Statistical Distributions and what you can do with them

INFN Statistics School, Ischia,

Roger Barlow

Huddersfield University

7th May 2017

Contents

- General properties
- The main distributions: Poisson and Gaussian
- A quick look at lot of others
  with one vital fact per distribution
- From Random Numbers to Random distributions

Some General Properties

Random variable: integer (usually called r) or real (usually called x)
P_r is probability of r. Dimensionless numbers. $\sum P_r = 1$
$P(x)$ is probability density for x. $[P(x)] = [x]^{-1}$. $\int P(x)\,dx = 1$
Expectation values $< f >= \sum f(r) P_r$ or $\int f(x) P(x)\,dx$

Measures of Location

Mean: $\mu =< x >$

Mode: $P(mode) = max(P(x))$

Median: $\int^{median} P(x)\,dx = 0.5$

Measures of Scale

$\sigma = \sqrt{< (x-\mu)^2 >} = \sqrt{< x^2 > - < x >^2}$

FWHM=Full Width Half Max

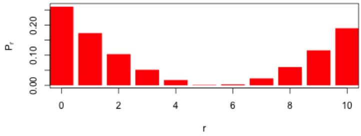
Inter-quartile range

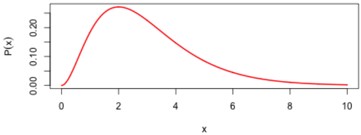
Other stuff

Skew: $\gamma = \frac{<(x-\mu)^3>}{\sigma^3}$

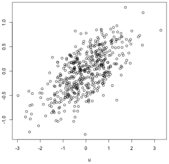
Kurtosis: $K = \frac{<(x-\mu)^4>}{\sigma^4} - 3$

Moments: $M_N =< x^N >, \mu_N =< (x-\mu)^N >$





More than one variable: Joint distributions



Two variables

Covariance $Cov(x, y) = < xy > - < x >< y >$

Correlation $\rho = \frac{Cov(x,y)}{\sigma_x \sigma_y}$

Several variables

Covariance $C_{ij} = < x_i x_j > - < x_i >< x_j >$

Correlation $\rho_{ij} = \frac{C_{ij}}{\sigma_i \sigma_j}$

Diagonals: $C_{ii} = \sigma_i^2$, $\rho_{ii} = 1$

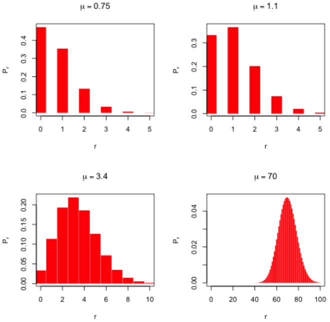
Can be shown that: $|\rho| \leq 1$

The Poisson

Memoryless random source. Mean number μ. Actual number r

$$P(r;\mu) = e^{-\mu}\frac{\mu^r}{r!}$$

Classic example: Geiger counter clicks
Also: Prussian soldiers killed by horses. Photomultipliers. Rare decays
Counterexamples: photons from lasers. Traffic (especially buses).



Vital fact: $\sigma = \sqrt{\mu}$
Small μ: 0 is mode
$\mu > 1$: peak develops
Distribution has positive skew - tail to high values
Large μ: shape becomes Gaussian

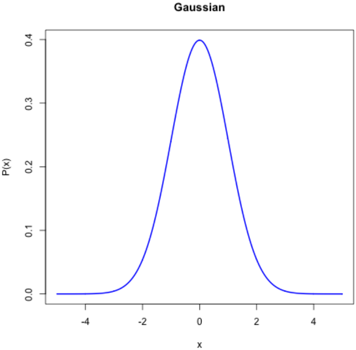
The Gaussian

or Normal Distribution

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

(Inaccurately) called the 'Bell curve'

μ is mean and mode and median
σ is standard deviation
68.27% of area within 1 σ
so 1/3 of error bars should miss!
95.45% of area within 2 σ
99.73% of area within 3 σ

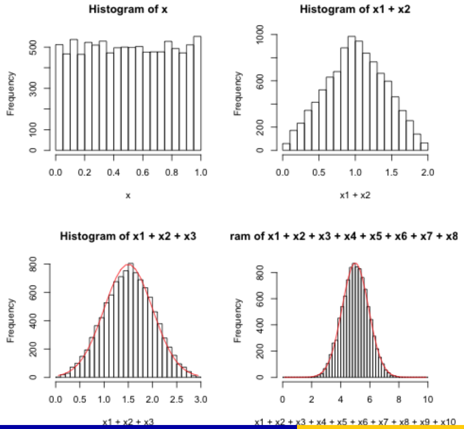


Describes: large μ Poisson, measurement errors, height, IQ...

Vital fact: Thanks to **Central Limit Theorem**: convolution of N random variables $P(x)$ tends to Gaussian for large N, irrespective of $P(x)$.

Demonstrating the CLT

Exercise: using your favourite package (ROOT, Python, Matlab, R, whatever) generate many uniform random numbers and histogram them. Get flat plot, very non-Gaussian. Then generate pairs and add them - get triangular shape. Then triples. Then tens, Looks pretty Gaussian...



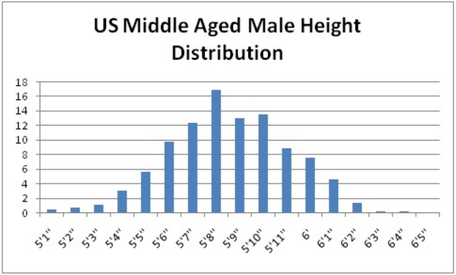
Central Limit Theorem: the proof

Optional: skip this slide if you're lazy or stupid...

Show: if you convolute $P(x)$ with itself $N(\to \infty)$ times you get a Gaussian
Given $P(x)$, Fourier Transform is $\tilde{P}(k) = \int P(x) e^{ikx}\, dx =< e^{ikx} >$
Expand and separate: $1 + ik < x > + \frac{(ik)^2}{2!} < x^2 > + \frac{(ik)^3}{3!} < x^3 > ...$
Take the logarithm, and use $ln(1+z) = z - \frac{z^2}{2} + \frac{z^3}{3} ...$
Get series in k: $ln\tilde{P}(k) = (ik)\kappa_1 + \frac{(ik)^2}{2!}\kappa_2 + \frac{(ik)^3}{3!}\kappa_3 + ...$ where the κ_r ("semi-invariant cumulants of Thiele") are made of expectation values of x to the r^{th} power. $\kappa_1 =< x >= \mu$,$\kappa_2 =< x^2 > - < x >^2 = \sigma^2$, etc
Semi-invariant? Location only changes κ_1, scaling by factor α, $\kappa_r \to \alpha^r \kappa_r$
Fact: The FT of a convolution is the product of the individual FTs.
So the log of the FT of a convolution is the sum of the logs and $K_r = N\kappa_r$.
To discuss shape, scale by standard deviation $\sqrt{K_2}$
$K_2' = 1$, $K_r' = K_r / \sqrt{K_2}^r = N\kappa_r/(N\kappa_2)^{r/2}$, vanishes as $N \to \infty$ for $r > 2$.
So in the large N limit all K_r with $r \geq 3$ vanish, the log of the FT is quadratic: the FT itself is the exponential of a quadratic, i.e. a Gaussian.
Transforming, the (back) FT of a Gaussian is also a Gaussian. QED.

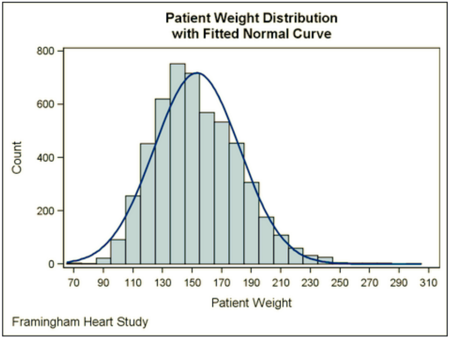
Real world Gaussians(1)

Distribution of heights
Nice Gaussian distribution



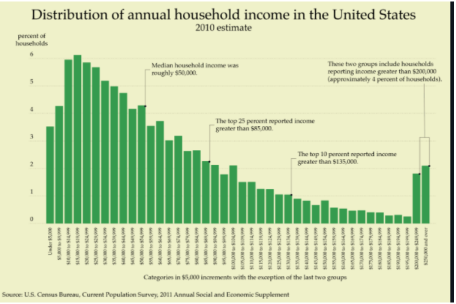
Real world Gaussians(2)

Distribution of weights .Not really Gaussian. Definite positive skew.



Real world Gaussians(3)

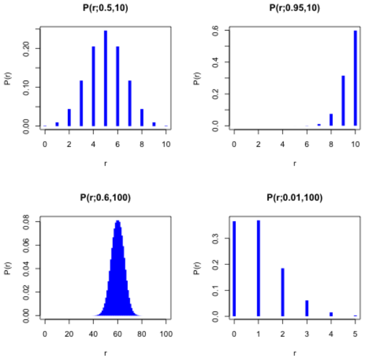
Distribution of income (per household, in US. Other examples are similar). Totally non-Gaussian.



The Binomial

Probability of r 'successes' from n trials, each with probability p.

$P(r; n, p) = \frac{n!}{r!(n-r)!} p^r q^{n-r} \qquad with\ q \equiv 1 - p$



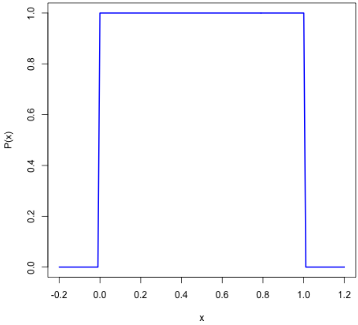
$\mu = np \qquad \sigma = \sqrt{npq}$

Limit: n large, p small, $np = \mu$ fixed $P(r) \rightarrow$ Poisson

Vital Fact: Basically just like tossing coins

The Uniform

or Top Hat

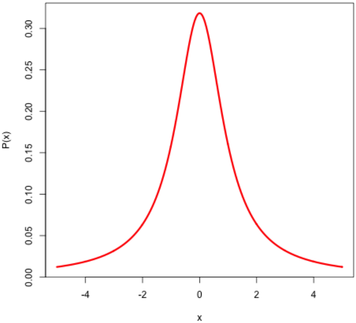


Generally, $P(x) = \frac{1}{a}$ between $\mu - a/2$ and $\mu + a/2$

Vital fact: Standard Deviation $\sigma = \frac{a}{\sqrt{12}}$

The Breit-Wigner

or Cauchy or Lorentzian



$$P(x) = \frac{1}{\pi}\frac{1}{1+x^2}$$
$$P(E; M, \Gamma) = \frac{1}{2\pi}\frac{\Gamma}{(E-M)^2+(\Gamma/2)^2}$$

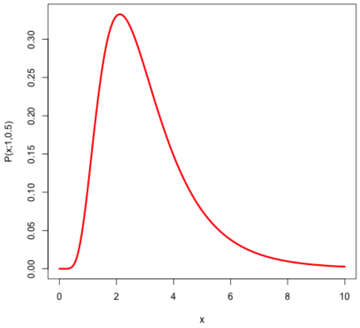
Does not have a standard deviation!
integral diverges

FWHM=Γ
for a Gaussian, FWHM=2.35 σ,
hence use of 'σ'=$\Gamma/2.35$

Vital fact: Useful for describing measurements that should be Gaussian but aren't

The log-normal

$P(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$
where μ and σ are mean and sd of $\ln x$

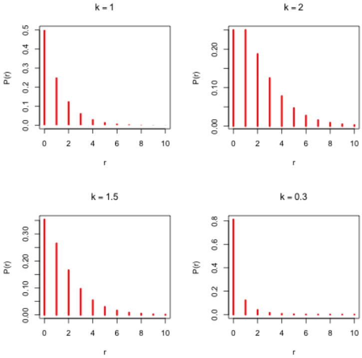


Vital fact: Applies(thanks to CLT) when effect of many factors combine multiplicatively

Example: energy measured in calorimeter with $x = E_0 - E$

The Negative Binomial

Coin-tossing again. This time ask 'How many successes before k failures?'



$$P(r;k,p) = \frac{(k+r-1)!}{r!(k-1)!} p^r q^k$$

Can write factor as $(-1)^r {}_rC_{-k}$

(hence unhelpful name)

or as $\frac{\Gamma(k+r)}{\Gamma(k)r!}$

generalise to non-integer k

don't ask what that means

All plots here have $p = 0.5$

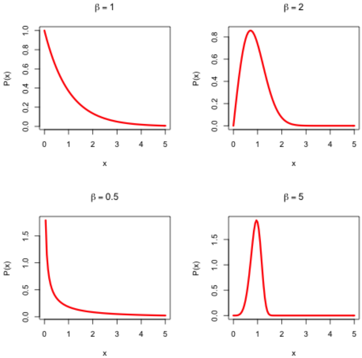
Vital fact: Used to describe events where $\sigma > \sqrt{N}$

i.e. more spread out than Poisson.

The Weibull

$P(x;\alpha,\beta) = \alpha\beta(\alpha x)^{\beta-1} e^{-(\alpha x)^\beta}$

Devised to describe the lifetime of lightbulbs



'Failure rate' $\propto x^{\beta-1}$

- $\beta < 1$: weak die early ('burn in')
- $\beta = 1$: constant rate (rad. decay)
- $\beta > 1$: aging process

α is just a scale factor

Vital fact: Handy as a way of parametrising rise-and-fall shapes

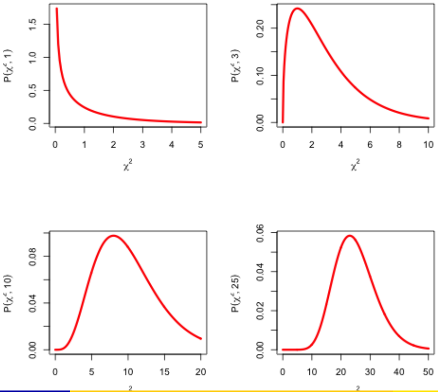
The χ^2

Much more on this in later lectures!

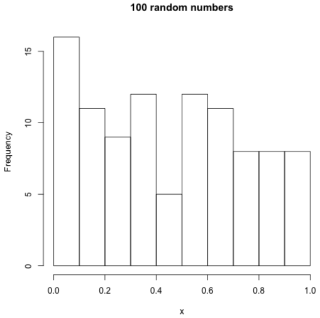
$\chi^2 = \sum_1^N \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2$ (x_i Gaussian)

Measures agreement between x_i and μ_i

Vital fact: $\overline{\chi^2} = N$, but big spread



Need (pseudo) random number generator for simulations: from Geant4 to Toy Monte Carlo.
All systems seem to contain a function that produces uniform random numbers between 0 and 1 - may be called
`ran(),ranf(),rnd(),runif(),?,TRandom,TRandom3...`



Doesn't look random, does it?
Very easy to see structure!
(Hence the need for Blind Analyses)
Try it yourself!
Extension to other uniform distributions is trivial

Technical detail

Such functions all based on generator of random integers, then mapped into $[0, 1]$.
Classical Method: Linear Congruential Generator (TRandom)
$R_{n+1} = (aR_n + b)|c$
with a, b, c suitably chosen. (c generally 2^{64} or 2^{32})
Start with some 'seed' R_0
(If you want a really random number, use the clock as the seed.)

Drawbacks: repeats with cycle of 2^{64} or 2^{32} - large but not always large enough. Particular R will never recur till the cycle repeats.
Modern methods: Mersenne Twister(TRandom3) (and its successors).
Large random state from which 62 or 32 bit number extracted.
Even more complicated random numbers used and needed for encryption.

Other distributions

Suppose you've got a $[0,1]$ random number U

For random direction:

$\phi = 2\pi U_1 \qquad \theta = acos(2U_2 - 1)$

The Exponential

Needed to generate decays (with time) and interactions (with distance). If the rate is r then $x = -r\ln(U)$

The Gaussian To get a 'unit Gaussian' $(\mu = 0, \sigma = 1)$

Lazy way: Add 12 instances of U and subtract 6

Why does this work?

Smart way: Generate U_1 and U_2. Form $R = -\ln U_1, \theta = 2\pi U_2$

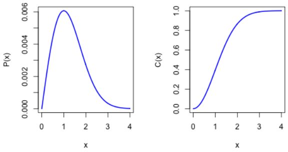
Then $R\cos\theta$ and $R\sin\theta$ are both Gaussian random numbers (and uncorrelated!)

Best way: Use the Gaussian generator provided by the system.

General functions

1: Inversion

From desired $P(x)$, form cumulative distribution $C(x) = \int^x P(x')\,dx'$
Generate uniform u in $[0, 1]$ and find x such that $C(x) = u$

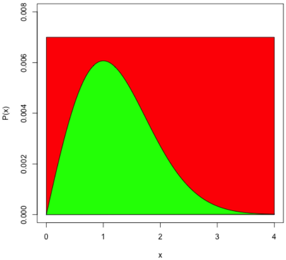


Example: To generate pdf $P(x) = 0.4 + 0.1x$ with $x \epsilon [0, 2]$
$C(x) = 0.4x + 0.05x^2$
$.05x^2 + 0.4x - u = 0$
$x = \frac{-.4 + \sqrt{.16 + .2u}}{0.1}$
Works great if you can (1) integrate $P(x)$ to get $C(x)$ and (2) invert $u = C(x)$ to get $x = C^{-1}(u)$
If not possible analytically, numerical methods may be used

General functions

2: von Neumann sampling

Generate x uniformly over range
Generate r uniformly between 0 and M, where $M \geq max(P(x))$
If $P(x) > r$, accept.
Else reject and try again



Works easily for multidimensional functions.
If M is overestimated, method is still valid (just a hit in the efficiency)
Can be very inefficient if $P(x)$ has sharp peaks-
may be improved by generating x according to some P_0 and using $P(x)/P_0(x)$ in the acceptance comparison

General functions

3: Weighting

Not all events need to be equal!
Generate x uniformly and weight the event by $P(x)$
when filling histograms, forming sums, etc, include the weight.
Can be effective when simulating low-probability processes that reject a lot of events.
More work, but not as hard as it looks.

Doesn't always help...
Poisson error on a weighted number is $\sqrt{N\overline{w^2}}$, always bigger than $\sqrt{N}\overline{w}$, i.e. error worse than pure Poisson $\sigma = \sqrt{N}$.
If weights all much the same, not a problem.
If a few events with enormous weights dominate, get big statistical errors.

Conclusions

There are many distribution for you to use - very big toolkit

Sometimes founded on dynamics of the problem

Sometimes empirical, found by experience to have useful behaviour in particular circumstances

Be open-minded and on the lookout for new ones!