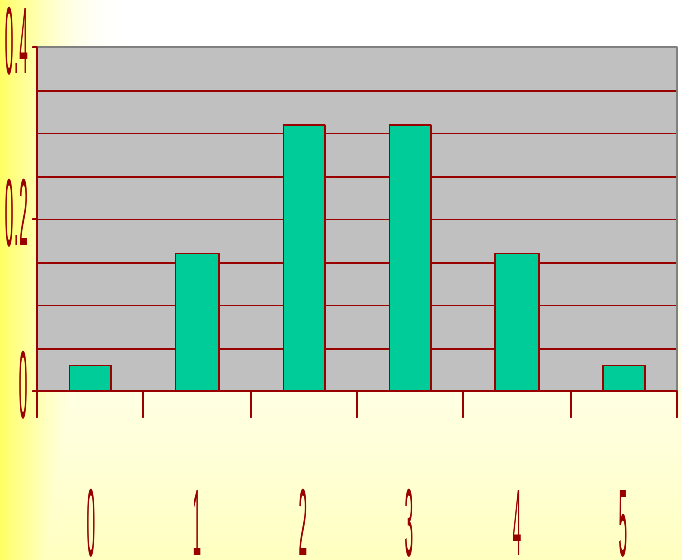
Statistics for HEP

Roger Barlow
Manchester University

Lecture 2: Distributions

The Binomial



n trials   r successes

Individual success probability p

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

Variance

$V\equiv\sigma^2=\langle(r-\mu)^2\rangle=\langle r^2\rangle-\langle r\rangle^2$

$=np(1-p)$

Mean

$\mu=\langle r\rangle=\sum rP(r)$

$= np$

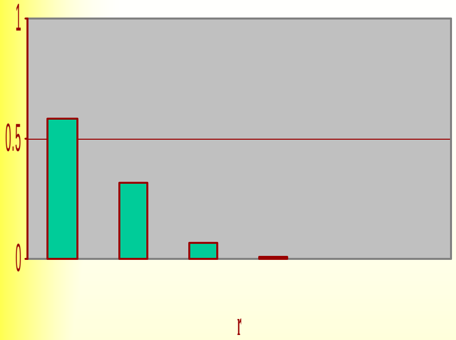
$1-p\equiv\bar{p}\equiv q$

Met with in Efficiency/Acceptance calculations

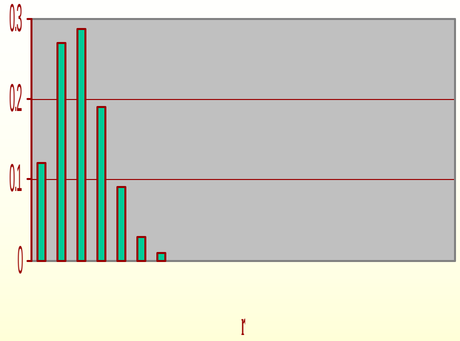
Binomial Examples

p=0.1

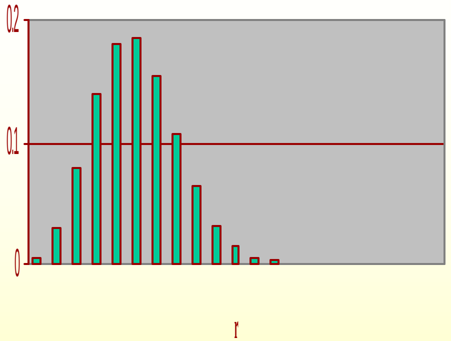
n=5



n=20

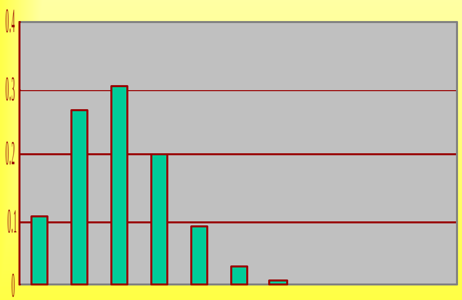


n=50

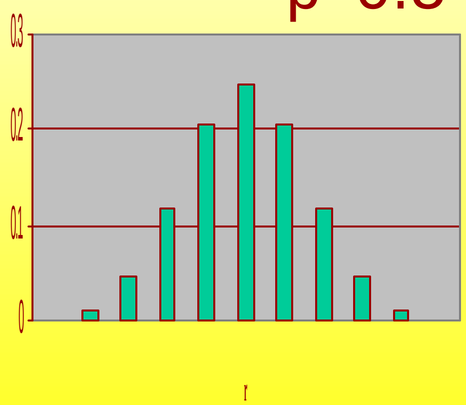


n=10

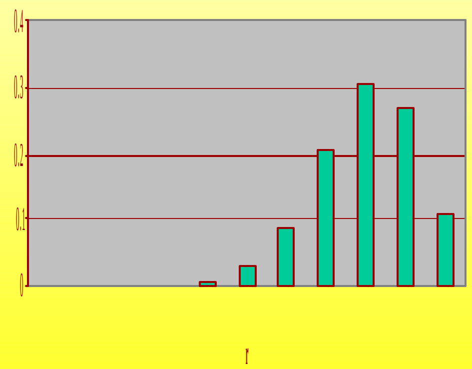
p=0.2



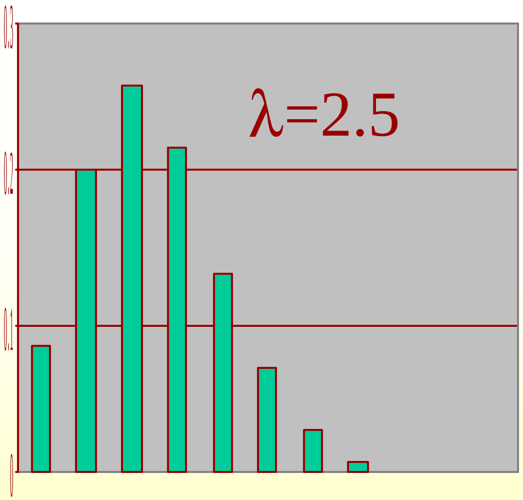
p=0.5



p=0.8



Poisson



‘Events in a continuum’

e.g. Geiger Counter clicks

Mean rate λ in time interval

Gives number of events in data

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

Mean

$\mu = \langle r \rangle = \sum r P(r)$

$= \lambda$

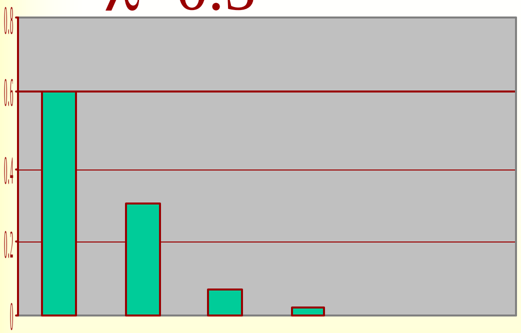
Variance

$V \equiv \sigma^2 = \langle (r - \mu)^2 \rangle = \langle r^2 \rangle - \langle r \rangle^2$

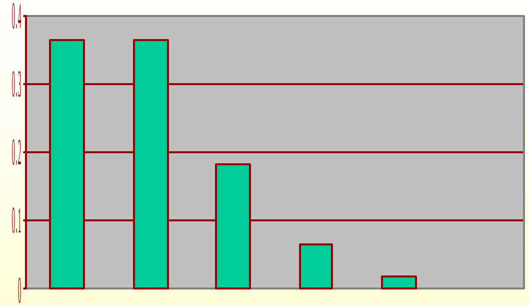
$= \lambda$

Poisson Examples

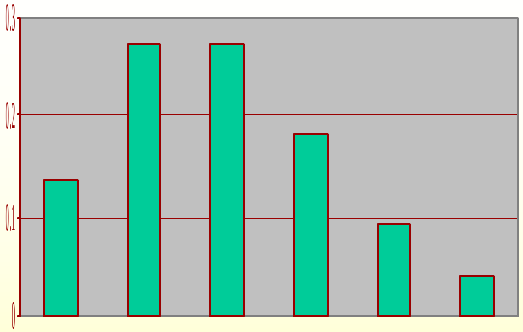
λ=0.5



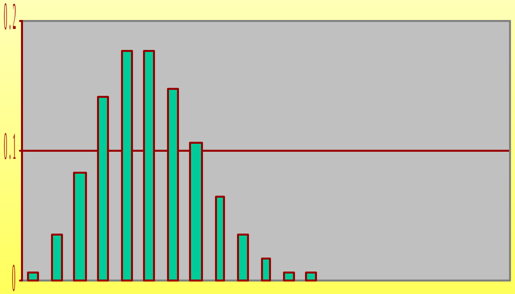
λ=1.0



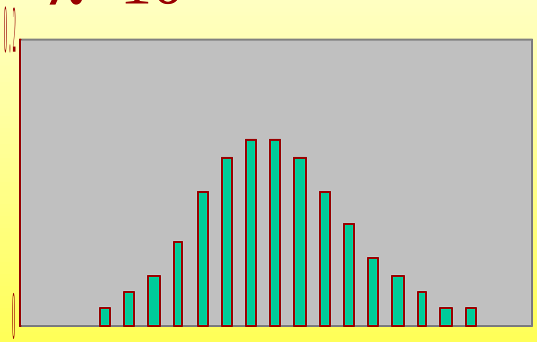
λ=2.0



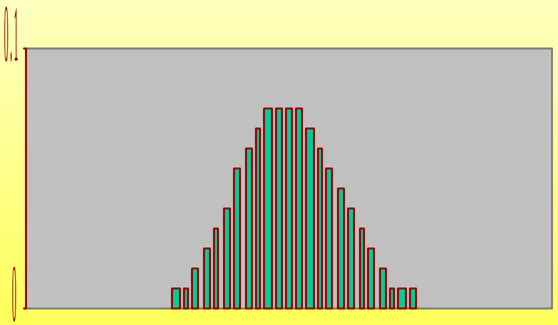
λ=5.0



λ=10



λ=25



Binomial and Poisson

From an exam paper

A student is standing by the road, hoping to hitch a lift. Cars pass according to a Poisson distribution with a mean frequency of 1 per minute. The probability of an individual car giving a lift is 1%. Calculate the probability that the student is still waiting for a lift

(a) After 60 cars have passed

(b) After 1 hour

a) $0.99^{60}=0.5472$

b) $e^{-0.6}=0.5488$

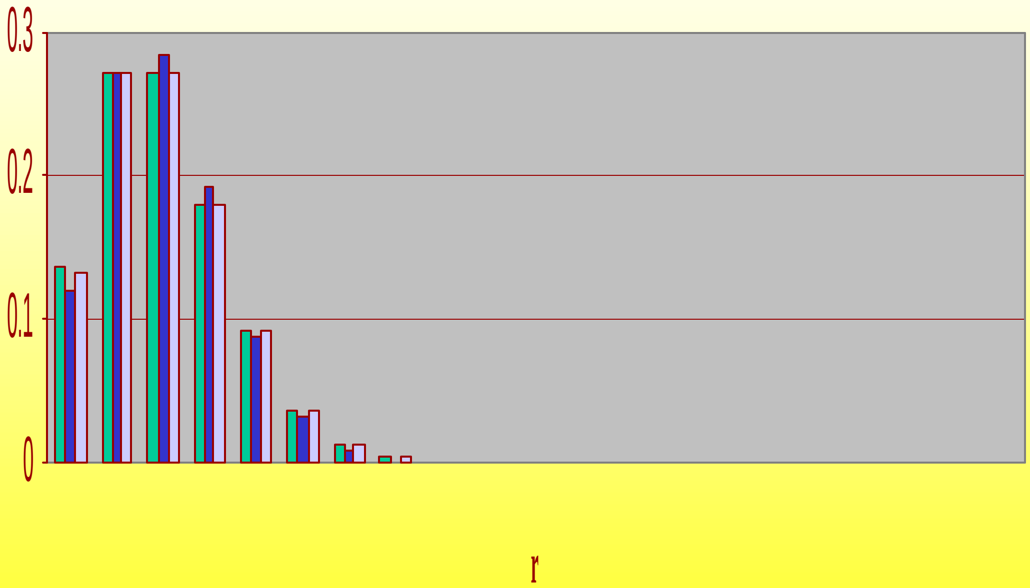
Poisson as approximate binomial

Poisson mean 2

Binomial: p=0.1, 20 tries

Binomial: p=0.01, 200 tries

Use: MC simulation (Binomial) of Real data (Poisson)



Two Poissons

2 Poisson sources, means λ_1 and λ_2

Combine samples

e.g. leptonic and hadronic decays of W

Forward and backward muon pairs

Tracks that trigger and tracks that don't

What you get is a *Convolution*

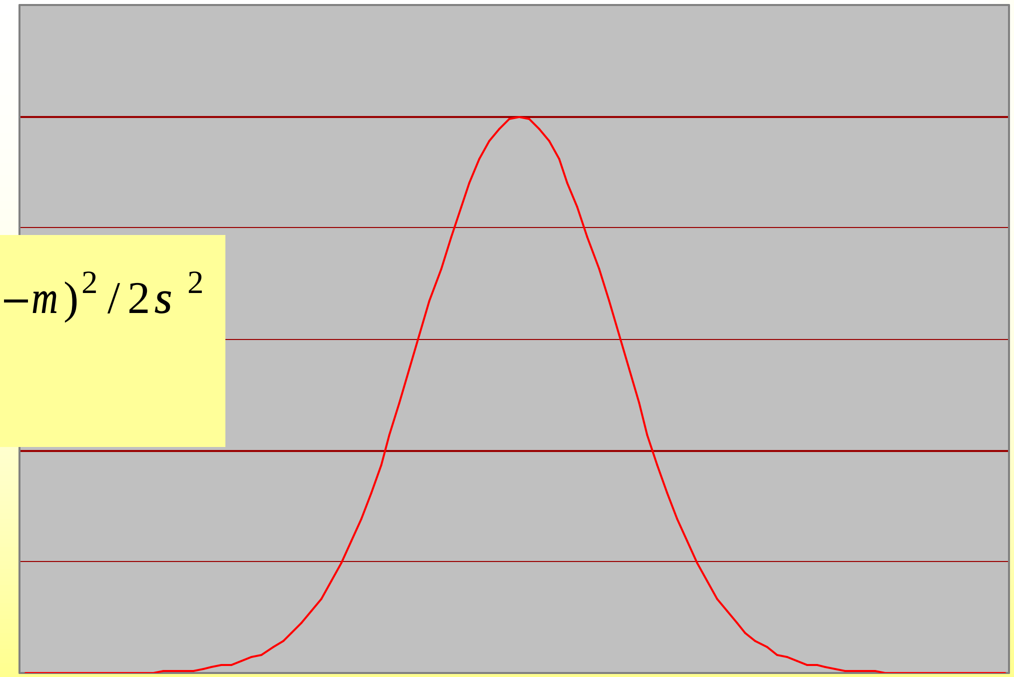
$$P(r) = \sum P(r'; \lambda_1)\, P(r-r'; \lambda_2)$$

Turns out this is also a Poisson with mean $\lambda_1 + \lambda_2$

Avoids lots of worry

The Gaussian

Probability Density



$$P(x;m,s) = \frac{1}{s\sqrt{2p}}\, e^{-(x-m)^2/2s^2}$$

Mean

$\mu = \langle x \rangle = \int x P(x)\, dx$

$= \mu$

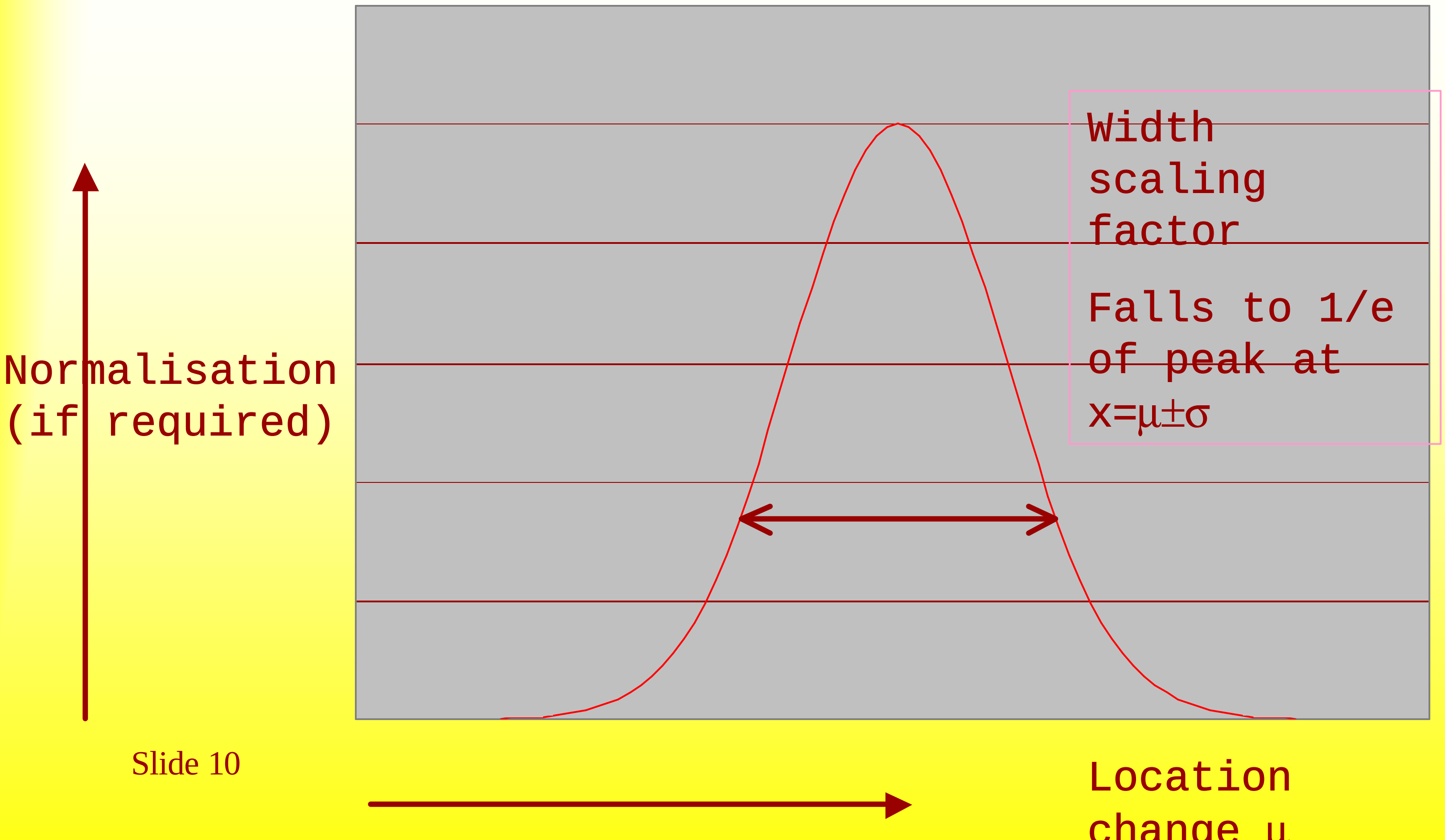
Variance

$V \equiv \sigma^2 = \langle (x-\mu)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2$

$= \sigma^2$

Different Gaussians

There's only one!



Probability Contents

68.27% within 1σ

95.45% within 2σ

99.73% within 3σ

90% within 1.645 σ

95% within 1.960 σ

99% within 2.576 σ

99.9% within 3.290σ

These numbers apply to Gaussians and only Gaussians

Other distributions have equivalent values which you could use of you wanted

Central Limit Theorem

Or: why is the Gaussian Normal?

If a Variable x is produced by the convolution of variables $x_1, x_2 \ldots x_N$

I) $\langle x \rangle = \mu_1 + \mu_2 + \ldots \mu_N$

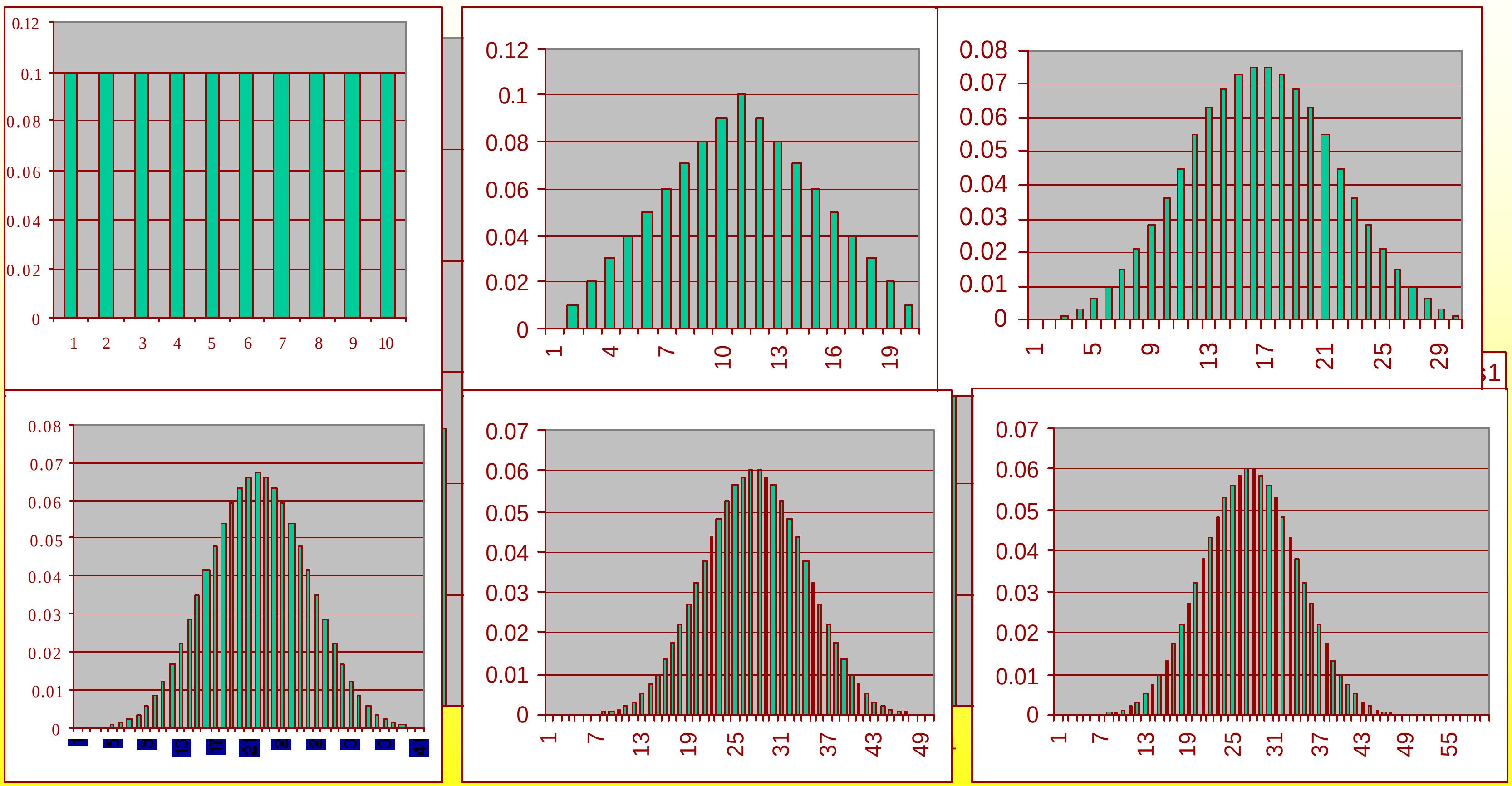
II) $V(x) = V_1 + V_2 + \ldots V_N$

III) P(x) becomes Gaussian for large N

There were hints in the Binomial and Poisson examples

CLT demonstration

Convolute Uniform distribution with itself



CLT Proof (I) Characteristic functions

Given P(x) consider

$$\left\langle e^{ikx} \right\rangle = \int e^{ikx} P(x) dx = \tilde{P}(k) = \Phi(k)$$

The Characteristic Function

For convolutions, CFs multiply

If $f(x) = g(x) \otimes h(x)$ then $\tilde{f}(k) = \tilde{g}(k)\tilde{h}(k)$

Logs of CFs Add

CLT proof (2) Cumulants

CF is a power series in k

$\langle 1\rangle + \langle ikx\rangle + \langle (ikx)^2/2!\rangle + \langle (ikx)^3/3!\rangle + \ldots$

$1 + ik\langle x\rangle - k^2\langle x^2\rangle/2! - ik^3\langle x^3\rangle/3! + \ldots$

Ln CF can then be expanded as a series

$ikK_1 + (ik)^2K_2/2! + (ik)^3K_3/3! \ldots$

K_r : the "semi-invariant cumulants of Thiele"

Total power of x^r

If $x \to x+a$ then only $K_1 \to K_1 + a$

If $x \to bx$ then each $K_r \to b^r K_r$

CLT proof (3)

- The FT of a Gaussian is a Gaussian

$$e^{-x^2/2s^2} \to e^{-k^2 s^2/2}$$

Taking logs gives power series up to k^2

$K_r=0$ for $r>2$ defines a Gaussian

- Selfconvolute anything n times: $K_r'=n\,K_r$

Need to normalise – divide by n

$K_r''=n^{-r}\,K_r'= n^{1-r}\,K_r$

- Higher Cumulants die away faster

If the distributions are not identical but similar the same argument applies

CLT in real life

Examples

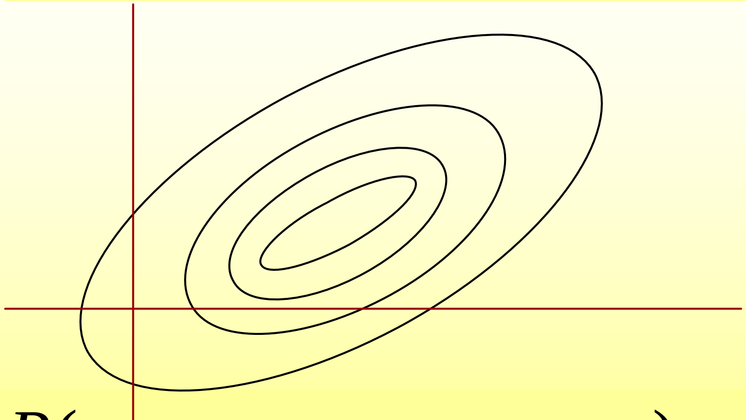
- Height
- Simple Measurements
- Student final marks

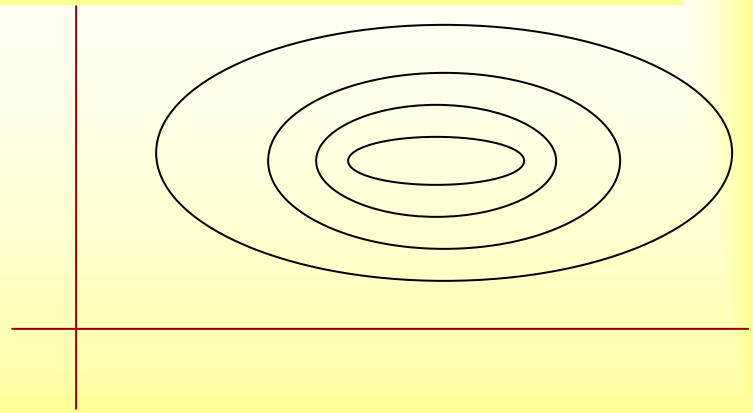
Counterexamples

- Weight
- Wealth
- Student entry grades

Multidimensional Gaussian

$$P(x,y;m_x,m_y,s_x,s_y) = \frac{1}{s_x s_y 2p} e^{-(x-m_x)^2/2s_x^2} e^{-(y-m_y)^2/2s_y^2}$$





$$P(x,y;m_x,m_y,s_x,s_y,r)$$

$$= \frac{1}{s_x s_y 2p\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)}\left((x-m_x)^2/s_x^2 + (y-m_y)^2/s_y^2 - 2r(x-m_x)(y-m_y)/s_x s_y\right)}$$

Chi squared

$$c^2 = \sum_{i=1}^{n} \left(\frac{x_i - m_i}{s_i} \right)^2$$

Sum of squared discrepancies, scaled by expected error

Integrate all but 1-D of multi-D Gaussian

$$P(c^2; n) = \frac{2^{-n/2}}{\Gamma(n/2)} c^{n-2} e^{-c^2/2}$$

Mean n

Variance 2n

CLT slow to operate

Generating Distributions

Given `int rand()`in `stdlib.h`

```
float Random()
                {return ((float)rand())/RAND_MAX;}
float uniform(float lo, float hi)
                     {return lo+Random()*(hi-lo);}
float life(float tau)
                      {return -tau*log(Random());}
float ugauss()        // really crude. Do not use
    {float x=0; for(int i=0;i<12;i++) x+=Random();
                                      return x-6;}
float gauss(float mu, float sigma)
                       {return mu+sigma*ugauss();}
```

A better Gaussian Generator

```
float ugauss(){
  static bool igot=false;
  static float got;
  if(igot){igot=false; return got;}
  float phi=uniform(0.0F,2*M_PI);
  float r=life(1.0f);
  igot=true;
  got=r*cos(phi);
  return r*sin(phi);}
```

More complicated functions

Find P_0=max[P(x)].

Overestimate if in doubt

Repeat :

Repeat:

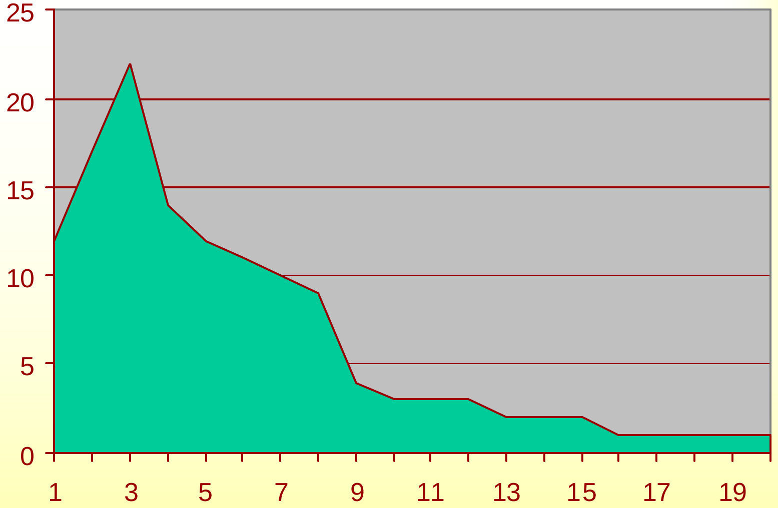
Generate random x

Find P(x)

Generate random P in range 0-P_0

till P(x)>P

till you have enough data



If necessary make x non-random and compensate

Other distributions

Uniform(top hat)

σ=width/$\sqrt{12}$

Breit Wigner (Cauchy)

Has no variance – useful for wide tails

Landau

Has no variance or mean

Not given by $e^{e^{-l}}$ .Use CERNLIB

Functions you need to know

- Gaussian/Chi squared
- Poisson
- Binomial
- Everything else