Statistics for counting experiments

R. J. Wilkes
Dept. of Physics, UW
8/8/02

Probability

- Frequency theory of probability
  - $\text{Prob(event)} = \dfrac{\text{How many times event happened}}{\text{How many opportunities for it to happen}}$

  - Unless denominator is large (*high statistics experiment*), we have only a relatively poor estimate of the "true" probability -- assumed to be due to some underlying "law"

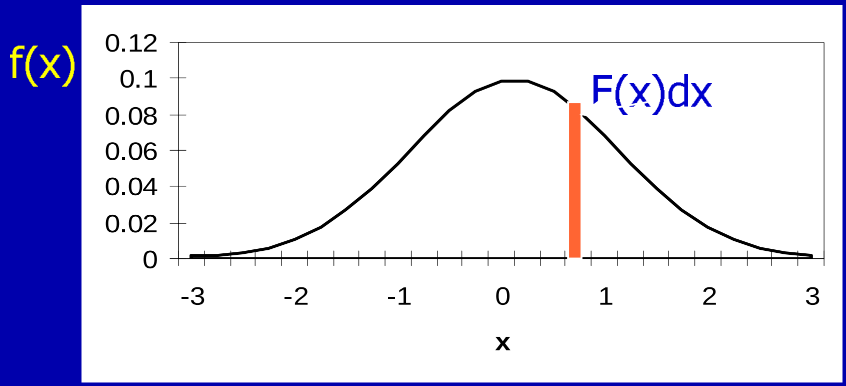
Man-in-the-Street views of probability

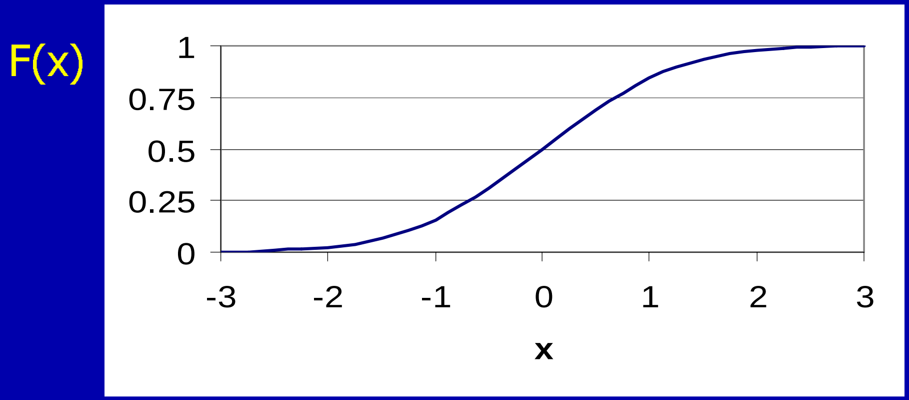
- Fallacies about denominators
  - "90% of our flights arrive on time"
    - » correct statement: "flights delayed several hours are cancelled, not 'delayed', so they get excluded from our average"
  - "The average worker is making 10% more now than he was 10 years ago"
    - » correct statement: "the minimum wage has risen, and more low-income people are unemployed"
- Fallacies about independence
  - "This slot machine hasn't paid off in a long time, so I'm sure to win soon"
    - » correct statement: "If this slot machine is truly random, i am no more likely to win on the next try as at any other time"
  - "Nobody's won the state lottery in a long time, so it is more likely to happen this week"
    - » correct statement: "Nobody's won the state lottery in a long time, so the payoff is bigger"
- ...or both combined
  - "Our survey shows most people lose 10 pounds in a month on this diet"
    - » correct statement: "happy customers who lost weight were most likely to respond to our survey; the ones who gained weight most likely threw away our postcard..."

Probability distributions and PDFs

- Probability Density Function (PDF) = f(x)
  - probability of x in range x′ to x′+dx
- "Probability distribution" = F(x)
  - *cumulative* or *integral* distribution = probability of x<x′

$$F(x) = \int_{x_{MIN}}^{x} f(x)dx \quad \text{(where } x_{MIN} \text{ could be } -\infty\text{)}$$





Descriptive parameters for PDFs

- Measures of central location:
  mean <x> = $\Sigma x_i / N$ (*sample* mean)
  median = x at which F(x)=0.5
  mode = x at which f(x)=maximum
  for symmetrical distributions, mean=median

- Measures of width of distributions:
  variance σ^2 (σ = *standard deviation*)
  $\sigma^2 = \Sigma(x_i - \mu_1)^2 / N$
  but μ_1 = mean of *true* PDF
  we can only *estimate* μ_1 with <x>
  Best estimator for σ^2 is
  $s^2 = \Sigma(x_i - \langle x \rangle)^2 / (N-1)$ = *sample variance*

Counting statistics

- We have a set of data = N measurements of some sort: $\{ x_1\ x_2\ x_3\ \dots\ x_N \}$

- Statistic = a function of the data only - no unknown parameters

  examples:
  - Sample mean (experimental mean)

    $$\overline{x} = \frac{1}{N}\sum_{1}^{N} x_i$$

  - Median

    sort the data in ascending or descending order

    median = the (N/2)th entry in this list $x_{MED} = x_{\frac{N}{2}}$ in $sort_{\uparrow}(\{x_i\})$
  - Mode
    - » Value with maximum probability density: location of peak of PDF

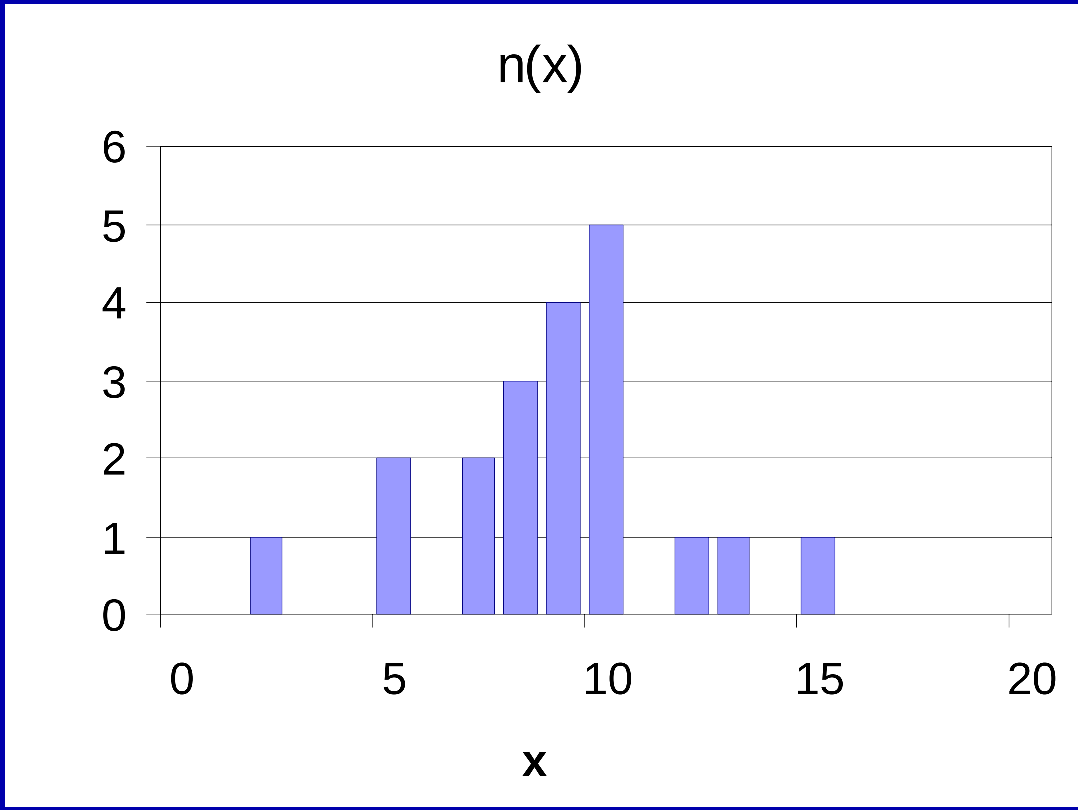
$$x_i \text{ such that } P(x_i) = \max P(x)$$

Example: 20 sets of 1 minute counts

x_k, k=0…20:

k	x_k
0	0
1	9
2	10
3	13
4	10
5	9
6	9
7	9
8	15
9	2
10	10
11	12
12	10
13	8
14	5
15	5
16	10
17	7
18	7
19	8
20	8

Histogram of the data:
A bar graph showing how often each possible count value occurred

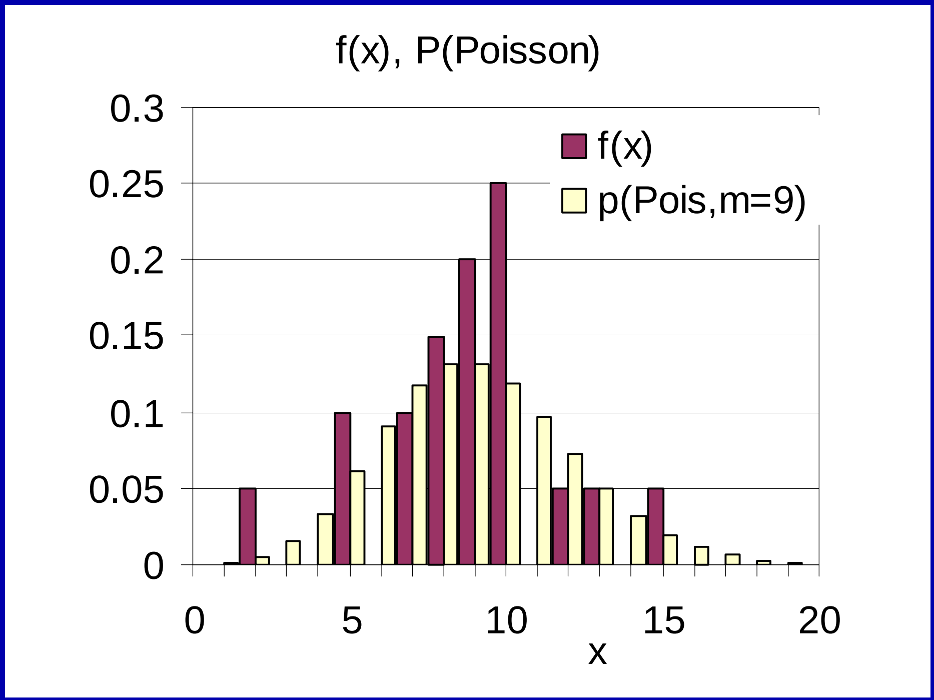


Frequency distribution

x	n(x)	f(x)
0	0	0
1	0	0
2	1	0.05
3	0	0
4	0	0
5	2	0.1
6	0	0
7	2	0.1
8	3	0.15
9	4	0.2
10	5	0.25
11	0	0
12	1	0.05
13	1	0.05
14	0	0
15	1	0.05
16	0	0
17	0	0
18	0	0
19	0	0
20	0	0

- Use the histogram to *estimate* probability of each possible x value: f(x)=n(x)/N
- This is the Probability Density Function (PDF) or differential probability distribution

(also shown below is the Poisson probability density function for mean value = 9 -- more on this later)



Statistics of the data set

- sample mean:

sum of data: 176

sample mean = sum/20: 8.8

- sample variance:

sorted data

k	x_k
0	0
1	2
2	5
3	5
4	7
5	7
6	8
7	8
8	8
9	9
10	9 ← median
11	9
12	9
13	10
14	10
15	10
16	10
17	10
18	12
19	13
20	15

- median=9

Some famous probability distributions and their applications

- Uniform
  - basis for generating numbers for simulations (computer pseudo-random number generators)
- binomial
  - Yes/No situations
- Poisson
  - Many physics applications
  - Applies when P(event) is "small" and "independent of previous history"
- Gaussian (Normal)
  - Applies to results produced a series of random processes
    - » Most scientific data are acquired through a series of processes, each with some random error contribution!

Uniform distribution

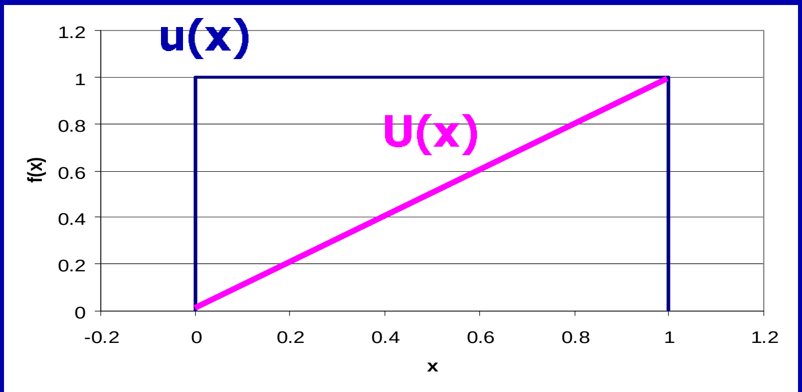
- Uniform PDF: u(x) = constant = $1/(x_{max} - x_{min})$
  - basic PDF supplied on computers: u(0;1)=1
  - Properties: $\langle x \rangle = (x_{max} + x_{min})/2$, $\sigma^2 = (x_{max} + x_{min})^2 / 12$
  - Any PDF can be obtained from u(x) by inverting its integral distribution F(x)
    - » Can use this to generate random numbers for simulations, etc

Choose uniform random number on [0,1] and use it to select x from F(x)

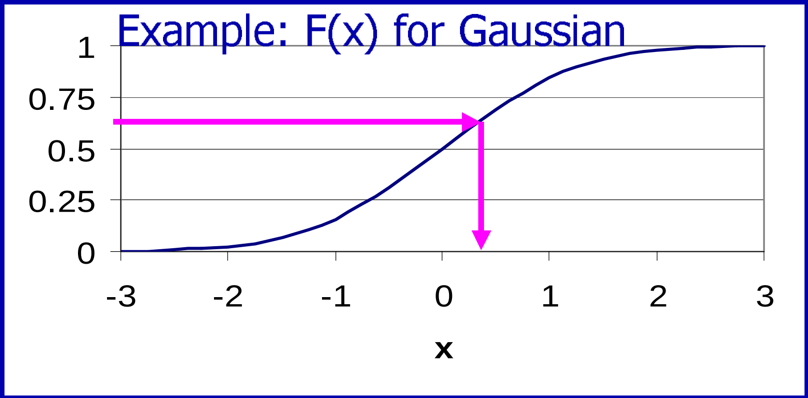
Example: Exponential distribution f(y)=exp(-y)

Exercise: show y= - ln(1-x) (with x uniformly distributed)

is exponentially distributed.



F(x)



Binomial Distribution

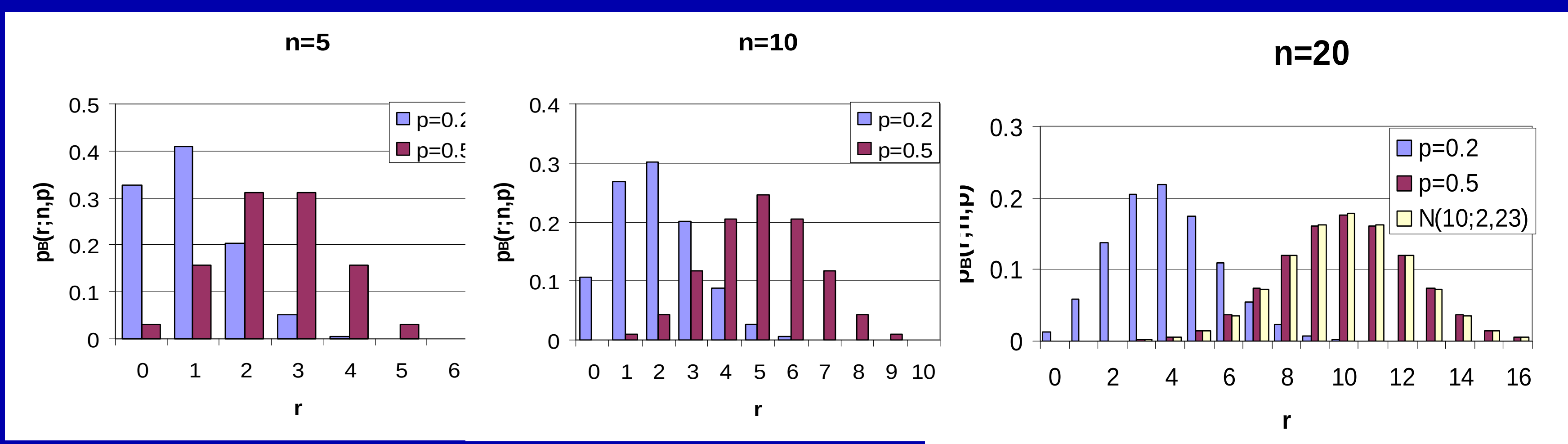
- Applies to cases with binary outcomes like coin flips:
  - 0/1, heads/tails, T/F, yes/no, win/lose, success/failure
- *Discrete-valued* PDF gives $P(n_{SUCCESSES} = \text{integer})$
- 2 parameters: p(success per trial = real), N_{TRIALS}
  - P(n successes followed by (N-n) failures)

  $= p^n (1-p)^{N-n}$ (independent trials: multiply trial probs.)
  - But we don't care about order in which they occur:

  number of permutations is $N!/(n!(N-n)!)$

  so $P(n; p,N) = \{N!/(n!(N-n)!)\}\, p^n (1-p)^{N-n}$

- Properties: $\mu = Np$, $\sigma^2 = Np(1-p) = \mu(1-p)$, ~ Gaussian for large Np



Poisson distribution

- Limiting case of binomial distribution for $p \to 0$
- only 1 parameter: mean value μ

  P(n successes | μ expected) = $(1/\,n!)\,\mu^n \exp(-\mu)$

  n is integer; μ can be real
- Properties:

  variance $\sigma^2 = \mu$, so standard deviation $\sigma = \mathrm{sqrt}(\mu)$
- Applies when *Poisson assumptions* are valid:
  1. P(event) in interval δx is *proportional to* δx: $p=g\delta x$
  2. Occurrence of an event in an interval δx_j is *independent* of events or absence of events in any other non-overlapping interval δx_k
  3. For sufficiently small δx, there can be at most 1 event in δx

Example of a Poisson Process

- Bubbles in a bubble chamber track

Prob of 1 bubble in $\delta x: p_1(\delta x) = g\delta x \quad (from \quad \#1)$

Prob of 0 bubbles in $\delta x: p_0(\delta x) = 1 - p_1 = 1 - g\delta x \quad (from \quad \#3)$

$$p_0(x+\delta x) = p_0(x) \bullet p_0(\delta x) = p_0(x)(1-g\delta x) \quad (from \quad \#2)$$

$$\therefore \frac{p_0(x+\delta x) - p_0(x)}{\delta x} = -g$$

$$p_0(x) \rightarrow \frac{dp_0}{dx} = -gp_0$$

$Solution: \ p_0(x) = e^{-gx}$ So $p_0(x)$= *exponential distribution*

Prob of exactly r bubbles in $x + \delta x:$

$$p_r(x+\delta x) = p_r(x) \bullet p_0(\delta x) + p_{r-1}(x) \bullet p_1(\delta x) \quad (from \quad \#3)$$

$$\therefore \frac{p_r(x+\delta x) - p_r(x)}{\delta x} \rightarrow \frac{dp_r}{dx} = -gp_r(x) + gp_{r-1}(x)$$

$Solution: \ p_r(x) = \frac{1}{r!}(gx)^r e^{-gx} =$ Poisson distribution $(\mu = gx)$

Gaussian (Normal) distribution

- Gaussian = famous "bell-shaped curve"
  - Describes IQ scores, number of ants in a colony of a given species, wear profile on old stone stairs…

  All these are cases where:
  - deviation from norm is equally probable in either direction
  - Variable is continuous (or large enough integer to look continuous - far from the "wall" at zero)
- *Real-valued* PDF: $f(x) \rightarrow \quad -\infty < x < +\infty$

  $N(x;\mu,\sigma) = (1/\mathrm{sqrt}[2\pi\sigma^2]) \exp[-(x-\mu)^2/2\sigma^2]$
- 2 independent parameters: μ , σ (central location and width)
- Properties:

  Symmetrical, mode at μ ,

  median=mean=mode, Inflection points at $\pm\sigma$

  Cumulative distribution :

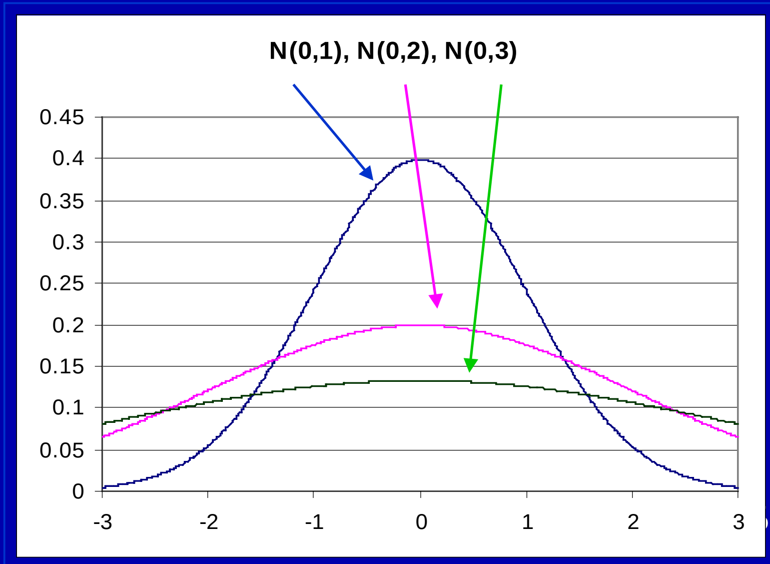
  $\int_{-\infty}^{x} n(x;0,1)dx = \mathrm{erf}(x)$

  Area (probability of observing event) within:

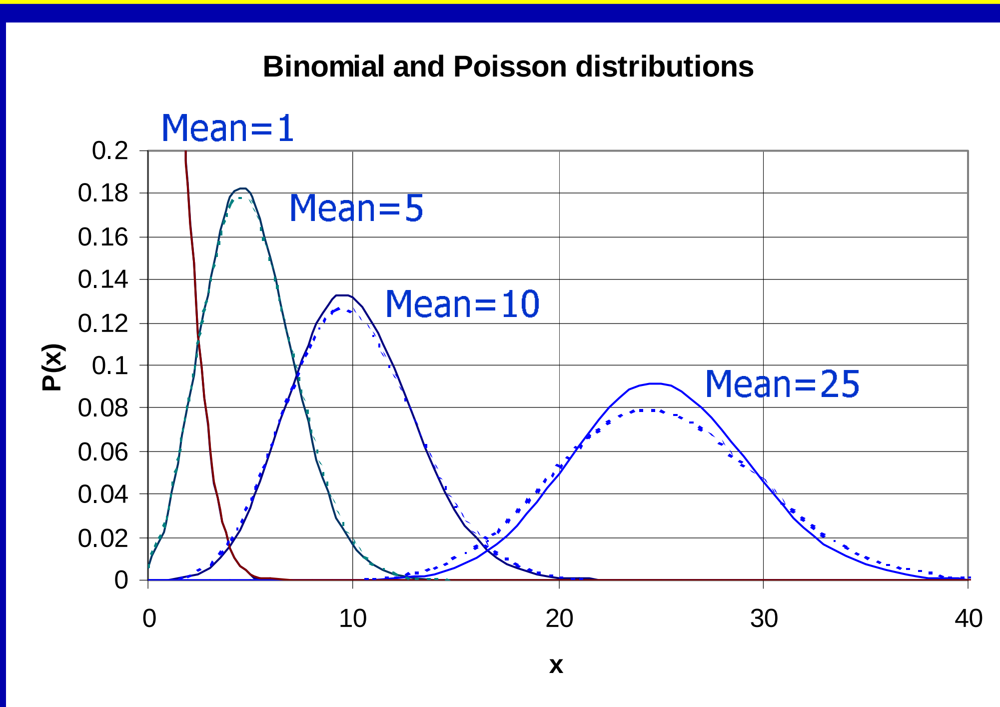
  $\pm 1\sigma = 0.683$ = erf(1)-erf(-1)

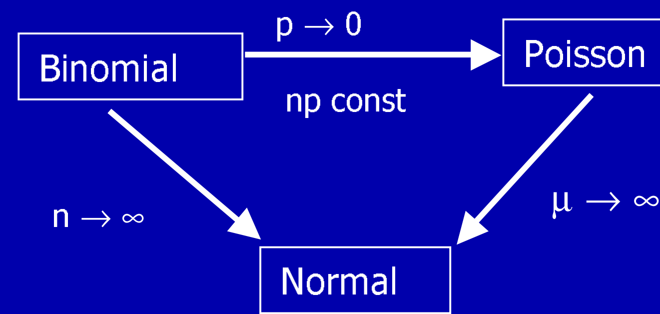
  $\pm 2\sigma = 0.955$ = erf(2)-erf(-2)

For larger σ, bell shaped curve becomes wider and lower (since area =1 for any σ)



Binomial, Poisson, Gaussian





Shown above:

- Binomial for 100 trials, p=0.01, 0.05, 0.10, 0.25 (solid)
- Poisson for μ = 1, 5, 10, 25 (dashed line)

Poisson is broader and has peak slightly below μ
Both become similar to Gaussian N(μ, $\sigma=\sqrt{\mu}$) as mean value gets larger
(Gaussian would be indistinguishable from Poisson for mean=25 on this plot)

Why the Normal Distribution is important...

- *Central Limit Theorem:*

  Given N independent random variables x_k, each with mean μ_k and variance σ_k specified (but *not* details of individual PDF's), the random variable $z = \Sigma x_k$ has

  $\mu_Z = \Sigma \mu_k$ and $\sigma_Z^2 = \Sigma \sigma_k^2$,

  and for $N \to \infty$, its PDF will be Gaussian, i.e. $p(z)=N(\mu_Z , \sigma_Z)$

  $(\Sigma x_k - \Sigma \mu_k) / \mathrm{sqrt}[\Sigma \sigma_k^2] = n(x;0,1)$

- Applies to: any situation with real-valued result where several *independent* processes *add:* additive errors. Examples:
  - Random walk of 100 steps. Each step is independent of others, any probability distribution for direction and length of each step (but μ, σ^2 known).
  - To make a simple Gaussian random number generator, just take sum of 12 standard uniformly distributed numbers:

    $x=\Sigma (u_k - 6)$; x will be distributed ~ $n(x;0,1)$

    (recall: $u(0;1)$ has $\mu= 0.5$, $\sigma^2= 1/12$)

- Parameters μ,σ are independent (and converse: if a random variable has μ,σ independent, it is normal).

  Given N random numbers x_k drawn from a normal distribution,

  the sample mean $\mu = (1/N)\Sigma x_k$

  and sample variance $s^2 = \Sigma \sigma_k^2 / (N-1)$

  are *independent statistics*

Applications to counting

- Errors in single counts
  - CR counts are a Poisson process, so $\sigma_k^2 = N$, $\sigma_k = \sqrt{N}$
- Errors on histogram bins contents
  - In/out of bin = binomial process, so $\sigma_k^2 = Np_k(1-p_k)$ where $p_k = n_k/N$
  - Poisson approximation $\sigma_k = \sqrt{n_k}$ is valid for $n_k > 10$
- Significance of deviations from expectation

Significance of deviations from expectation

Example: counting statistics and limits of detectability

- How can we tell if a significant signal exists in the presence of background?

  N_T = observed counts in time T

  N_B = background counts (separate experiment)

  Then $N_T = N_S + N_B$ where N_S = true signal counts

  Assume T is long enough so all counts are "not small" (>>5)

  Then expect N's to be Poisson distributed (~ Gaussian-distributed), with $\sigma = \sqrt{N}$

  $N_S = N_T - N_B$, so $\sigma_S^2 = \sigma_T^2 + \sigma_B^2$

  – Suppose there is *no real activity* present, N_S actually = 0

  $\sigma_T^2 = \sigma_B^2$ so $\sigma_S^2 = 2\sigma_B^2$ or $\sigma_S = \sqrt{(2N_B)}$

  So we expect N_S to be drawn from a Gaussian distribution $N(0, \sqrt{(2N_B)})$

- Define H_0 = hypothesis that there is no activity present, all we are seeing is background

  – Reject H_0 if $N_T > N_C$ = "cut level" for decision

  How do we define N_C ?

Significance of deviations from expectation

- Decide on a *significance level* = acceptable probability for being fooled by a random fluctuation

  If we want, eg, <5% probability of false positive result, we must set N_C at the 5% tail of the Gaussian distribution.

- Example: H_0 = "no radioactive decays from this sample"

  No-sample run gives 6 counts, assumed to be background

  $\sigma_S = \sqrt{(2N_B)} = 3.5$

  Therefore if H_0 = true, and we count the sample many times,

  we would get fewer than:

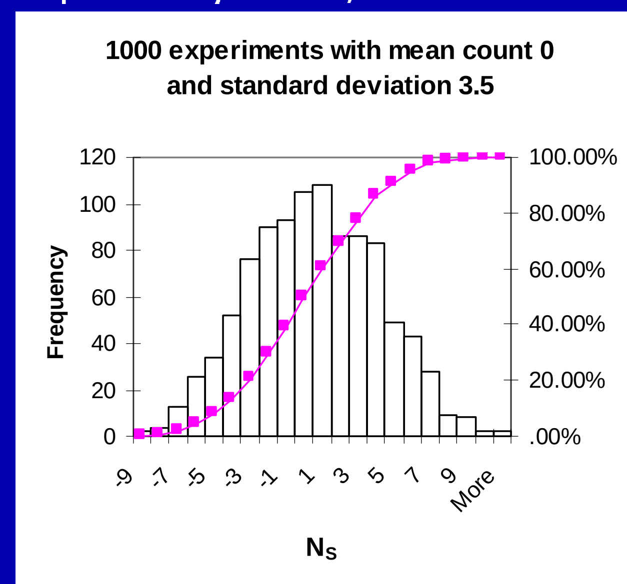
  3.5 counts 68% of the time

  7 counts 95% of the time

  10.5 counts 99.7% of the time

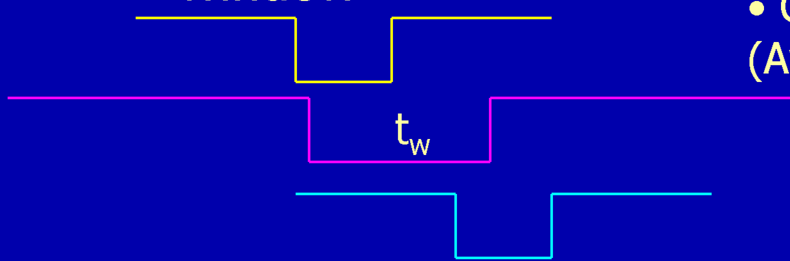
  Another way to say it:

  we can reject H_0 at the 95% confidence level if we observe N>7



"Accidentals"

- Accidentals = Chance coincidences due to uncorrelated noise pulses which happen to arrive within the logic gate's time window



- Counter 1's pulse arrives (Average spacing is $1/r_1$ sec)
- Logic gate opens a window (note delay)
- Counter 2's pulse arrives Average spacing is $1/r_2$ sec

- If noise is truly random, then the fraction of each second occupied by available coincidence windows is

  $f_{OCCUPIED} = r_1 * t_W$

  where r_1=singles rate of counter 1, Hz; t_w=window width, sec

  (This is equal to the probability that a randomly selected time lies within a coincidence window)

- The rate of 2-fold accidentals will thus be

  $r_{12}=r_2*f_{OCCUPIED} = r_2*r_1*t_W$ (for $r_{1,2}*t<<1$)