Biostat 200: Today's topics

- Review of some probability
- Check in on what you should have learned so far
- Look at some probability distributions

Review: Independence vs. Mutually Exclusive

- Mutually Exclusive: $P(A \cap B) = 0$
  - A and B cannot occur together
- Independence: If A and B are independent:

  $P(B \mid A) = P(B)$

  $P(A \mid B) = P(A)$

Review: Independence vs. Mutually Exclusive

- Mutually Exclusive: $P(A \cap B) = 0$
  - A and B cannot occur together
  - Mutually exclusive gives us the additive rule:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= P(A) + P(B)$$

- Independence gives us the multiplication rule
  - If A and B are independent, then

$$P(A \cap B) = P(A)P(B)$$

Review: Independence vs. Mutually Exclusive

- Examples of non-independence of an event
  - Repeated measurement of the same study participant
  - Measurements of persons within the same family or some other clustering unit
- We also talk about two random variables being independent, i.e. if one is related to the other. You will learn the chi-square test which is often called a test of independence but you will also use correlation this way (only measures linear dependence)

What has been covered so far:

- Types of variables
- Basic summaries relevant to different types of variables
- Basic graphical analyses of different types of variables
- Basic probability concepts, especially conditional probability, mutual exclusivity, and independence

What's next

- Use probability concepts to discuss theoretical distributions
- Assuming that a variable follows a certain distribution, you can calculate the probability of observing a certain value for that variable
- Sampling distributions: Knowing the distribution of a sample mean allows us to calculate the probability of observing a particular sample mean

Probability distributions

- Variables whose outcome can occur by chance, i.e. are not fixed, are called random variables
- Probability distributions describe the possible values of the random variable
- Important because p-values and confidence intervals calculated when running statistical tests are based on probability distributions

Probability distributions

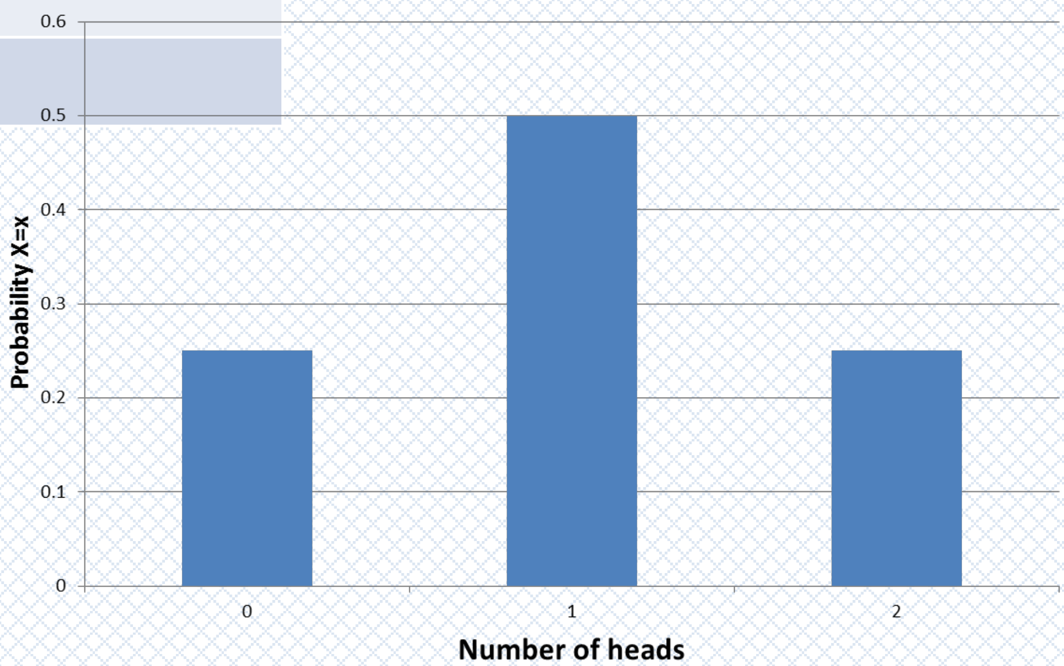
- For categorical random variables the probability distribution describes the probability of each possible value
- For example, consider the experiment in which you flip a fair coin 2 times and count the number of heads.
  - The possible outcomes of the experiment are: HH, TH, HT, TT.
  - You want to focus on the number of heads, which could be 0, 1, or 2. The probability of each outcome is:

Number of heads	Probability
0	.25
1	.5
2	.25

Probability distributions

- The table looks similar to a frequency table of the data, but it is actually the theoretical distribution
- If you perform an infinite number of experiments, your data may look like this table

Number of heads	Relative frequency
0	.25
1	.5
2	.25



Probability distributions

- Note that the probabilities add to 1. This is true of all probability distributions.
  - We are assuming that the probability of a head on each toss is .5
  - We are assuming that probability of heads on the first toss is independent of the second toss
  - This is the <u>Binomial distribution</u>

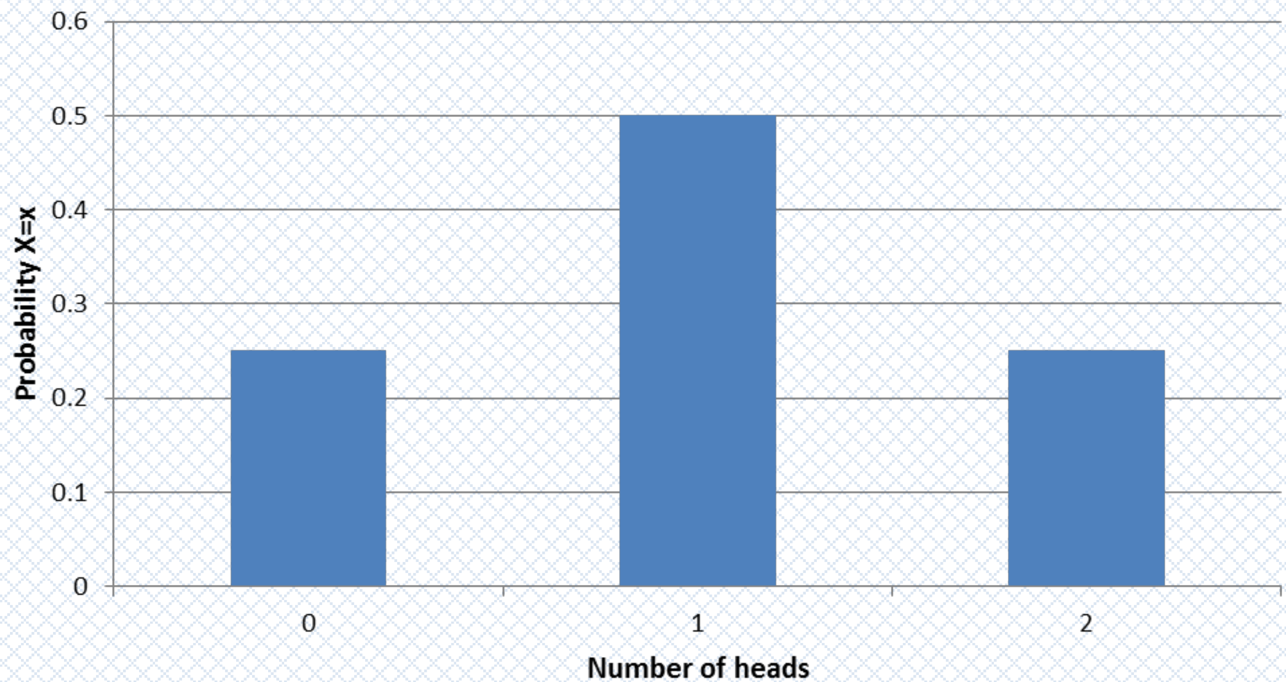
Probability distributions

- We can write down a formula for the probability of event X, which is observing a certain # of heads
- This is written in general form $P(X=x)$
  - For 0 heads: $P(X=0)$
  - For 1 head: $P(X=1)$
  - For 2 heads: $P(X=2)$
  - For 1 or more heads: $P(X\geq 1)$

Probability distributions

- Use this theoretical distribution to make predictions about future experiments
- e.g. The probability that there will be at least 1 head in a trial of 2 coin tosses

$$P(X \geq 1) = P(X=1) + P(X=2) = .5 + .25 = .75$$



Probability distributions

- If you performed the experiment (i.e. 2 coin tosses) once, you'd get 0,1, or 2 heads
- I performed the experiment 10 times and got: 2, 1, 1, 1, 1, 0, 0, 0, 1, 1 heads in 2 tosses

Number of heads	Frequency (%)
0	3 (30)
1	6 (60)
2	1 (10)

- What if we did the experiment 100 times? 1000 times? What would the frequency distribution for the outcomes look like?

Probability distributions

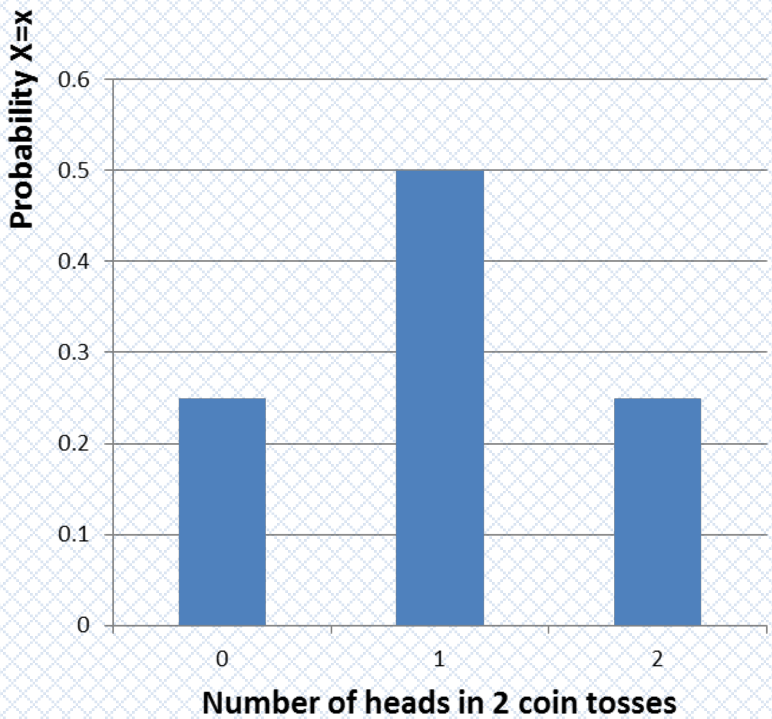
- Last class we looked at two independent diagnostic tests with 2% false positive probability.
  - 2% of the time the test randomly gives a positive result when the true result is negative
- The possible outcomes were NegNeg, NegPos, PosNeg, PosPos
  - Calculating the probability 0, 1, or 2 positive tests:

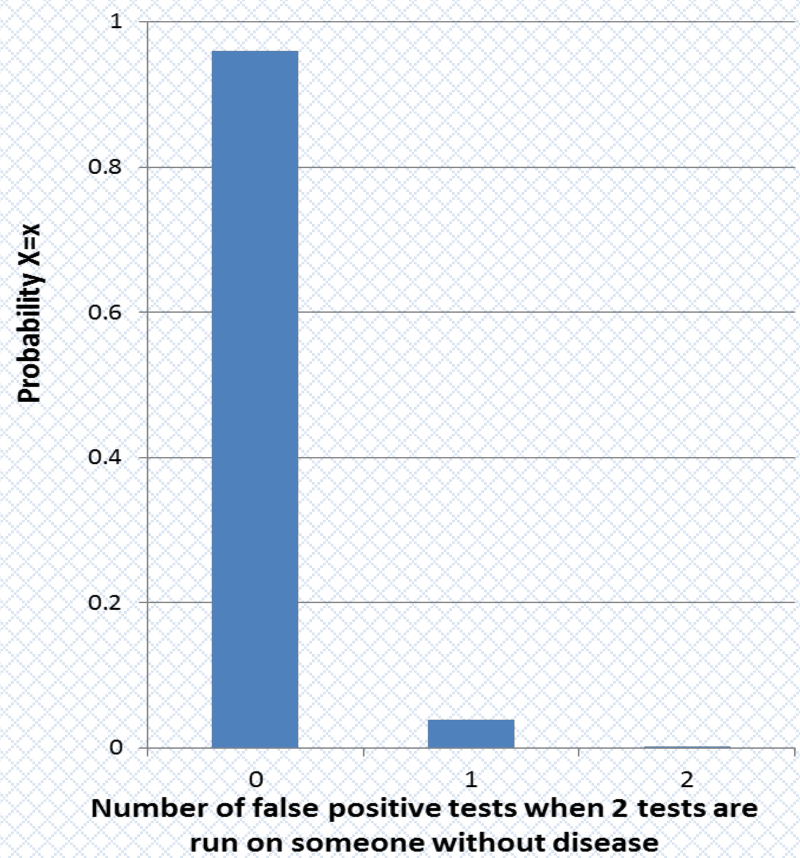
Number of positive tests	Probability
0	$.98^2 = .9604$
1	$.98*.02*2 = .0392$
2	$.02^2 = .0004$

Probability distributions

- The graphical representation of the probability distribution for number of false positive tests is:

Probability distributions





These both represent distributions that show the probability of 0, 1, or 2 events occurring in 2 independent trials. Why do they look so different?

Empirical Probability distributions

- Empirical probability distributions are based on real data
- They are usually based on a large sample or complete enumeration of a population
- The probabilities are calculated from the relative frequencies of the data
  - e.g. proportion of live births by age of mother in Michigan

Probability distributions

- For <u>discrete (or categorical) variables</u> the probability distribution describes the probability of <u>each possible value</u>

- For <u>continuous variables</u>, the distribution describes the probability of a <u>*range* of values</u>

Bernoulli (who was he?) random variables

- If you have a variable that can take on one of two values with a constant probability p, then it is a Bernoulli random variable
- If the proportion of people in the population with a disease (the prevalence) is 15%, then when you randomly select one person, the probability that he/she has the disease is $P(Y=1)=p= 0.15$
- The probability that a randomly selected person does not have the disease is

$$P(Y=0)=1-p =0.85$$

Bernoulli distribution:

a Binomial distribution with only one trial

- A Bernoulli random variable is said to follow the Binomial distribution
- p is the <u>parameter</u> that characterizes the distribution
- The Binomial distribution is a discrete distribution – the outcome is either 0 or 1
- As defined by Bernoulli, it describes only one trial – so really is more theoretical than practical – it is the building block to describe the distribution of more than one trial

Binomial distribution

- Example: The proportion of people in the population with the disease (the prevalence) is 15%, then P(Y=1)=0.15 and P(Y=0)=0.85.
- One random draw will produce either 0 or 1 person with disease
- If we take a random sample of 5 people from this population, there will be 0,1,2,3,4, or 5 people with the disease.

Binomial distribution

If the probability of disease in each person is independent, then we can write down the probability of each of these outcomes even before we draw the sample.

For example, the probability that ALL of them will have the disease is P(X=5):

$$=P(X_1=1)*\ P(X_2=1)*\ P(X_3=1)*\ P(X_4=1)*\ P(X_5=1)$$

$$= 0.15 \times 0.15 \times 0.15 \times 0.15 \times 0.15 = 0.00008$$

by the multiplication rule for independence
P(A and B)=P(A)P(B)

Binomial distribution

The probability that NONE of them will have the disease is $P(X=0)$:

$$=P(X_1=0)* P(X_2=0)* P(X_3=0)* P(X_4=0)* P(X_5=0)$$

$$=0.85 \times 0.85 \times 0.85 \times 0.85 \times 0.85 = 0.444$$

Binomial distribution

The probability that <u>exactly one person</u> $P(X=1)$ has the disease

$= P(X_1=1) * P(\text{the other } 4=0)$
$+ P(X_2=1) * P(\text{the other } 4=0)$
$+ P(X_3=1) * P(\text{the other } 4=0)$
$+ P(X_4=1) * P(\text{the other } 4=0)$
$+ P(X_5=1) * P(\text{the other } 4=0)$

$= \mathbf{0.15} \times 0.85 \times 0.85 \times 0.85 \times 0.85$
$+ 0.85 \times \mathbf{0.15} \times 0.85 \times 0.85 \times 0.85$
$+ 0.85 \times 0.85 \times \mathbf{0.15} \times 0.85 \times 0.85$
$+ 0.85 \times 0.85 \times 0.85 \times \mathbf{0.15} \times 0.85$
$+ 0.85 \times 0.85 \times 0.85 \times 0.85 \times \mathbf{0.15}$

$= 5 * .15 * .85^4 = 0.392$

Binomial distribution

The probability that exactly two people $P(X=2)$ of 5 have the disease

= **0.15** x **0.15** x 0.85 x 0.85 x 0.85

+ **0.15** x 0.85 x **0.15** x 0.85 x 0.85

+ **0.15** x 0.85 x 0.85 x **0.15** x 0.85

+ **0.15** x 0.85 x 0.85 x 0.85 x **0.15**

+ 0.85 x **0.15** x **0.15** x 0.85 x 0.85

+ 0.85 x **0.15** x 0.85 x **0.15** x 0.85

+ 0.85 x **0.15** x 0.85 x 0.85 x **0.15**

+ 0.85 x 0.85 x **0.15** x **0.15** x 0.85

+ 0.85 x 0.85 x **0.15** x 0.85 x **0.15**

+ 0.85 x 0.85 x 0.85 x **0.15** x **0.15**

$= 10 * .15^2 * .85^3 = 0.138$

Binomial distribution

The probability that <u>no people</u> P(X=0) of 5 have the disease = .444

The probability that <u>exactly one person</u> P(X=1) of 5 has the disease = .392

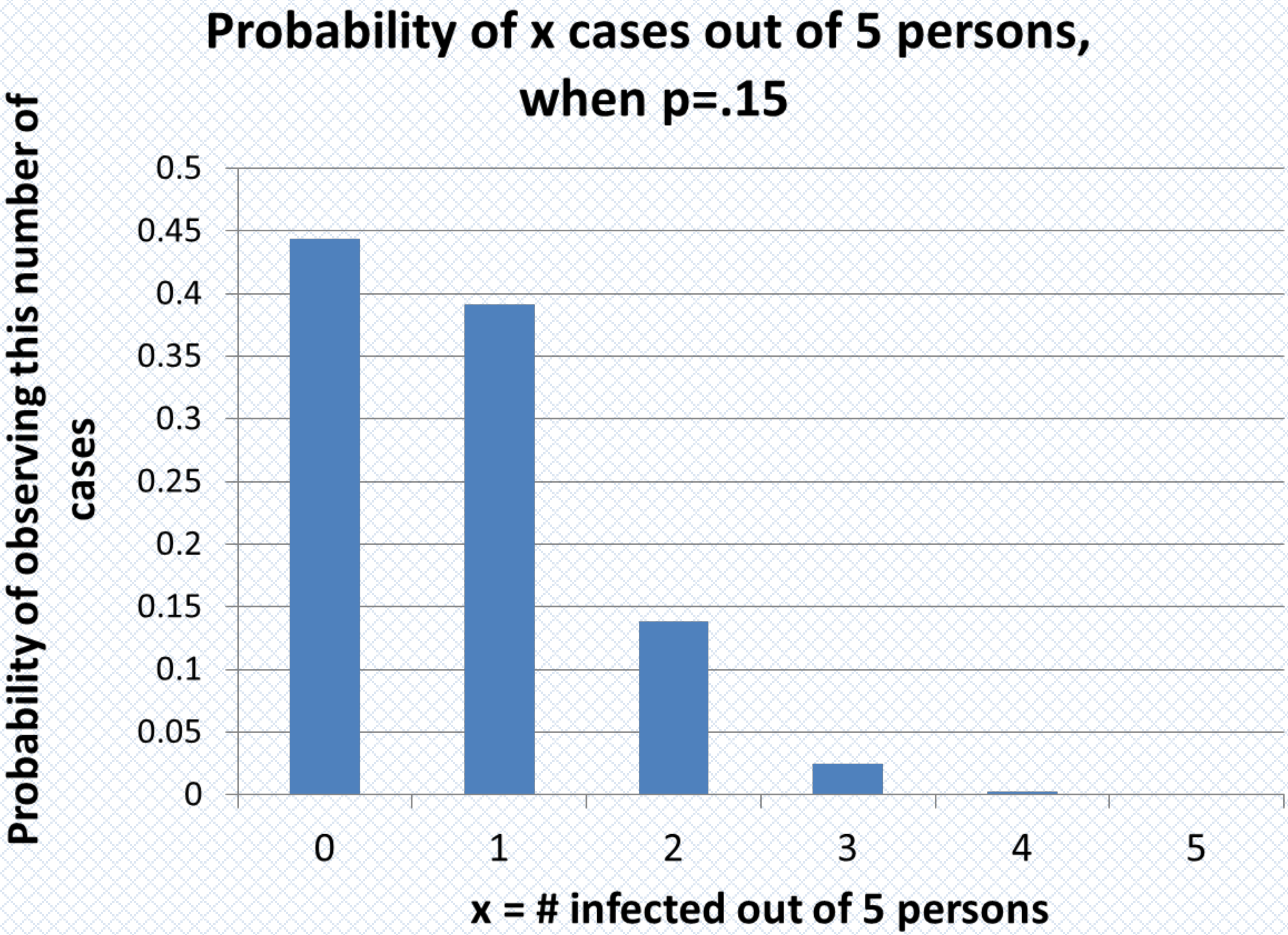
The probability that <u>exactly two people</u> P(X=2) of 5 have the disease = .138

The probability that <u>exactly three people</u> P(X=3) of 5 have the disease = .024

The probability that <u>exactly four people</u> P(X=4) of 5 have the disease = .002

The probability that <u>exactly five people</u> P(X=5) of 5 have the disease = .00008

Binomial distribution



Binomial distribution

The probability that exactly one person $P(X=1)$ has the disease

$P(X=1; n=5, p=0.15)$

$= \mathbf{0.15} \times 0.85 \times 0.85 \times 0.85 \times 0.85$

$+ 0.85 \times \mathbf{0.15} \times 0.85 \times 0.85 \times 0.85$

$+ 0.85 \times 0.85 \times \mathbf{0.15} \times 0.85 \times 0.85$

$+ 0.85 \times 0.85 \times 0.85 \times \mathbf{0.15} \times 0.85$

$+ 0.85 \times 0.85 \times 0.85 \times 0.85 \times \mathbf{0.15}$

$= 0.392$

$= 5 * .15^1 * .85^4$

$= 5 * p^1 * (1-p)^4$

5 is the number of different ways you could get one "success" in the 5 "trials"

Binomial distribution

This generalizes to: $$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

- *p* is probability of “success” in each “trial”
- *n* is the number of “trials” (e.g., coin flips, persons assessed for disease status, etc.)
- *n* and *p* are the parameters of the binomial distribution, i.e. the values that summarize the distribution
- x is the number of “successes” (e.g. heads, numbers with the disease, etc.)

Binomial distribution

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

This is the formula for the binomial distribution

- Note that Stata uses the symbol k for x in its documentation

Binomial distribution

- Assumptions:
  - There are a fixed number of trials n
  - Within each trial, there are two mutually exclusive outcomes (heads vs. tails, disease vs. no disease)
  - The outcomes of the n trials are independent
  - The probability of success p is constant for each trial

Binomial distribution

Combinations

$\binom{n}{x}$ is called "n choose x" and is the number of different ways to get x successes in n trials

There are 5 ways that there could be 1 success in 5 trials

There are 10 ways there could be 2 successes in 5 trials

Binomial distribution

- The formula for n choose x is

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \quad \text{where } n! = n * \ldots * 3 * 2 * 1$$

5 choose 1 = 5! / (1! * 4!)
= (5*4*3*2*1) / (1*4*3*2*1)
= 5

5 choose 2 = 5! / (2! * 3!)
= (5*4*3*2*1) / (2*1*3*2*1)
= 5*4/2 = 10

5 choose 3 = 5! / (3! * 2!) = 10

In Stata: display comb(n,k)

```
. display comb(5,3)
10
```

Binomial distribution

$$\binom{n}{0} = 1$$

$$\binom{n}{1} = n$$

$$\binom{n}{n} = 1$$

- Ways to find binomial probabilities
  - The previous equations
  - Stata: Binomialp(n,k,p)

Binomial distribution

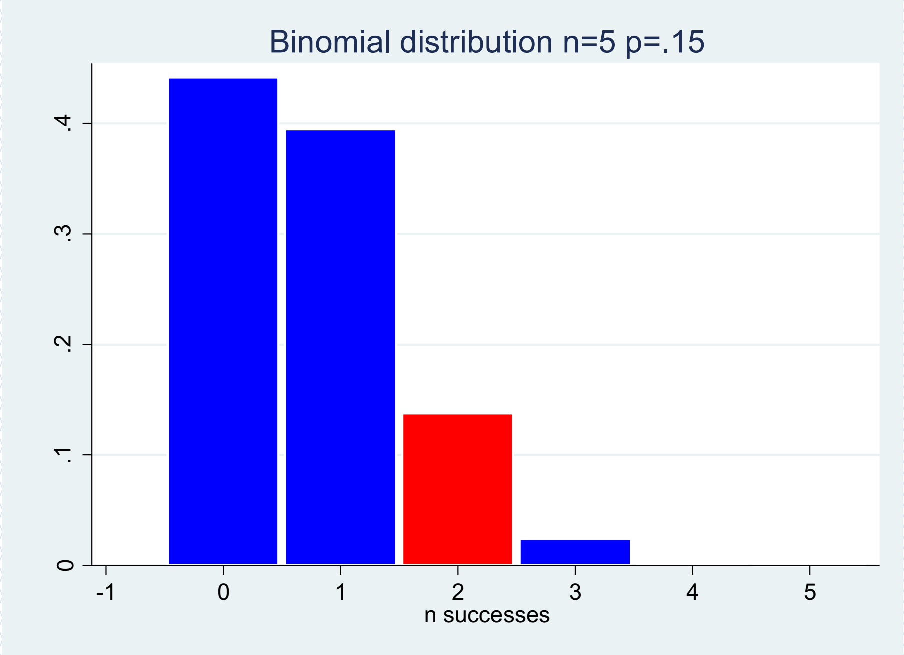
The binomial formula

What is the probability of exactly 2 cases of disease in a sample of n=5 where p=0.15?

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$P(X = 2) = \binom{5}{2} .15^2 (.85)^3$$

```
. di comb(5,2)*.15^2*.85^3
.13817812
```



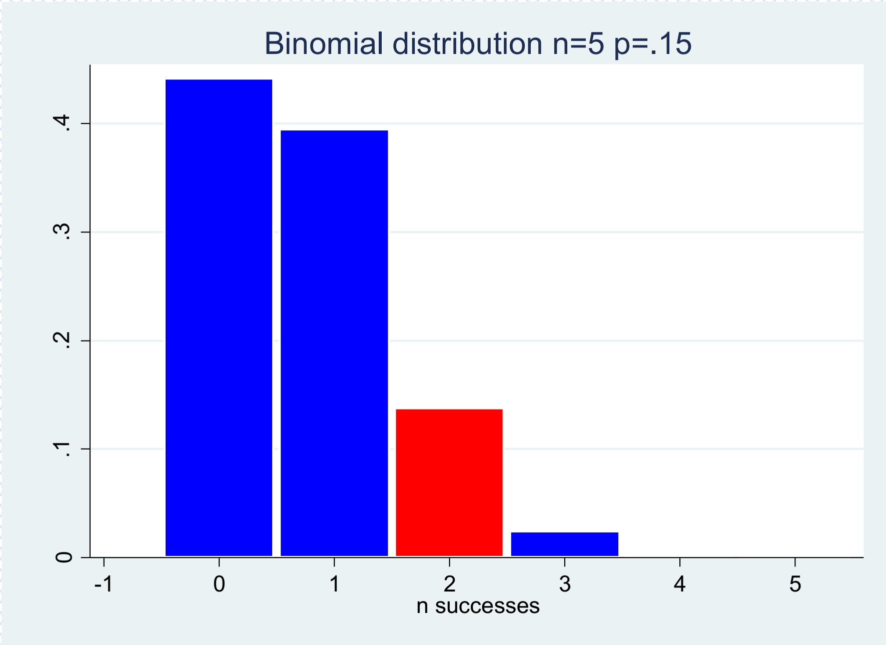
Binomial distribution

Stata binomialp()

What is the probability of exactly 2 cases of disease in a sample of n=5 where p=0.15?

–Use binomialp(n,k,p) to get P(X=k) in n trials with probability of success in each trial=p

- `di binomialp(5,2,.15)`
  `.13817813`

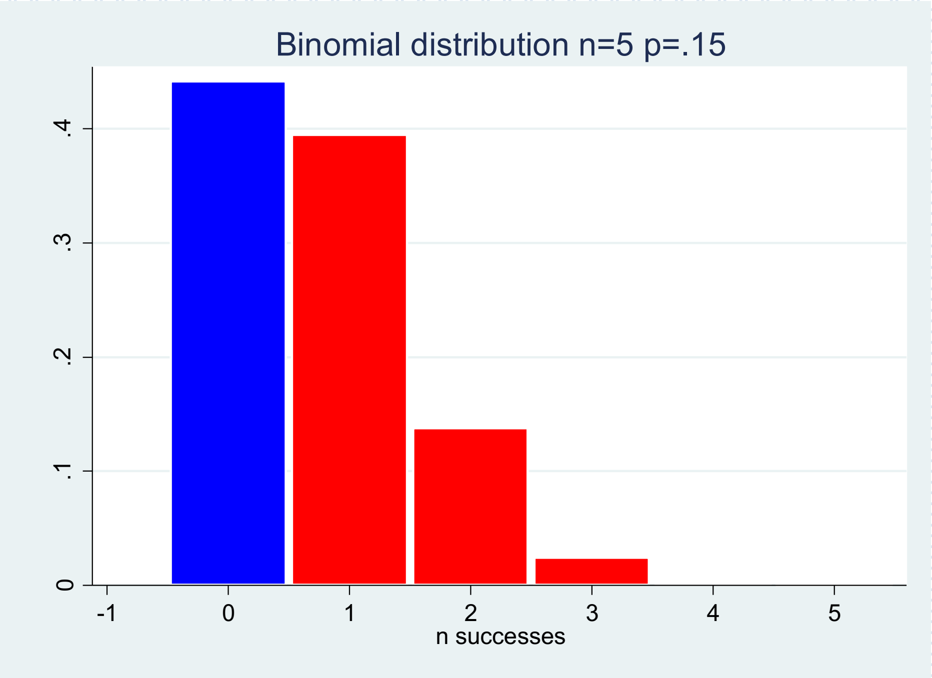


Binomial distribution

What is the probability of 1 or more cases of disease in a sample of n=5 where p=0.15?

- We want $P(X \geq 1)$
- One way would be to calculate all the probabilities: $P(X=1)+P(X=2)+ \ldots +P(X=5)$
- But remember:

$$P(X \geq 1) = 1-P(X=0)$$



Binomial formula

What is the probability of 1 or more cases of disease in a sample of n=5 where p=0.15?

- $P(X \geq 1) = 1 - P(X=0)$

$$P(X=0) = \binom{5}{0} .15^0 (.85)^5 = 1 * .85^5$$

```
    di comb(5,0)*.85^5
.44370531
```

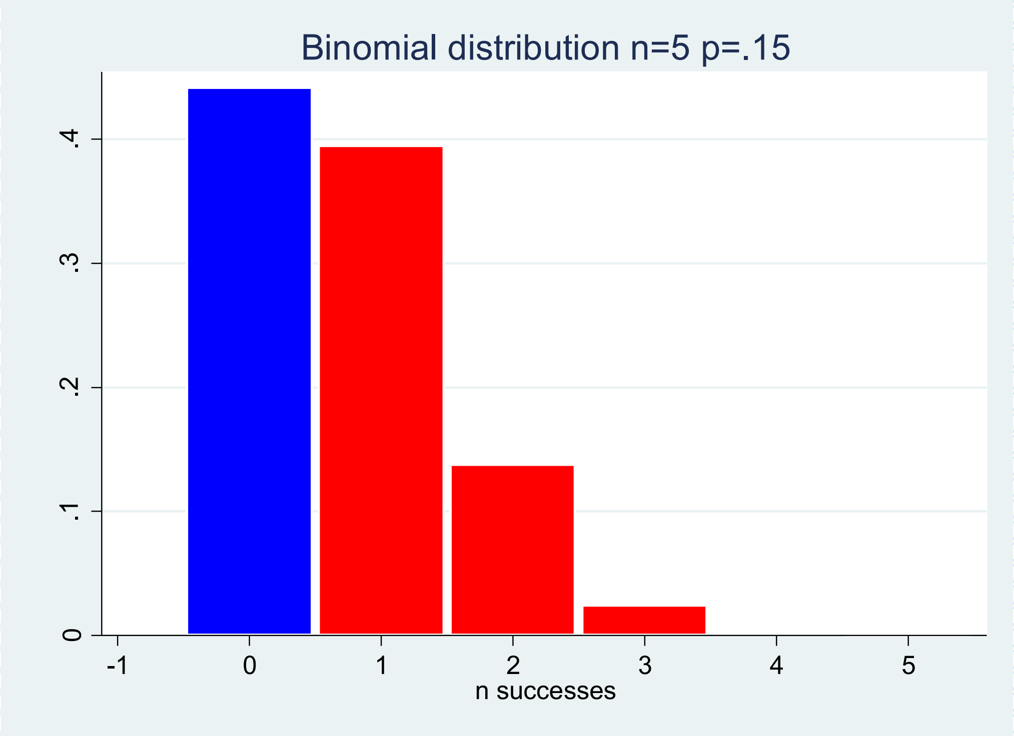
- So 1-P(X=0) = 1- 0.4437 = 0.5563
- Or use

```
display 1-binomialp(5,0,.15)
.55629469
```

Stata binomialtail()

- The function <u>binomialtail(n,k,p)</u> gives us P(X≥k) so we can use it without manipulation
- `display binomialtail(5,1,.15)`

```
    .55629469
```



Means and Variances

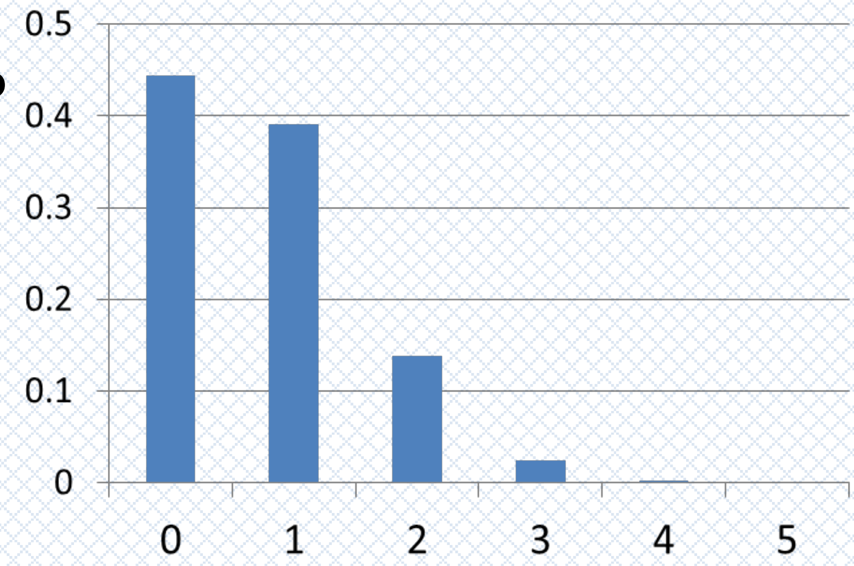
- In Lectures 1 & 2 we learned how to calculate the <u>sample</u> mean, variance, and standard deviation
- Theoretical distributions have <u>theoretical</u> means, variances, and standard deviations
- For a discrete distribution, the mean of the distribution is $\sum_{i=0}^{n} x_i \, P(X = x_i)$ over all the possible values of x.

Binomial distribution

- The <u>mean</u> of a binomially distributed random variable X is *np*
- This means that over an large number of samples of size n with probability p of success, the mean number of successes ($\bar{x}$) over the samples will be approximately *np*
- In shorthand, *np* is the mean of the binomial distribution

Binomial distribution

- For our example with n=5 and p=P(disease)=.15 the mean of the distribution of the number of cases is 5*.15 = .75
  - If you took a sample of n=5, repeatedly, the mean number of cases over these samples would be 0.75
  - Does this fit with the
    - graph of the distribution?



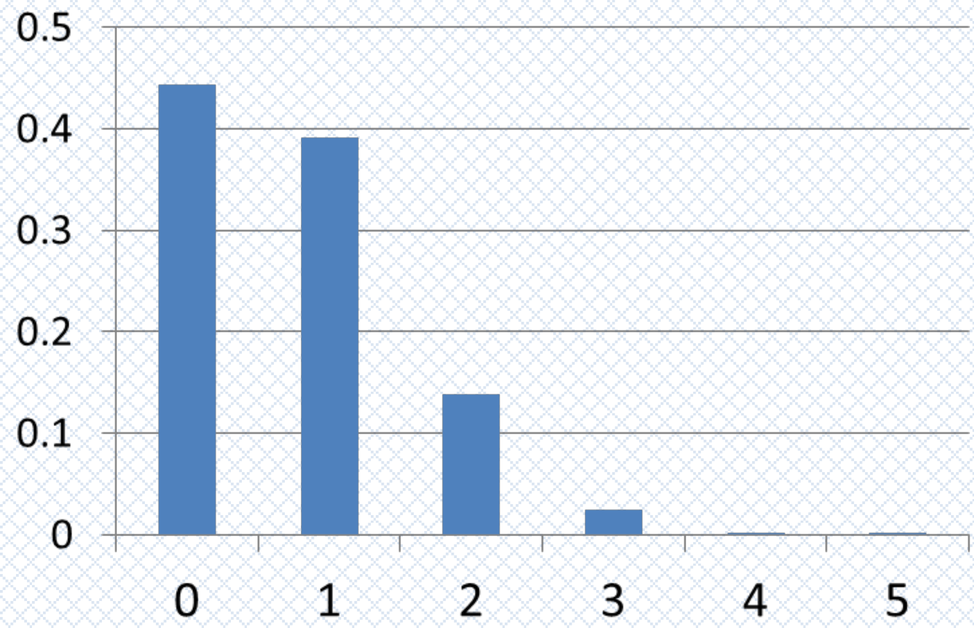
Binomial distribution

- The variance of a binomially distributed random variable X is *n*p*(1-p)*
- This means that over a large number of samples of size n, the *sample variance* of the X's, the number of successes, will be approximately *n*p*(1-p)*
- The standard deviation is

$$\sqrt{np(1-p)}$$

Binomial distribution

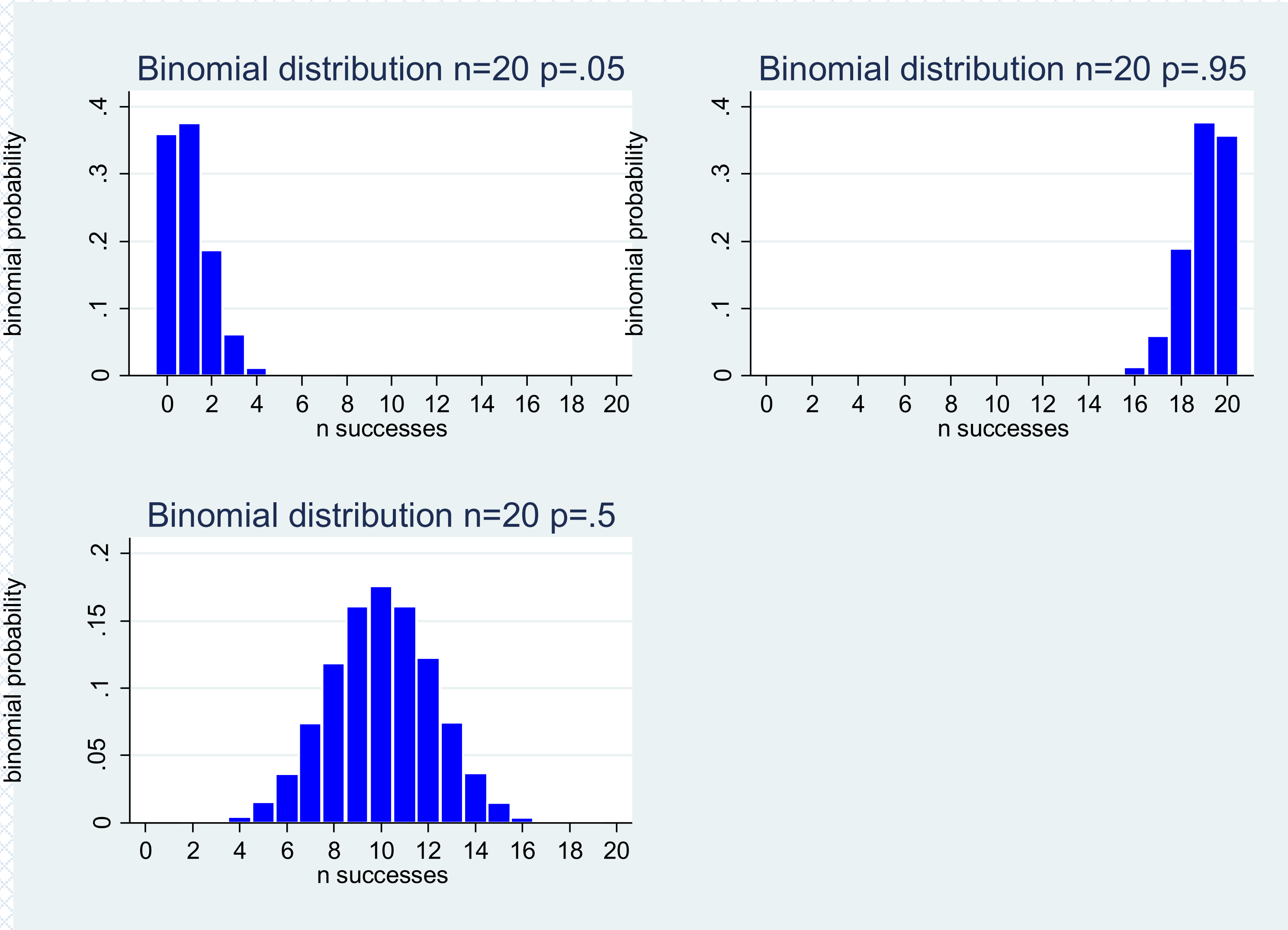
- For our example with n=5 and p=P(disease)=.15, the variance is n*p*(1-p) = 5*.15 *.85
  - If you took a sample of n=5, repeatedly, the sample variance in the number of cases over these samples would be this number



Binomial distribution

- Binomial mean = np
- Binomial variance= np(1-p)
  - Variance is largest when p=0.5, smaller when p closer to 0 or 1
  - The distribution is symmetric when p=0.5
  - The distribution is a mirror image for 1-p
    - e.g. the distribution for p=0.05 is the mirror image of the one for p=0.95

Binomial distribution



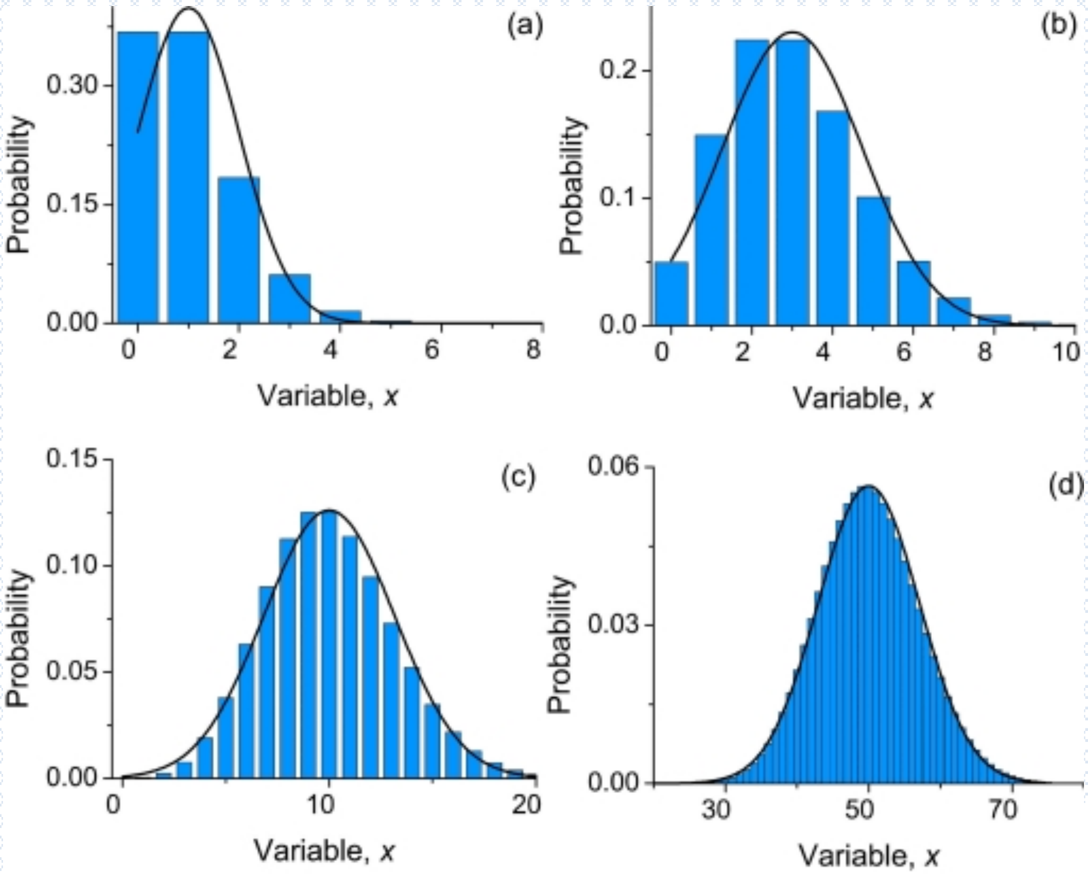
Poisson distribution (who was he)

- A discrete distribution to model **rare events** occurring in time or space
- Unlike the binomial distribution, which is based on a series of n trials, there is no theoretical limit to the number of events that can occur
- However, when n is large and p is small, it does act like the binomial
- The Poisson has only one parameter, λ, that is the mean number of events (and also the variance)

Poisson distribution

- Used to model counts – whole numbers
- It is possible to model counts but meaningless to consider how many have not occurred (so not binomial)
- Occurrences are independent – the events don't depend on each other

Poisson Distribution with varying means



$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

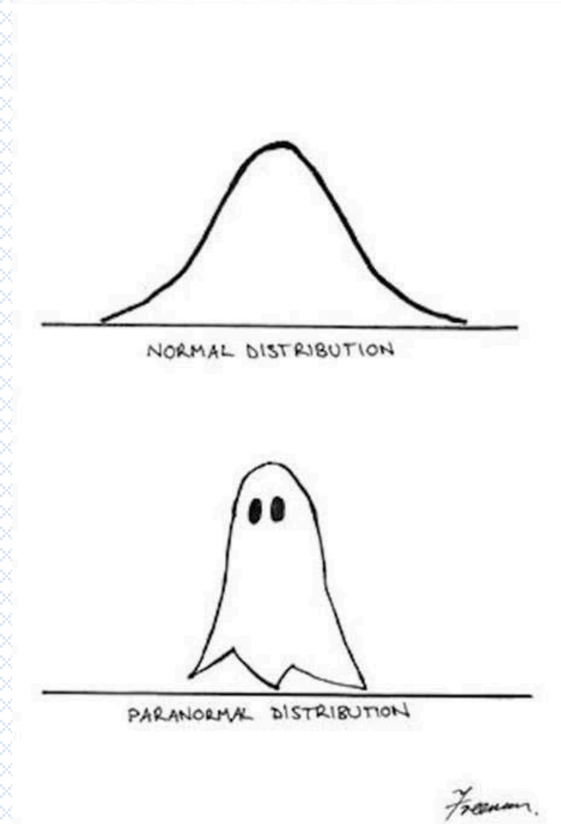
Other Important Related Distributions

- The companion distribution to the Poisson is the Exponential distribution (continuous). It calculates the probability of the length of time until a certain number of events. You will see this
- And extension of the Binomial distribution is the Hypergeometric distribution. It calculates probabilities when additional information about the outcomes is available.

You will see these illustrated in Stata in the lab.

Normal distribution

- Used for <u>continuous variables</u> that cover the entire range, i.e. values can take on any negative or positive number
- Classic bell shaped curve
- Values can span from -∞ to ∞
- Uni-modal and symmetric, so the mean is also equal to the median and mode



Normal distribution

- The probability density function is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{where } -\infty < x < \infty$$

- Yes calculus is part of its theoretical definition
- μ is the mean and σ is the standard deviation of a normally distributed random variable
  - They are the parameters of the normal distribution
  - π is the constant that is approximately 3.14159

Normal distribution

- Note that the left hand side of the equation is f(x) and not P(X=x)
- Why?
  - For a discrete distribution, the sum of the bars equals 1
  - For a continuous distribution, the area under the curve equals one (uses integration but we won't)
  - A continuous variable X can take on an infinite number of values, therefore P(X=x)=0

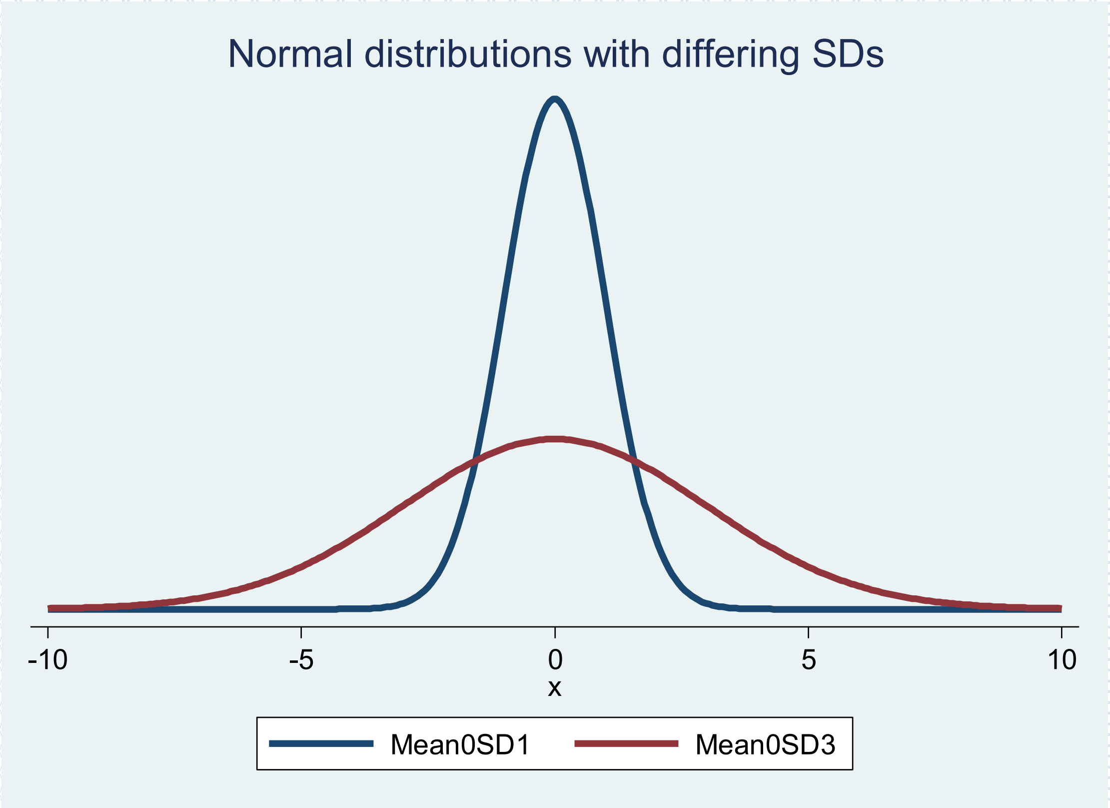
Normal distribution

- If X has a normal distribution with mean μ and standard deviation σ we write

$$X \sim N(\mu, \sigma)$$

- Many variables are approximately normally distributed
- We can use the distribution to calculate probabilities associated with such variables

Normal distribution

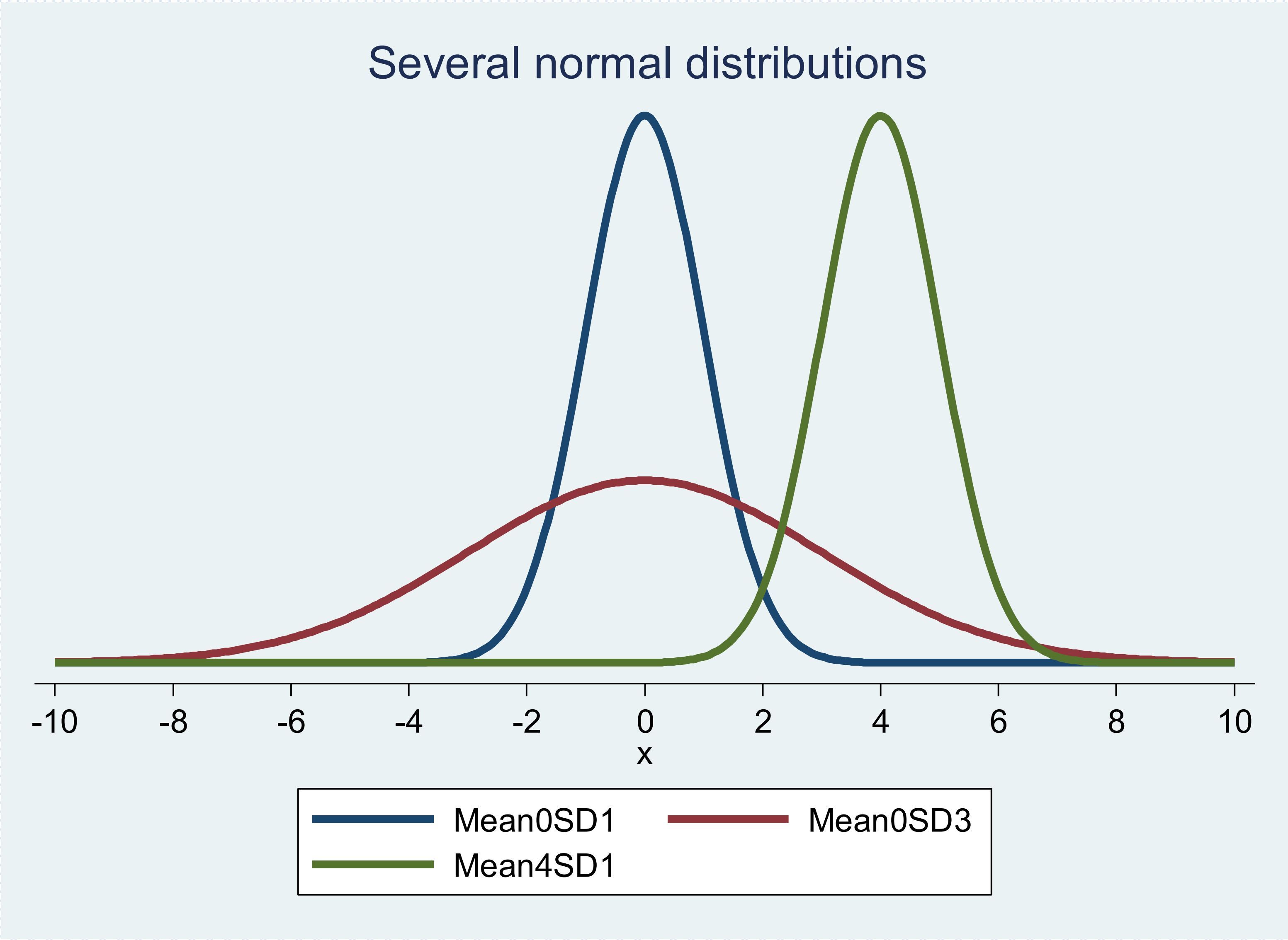


The standard deviation defines the amount of spread around the mean

Small standard deviation – little spread around the mean (blue line SD=1)

Large standard deviation – greater spread around the mean (red line SD=3)

Normal distribution



The Standard Normal Distribution

- μ and σ can take on an infinite number of values so we have an infinite number of curves
- For simplicity, we have a standard curve that we use as a reference

The Standard Normal Distribution

- This one curve has mean $\mu = 0$ and standard deviation $\sigma = 1$ (and variance $\sigma^2 = 1$).
- Denoted N(0,1)
- The formula for the distribution simplifies to

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \text{ where } -\infty < x < \infty$$

The Standard Normal Distribution

- If X is a normally distributed random variable with mean μ and standard deviation σ then

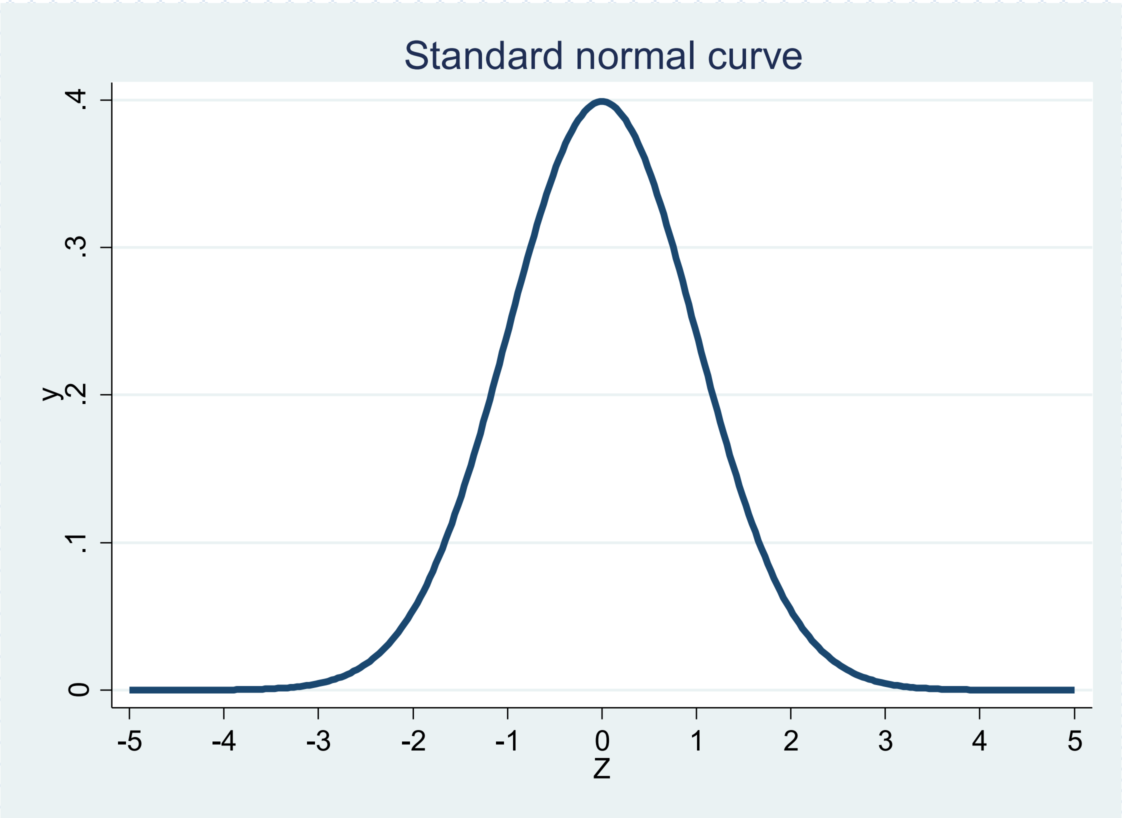
  $Z = (X - \mu)/\sigma$

  is a *standard* normal random variable

- That is, any normally distributed random variable, subtract it's mean & divide by its standard deviation, it becomes a standard normal random variable with mean=0 and standard deviation=1

The Standard Normal Distribution

- If $X \sim N(\mu, \sigma)$ then
- $Z = (X - \mu) / \sigma \sim N(0, 1)$



Normal distributions

- We can use theoretical distributions to determine the probability of particular values of random variables

- For the binomial distribution, we added probabilities of the assumed distribution to calculate the probability of observing a certain number (k) of events (or more).

- Remember the probability of observing 1 or more disease cases in a sample of 5 was equal to

$$P(X=1) + P(X=2) + P(X=3) + P(X=4) + P(X=5)$$

Normal distributions

• However, for a continuous variable, because there are an infinite number of values, we can't calculate P(X=x).

• However, we can calculate P(X ≥ x), which is the area under the normal curve from x to infinity

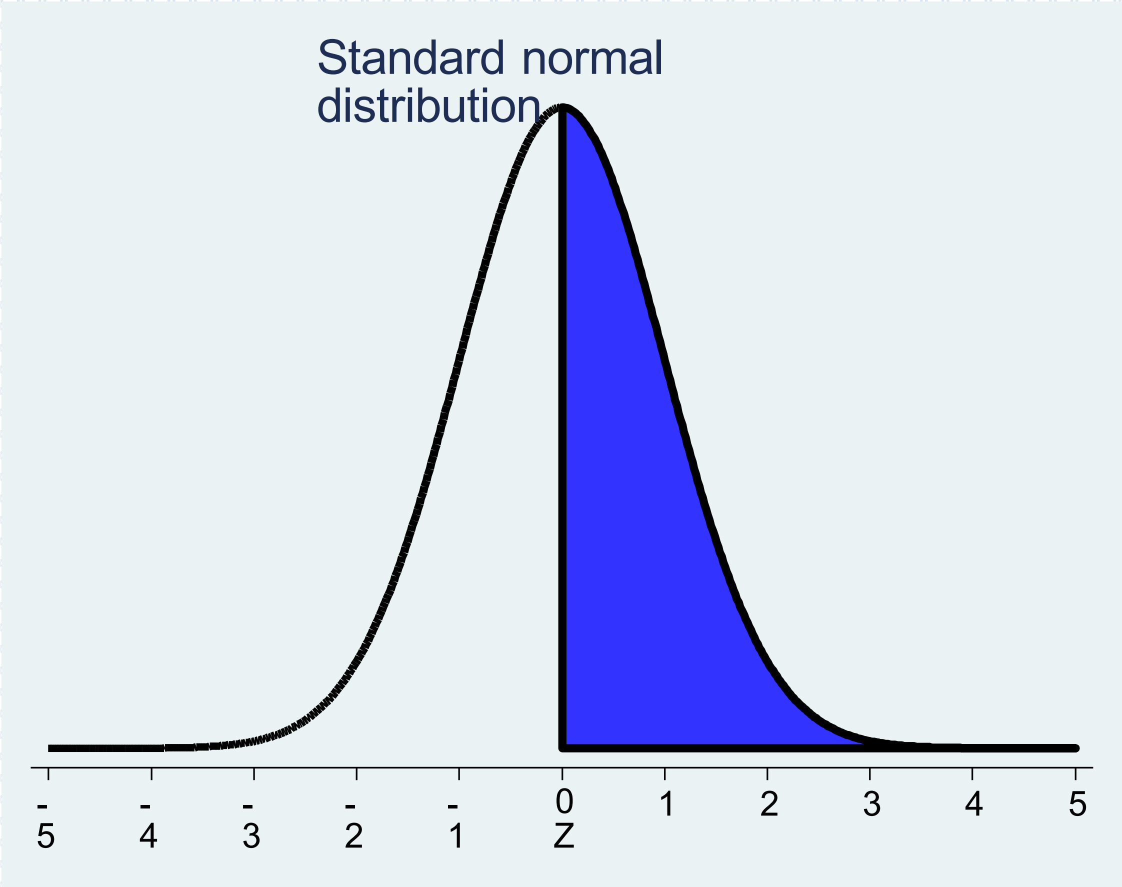
• The area under curves is calculated by taking the integral

$$P(X \geq x) = \int_{x}^{\infty} f(x) = \int_{x}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^{2}} dx$$

Normal distribution

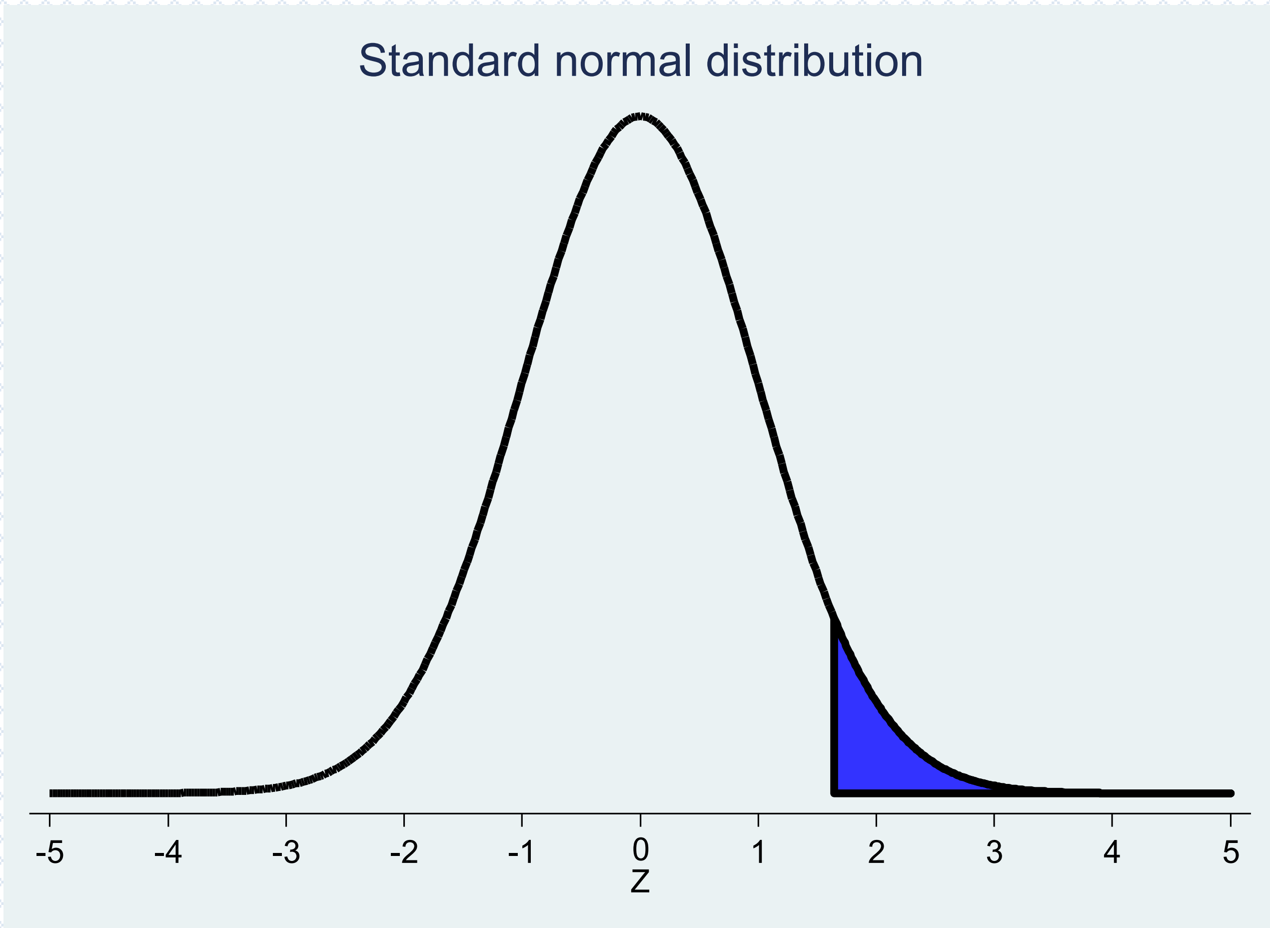
For Z ~ N(0,1), P(Z≥0) = 0.50

Remember that for the standard normal distribution that 0 is not only the mean but also the median of the standard normal distribution. Therefore the probability of observing the median value or more is automatically 0.5 .



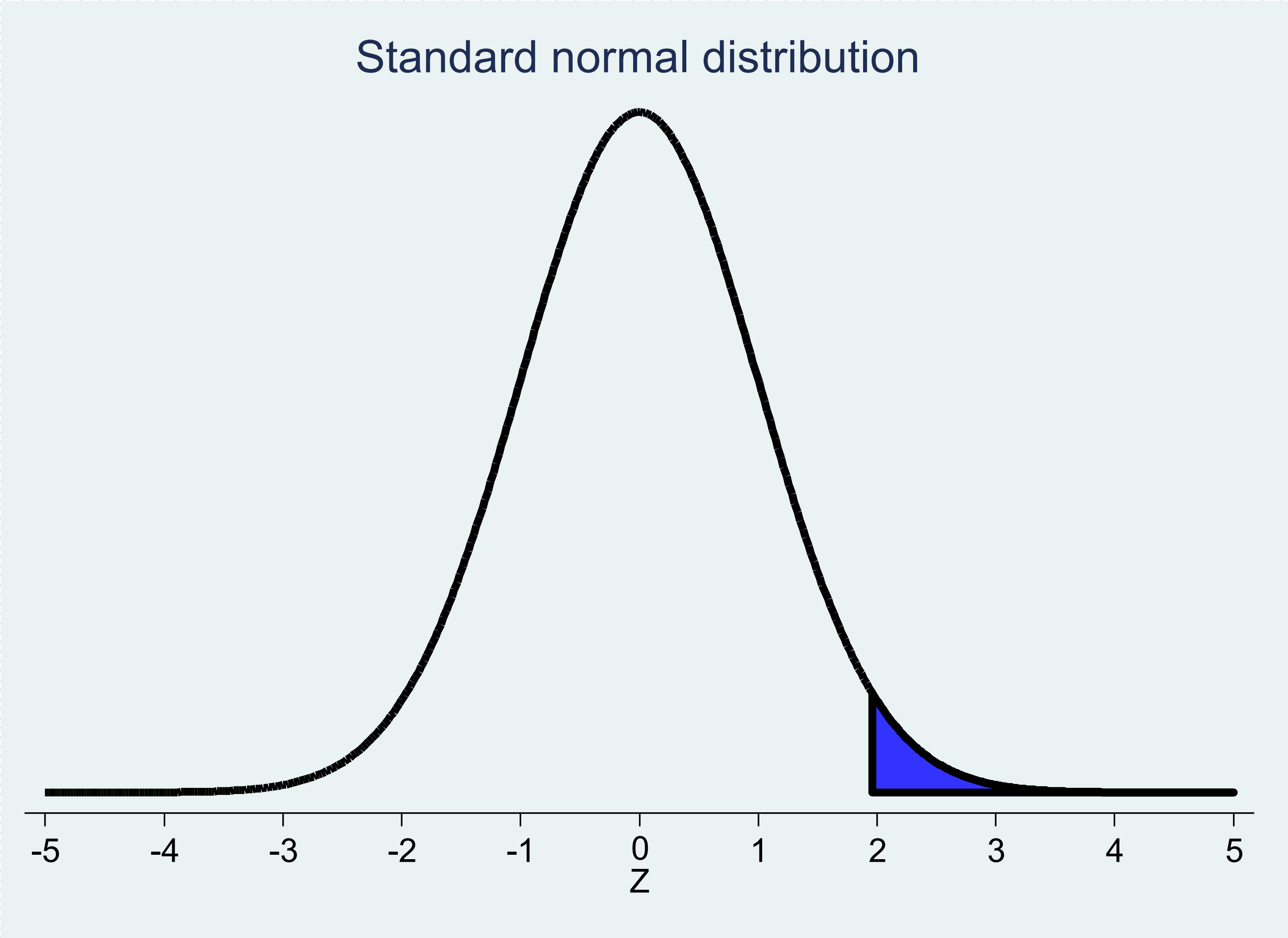
Normal distribution

For $Z \sim N(0,1)$ $P(Z \geq 1.65) = 0.049$



Normal distribution

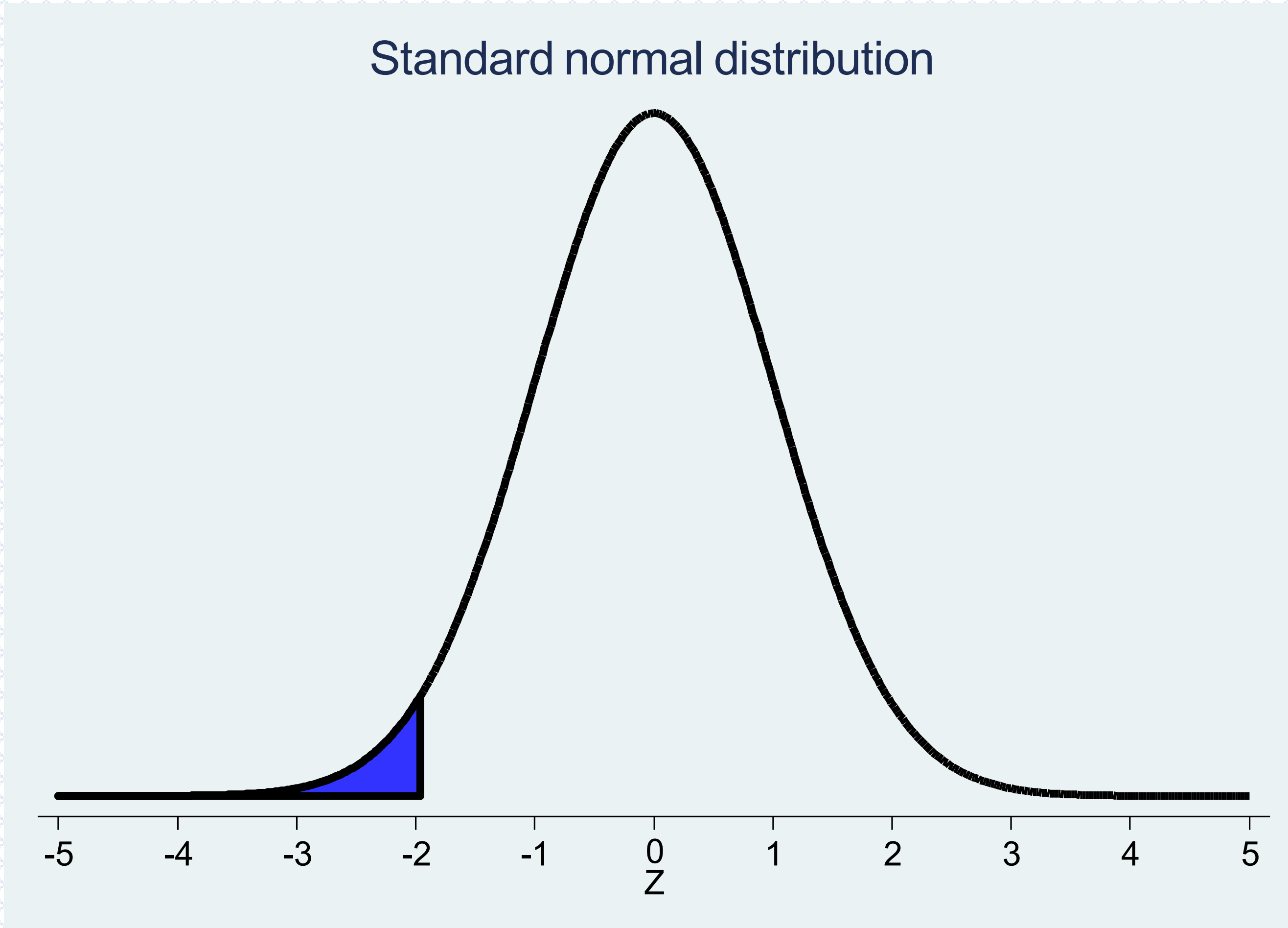
For Z ~ N(0,1) $P(Z \geq 1.96) = 0.025$



Normal distribution

For Z ~ N(0,1) $\quad$ P(Z<-1.96) = 0.025

Z is symmetric



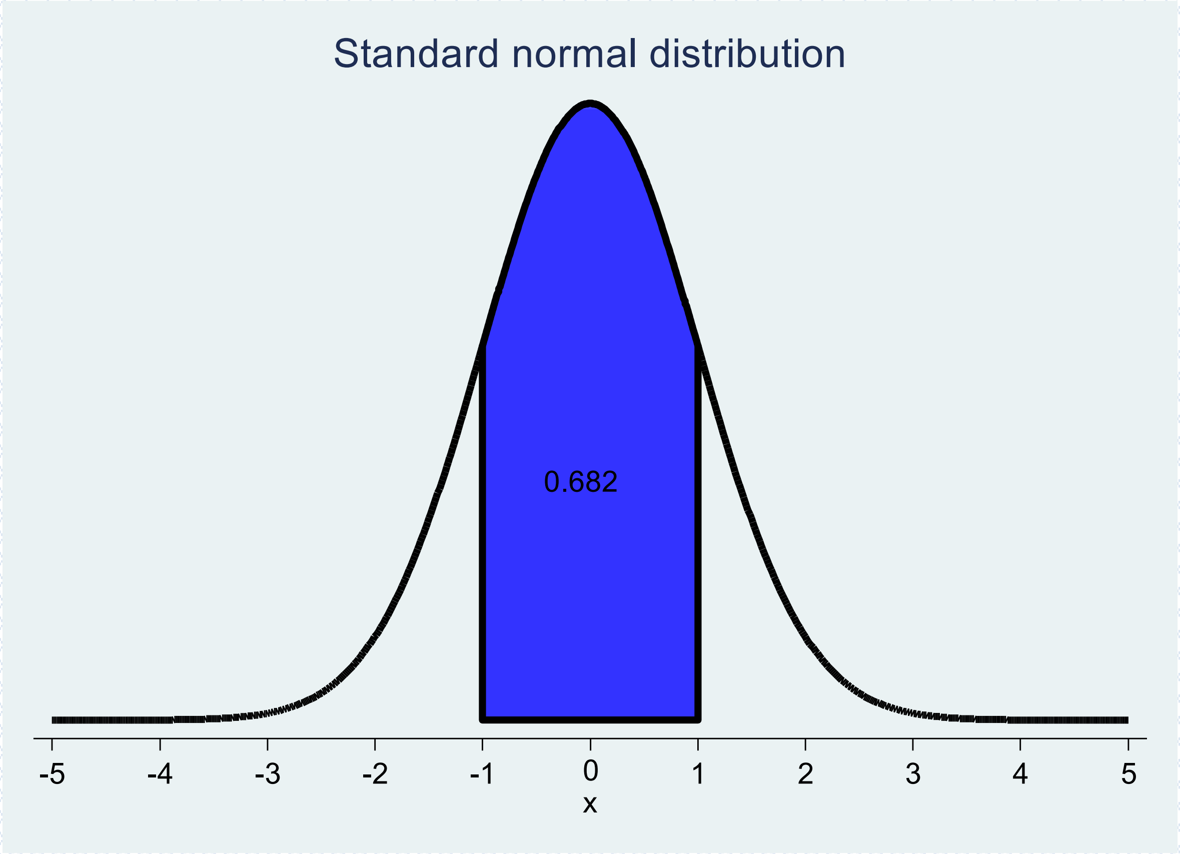
Normal distribution

P(μ-1σ ≤ Z ≤ μ+1σ)

Remember μ=0 and σ=1, so this is

P(-1 < Z < 1) = 0.682

Therefore, approximately 68.2% of the area of the standard normal is within 1 SD of the mean.



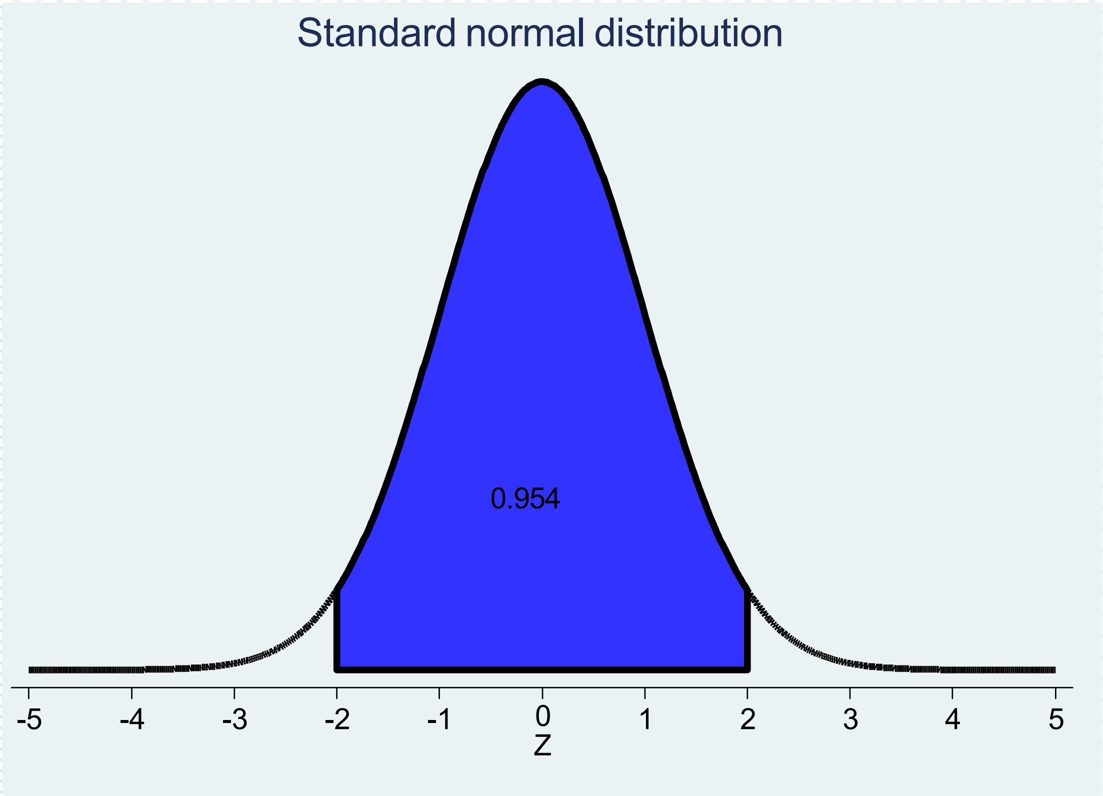
Normal distribution

$P(\mu-2\sigma \leq Z \leq \mu+2\sigma)$

Remember $\mu=0$ and $\sigma=1$, so this is

$P(-2 < Z < 2) = 0.954$

Therefore, approximately 95.4% of the area of the standard normal is within 2 SD of the mean.



Normal distribution

- Stata will calculate standard normal probabilities for you

- In Stata, the left portion of the curve P(Z<z) is calculated for you.

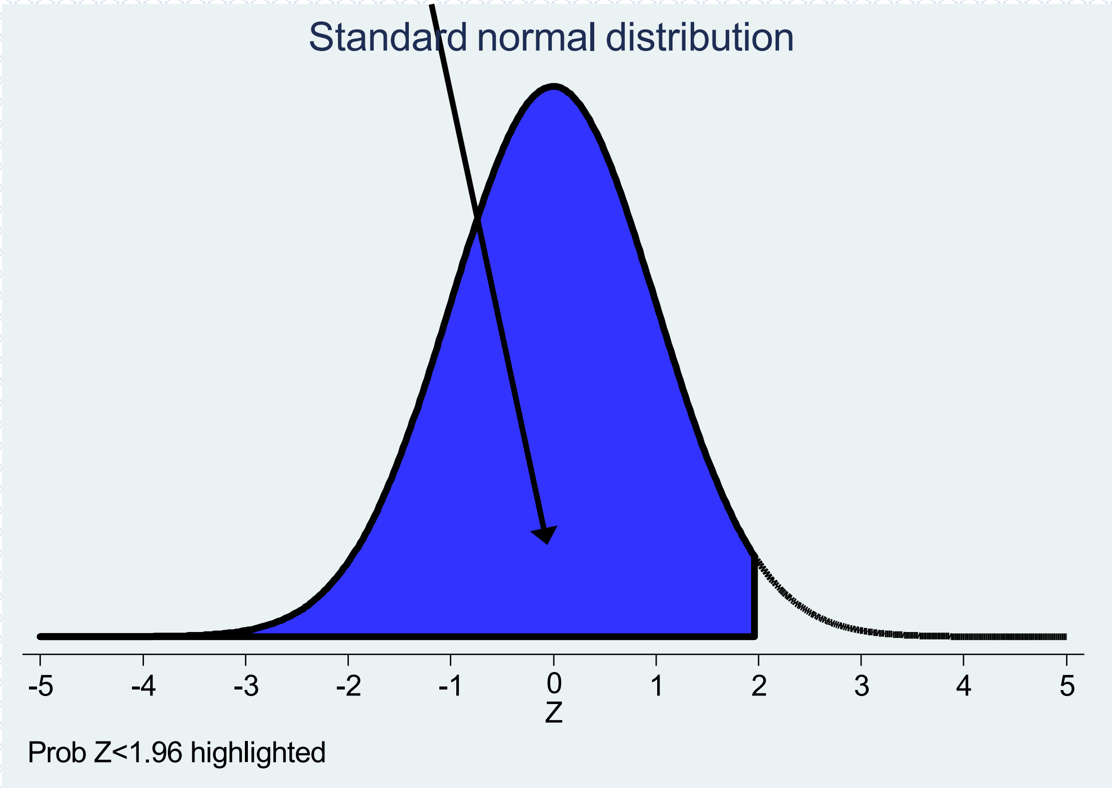
```
*** this gives you P(Z<1.96)
display normal(1.96)
.9750021
```

- If you want the right hand portion of the curve, P(Z≥z), you subtract your answer from 1

```
** this gives you P(Z<1.96)
display 1-normal(1.96)
.0249979
```

- If you want the middle:

```
 display normal(1.96)
             -normal(-1.96)
.95000421
```



Normal distribution: Example

- X is a random variable representing the distribution of systolic blood pressure in 18-74 year old U.S. males
  ~N(129, 19.8)
- What is the proportion in the population with systolic blood pressure of over 150 mm Hg? P(X>150) ?

1. Convert the variable of interest to a standard normal variable N(0,1) to get the probability. I.e., subtract off the mean and divide by the SD.

   z=(150-129)/19.8 = 1.06 This is the z-score or z-statistic

2. Find P(Z>z) = P(Z>1.06) using Stata

```
. di 1-normal(1.06)
.1445723
```

Normal Distribution: Example

- The distribution of systolic blood pressure in 18-74 year old US males is ~N(129, 19.8)

- What is the upper 2.5% value for SBP in this population?

1. Find the z value that satisfies this, and then convert back to the original distribution.

   What is the value of z for which $P(Z \geq \mathbf{z}) = 0.025$?

   ```
   display invnormal(.975)
    1.959964
   ```

   Answer z=1.96

2. Transform back to the original units

   $z=1.96=(x-129)/19.8$

   $x=1.96*19.8 +129 =167.8$ mm Hg

Normal distribution

- If you have a variable that is normally distributed and you know the mean and variance, you can find the values that comprise the middle 95% (or 99% or 90%) of the population Note that this is not a confidence interval

- For the middle 95%, the interval is

  $\mu - 1.96 * \sigma, \ \mu + 1.96 * \sigma$

- For the middle 99%, the interval is

  $\mu - 2.58 * \sigma, \ \mu + 2.58 * \sigma$

- Note that to include a higher percentage, the interval gets wider!

Major Points

- For discrete probability distributions, you can calculate $P(X=x)$
- The binomial distribution gives the probability of the number of successes in n trials $P(X=x)$
- For continuous probability distributions, you can only calculate $P(X>x)$ or $P(X<x)$
- The normal distribution describes some continuous data – we'll see some very useful properties next week
- We transform to the standard normal distribution in order to work with the probabilities