrm{V}[Y] = \sum_{i=1}^{n} a_i^2 \sigma_i^2$

- E.g., from these, we can derive the results for sample mean and standard error…

Distributions of discrete random variables: Bernoulli distribution

Jacob Bernoulli (1655-1705)

- Imagine drawing randomly (with replacement) a single sample with two possibilities, e.g. red or green sweets from a bag. This is called a *Bernoulli trial*. We can assign a variable x = 1 or 0 to describe the (binary) outcome and a probability θ:

$$p(x) = \begin{cases} \theta & \text{for } x = 1 \\ 1 - \theta & \text{for } x = 0 \end{cases}$$

- The corresponding Bernoulli distribution function can be written as:

$$p(x|\theta) = \theta^x (1-\theta)^{1-x} \quad \text{for } x = 0, 1$$

- If a variable X follows the Bernoulli distribution we can write:

$$X \sim \text{Bern}(\theta)$$

(the tilde '~' means 'is distributed as')

It can be shown that this variable has:

$$\text{E}(X) = \theta \qquad \text{V}(X) = \theta(1-\theta)$$

Distributions of discrete random variables: Binomial distribution

- What if we have repeated Bernoulli trials? Since we sample with replacement, the probabilities in successive draws are independent, so we can simply multiply probabilities for a single draw:

$$\Pr(red)\Pr(red)\Pr(green) = \theta^2(1-\theta)$$

Which is the probability for getting red, red, then green sweets (in that order).

- What if we don't care about the order, just about the probability of getting a certain number of red (or green) sweets in the total sample of trial results? Probability to get *x* 'successes' (with probability *θ)* from *n* trials:

$$p(x|n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \frac{n!}{(n-x)!x!}\theta^x(1-\theta)^{n-x}$$

Binomial coefficient

$$(\text{where } x = 0, 1, 2, \ldots, n)$$

For $X \sim \text{Binom}(n,\theta)$ $\quad \text{E}[X] = n\theta \quad \text{V}[X] = n\theta(1-\theta)$

Distributions of discrete random variables: Poisson distribution

- Now imagine detecting photons with a mean rate per time interval, λ. Split the interval into n sub-intervals. Each sub-interval is equivalent to a trial with probability of success $\theta = \lambda/n$. Thus we can rewrite the binomial distribution:

$$p(x|n, \lambda/n) = \frac{1}{x!} \frac{n!}{(n-x)!} \frac{\lambda^x}{n^x} \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

  - In reality, the distribution of possible arrival times in an interval is continuous with infinite possibilities, so we can take the limit $n \to \infty$

The 2nd term $\to n^x$ and we use: $\mathrm{e}^x = \lim_{n\to\infty} (1 + x/n)^n$

So: $\lim_{n\to\infty} (1 - \lambda/n)^{n-x} = \lim_{n\to\infty} (1 - \lambda/n)^n = (\mathrm{e}^{-1})^\lambda = \mathrm{e}^{-\lambda}$

And we get the Poisson distribution: $p(x|\lambda) = \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!}$

For $X \sim \mathrm{Pois}(\lambda)$ $\mathrm{E}[X] = \lambda$ $\mathrm{V}[X] = \lambda$

Distributions of continuous random variables: Normal (Gaussian) distribution

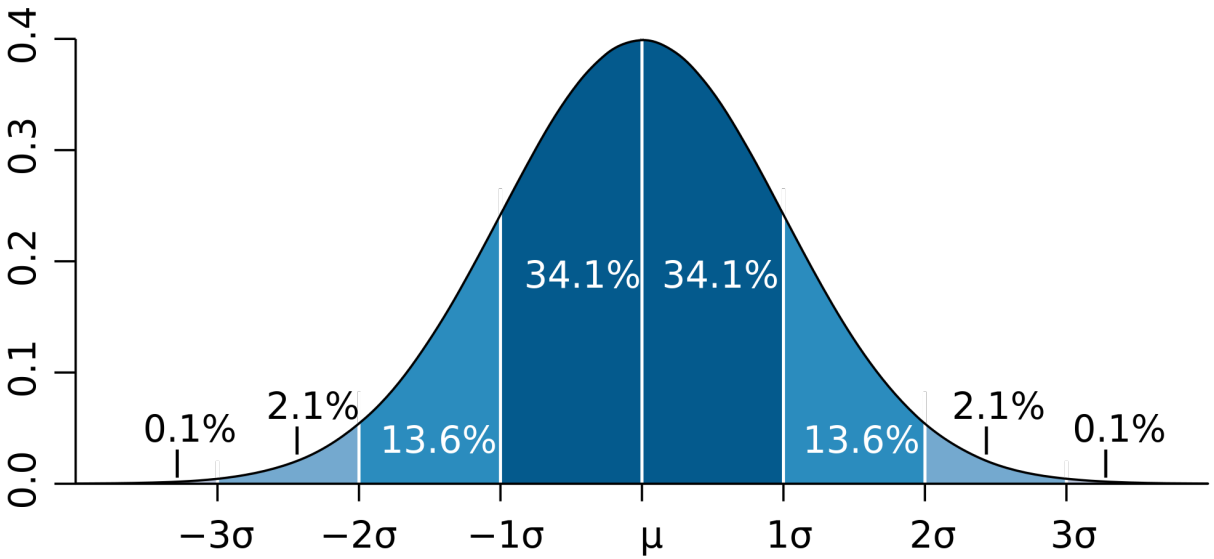
The normal (or Gaussian) distribution:

$$p(x|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-(x-\mu)^2/(2\sigma^2)}$$

The ***standard normal*** distribution:

$$p(z|0,1) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-z^2/2}$$

For $X \sim N(\mu,\sigma^2)$ $\quad \mathrm{E}[X] = \mu$ $\quad \mathrm{V}[X] = \sigma^2$



The Central Limit Theorem: where the magic happens

- Consider a sequence of *n* random variables X_i, each from a distribution with mean μ_i and variance σ^2_i
- From our expressions for linear combinations of variables, we can see that the sum:

$$Y = \sum X_i$$

has a mean: $\sum \mu_i$ and variance: $\sum \sigma_i^2$

- The ***central limit theorem*** (CLT) states that the distribution of *Y* becomes more normal as *n* increases. This is quite general and regardless of the distributions of X_i.
- Thus a sample mean formed from a large number of values can usually be assumed to be distributed close to normal, and many well-known statistical results can be applied….
- Binomial and Poisson distributions also converge on normal for large *n* or λ respectively

Distributions of continuous random variables: Chi-squared distribution

- Consider a set of independent standard normal variables [*N*(0,1)]:

$$X_1, X_2, \ldots, X_n$$

- We form a new variable by squaring and summing our normal variables:

$$X = \sum_{i=1}^{n} X_i^2$$

- The resulting variable is said to have a χ^2 (chi-squared) distribution:

$$p(x|\nu) = \frac{(1/2)^{\nu/2}}{\Gamma(\nu/2)} x^{\nu/2-1} \mathrm{e}^{-x/2}$$

Gamma function

ν '*degrees of freedom*' = *n* if we sum n independent squared standard normal variables

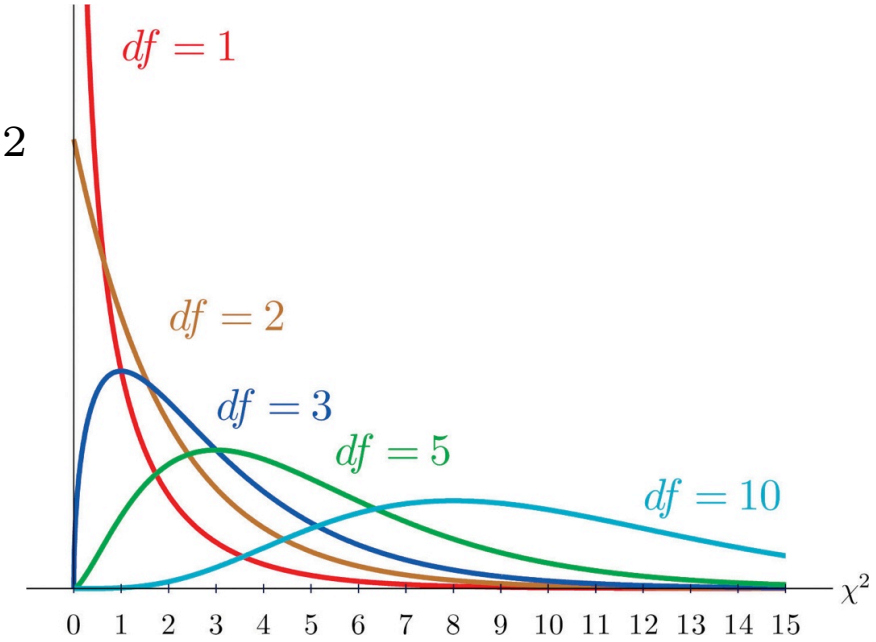
For $X \sim \chi^2_n$ $\quad \mathrm{E}[X] = \nu$ $\quad \mathrm{V}[X] = 2\nu$

Chi-squared distribution examples

$$\Gamma(1/2) = \sqrt{\pi},\ \Gamma(1) = 1,\ \Gamma(3/2) = \sqrt{\pi}/2,\ \Gamma(2) = 1\ldots$$
$$\Gamma(m) = (m-1)!$$ where m is a positive integer

So: $\chi_1^2 \to p(x) = \dfrac{1}{\sqrt{2\pi}} x^{-1/2} \mathrm{e}^{-x/2}$

$$\chi_2^2 \to p(x) = \frac{1}{2} \mathrm{e}^{-x/2}$$



- Sample variance is a sum of squared deviations, so is chi-squared distributed if these are normal and weighted (e.g. by dividing by errors) to be standard normal – chi-squared is an important statistic for model-fitting
- Kinetic energies of particles in thermal equilibrium

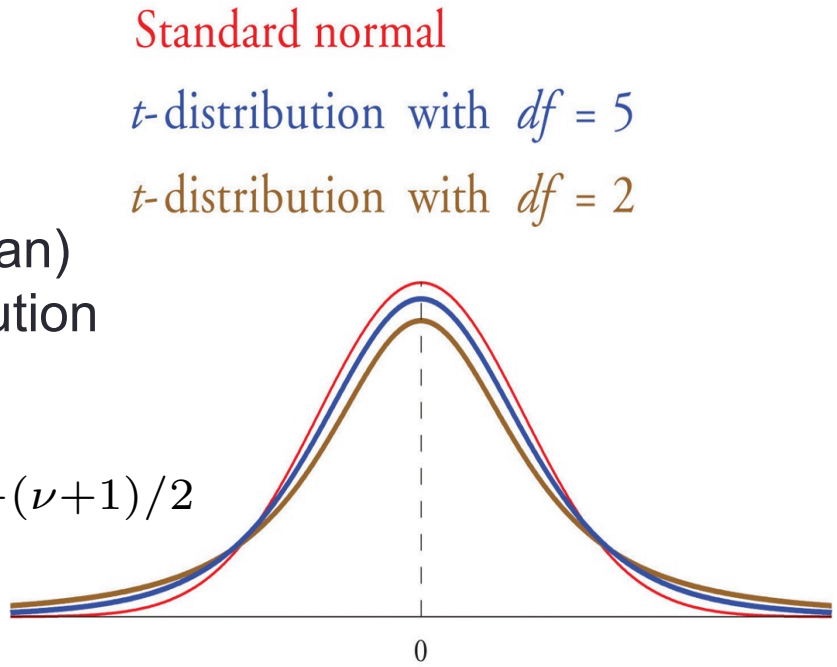
Distributions of continuous random variables: Student's t-distribution

- Consider variables which are normally distributed with mean μ and variance σ^2: $X_1, X_2, \ldots, X_n$
- Their mean is therefore also normally distributed. If we also take the sample variance s_x^2 we can obtain the t-statistic:

$$t = \frac{\overline{X} - \mu}{\sqrt{s_x^2/n}}$$

- If μ is correct (i.e. the true population mean) then t is distributed as Student's t-distribution with $\nu = n - 1$ degrees of freedom:

$$p(x|\nu) = \frac{\Gamma\left([\nu+1]/2\right)}{\sqrt{\nu\pi}\Gamma(\nu/2)}(1 + x^2/\nu)^{-(\nu+1)/2}$$



For $X \sim t_\nu$ $\quad \mathrm{E}[X] = 0$ $\quad \mathrm{V}[X] = \nu/(\nu - 2)$

Distributions of continuous random variables: Uniform distribution

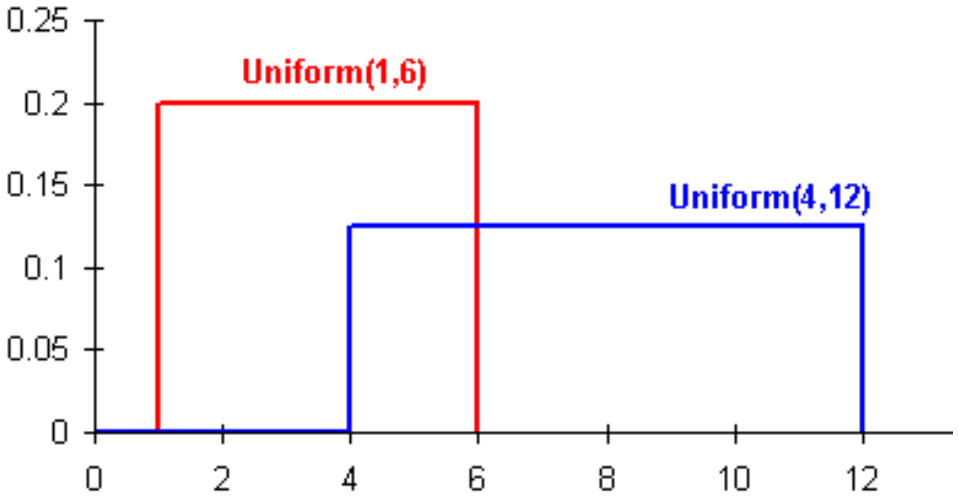
- A uniform distribution has equally probable values over some finite range [*a*, *b*]:

$$p(x|a,b) = 1/(b-a) \text{ for } a \leq x \leq b$$

For $X \sim U(a,b)$

$$\mathrm{E}[X] = (b-a)/2$$

$$\mathrm{V}[X] = (b-a)^2/12$$



Recall the relationship between the cdf and the value of a variable corresponding to a quantile *α*:

$$F(x_\alpha) = \int_{-\infty}^{x_\alpha} p(x)dx = \alpha \Longleftrightarrow x_\alpha = F^{-1}(\alpha)$$

Since *α* lies between 0 and 1, random numbers drawn from *U*(0,1) can be used to randomly select variables for any invertible distribution!