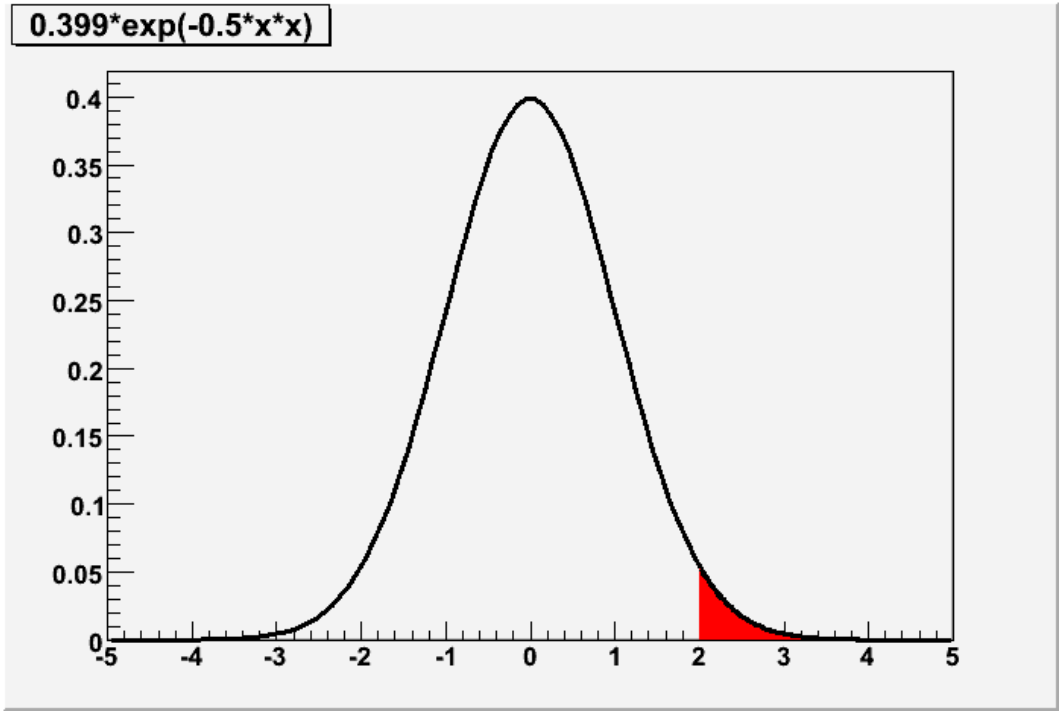
Basic Descriptive Statistics & Probability Distributions

Scott Oser
Lecture #2

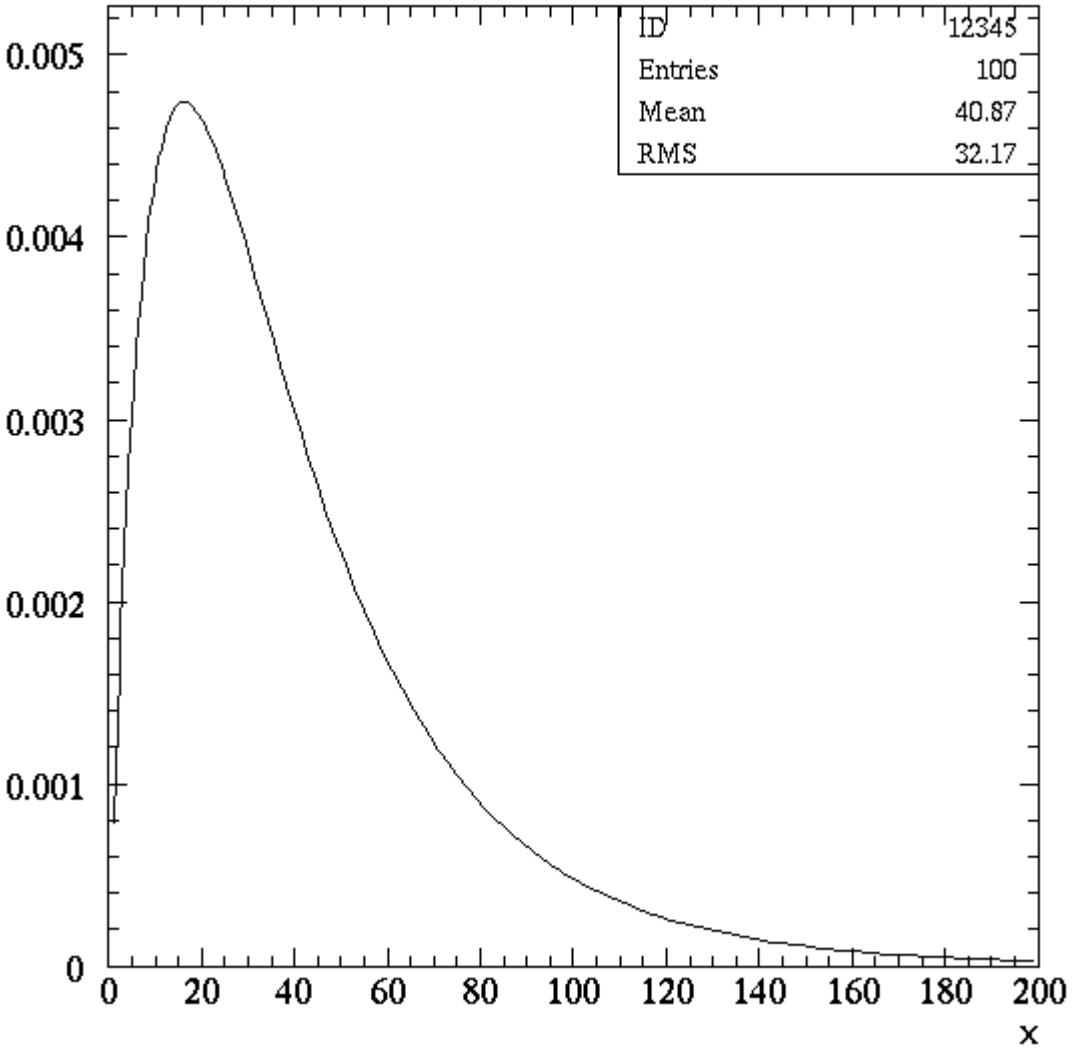


Outline

Last time: we discussed the meaning of “probability”, did a few warmups, and were introduced to Bayes theorem. Today we cover more basics.

- Basic descriptive statistics
- Covariance and correlation
- Properties of the Gaussian distribution
- The binomial distribution
- Application of binomial distributions to sports betting
- The multinomial distribution

Basic Descriptive Statistics



WHAT IS THIS DISTRIBUTION?

Often the probability distribution for a quantity is unknown. You may be able to sample it with finite statistics, however.

Basic descriptive statistics is the procedure of encoding various properties of the distribution in a few numbers.

The Centre of the Data: Mean, Median, & Mode

Mean of a data set:

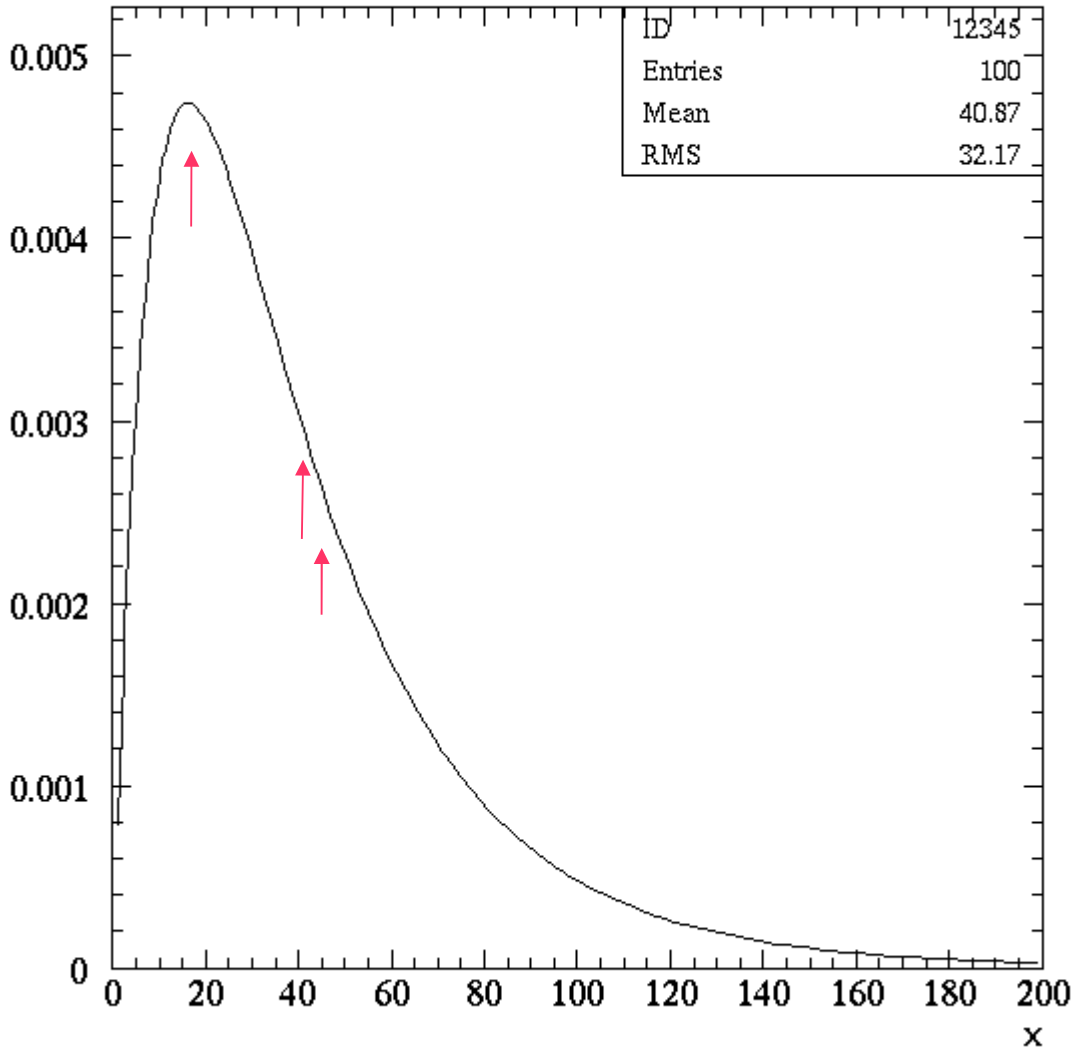
$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Mean of a PDF = expectation value of x

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

Median: the point with 50% probability above & 50% below. (If a tie, use an average of the tied values.) Less sensitive to tails!

Mode: the most likely value



Variance *V* & Standard Deviation σ (*a.k.a.* RMS)

Variance of a distribution: $V(x)=\sigma^2=\int dx\,P(x)(x-\mu)^2$

$$V(x)=\int dx\,P(x)\,x^2-2\mu\int dx\,P(x)\,x+\mu^2\int dx\,P(x)=\langle x^2\rangle-\mu^2=\langle x^2\rangle-\langle x\rangle^2$$

Variance of a data sample (regrettably has same notation as variance of a distribution---be careful!):

$$V(x)=\sigma^2=\frac{1}{N}\sum_i (x_i-\bar{x})^2=\overline{x^2}-\bar{x}^2$$

An important point we'll return to: the above formula underestimates the variance of the underlying distribution, since it uses the mean calculated from the data instead of the true mean μ of the true distribution.

$$\hat{V}(x)=\sigma^2=\frac{1}{N-1}\sum_i (x_i-\bar{x})^2$$

This is unbiased if you must estimate the mean from the data.

$$V(x)=\sigma^2=\frac{1}{N}\sum_i (x_i-\mu)^2$$

Use this if you know the true mean of the underlying distribution.

FWHM & Quartiles/Percentiles

FWHM = Full Width Half Max. It means what it sounds like--- measure across the width of a distribution at the point where $P(x)=(1/2)(P_{max})$. For Gaussian distributions, FWHM=2.35σ.

Quartiles, percentiles, and even the median are “rank statistics”. Sort the data from lowest to highest. The median is the point where 50% of data are above and 50% are below. The quartile points are those at which 25%, 50%, and 75% of the data are below that point. You can also extend this to “percentile rank”, just like on a GRE exam.

FWHM or some other width parameter, such as “75% percentile data point – 25% data point”, are often robust in cases where the RMS is more sensitive to events on tails.

Higher Moments

Of course you can calculate the r^{th} moment of a distribution if you really want to. For example, the third central moment is called the skew, and is sensitive to the asymmetry of the distribution (exact definition may vary---here's a unitless definition):

$$\text{skew} = \gamma = \frac{1}{N\sigma^3}\sum_i (x_i - \bar{x})^3$$

Kurtosis (or curtosis) is the fourth central moment, with varying choices of normalizations. For fun you are welcome to look up the words “leptokurtotic” and “platykurtotic”, but since I speak Greek I don't have to.

Warning: Not every distribution has well-defined moments. The integral or sum will sometimes not converge!

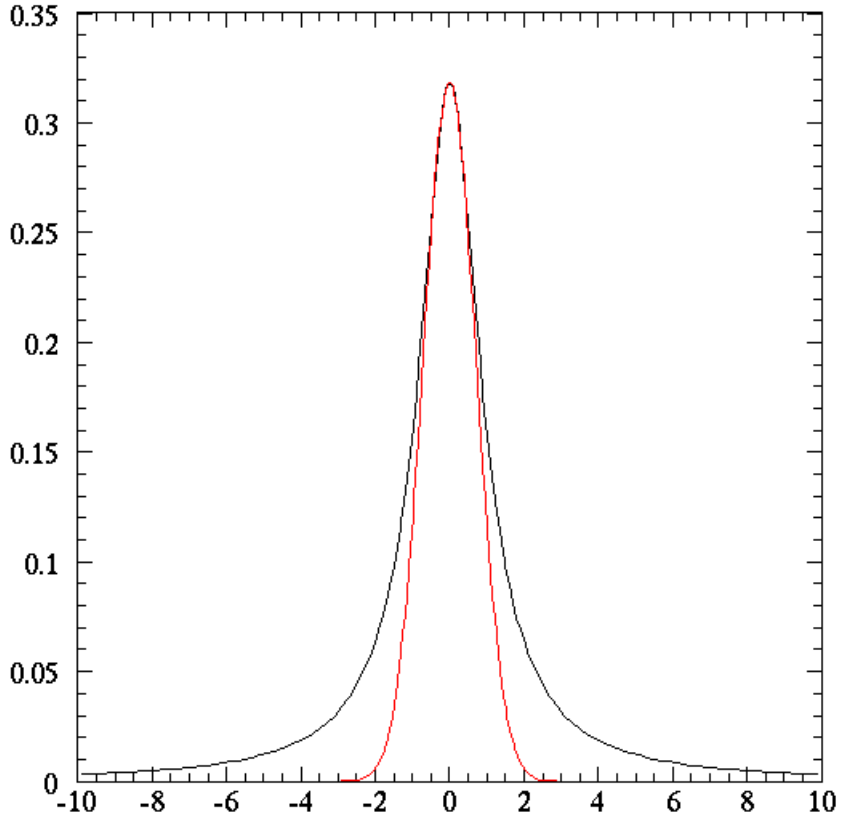
A “bad” distribution: the Cauchy distribution

Consider the Cauchy, or Breit-Wigner, distribution. Also called a “Lorentzian”. It is characterized by its centroid M and its FWHM Γ.

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

A Cauchy distribution has infinite variance and higher moments!

Unfortunately the Cauchy distribution actually describes the mass peak of a particle, or the width of a spectral line, so this distribution actually occurs!



Cauchy (black) vs. Gaussian (red)

Covariance & Correlation

The covariance between two variables is defined by:

$$\mathrm{cov}(x,y)=\langle(x-\mu_x)(y-\mu_y)\rangle=\langle xy\rangle-\langle x\rangle\langle y\rangle$$

This is the most useful thing they never tell you in most lab courses! Note that cov(x,x)=V(x).

The correlation coefficient is a unitless version of the same thing:

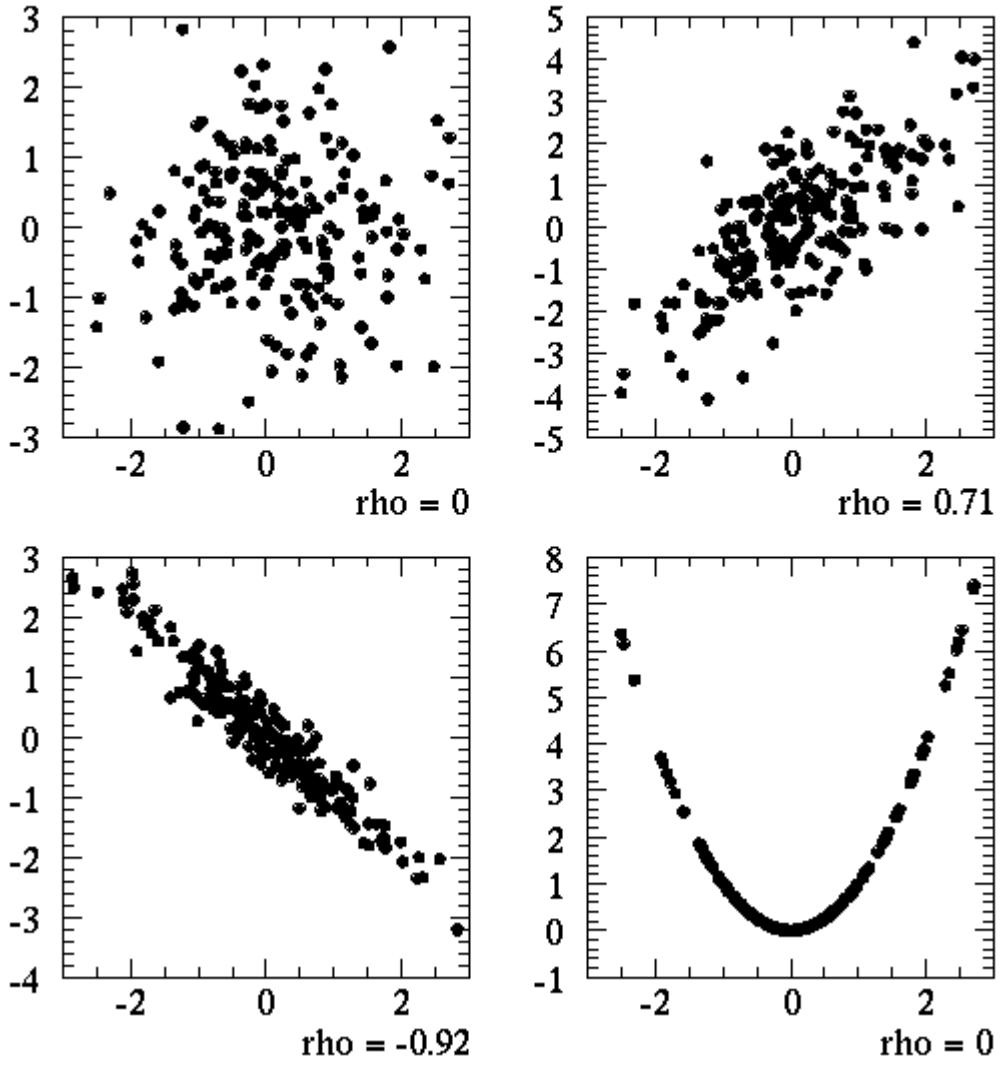
$$\rho=\frac{\mathrm{cov}(x,y)}{\sigma_x\sigma_y}$$

If x and y are independent variables *(P(x,y) = P(x)P(y))*, then

$$\mathrm{cov}(x,y)=\int dx\,dy\,P(x,y)\,xy\;-\;\left(\int dx\,dy\,P(x,y)\,x\right)\left(\int dx\,dy\,P(x,y)\,y\right)$$

$$=\int dx\,P(x)\,x\int dy\,P(y)\,y\;-\;\left(\int dx\,P(x)\,x\right)\left(\int dy\,P(y)\,y\right)=\;0$$

More on Covariance



Correlation coefficients for some simulated data sets.

Note the bottom right---while independent variables must have zero correlation, the reverse is not true!

Correlation is important because it is part of the error propagation equation, as we'll see.

Variance and Covariance of Linear Combinations of Variables

Suppose we have two random variable X and Y (not necessarily independent), and that we know cov(X,Y).

Consider the linear combinations W=aX+bY and Z=cX+dY. It can be shown that

$$\begin{aligned}\mathrm{cov}(W,Z) &= \mathrm{cov}(aX+bY,cX+dY)\\ &= \mathrm{cov}(aX,cX) + \mathrm{cov}(aX,dY) + \mathrm{cov}(bY,cX) + \mathrm{cov}(bY,dY)\\ &= ac\,\mathrm{cov}(X,X) + (ad+bc)\,\mathrm{cov}(X,Y) + bd\,\mathrm{cov}(Y,Y)\\ &= ac\,V(X) + bd\,V(Y) + (ad+bc)\,\mathrm{cov}(X,Y)\end{aligned}$$

Special case is V(X+Y):

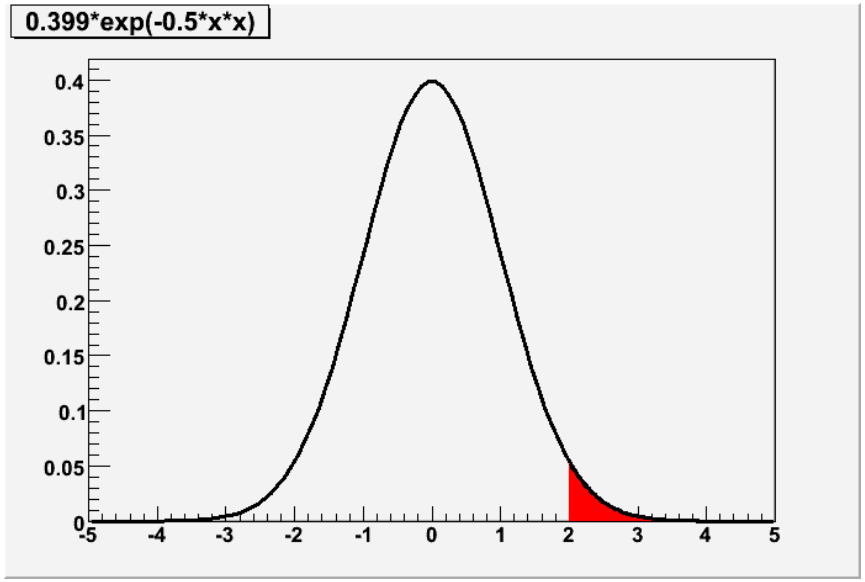
$$V(X+Y) = \mathrm{cov}(X+Y,X+Y) = V(X) + V(Y) + 2\mathrm{cov}(X,Y)$$

Very special case: variance of the sum of independent random variables is the sum of their individual variances!

Gaussian Distributions

By far the most useful distribution is the Gaussian (normal) distribution:

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



Mean = μ, Variance=σ^2

Note that width scales with σ.

Area out on tails is important---use lookup tables or cumulative distribution function.

In plot to left, red area (>2σ) is 2.3%.

68.27% of area within ±1σ
95.45% of area within ±2σ
99.73% of area within ±3σ

90% of area within ±1.645σ
95% of area within ±1.960σ
99% of area within ±2.576σ

Why are Gaussian distributions so critical?

- They occur very commonly---the reason is that the average of several independent random variables often approaches a Gaussian distribution in the limit of large N.
- Nice mathematical properties---infinitely differentiable, symmetric. Sum or difference of two Gaussian variables is always itself Gaussian in its distribution.
- Many complicated formulas simplify to linear algebra, or even simpler, if all variables have Gaussian distributions.
- Gaussian distribution is often used as a shorthand for discussing probabilities. A “5 sigma result” means a result with a chance probability that is the same as the tail area of a unit Gaussian:

$$2\int_5^\infty dt\, P(t|\mu=0,\sigma=1)$$

This way of speaking is used even for non-Gaussian distributions!

Why you should be very careful with Gaussians ..

The major danger of Gaussians is that they are overused. Although many distributions are approximately Gaussian, they often have long non-Gaussian tails.

While 99% of the time a Gaussian distribution will correctly model your data, many foul-ups result from that other 1%.

It's usually good practice to simulate your data to see if the distributions of quantities you think are Gaussian really follow a Gaussian distribution.

Common example: the ratio of two numbers with Gaussian distributions is itself often not very Gaussian (although in certain limits it may be).

A slightly non-trivial example

Two measurements (X & Y) are drawn from two separate normal distributions. The first distribution has mean=5 & RMS=2. The second has mean=3 & RMS=1. The correlation coefficient of the two distributions is ρ=-0.5. What is the distribution of the sum Z=X+Y?

A slightly non-trivial example

Two measurements (X & Y) are drawn from two separate normal distributions. The first distribution has mean=5 & RMS=2. The second has mean=3 & RMS=1. The correlation coefficient of the two distributions is ρ=-0.5. What is the distribution of the sum Z=X+Y?

First, recognize that the sum of two Gaussians is itself Gaussian, even if there is a correlation between the two. To see this, imagine that we drew two truly independent Gaussian random variables X and W. Then we could form a linear combination Y=aX+bW. Y would clearly be Gaussian, although correlated with X. Then Z=X+Y=X+aX+bW=(a+1)X+bW is the sum of two truly independent Gaussian variables itself. So Z must be a Gaussian.

A slightly non-trivial example

Two measurements (X & Y) are drawn from two separate normal distributions. The first distribution has mean=5 & RMS=2. The second has mean=3 & RMS=1. The correlation coefficient of the two distributions is ρ=-0.5. What is the distribution of the sum Z=X+Y?

Now, recognizing that Z is Gaussian, all we need to figure out are its mean and RMS. First the mean:

$$\langle X+Y\rangle = \int dX\,dY\,P(X,Y)(X+Y) = \int dX\,dY\,P(X,Y)X + \int dX\,dY\,P(X,Y)Y = \langle X\rangle + \langle Y\rangle$$

This is just equal to 5+3 = 8.

A slightly non-trivial example

Two measurements (X & Y) are drawn from two separate normal distributions. The first distribution has mean=5 & RMS=2. The second has mean=3 & RMS=1. The correlation coefficient of the two distributions is ρ=-0.5. What is the distribution of the sum Z=X+Y?

Now for the RMS. Use V(Z)=cov(Z,Z)=cov(X+Y,X+Y)

$$V(Z) = \mathrm{cov}(X,X) + 2\,\mathrm{cov}(X,Y) + \mathrm{cov}(Y,Y)$$
$$= \sigma_x^2 + 2\sigma_x\sigma_y\rho + \sigma_y^2$$
$$= (2)(2) + 2(2)(1)(-0.5) + (1)(1) = 3$$

So Z is a Gaussian with mean=8 and RMS of σ=sqrt(3)

Binomial Distributions

Many outcomes are binary---yes/no, heads/tails, etc.

Ex. You flip N unbalanced coins. Each coin has probability p of landing heads. What is the probability that you get m heads (and N-m tails)?

The binomial distribution:

$$P(m|p,N) = p^m (1-p)^{N-m} \frac{N!}{m!(N-m)!}$$

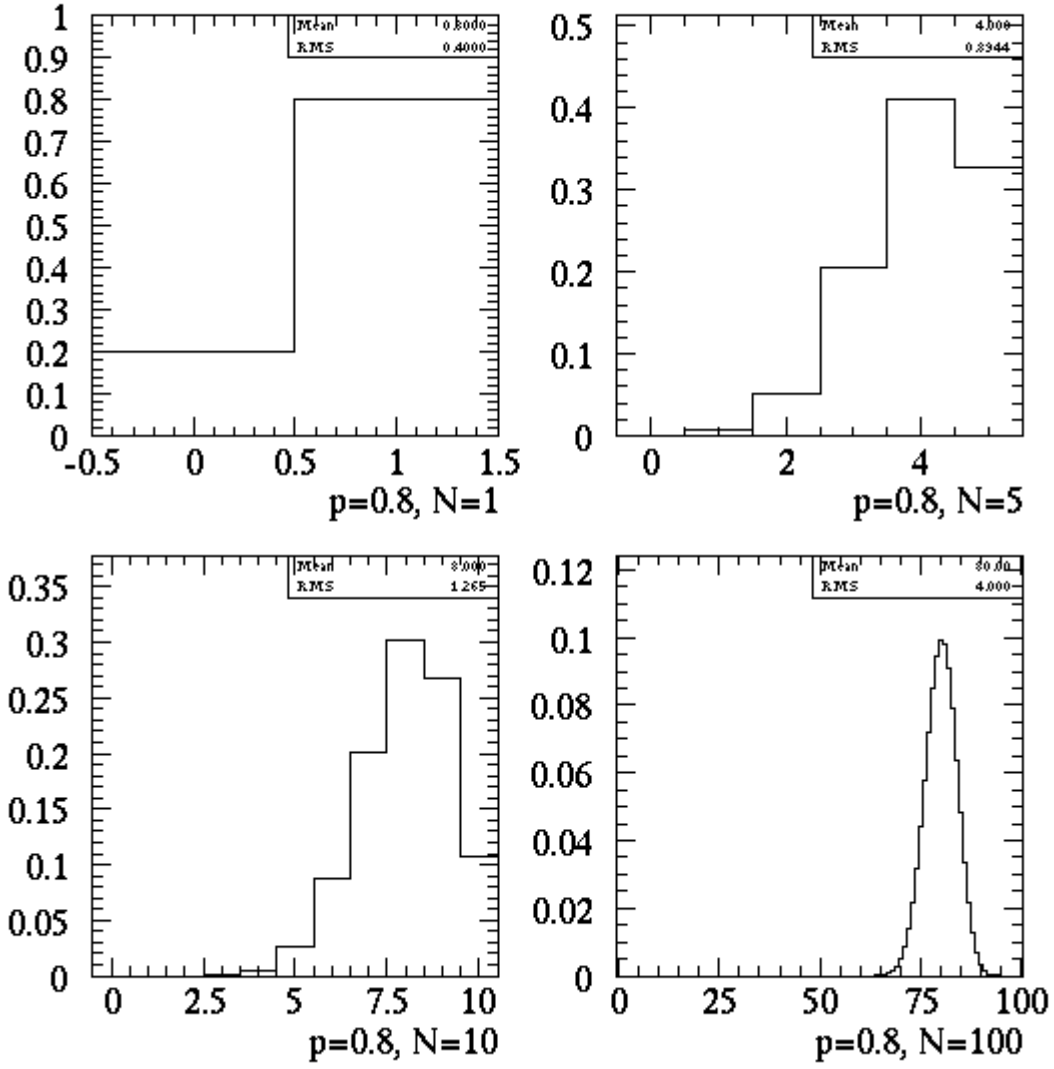
First term: probability of m coins all getting heads

Second term: probability of N-m coins all getting tails

Third term: number of different ways to pick m different coins from a collection of N total be to heads.

Binomial distributions

$$P(m|p,N) = p^m (1-p)^{N-m} \frac{N!}{m!(N-m)!}$$



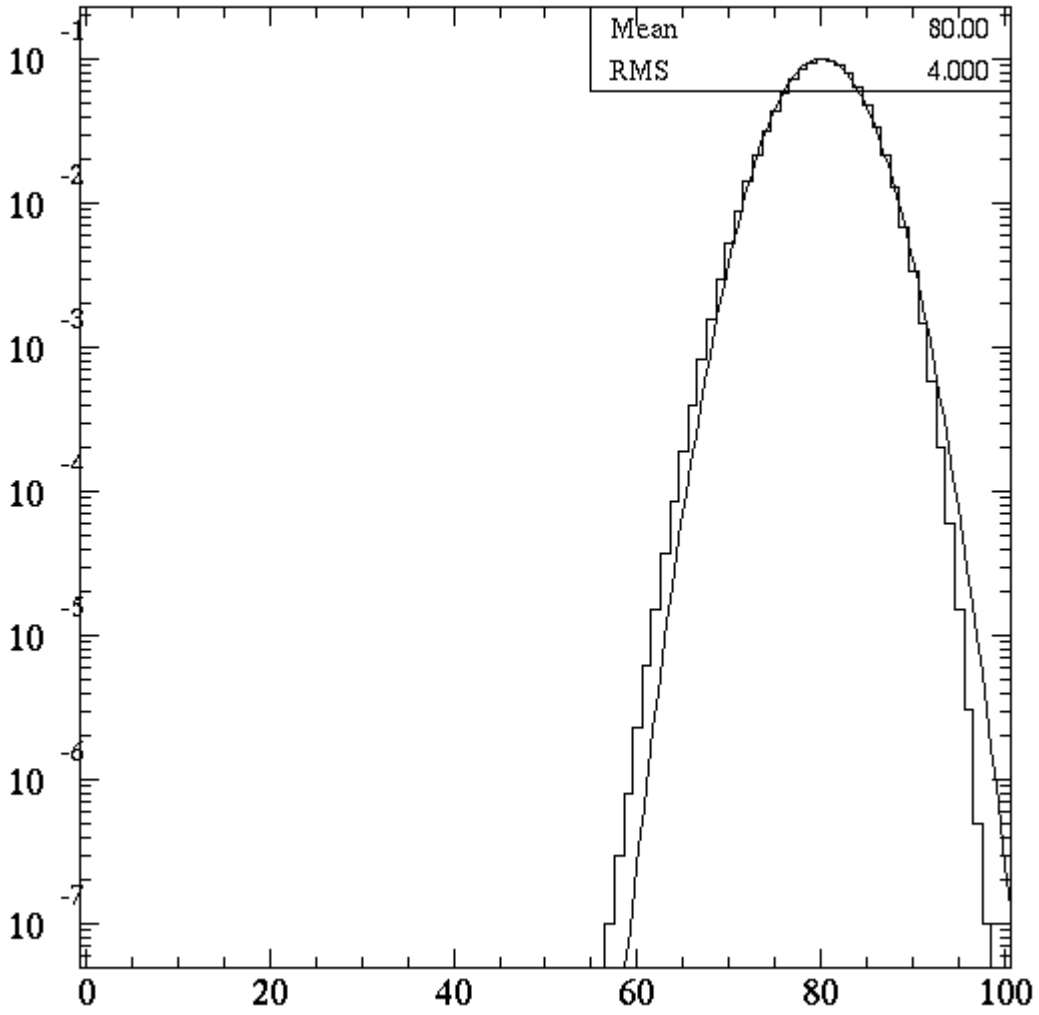
Mean = *Np*

Variance = *Np(1-p)*

Notice that the mean and variance both scale linearly with N. This is understandable---flipping N coins is the sum of N independent binomial variables.

When N gets big, the distribution looks increasingly Gaussian!

But a binomial distribution isn't a Gaussian!



Gaussian approximation fails out on the tails ...

More on the binomial distribution

In the limit of large *Np*, Gaussian approximation is decent so long as P(m=0) ≈ P(m=N) ≈ 0, provided you don't care much about tails.

Beware a common error: $\sigma = \text{sqrt}(Np(1-p))$, not $\sigma = \text{sqrt}(m) = \text{sqrt}(Np)$. The latter is only true if $p \ll 1$.

The error is not always just the simple square root of the number of entries!

Use a binomial distribution to model most processes with two outcomes:

- Detection efficiency (either we detect or we don't)
- Cut rejection
- Win-loss records (although beware correlations between teams that play in the same league)

An example from the world of sports ...

Consider a best-of-seven series ... the first team to win four games takes the prize.

We have a model which predicts that Team A is favoured in any game with p=0.6. What is the probability that A wins the series?

How could we approach this problem?

Best of 7 series: brute force

Easiest approach may be simply to list the possibilities:

A. Win in 4 straight games. Probability = p^4

B. Win in 5 games. Four choices for which game the team gets to lose. Probability = $4p^4(1-p)$

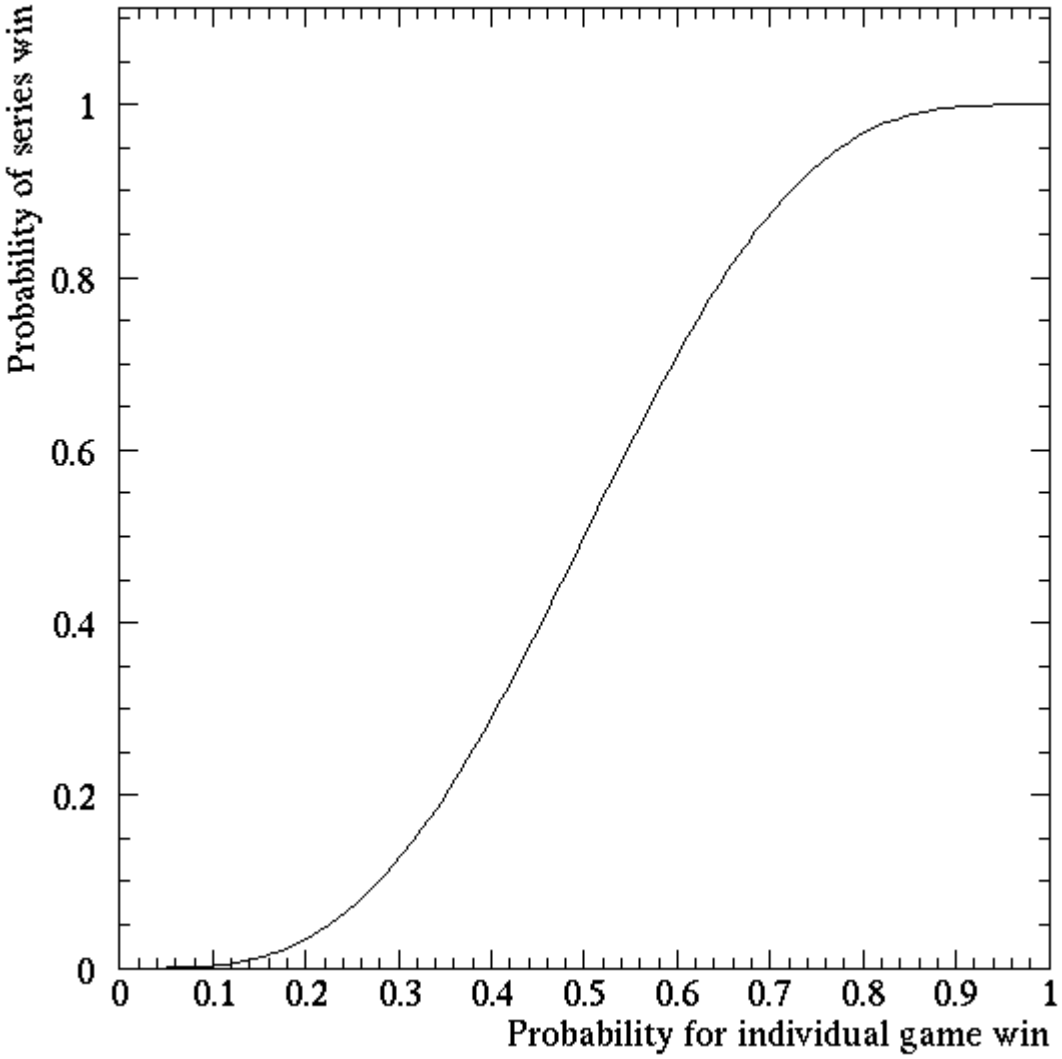
C. Win in 6 games. Choose 2 of the previous five games to lose. Probability = $C(5,2)p^4(1-p)^2 = 10p^4(1-p)^2$

D. Win in 7 games. Choose 3 of the previous six games to lose. Probability = $C(6,3)p^4(1-p)^3 = 20p^4(1-p)^3$

$$Prob(p) = p^4\left(1+4(1-p)+10(1-p)^2+20(1-p)^3\right)$$

Best of 7 series: outcomes

$$Prob(p) = p^4(1 + 4(1-p) + 10(1-p)^2 + 20(1-p)^3)$$



Symmetry evident between *p* and *1-p*, which makes good logical sense

For *p*=0.6, probability of series win is only 71%

Best of 7 series: online betting studies

Efficient market hypothesis: if a market mis-estimates a risk, smart investors will figure this out and bet accordingly, driving the odds back to the correct value. There is significant evidence that this hypothesis (almost) holds in many real-life markets.

See “A Random Walk Down Wall Street” for details.*

Does this work for online sports betting?

* Warning: reading this book may endanger your career in physics by getting you interested in quantitative analysis of financial markets.

Best of 7 series: online betting studies

I got interested in this during the 2006 baseball playoffs, as my beloved Cardinals came very close to collapsing entirely, yet went on to win the World Series.

I used a “coin flip” model to predict series odds:

- All games treated as independent, with equal probability.
- In simplest case, assume *p=0.5*
- More complicated case: using Bill James' “Pythagorean Theorem” to predict winning percentage of matchup:

$$p = \frac{(\text{Runs Scored})^2}{(\text{Runs Scored})^2 + (\text{Runs Allowed})^2}$$

My brother is stupid.

Younger brothers always are.

He objected to my coin flip model:

- Assigning 50/50 odds is ludicrous when you know the Astros will start Clemens.
- If you want to estimate *p*, you should only look at recent records to estimate odds.

How dare he deny my math!

But the proof is in the pudding ...

What do the markets say?

Odds for St. Louis to win NL Central title (going into final weekend): Coin flip model says 74.6%. Betting market said 74%.

After St. Louis loses some ground to Houston, coin flip model says 59%. The betting markets predicted 61%. My brother's "recency prior" for *p* predicts 45%.

Going into next to last day of season, my coin flip model says 89% for Cards. Betting market is mixed: odds for Cards to win are all over the map, but odds for Houston to win is right at 11%. *Opportunity for arbitrage.*

Last day: coin flip and markets both predict ~93%

In the middle of the first round of playoffs

	Coin Flip	Betting Markets
St Louis over San Diego	68.8%	66.1%
Dodgers over Mets	50.0%	44.6%
Twins over A's	31.3%	36.0%
Yanks over Detroit	68.8%	85.5%

One way to view the betting market odds is actually as an estimator of the *p* value for a matchup. For example, the market felt (wrongly) that Detroit was badly overmatched.

In the end, CARDS WIN!

Negative binomial distribution

In a regular binomial distribution, you decide ahead of time how many times you'll flip the coin, and calculate the probability of getting *k* heads.

In the negative binomial distribution, you decide how many heads you want to get, then calculate the probability that you have to flip the coin *N* times before getting that many heads. This gives you a probability distribution for N:

$$P(N|k,p)=\binom{N-1}{k-1}p^k(1-p)^{N-k}$$

Multinomial distribution

We can generalize a binomial distribution to the case where there are more than two possible outcomes. Suppose there are k possible outcomes, and we do N trials. Let n_i be the number of times that the i^{th} outcome comes up, and let p_i be the probability of getting outcome i in one trial. The probability of getting a certain distribution of n_i is then:

$$P(n_1, n_2, \ldots, n_k | p_1 \ldots p_k) = \frac{N!}{n_1! n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

Note that there are important constraints on the parameters:

$$\sum_i^k p_i = 1 \qquad \sum_i^k n_i = N$$

What is the multinomial distribution good for?

Any problem in which there are several discrete outcomes (binomial distribution is a special case).

Note that unlike the binomial distribution, which basically predicts one quantity (the number of heads---you get the number of tails for free), the multinomial distribution is a joint probability distribution for several variables (the various n_i, of which all but one are independent).

If you care about just one of these, you can marginalize over the other (sum them over all of their possible values) to get the probability distribution for the one you care about. This obviously will have a binomial distribution.

A common application: binned data! If you sample independent trials from a distribution and bin the results, the numbers you predict for each bin follow the multinomial distribution.

Dealing with binned data

Very often you're going to deal with binned data. Maybe there are too many individual data points to handle efficiently. Maybe you binned it to make a pretty plot, then want to fit a function to the plot. Some gotchas:

- Nothing in the laws of statistics demands equal binning. Consider binning with equal statistics per bin.
- Beware bins with few data points. Many statistical tests implicitly assume Gaussian errors, which won't hold for small numbers. General rule of thumb: rebin until every bin has >5 events.
- Always remember that binning throws away information. Don't do it unless you must. Try to make bin size smaller than any relevant feature in the data. If statistics don't permit this, then you shouldn't be binning, at least for that part of the distribution.