Marcello Pagano

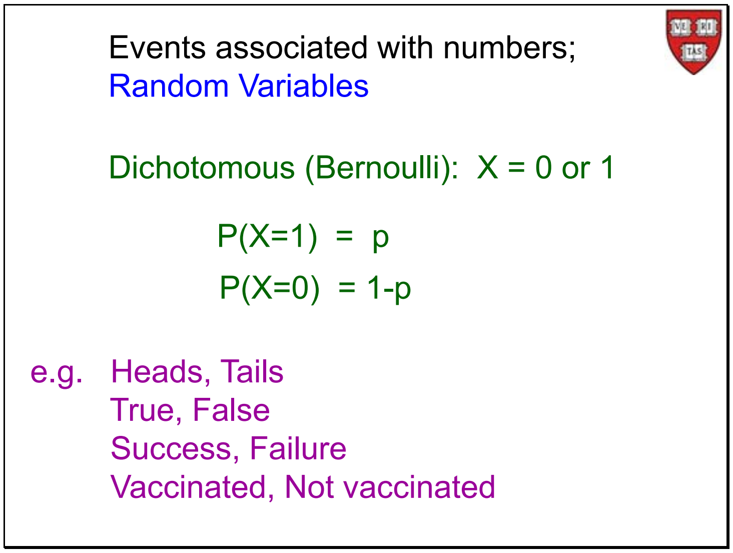
# [JOTTER 4 – PROBABILITY MODELS]

Binomial Model, Central Limit Theorem, Poisson, and Normal Models

Life’s most important questions are, for the most part, nothing but probability problems.

Pierre-Simon Laplace

Probability Models



Events associated with numbers;
Random Variables

Dichotomous (Bernoulli): X = 0 or 1

$P(X=1) = p$

$P(X=0) = 1-p$

e.g. Heads, Tails
True, False
Success, Failure
Vaccinated, Not vaccinated

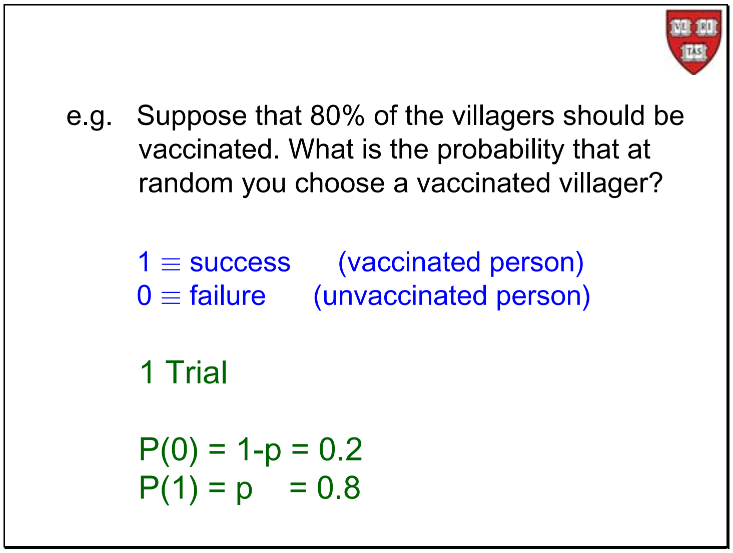
Now we are going to start applying what we learned about probability, and we start by applying probability to numbers and models. Mathematical models are an idealization, an idealization where there is right, wrong, exact, et cetera, and we now search how to use such models to approximate, or model, reality.

We start with random variables—random because we do not yet know exactly what value these variables take until we observe them. We know what values they potentially can take.

The simplest random variable is the dichotomous, or binary variable, which is also called the Bernoulli variable. It takes on one of two values; let us call them 0, and 1. We could call them 1, and 2, or any other two distinct values (nominal variable), but for the sake of definiteness, use 0 and 1.

To describe this variable probabilistically, we need to state the probability that it takes on the value 1. This also fixes the complement, namely the probability that it would take the value 0. The probability that it takes the value 1 is p. Therefore the probability that it takes the value zero is 1- p and that then covers the whole spectrum (exhaustive).

And this is what we talk about with heads, tails; true, false; success, failure etc. This is where, for example, we spin a fair coin, and thus p=0.5. We also use this simple model in the very serious application to clinical trials, when we need to randomize a patient to one of two treatments in such a way as to not show any favoritism for one treatment or the other. We return to clinical trials later.



e.g. Suppose that 80% of the villagers should be vaccinated. What is the probability that at random you choose a vaccinated villager?

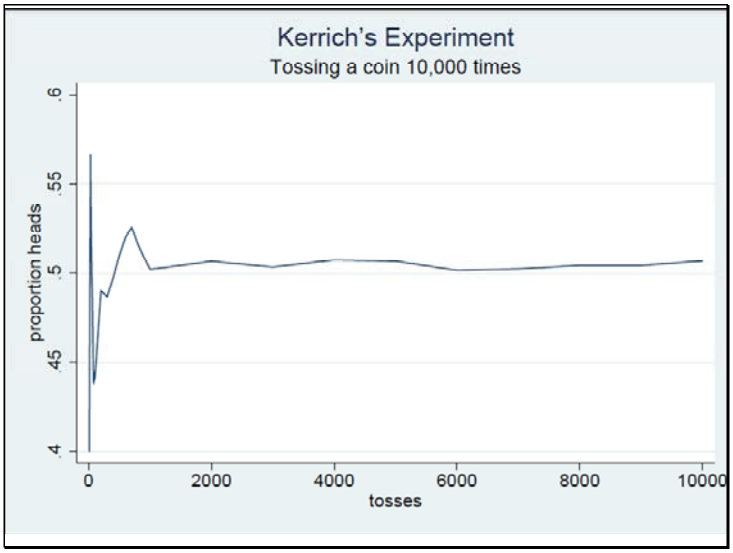
1 ≡ success (vaccinated person)
0 ≡ failure (unvaccinated person)

1 Trial

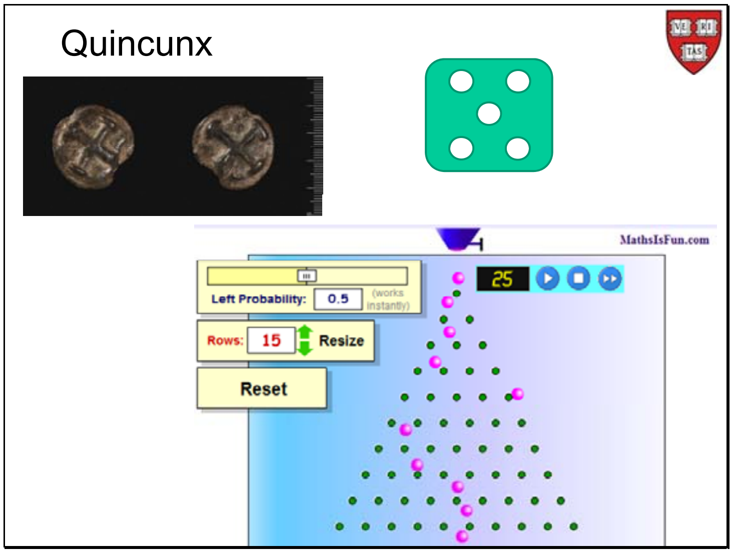
$P(0) = 1-p = 0.2$
$P(1) = p \quad = 0.8$

In our model we do not always have to have p = 0.5. For example, suppose we are concerned with vaccination coverage, then we might be aiming at a higher percentage, possibly such that 80% of the villagers should be vaccinated.

The way we model this is we say, what is the probability that at random you choose a vaccinated villager? You go into the village and choose a villager at random. What is the probability that that villager is vaccinated? We thus want to know what the probability is that we get a 1, if we label those vaccinated as 1's and those not vaccinated as 0's. So in this situation, if 80% of the villagers are vaccinated, then the probability of 1 would be 0.8.



Kerrich was faced with precisely such a situation. Remember, he spun a coin 10,000 times and he got 5,067 heads. Initially he saw a lot of variability in the ratio of heads to the number of tosses, but as time progressed it stabilized around approximately 0.5. We conceivably could repeat what Kerrich did, but remember, it took him years to perform this experiment.

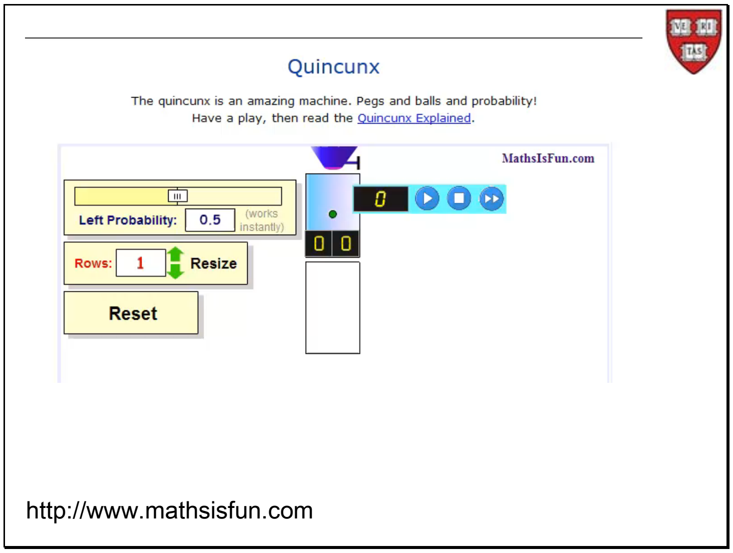


Quincunx

Left Probability: 0.5 (works instantly)

Rows: 15

Alternatively, today we can do a similar experiment on a device that makes use of computer simulations and is called a quincunx. The name comes from an old Roman coin, shown above—quinc, of course, is from the Latin quinque, meaning five. The name is now attached to a design that puts five dots on a surface, just like the face that shows a five on your typical die. The same style triangular patterns occur all over the quincunx, so Francis Galton, who invented the quincunx, so named it. (We revisit Galton when we get to correlation.)



Quincunx

The quincunx is an amazing machine. Pegs and balls and probability!
Have a play, then read the Quincunx Explained.

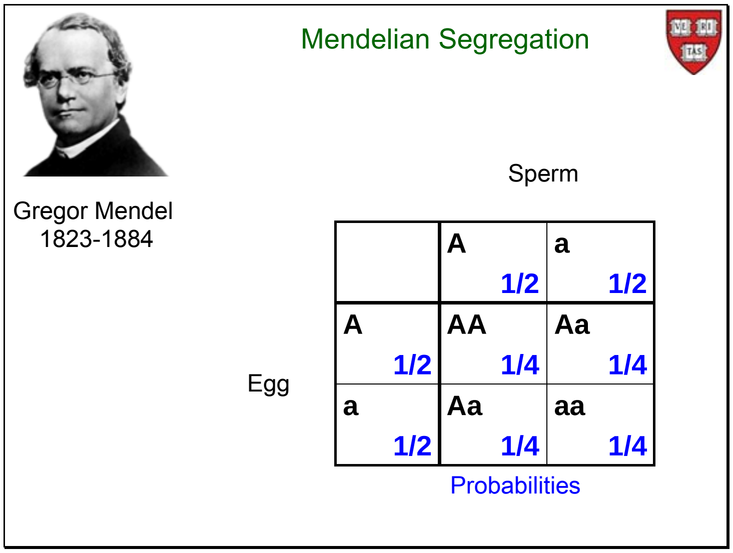
Left Probability: 0.5 (works instantly)

Rows: 1

http://www.mathsisfun.com

We can choose how many rows we want in our quincunx. Let us start with one row. Balls come out of the funnel at the top and bounce on the peg, either to the left or the right, and then collect in the wells at the bottom.

The way we've set it up is to choose “Left Probability” to be 0.5. That means that half the time the ball will bounce to the left, and half the time to the right, on average. What that means is that if you watch the wells at the bottom, then roughly half the balls gather in the left well, and half in the right well. The counter tells us exactly how many go in either well. This emulates tossing a fair coin.



Mendelian Segregation

Gregor Mendel
1823-1884

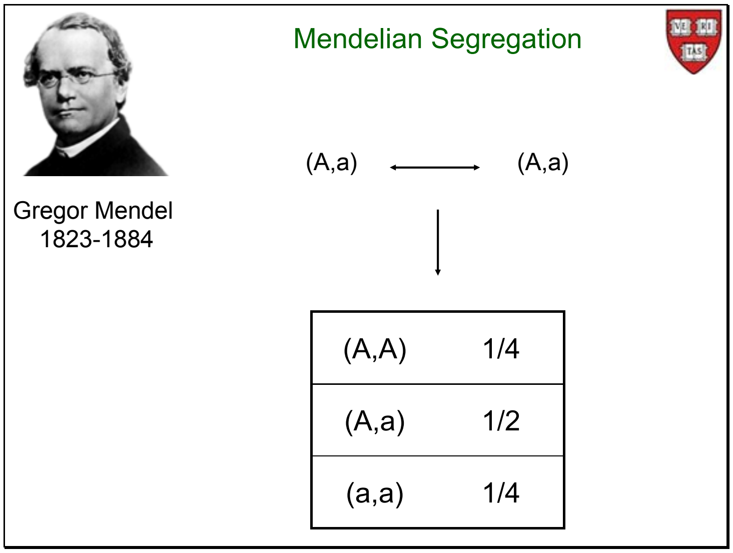
Sperm

Egg

| | A<br>1/2 | a<br>1/2 |
|---|---|---|
| A<br>1/2 | AA<br>1/4 | Aa<br>1/4 |
| a<br>1/2 | Aa<br>1/4 | aa<br>1/4 |

Probabilities

Returning to Mendelian segregation, we see that we would need to expand the Quincunx, if we wish to model that situation. Here we have two actions going on. First, what happens with the sperm and second, what happens with the egg. We can spin a fair coin to see whether the sperm carries the capital A or the little a, and similarly an independent fair coin to see whether the egg caries a capital A or a little a, before they join.



Mendelian Segregation

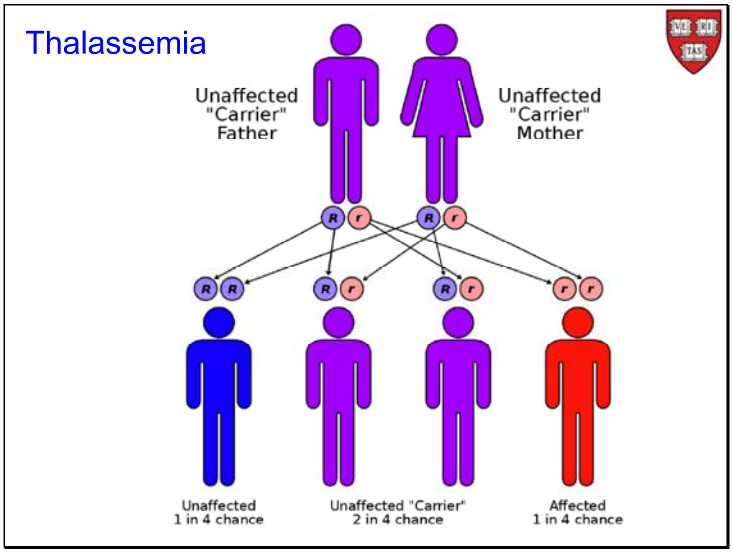
Gregor Mendel
1823-1884

(A,a) ↔ (A,a)

↓

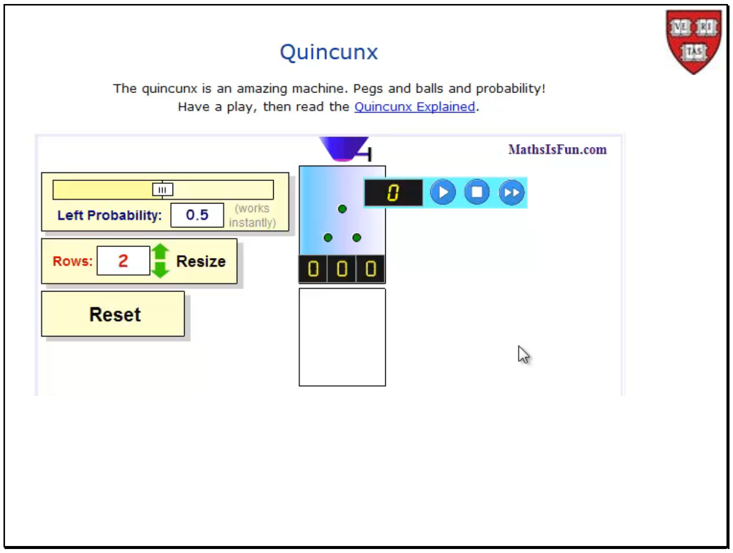
| | |
|---|---|
| (A,A) | 1/4 |
| (A,a) | 1/2 |
| (a,a) | 1/4 |

This is similar to spinning a coin twice (or equivalently spinning two coins) and associating AA with the outcome two heads, Aa with a head and a tail, and aa with two tails. We then end up with this probability distribution.



Thalassemia

And we see that the ratios are 1 to 2 to 1, for the three categories (or the three wells in the Quincunx), which is exactly the ratios we saw with the Thalassemia application.



Quincunx

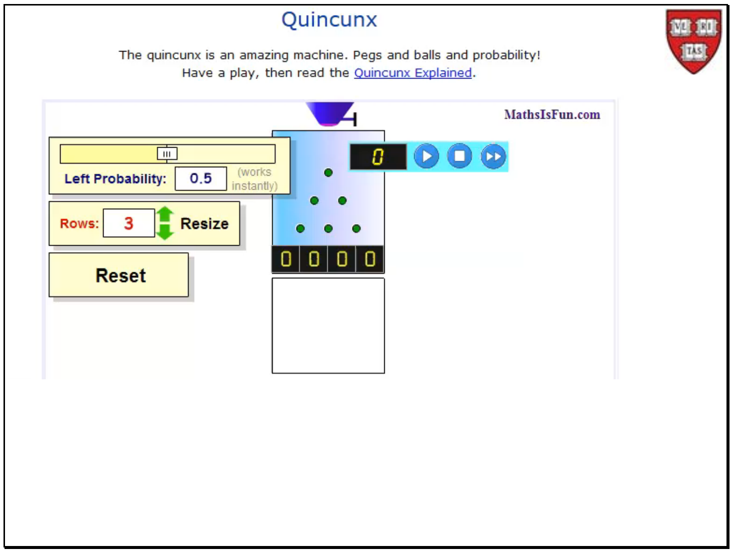
The quincunx is an amazing machine. Pegs and balls and probability!
Have a play, then read the Quincunx Explained.

Left Probability: 0.5 (works instantly)

Rows: 2

Returning to the Quincunx but this time set the number of rows to two. Now we see the ball first bounces on the top peg and then on a peg in the second row. To reach the leftmost well, the ball needs to bounce left and left.  To reach the rightmost well, the ball has to bounce right and right. Whereas, in order to reach the center well the ball must bounce either left, and then right, or right, and then left.  Since each path to the bottom is equally likely (because at each peg the ball goers left or right with equal probabilities, 0.5), that means that twice as many balls accumulate in the middle well (two ways of getting there) as in the outer two wells (only one way to get to either). So this looks like very much like the Mendelian example as well as theThalassemia situation.

Thus theory tells us that we should get roughly a 1 to 2 to 1 ratio in the number of balls in the wells, and in the long run, if we let the Quincunx run for a while, this is what we expect to see.



Quincunx

The quincunx is an amazing machine. Pegs and balls and probability!
Have a play, then read the Quincunx Explained.

Left Probability: 0.5 (works instantly)

Rows: 3

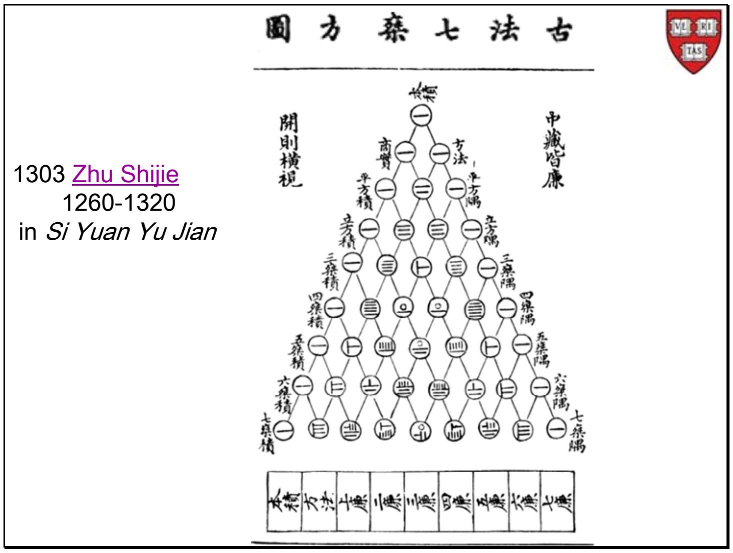
We can continue to build up the complexity of the Quincunx by adding more and more rows. For example, for the three rowed Quincunx, then there is only one way to get to the outermost wells, as usual. This time there are three ways of getting to the two middle wells (left-left-right or left-right-left or right-left-left to reach well number two from the left, for example). So the ratio for the number of balls in the four wells is going to be 1:3:3:1. You can have a lot of fun yourself when you go to mathsisfun.com (watch the s after math, this is an Australian site and they say, maths, not math) and play with the Quincunx and see what you get and how long "long-run" is.

This counting of paths to reach the bottom can quickly get tedious. But there is a device that has proven itself very useful over the years. We call it the Pascal triangle (after Blaise Pascal, 1623-1662), even though it was well known, even in Europe, before Pascal—we know that Tartaglia (1499-1557) knew about it.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Row 0 | | | | | | 1 | | | | | |
| Row 1 | | | | | 1 | | 1 | | | | |
| Row 2 | | | | 1 | | 2 | | 1 | | | |
| Row 3 | | | 1 | | 3 | | 3 | | 1 | | |
| Row 4 | | 1 | | 4 | | 6 | | 4 | | 1 | |
| Row 5 | 1 | | 5 | | 10 | | 10 | | 5 | | 1 |

We construct the triangle, which looks very much like a Quincunx, by having ones running down both edges either side. Then from top to bottom, all other entries are obtained by taking the sum of the two numbers in the row above on either side of the number you need to fill in. So 2=1+1, 3=1+2, 3=2+1, 4=1+3 etc.

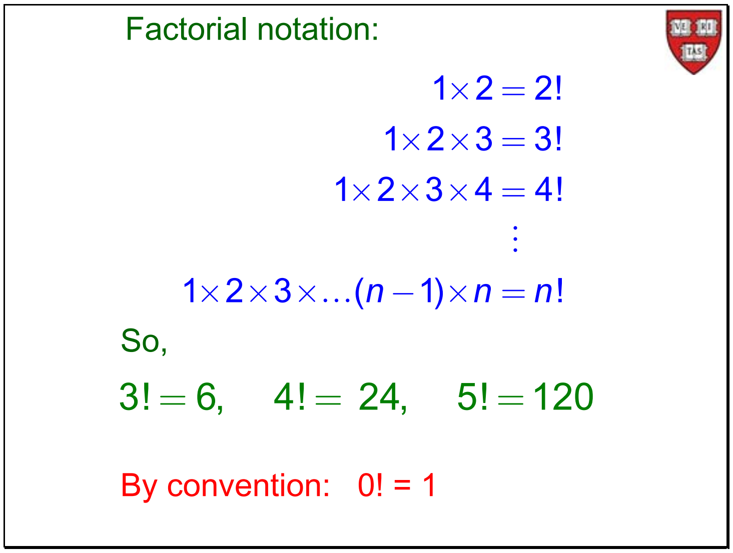
Any row in the Pascal triangle will give you the ratios of balls in the wells of the Quincunx if you let it run for long enough with "Left Probability" set at 0.5. For example, with tossing a coin we look to Row 1; for the Thalassemia (Mendel) example we look to Row 2.



1303 Zhu Shijie
1260-1320
in *Si Yuan Yu Jian*

We call it Pascal's triangle, even though it was apparently known in Persia in the 11th century or so. Here it is in a Chinese text in the 14th century.

This is all of historical interest, but it does help us understand the Quincunx.  In turn, the Quincunx helps us understand one of the most basic and beautiful models in statistics, the binomial model. Before we get there we need some notation.



Factorial notation:

$$1\times 2 = 2!$$
$$1\times 2\times 3 = 3!$$
$$1\times 2\times 3\times 4 = 4!$$
$$\vdots$$
$$1\times 2\times 3\times \ldots (n-1)\times n = n!$$

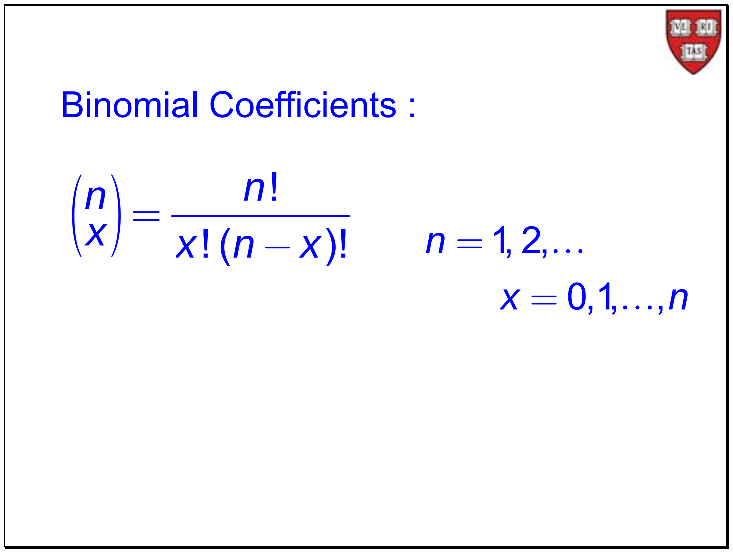
So,

$$3! = 6, \quad 4! = 24, \quad 5! = 120$$

By convention:   0! = 1

Consider the factorial notation: it is a positive integer followed by the exclamation mark.

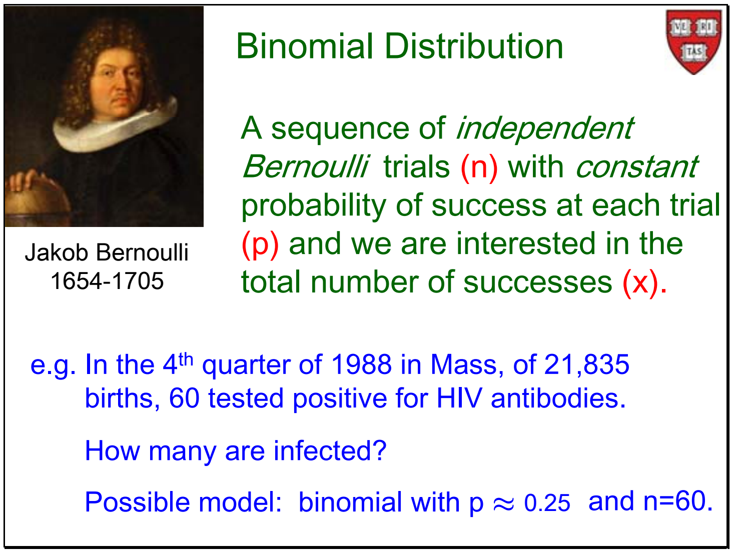
To evaluate the expression for a particular integer, we multiply that integer by all positive integers smaller than it. So for example, 2 factorial, is 1 times 2, or 2.  The product of 1, 2 and 3 is the same as 3 factorial; which, of course, equals 6. . By convention, we extend the factorial to include zero and define zero factorial to be one.



Binomial Coefficients :

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!} \qquad n = 1, 2, \ldots \qquad x = 0, 1, \ldots, n$$

This leads us to the binomial coefficients, so called because they provide the coefficients in the binomial expansion of $(a+b)^n$. For n=1, we look to Row 1 in the Pascal triangle and get 1a+1b. For n=2, we look to Row 2 and get $(a+b)^2 = 1a^2 + 2ab + 1b^2$. For n=3, we look to Row 3 and get $(a+b)^3 = 1a^3 + 3a^2b + 3\ a\ b^2 + 1\ b^3$, and so on.



Binomial Distribution

Jakob Bernoulli
1654-1705

A sequence of *independent Bernoulli* trials (n) with *constant* probability of success at each trial (p) and we are interested in the total number of successes (x).

e.g. In the 4th quarter of 1988 in Mass, of 21,835 births, 60 tested positive for HIV antibodies.

How many are infected?

Possible model: binomial with $p \approx 0.25$ and n=60.

We now have all the background we need to introduce one of the most important fundamental models in all of statistics; the binomial model. All such models, including the model of independence that we have studied and Sally Clark got hurt with, have conditions which have to be satisfied in order for the model to be correctly applied. The condition of independence was not satisfied in the Sally Clark situation, and that led to a miscarriage of justice.

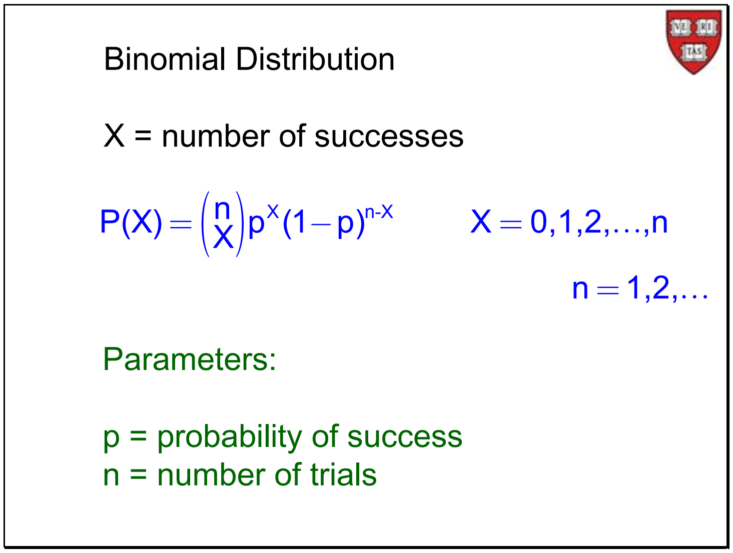
The conditions under which we can apply the binomial distribution, are: First, that we have a sequence of independent Bernoulli trials—named for the person Jakob Bernoulli, a brilliant mathematician from a family of brilliant mathematicians, who did the early work on this model. Second, that there are a fixed number, n, of such trials. Third, that the probability of success at every single trial is a constant, call it p. Fourth, that we have a fixed number of trials.

For example, our “independent trials” might well be patients. We need to ask ourselves, does it make sense to think of these patients as independent? Will each patient have the same chance of success? If we can answer these in the affirmative, then we can use the binomial distribution. And the answers we get from the model will only be as good as how well the assumptions are satisfied.

If the assumptions are not satisfied, then we should not use the model. Note that we also need to have a fixed number of trials, and the binomial model is not appropriate when you continue the trials until you have a success, for example—think of having children till you have one of a particular sex.

If the conditions are satisfied, then we can apply this model. For example, in the fourth quarter of 1988, there were 21,835 births in the Commonwealth of Massachusetts which babies, at birth were tested for HIV. That test was an antibody test and 60 of these babies tested positive, which meant that the mothers of those 60 babies were infected with the HIV.  The question then arises, in order to plan health services for these babies, how many of them can we expect to be infected with the HIV?  At that time the mother to infant transmission rate was about 25%, so one possibility is to model the situation as a Binomial with n=60 and p=0.25, if we assume the babies independent of each other.

This model is an idealization that hopefully yields some guidance. This is how we use models in biostatistics and epidemiology, in public health, in medicine et cetera.



Binomial Distribution

X = number of successes

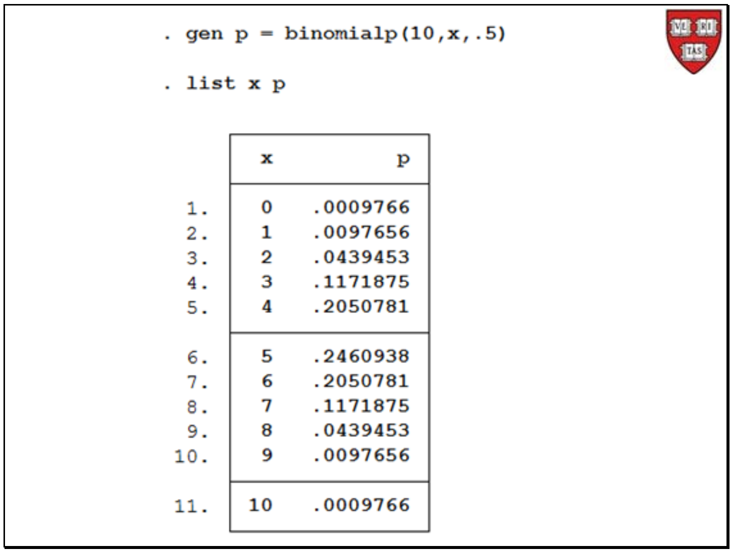
$$P(X) = \binom{n}{X} p^{X} (1-p)^{n-X} \qquad X = 0,1,2,\ldots,n$$

$$n = 1,2,\ldots$$

Parameters:

p = probability of success
n = number of trials

The binomial model yields the binomial distribution that we would use to calculate the probability of obtaining X successes in n independent trials when the probability of success at each trial is p. So with the babies above, we would put n=60 and p=0.25.

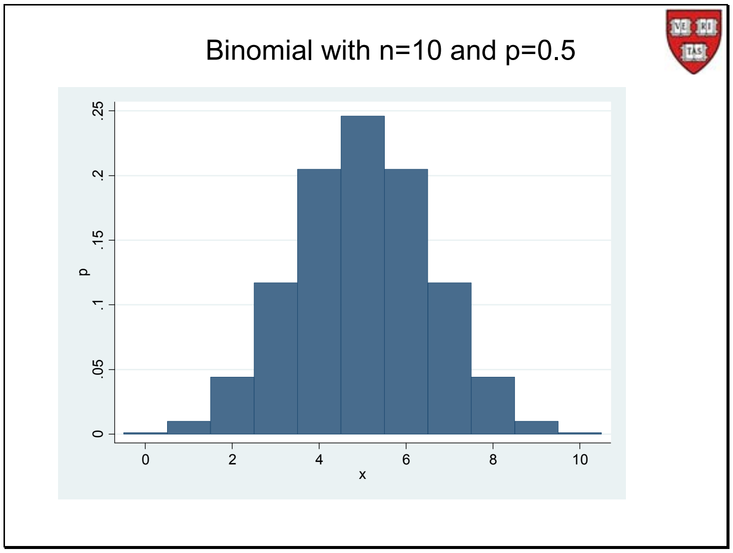


```
. gen p = binomialp(10,x,.5)

. list x p

     +--------------+
     |  x         p |
     |--------------|
  1. |  0   .0009766 |
  2. |  1   .0097656 |
  3. |  2   .0439453 |
  4. |  3   .1171875 |
  5. |  4   .2050781 |
     |--------------|
  6. |  5   .2460938 |
  7. |  6   .2050781 |
  8. |  7   .1171875 |
  9. |  8   .0439453 |
 10. |  9   .0097656 |
     |--------------|
 11. | 10   .0009766 |
     +--------------+
```

Of course you are not going to do these calculations yourself. You will ask Stata to do the hard work for you. In Stata it is a function called *binomialp* that has to be summoned. Here, for example are all 11 values it can take when n=10 and p=0.5. So for x=0 we get that p=0.0009766, which is also the probability of getting zero heads when tossing a fair coin 10 times.

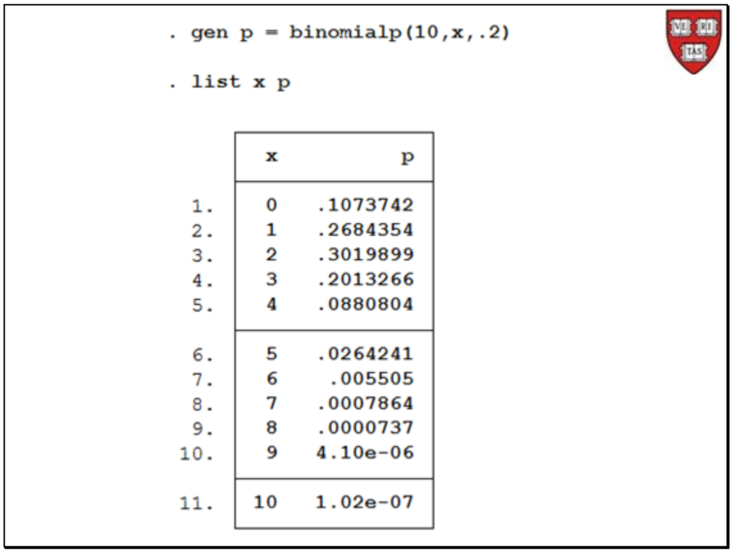
The first thing to notice is the up and down symmetry around x=5. So the probability of zero heads is the same as the probability of 10 heads. So too the probability of 1 head is the same as the probability of 9 heads. And so on. This is due to p=0.5 and the resultant symmetry and also in the arbitrariness in which side of the coin is called a head and which side is called a tail.

So that makes sense. Further, 5 is the most popular value (the mode), it is also actually the mean, which also fits in with intuition: if we spin a fair coin 10 times we expect 5 heads.



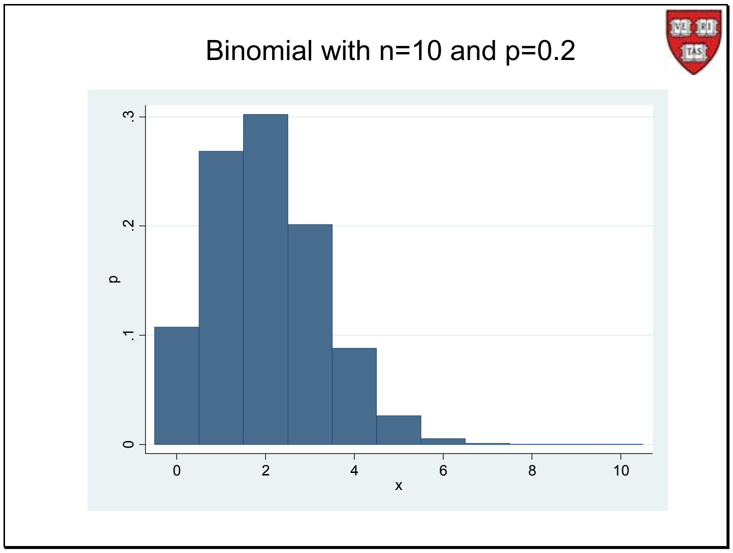
Let's plot it and this is what we get. So this is what your bin or the histogram underneath your bin in the Quincunx should look like if you run the Quincunx with 10 rows and p=0.5. Try it.

If you do, see how long it takes to get a shape that reasonably resembles this graph. How long is the “long run.”

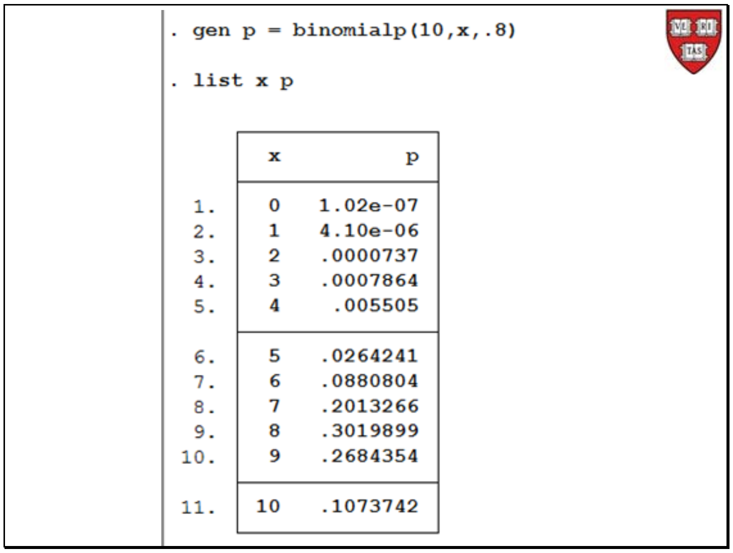


```
. gen p = binomialp(10,x,.2)

. list x p

     +---------------+
     |  x          p |
     |---------------|
  1. |  0   .1073742 |
  2. |  1   .2684354 |
  3. |  2   .3019899 |
  4. |  3   .2013266 |
  5. |  4   .0880804 |
     |---------------|
  6. |  5   .0264241 |
  7. |  6    .005505 |
  8. |  7   .0007864 |
  9. |  8   .0000737 |
 10. |  9   4.10e-06 |
     |---------------|
 11. | 10   1.02e-07 |
     +---------------+
```

If we change the p from 0.5 to 0.2, we would expect to see fewer successes and thus lose the symmetry. Here is the distribution. Now the mode and mean are at 2 (since 10x0.2=2).

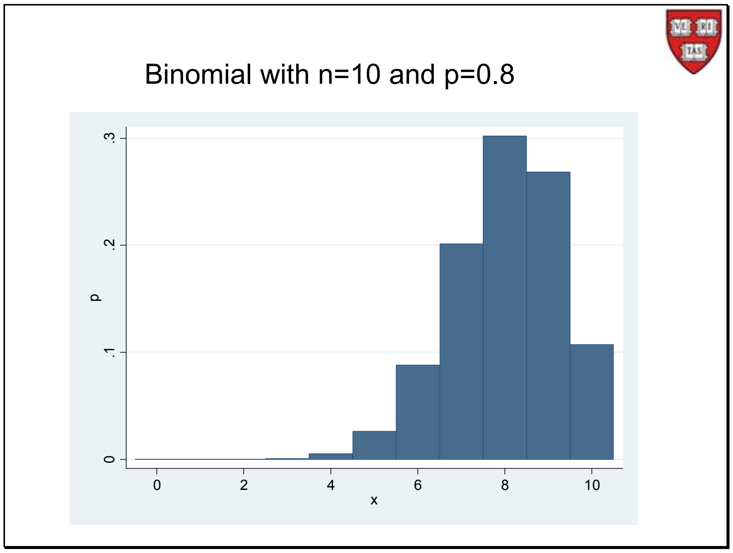


If we plot the distribution we get this graph. We see the symmetry gone and a tail appearing on the right.



```
. gen p = binomialp(10,x,.8)

. list x p

     +---------------+
     |  x          p |
     |---------------|
  1. |  0   1.02e-07 |
  2. |  1   4.10e-06 |
  3. |  2   .0000737 |
  4. |  3   .0007864 |
  5. |  4    .005505 |
     |---------------|
  6. |  5   .0264241 |
  7. |  6   .0880804 |
  8. |  7   .2013266 |
  9. |  8   .3019899 |
 10. |  9   .2684354 |
     |---------------|
 11. | 10   .1073742 |
     +---------------+
```

If we switch the probability of success from 0.2 to 0.8, then we should expect to see the mirror image of what we just saw—interchange your definitions of success and failure. So now our most popular value should be 8 (=10x0.8).

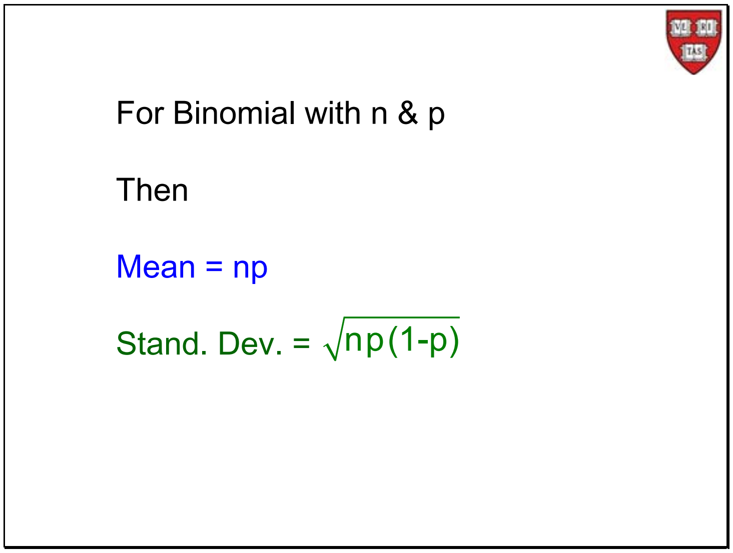


Binomial with n=10 and p=0.8

This is confirmed in the graph, as is the tail switching from the right to the left.

We can use this model directly if we ask about the 60 babies who tested positive for the antibodies to HIV where we said that the probability, of transmission, is 0.25 that any one of them actually has the virus. We have a fixed number of them, 60. Then if we assume the probability the same for each one of them and that the babies are independent of each other, we can use the binomial distribution to evaluate the probability for any number of them to be infected.

We can also reverse the logic. This graph gives us the probabilities for any particular number of successes. So let us ask, if we visit a village where the vaccination coverage is 80% (so p=0.8) and I choose 10 villagers at random, what do I expect to see. In answer we can see where most of the mass is, mostly around 8, so that is what we expect to see. But what if I tell you that of the ten people, none were vaccinated; or one was; or two were; or three were. In each one of these cases you would say to me, “That is virtually impossible. Look, there is no probability mass down at the left end of the curve.” So in such instances, what are we left with? Well we can argue that an extremely rare event has occurred, or we can seek an alternative explanation. One possibility is to question the validity of the assumption that p=0.8. In other words, is it possible that the vaccination crew missed this village in its last rounds? Or, was there a huge immigration of unvaccinated individuals into this village? This is how we use these models, namely to tell us what to expect and then to contrast that to what we observe, and then possibly question the assumptions so as to improve conditions.



For Binomial with n & p

Then

Mean = np

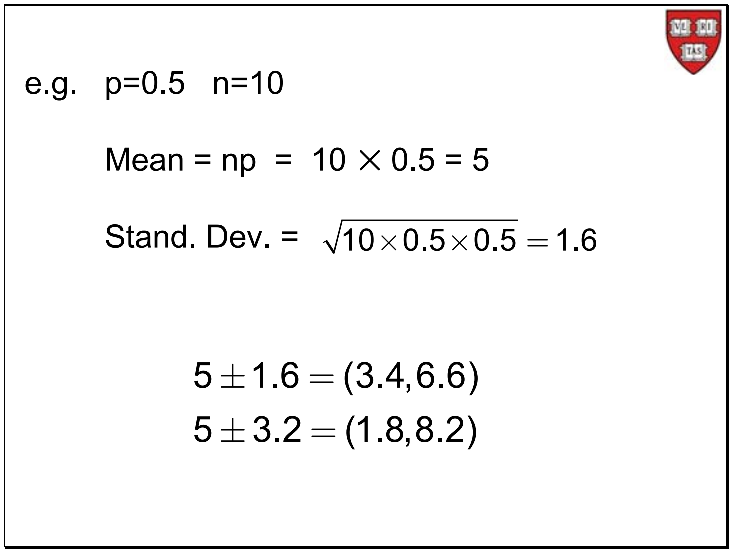
Stand. Dev. = $\sqrt{np(1-p)}$

The binomial has two parameters, n and p. Remember, when we looked at the empirical rule, we said look that the mean and the standard deviation of a variable summarizes the distribution. So, if the conditions for the empirical rule are obeyed, then we could use it.

We saw that for p=0.5 the binomial is perfectly symmetric. It also has finite tails, so they certainly go to zero very fast (they are thin). The mean for a binomial is np, and the variance is np(1-p). So if we spin a fair coin 10 times, the mean is 5 and the standard deviation is ½ $n^{1/2}$, or 1.6.

This is exactly what you might have suspected; for example, if p=0.5, and you spin a coin 10 times, you expect half of those values, namely 5, to be heads.

The formula for the standard deviation is not intuitive at all. This takes some mathematics to work this out.



e.g. p=0.5 n=10

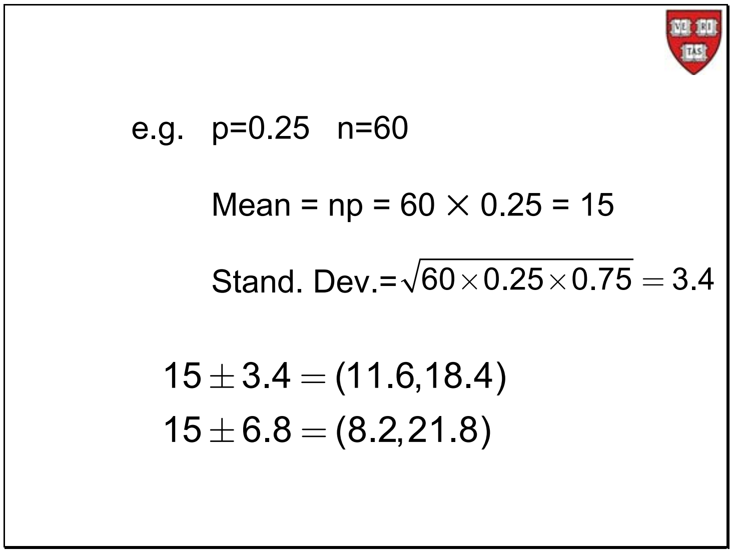
Mean = np = 10 × 0.5 = 5

Stand. Dev. = $\sqrt{10 \times 0.5 \times 0.5} = 1.6$

$$5 \pm 1.6 = (3.4, 6.6)$$
$$5 \pm 3.2 = (1.8, 8.2)$$

Applying the empirical rule to this, we get that two-thirds of the time we should get between 3.4 and 6.6 heads, and 95% of the time we should get between 1.8 and 8.2 successes.

These intervals are quite wide, and in fact are the widest since the worst variance (p(1-p)) occurs when p=0.5.



e.g. p=0.25 n=60

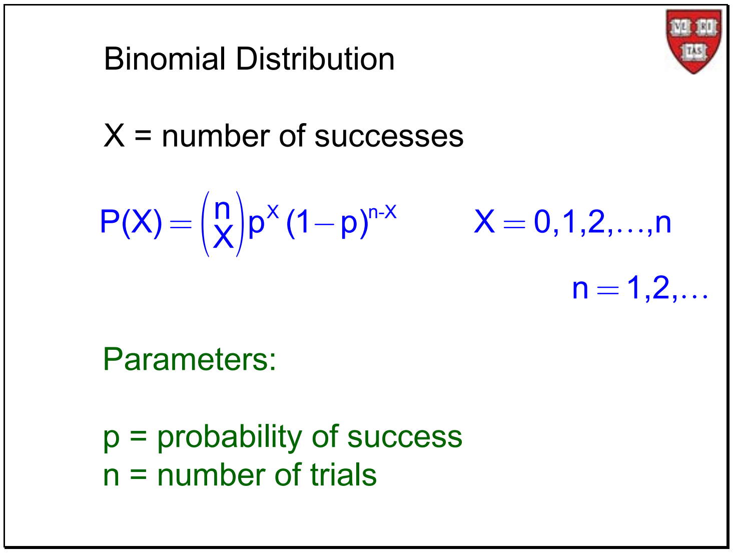
Mean = np = 60 × 0.25 = 15

Stand. Dev.=$\sqrt{60 \times 0.25 \times 0.75} = 3.4$

$$15 \pm 3.4 = (11.6, 18.4)$$
$$15 \pm 6.8 = (8.2, 21.8)$$

If we return to our babies tested for HIV, p=0.25 and n=60, and here are the two intervals.

We must be a little more cautious when applying the empirical rule here because p is no longer 0.5, so the symmetry no longer holds. We can, using Stata, check exactly what proportions fall into these intervals.



Binomial Distribution

X = number of successes

$$P(X) = \binom{n}{X} p^X (1-p)^{n-X} \qquad X = 0,1,2,\ldots,n$$

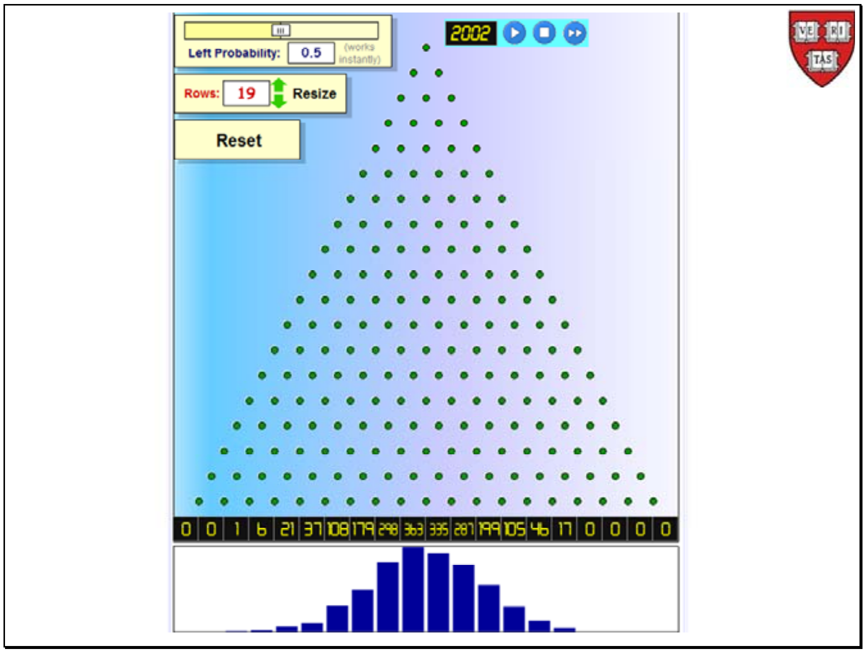
$$n = 1,2,\ldots$$

Parameters:

p = probability of success
n = number of trials

Something wonderful happens to the binomial when n becomes large. Today we have the computer to do all our calculations so we can be cavalier about hard calculations, relying on intelligent numerical analysts to do the right thing, but in the old days, all of these calculations had to be done by hand, and as a result, there were lots of errors. Also they were very difficult to do in this case when n was large, especially if p was small. The result was multiplying very big numbers by very small numbers and adding the resultant to small numbers and so on.  The calculations took a tremendous amount of time and their accuracy was not to be trusted.

We are fortunate because of the computer we can see what happens when n is large.  Make n as large as you can on the Quincunx; i.e. n=19. Now run it with p=0.5 for a long time.  Go make yourself a sandwich in the kitchen and come back. What do you see? Here is what I saw,



Left Probability: 0.5 (works instantly)

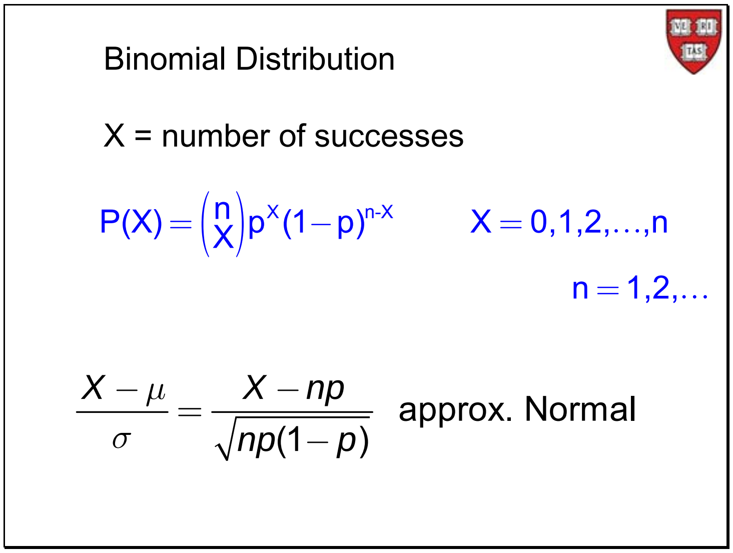
Rows: 19

2002

0 0 1 6 21 37 108 179 298 363 335 281 199 105 46 17 0 0 0 0

It looks lovely. We see all these pink balls in a cascade, bouncing from peg to peg and building up a wonderful design at the bottom. Here is the result after 2002 balls came down—I tried to stop it at exactly 2000 but I was not successful. It's a lovely shape. The magic is that no matter how often you restart it and repeat this experiment, you get the same shape at the bottom! This is the magic of mathematics, the magic of statistics.

And this is what de Moivre discovered—he did not have the Quincunx to point the way, so you can imagine how brilliant he must have been! As a sidebar, De Moivre, apparently shared a characteristic with Cardano before him; namely, both are said to have predicted their own deaths. De Moivre based his prediction on mathematics: he discovered that he was sleeping an *extra* 15 minutes each night, so he posited that he would die the day he slept for 24 hours. He did.



Binomial Distribution

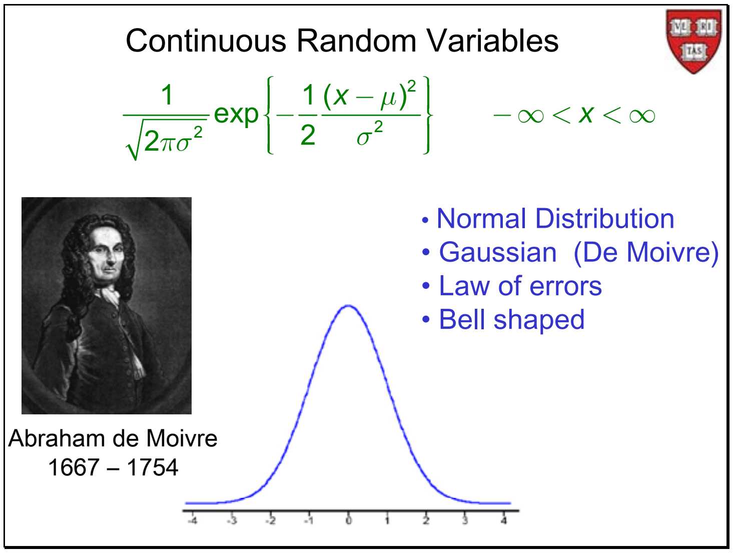
X = number of successes

$$P(X) = \binom{n}{X} p^{X}(1-p)^{n-X} \qquad X = 0,1,2,\ldots,n$$

$$n = 1,2,\ldots$$

$$\frac{X-\mu}{\sigma} = \frac{X-np}{\sqrt{np(1-p)}} \quad \text{approx. Normal}$$

The way to describe what De Moivre discovered is to say, if we look at our binomial variable, X, and standardize it by subtracting its mean, np, and dividing by its standard deviation, $\sqrt{np(1-p)}$, then this standardized variable, as n gets very large, can be treated like a standard normal variable.



## Continuous Random Variables

$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\} \qquad -\infty < x < \infty$$

- Normal Distribution
- Gaussian  (De Moivre)
- Law of errors
- Bell shaped

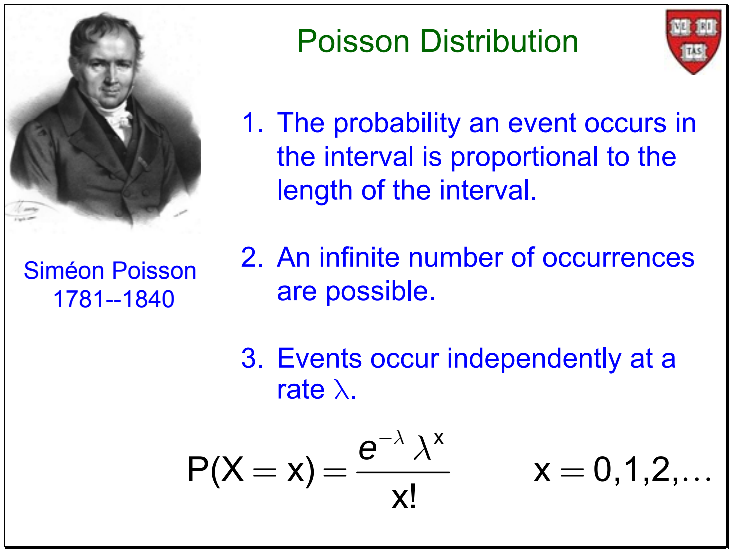
Abraham de Moivre
1667 – 1754

That means that the distribution, or density, will always have this shape, as defined by the mathematical expression above, as n becomes large.  This is the distribution of a continuous random variable.  A fantastic result.

Why this distribution is not called the Demoivrian and why it is called the Gaussian—after a person who came some one hundred years later—distribution is a puzzle the historians will have to unravel.  This is the first example of what we revisit shortly, a central limit theorem, that is central to the practice of statistics. This is also sometimes called the normal distribution, because at one time we thought that every variable would have this distribution. It is also called the bell shaped curve, presumably because bells look like this.

The normal distribution, has two parameters, mu and sigma squared. So we are back to summarizing a distribution, just as we did in our empirical rule, by looking at the mean and the standard deviation. In the normal case, we get the entire distribution of the entire population by specifying just those two parameters. We return to it shortly.

Poisson Distribution



Poisson Distribution

Siméon Poisson
1781--1840

1. The probability an event occurs in the interval is proportional to the length of the interval.
2. An infinite number of occurrences are possible.
3. Events occur independently at a rate $\lambda$.

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad x = 0,1,2,\ldots$$

Another wonderful result occurs when we let n get large in the Binomial model, but this time we also let p simultaneously get very small. If this were to happen, we arrive at what is sometimes called the model for rare events, or more commonly, the Poisson model.

Mathematically, Poisson stipulated not only that n should get large, and that p get extremely small, but in such a way as to have their product, np, approach a constant, we call lambda, λ. We know that the mean of the Binomila is np, so λ is the mean of our new distribution.

When might such a model be appropriate? Think of a traffic intersection in your town. The probability of there being an accident when you observe the intersection for a minute, say, is probably negligible. But if you look at it long enough, say for a year, you observe two accidents, something of that order.

So there we have a situation where if you observe the process for a long time, so your n is really big, your p, the probability that you have an accident in any one of those minutes is extremely small—for example there are approximately 60x24x365.25 = 525,960 minutes in a year. So if we say that an accident is equally likely to happen at any one of those minutes—probably a gross oversimplification—and two happen a year, then p = $4\text{x}10^{-6}$ ; a very small number. But the product, huge n times tiny p, to represent what happens in a year, can be a reasonable number—in our example it is 2.

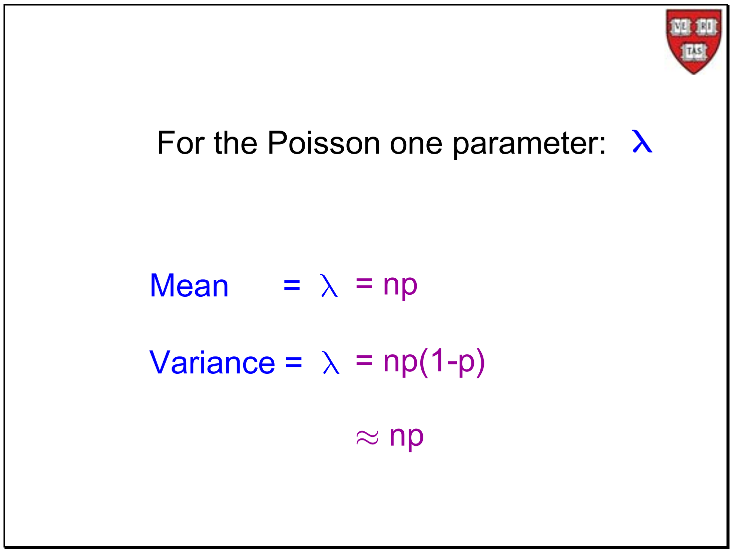
So Poisson set up the conditions under which his model is correct—remember, we continue to work under the premise that models are idealizations that we wish to use to approximate reality—and they are:

1. The probability that an event occurs in an interval is proportional to the length of the interval (So if I watch it for twice as long, then my probability is going to be twice as big that I will observe an accident).

2. To make the mathematics possible, an infinite number of occurrences are possible (We know that can't happen, of course. But we'll see in a minute that that's not that big an assumption.)
3. The third assumption is a big one, and that is that events occur independently (So if we're watching the intersection and there was an accident yesterday or the day before, that's not going to make any difference to whether there's an accident today, let's say. So there is a certain amount of independence.)

These are Poisson’s three conditions. If they hold, then the probability of having x events, or accidents, in a given year is given by the formula above.

Looking at the formula, we see a single parameter, λ. The conditions for the model are strict, but if you can apply the model, then you only need specify a single parameter.



For the Poisson one parameter: $\lambda$

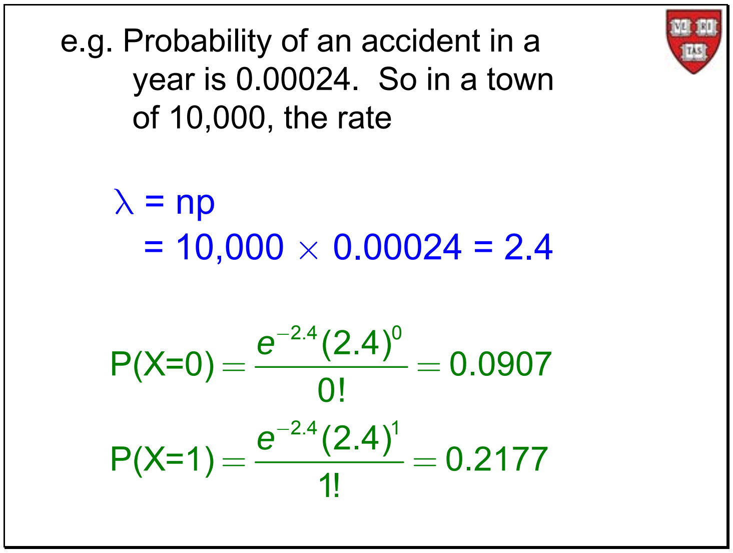
Mean $= \lambda = np$

Variance $= \lambda = np(1-p)$

$\approx np$

Let us look at this single parameter a little more closely.  We defined it as equal to np. From the binomial we know that that is its mean, so lambda is also the mean of the Poisson. Now consider the variance of the binomial.  As n gets large, p gets small, and np equals lambda, that means that the variance of the Poisson is also lambda. Thus lambda is both the mean and the variance.

This is why we often use the Poisson model when we study a phenomenon where the variance increases with the mean.



e.g. Probability of an accident in a year is 0.00024. So in a town of 10,000, the rate

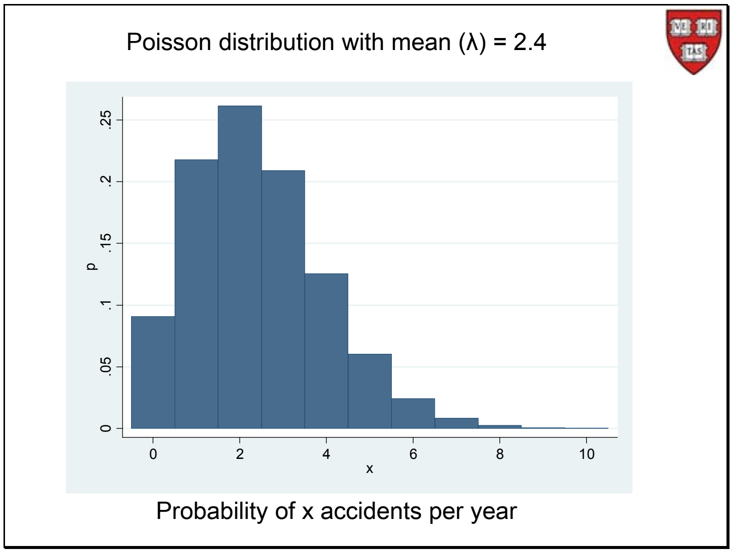
$\lambda = np$
$= 10{,}000 \times 0.00024 = 2.4$

$$P(X=0) = \frac{e^{-2.4}(2.4)^0}{0!} = 0.0907$$

$$P(X=1) = \frac{e^{-2.4}(2.4)^1}{1!} = 0.2177$$

Let us look at an example. Return to our accident example, except let us look at it from the perspective of each person. Suppose that the probability that any one person has an accident in a year is 0.00024; an extremely tiny probability. Suppose further that we are talking about a village that has 10,000 inhabitants. So np = λ = 2.4, and we expect a total of 2.4 accidents a year in this village.

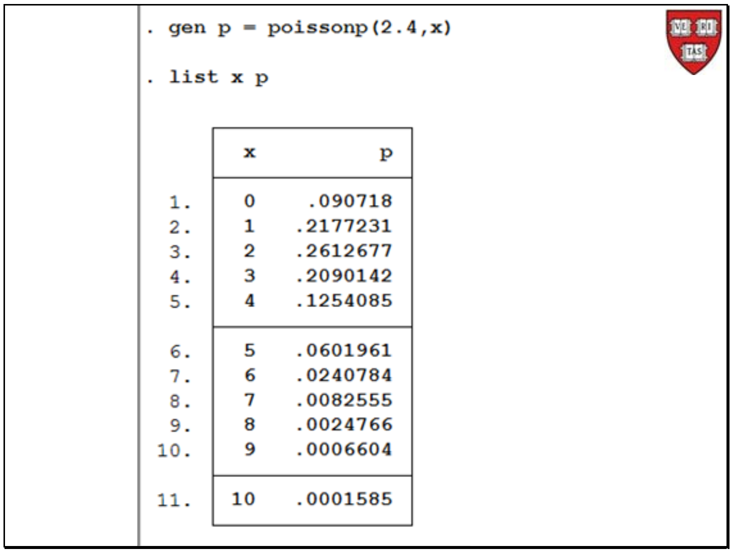
So if we fit a Poisson to this situation, we put λ = 2.4 into our Poisson formula to get that the probability of no accidents in a year is 0.0907, or approximately 0.1. So we expect that one in ten years we will not see any accidents in that village. The probability of one accident in a year is 0.2177. So we expect every five years to have a single accident in the village. And so on.



Poisson distribution with mean (λ) = 2.4

Probability of x accidents per year

We can plot this distribution and here is what it looks like. So the most popular value (mode) is 2 and then things tail off after two. And you can see it's essentially 0 by the time we get to 10. So the fact that theoretically we can go to infinity is not a big deal, because we have most of the probability mass before we reach 10.



```
. gen p = poissonp(2.4,x)

. list x p

     +--------------+
     |  x         p |
     |--------------|
  1. |  0   .090718 |
  2. |  1  .2177231 |
  3. |  2  .2612677 |
  4. |  3  .2090142 |
  5. |  4  .1254085 |
     |--------------|
  6. |  5  .0601961 |
  7. |  6  .0240784 |
  8. |  7  .0082555 |
  9. |  8  .0024766 |
 10. |  9  .0006604 |
     |--------------|
 11. | 10  .0001585 |
     +--------------+
```

Here is the Stata command we used to calculate the Poisson probabilities, and here are the first eleven values of the curve.

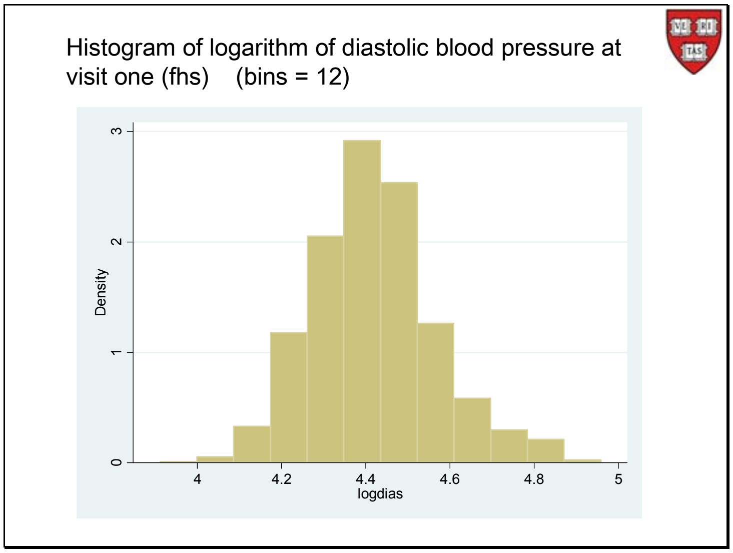
So the Poisson model is related to the Binomial by letting n go to infinity, and at the same time, let p go to zero in such a way that np = λ.  That is a way of reaching the Poisson, but it also stands on its own as a useful model.

Returning to the De Moivre result, we also saw that simply letting n go to infinity in the binomial, we get the normal distribution. Let us return to that and study it more closely.

The Normal Distribution

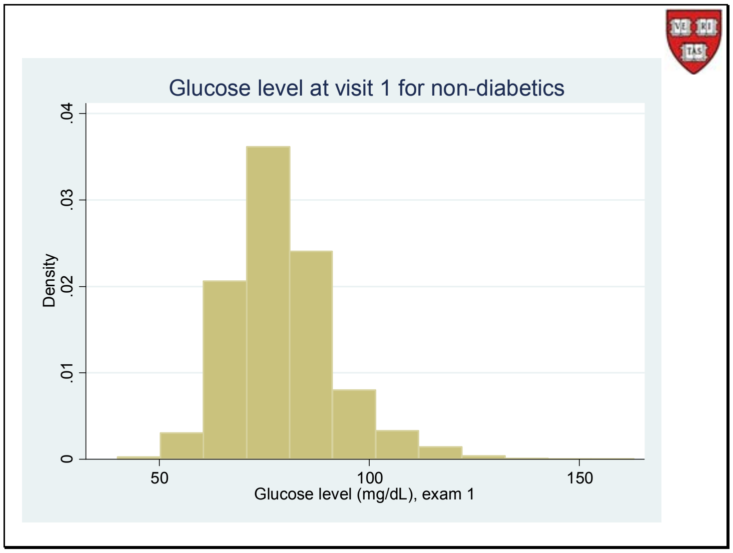
Here is the 10 Deutsche mark, the currency used in Germany before they started using the euro. We see a picture of Carl Friedrich Gauss (1777— 1855)—he of Gaussian distribution fame. In the middle of the mark, in the background, one can see the normal curve. We return to Gauss when we discuss regression theory and least squares.

The normal distribution stands on its own, and not just as an approximation to the binomial.

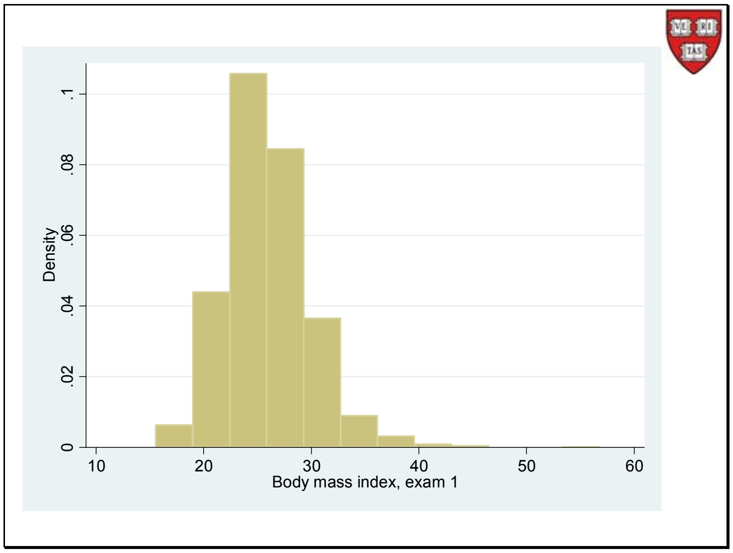


Histogram of logarithm of diastolic blood pressure at visit one (fhs) (bins = 12)

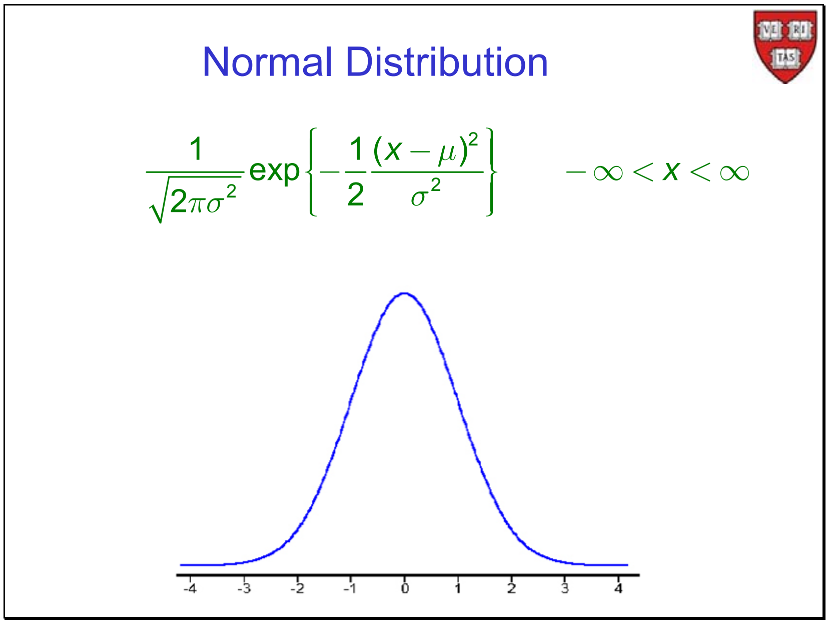
For example, here is the histogram for the logarithm of diastolic blood pressure at visit 1 in the Framingham Heart Study that we looked at earlier in this course. It seems approximately normally distributed.



Here is the glucose level at Visit 1 for the non-diabetics. Except for a small tail on the right, possibly pre-diabetics, they look approximately normally distributed.



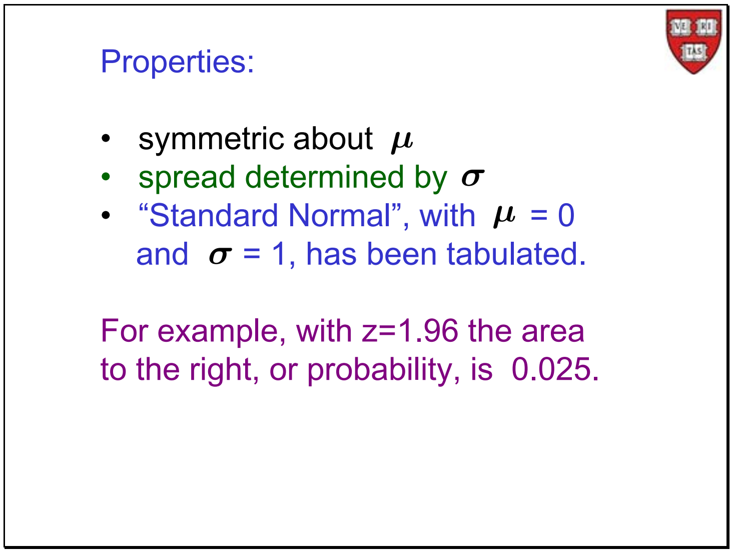
Here is another example, the BMI, the Body Mass Index, at Exam 1, also from the Framingham Heart Study. They too look approximately normally distributed. So let us look at the normal distribution more closely.



## Normal Distribution

$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\} \qquad -\infty < x < \infty$$

Here is the equation for the normal density function, and a plot of one such, namely the one called the standard normal; when the mean, μ = 0 and the standard deviation, σ = 1. You can see the symmetry around zero. All normals are symmetric around their mean (zero in this case). The spread is determined by σ, so making σ larger makes the curve flatter and making σ smaller makes the curve more peaked.

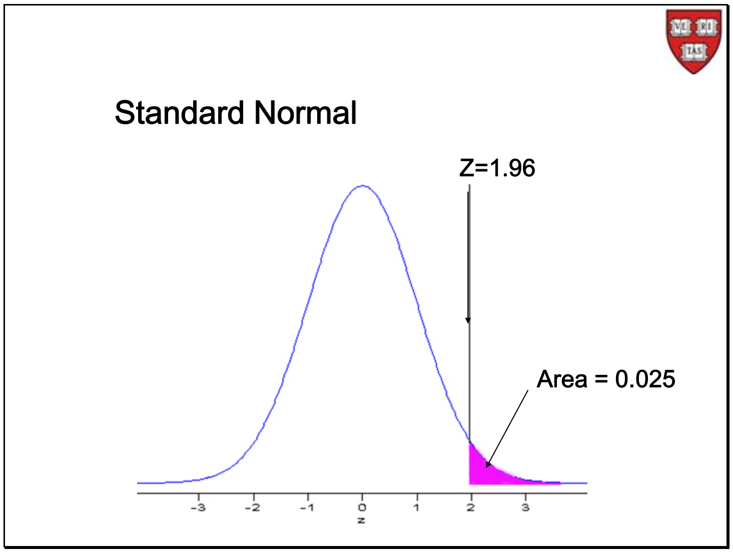


Properties:

- symmetric about $\mu$
- spread determined by $\sigma$
- "Standard Normal", with $\mu$ = 0 and $\sigma$ = 1, has been tabulated.

For example, with z=1.96 the area to the right, or probability, is 0.025.

If you look for tabulated values of the normal distribution—unnecessary for us since we have Stata—then the standard normal is the one that gets tabulated.

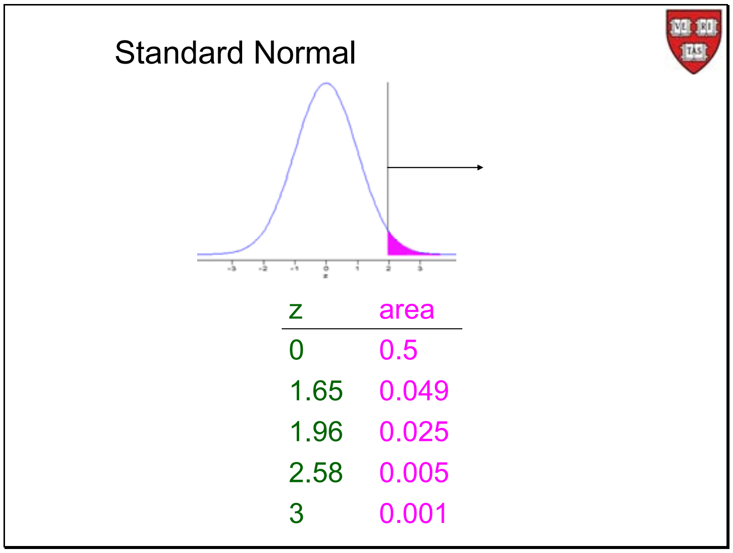


Standard Normal

Z=1.96

Area = 0.025

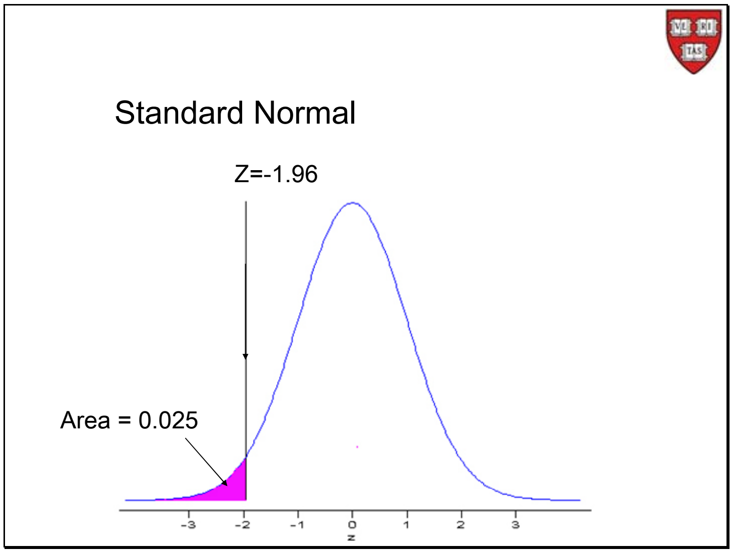
These tables give you the area under the curve, typically the area to the right of a particular z value. For example, in the standard normal, the area to the right of z=1.96 is 0.025—remember that the total area, very much like the Venn diagram, and for the same reason, is equal to one.



Standard Normal

| z | area |
|---|---|
| 0 | 0.5 |
| 1.65 | 0.049 |
| 1.96 | 0.025 |
| 2.58 | 0.005 |
| 3 | 0.001 |

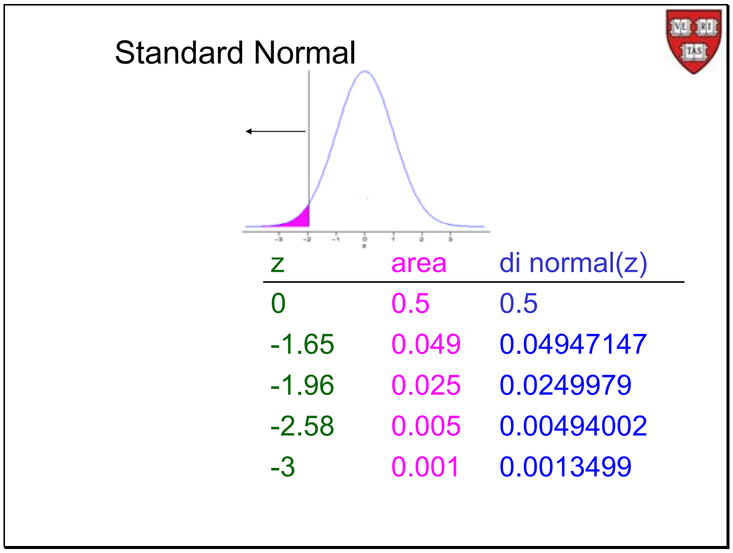
Other popular z values are: z = 0, then area to the right of 0 is 0.5, which makes sense since the curve is symmetric and the total area is one. The area to the right of z = 1.65 is 0.049. When z = 3, the area to the right is 0.001.



Standard Normal

Z=-1.96

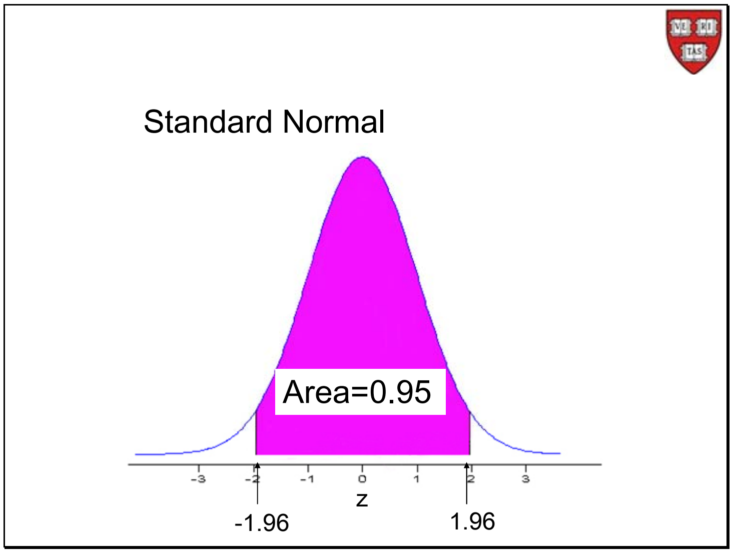
Area = 0.025



Standard Normal

| z | area | di normal(z) |
|---|---|---|
| 0 | 0.5 | 0.5 |
| -1.65 | 0.049 | 0.04947147 |
| -1.96 | 0.025 | 0.0249979 |
| -2.58 | 0.005 | 0.00494002 |
| -3 | 0.001 | 0.0013499 |

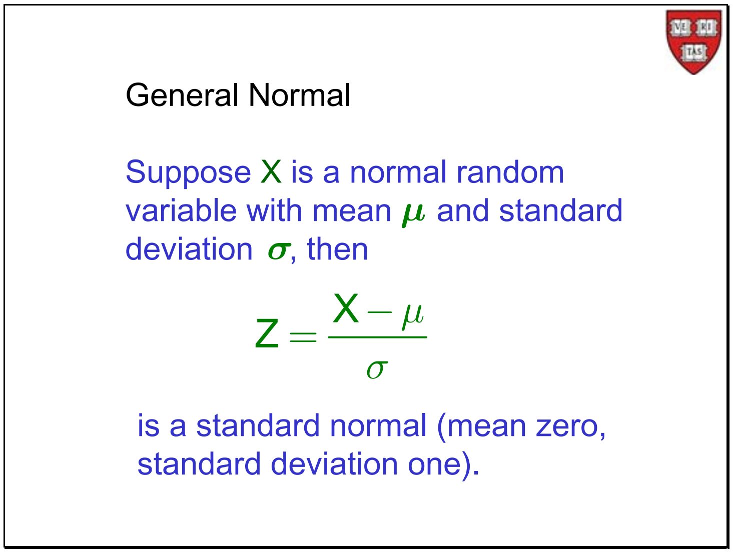
The curve is symmetric, so the area to the left of –z is the same as the area to the right of z. Of course, the same can be said for the area to the right of –z being the same as the area to the left of z. The values above are all obtained from the Stata function *normal*.



Standard Normal

Noting the symmetry, and that the total area is equal to one, we can also calculate the area between any two values. So, for example we have that the area between z = -1.96 and z = 1.96 is 0.95. (The area to the right of 1.96 is 0.025, therefore that is the area to the left of -1.96, and thus the area between -1.96 and 1.96 is 1-0.025-0.025=0.95.)

That is *exacly* 95%. No more approximation, like we had in the empirical rule. This is the idealization of the empirical rule. The mean plus or minus 1.96 standard deviations gives us exactly 95% of the data. I leave it to you to calculate the exact area for plus or minus one, and three standard deviations.



General Normal

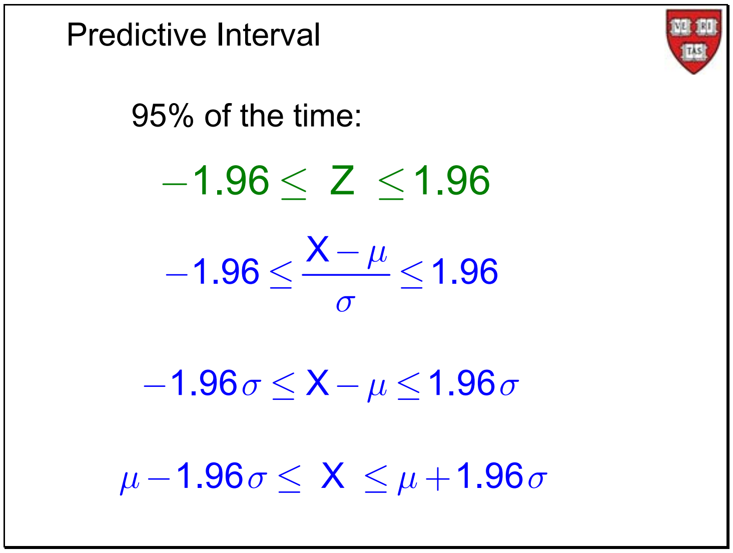
Suppose X is a normal random variable with mean $\mu$ and standard deviation $\sigma$, then

$$Z = \frac{X - \mu}{\sigma}$$

is a standard normal (mean zero, standard deviation one).

Let us look at a few ways of applying the normal distribution. When, in the past, we relied solely on tabulated values for the normal we reduced everything to the standard normal and proceeded from there—very similar to the idea of standardization we introduced a few weeks back. We can repeat that thinking here.

When dealing with a general normal variable that has a mean μ and a standard deviation σ, then standardize it by subtracting μ and dividing by σ:



Predictive Interval

95% of the time:

$$-1.96 \leq Z \leq 1.96$$

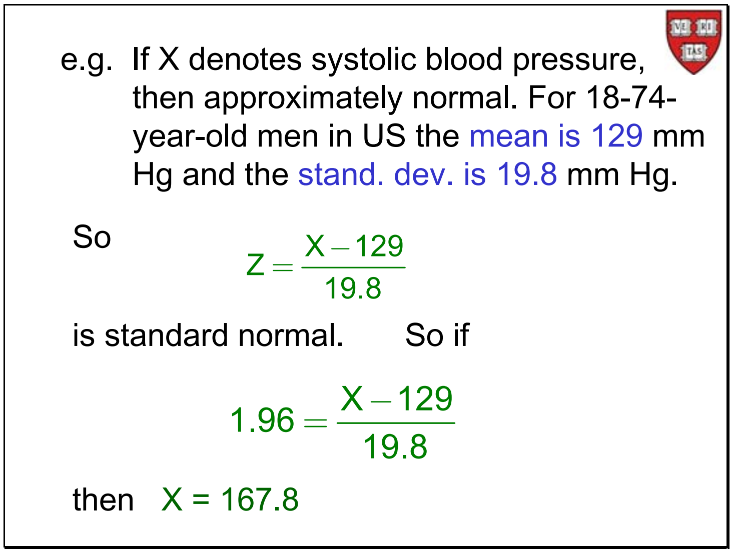
$$-1.96 \leq \frac{X - \mu}{\sigma} \leq 1.96$$

$$-1.96\sigma \leq X - \mu \leq 1.96\sigma$$

$$\mu - 1.96\sigma \leq X \leq \mu + 1.96\sigma$$

So let's look at an example of the use of this idea. We know that a standard normal will fall in the interval (-1.96, 1.96) 95% of the time; that is what we call a *predictive interval*. We cannot tell you exactly what value you are going to observe, but 95% of the time, it will be in this interval.

We can translate this into a predictive interval around X by using the standardization formula, to get that 95% of the time, X will take a value that is in the interval (μ-1.96σ, μ+1.96σ).



e.g. If X denotes systolic blood pressure, then approximately normal. For 18-74-year-old men in US the mean is 129 mm Hg and the stand. dev. is 19.8 mm Hg.

So

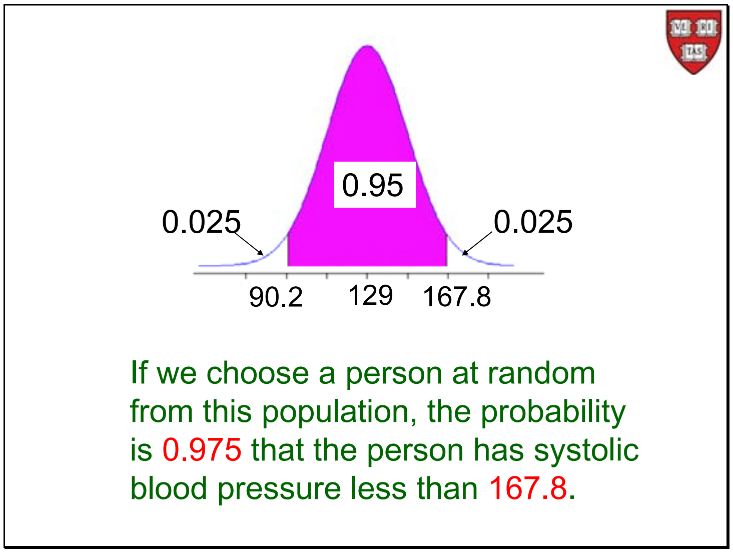
$$Z = \frac{X - 129}{19.8}$$

is standard normal. So if

$$1.96 = \frac{X - 129}{19.8}$$

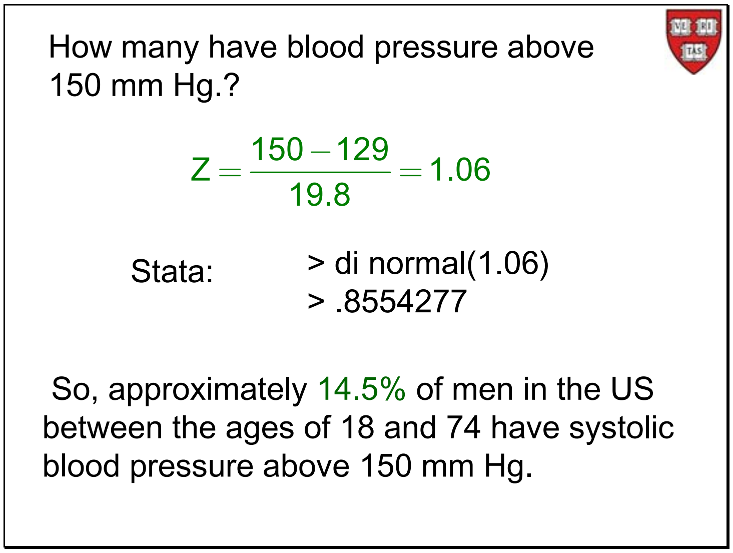
then $X = 167.8$

So for example, let us apply this to measuring systolic blood pressure on 18 to 74-year-old men in the US. We know that for this group, their systolic blood pressure is approximately normally distributed with mean μ = 129 mm Hg, and a standard deviation σ = 19.8 mm Hg. So from the standardization formula we get that X = 167.8. So in this population we are going to get a value bigger than 167.8, 2.5% of the time. Similarly we can calculate 90.2 to be the lower 2.5% cutoff.



If we choose a person at random from this population, the probability is 0.975 that the person has systolic blood pressure less than 167.8.

Thus we can make statements like: If we choose a person at random from this population, the probability is 0.975 that the person has systolic blood pressure less than 167.8 mm Hg; or, that 95% of all men in this age group in the US have a systolic blood pressure between 90.2 and 167.8 mm Hg.



How many have blood pressure above 150 mm Hg.?

$$Z = \frac{150 - 129}{19.8} = 1.06$$

Stata:

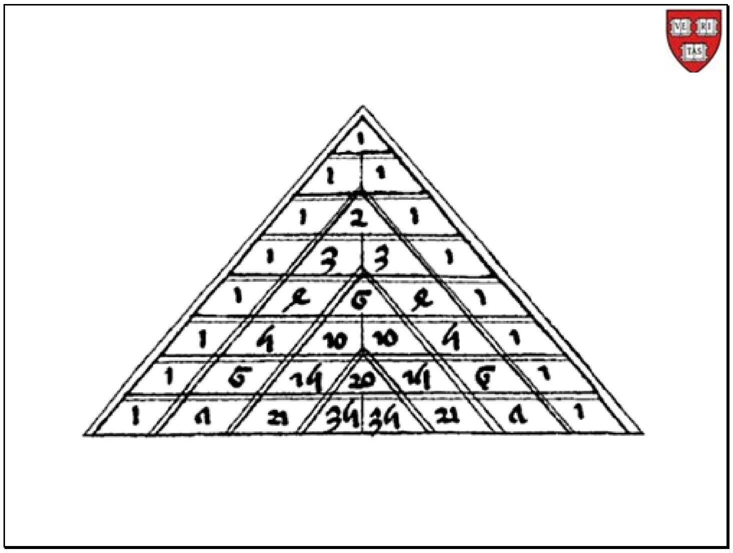
```
> di normal(1.06)
> .8554277
```

So, approximately 14.5% of men in the US between the ages of 18 and 74 have systolic blood pressure above 150 mm Hg.

We can also turn things slightly around and ask statements directly on the x scale. For example, how many have blood pressure above 150 mm Hg? In order to answer that question we have to go from our x scale of 150, subtract the mean, divide by the standard deviation, to a statement about our standardized Z. And it turns out to be 1.06.

So we can ask Stata what is the area to the left of 1.06, and it comes back with 0.855. One minus this would be 14.5%. So, approximately 14.5% of men in the US between the ages of 18 and 74 have systolic blood pressure above 150 millimeters off mercury.

These are some of the questions we can use the models to answer.



1

1 1

1 2 1

1 3 3 1

1 4 6 4 1

1 5 10 10 5 1

1 6 15 20 15 6 1

1 7 21 35 35 21 7 1

De Arithetica by Jordanus de Nemore