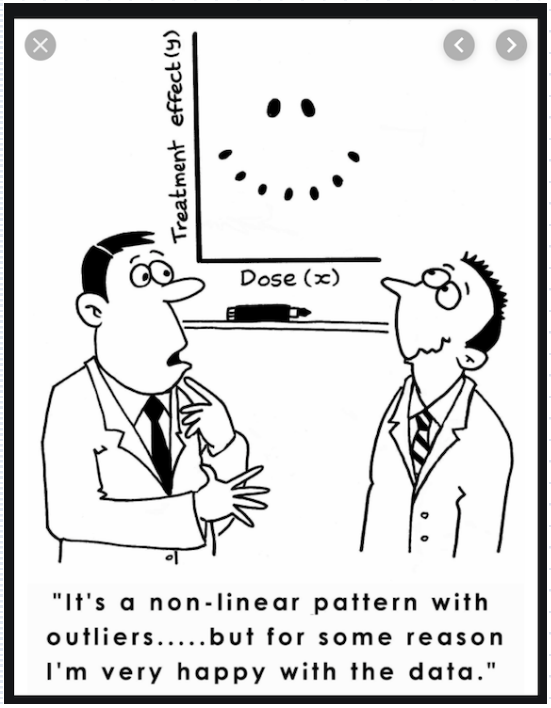
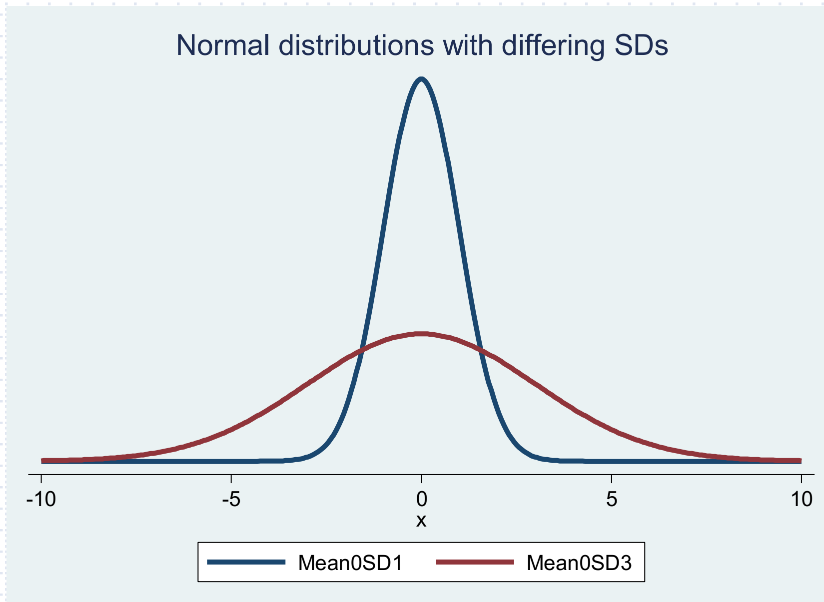
Distributions & the Central Limit Theorem 1



Normal distribution

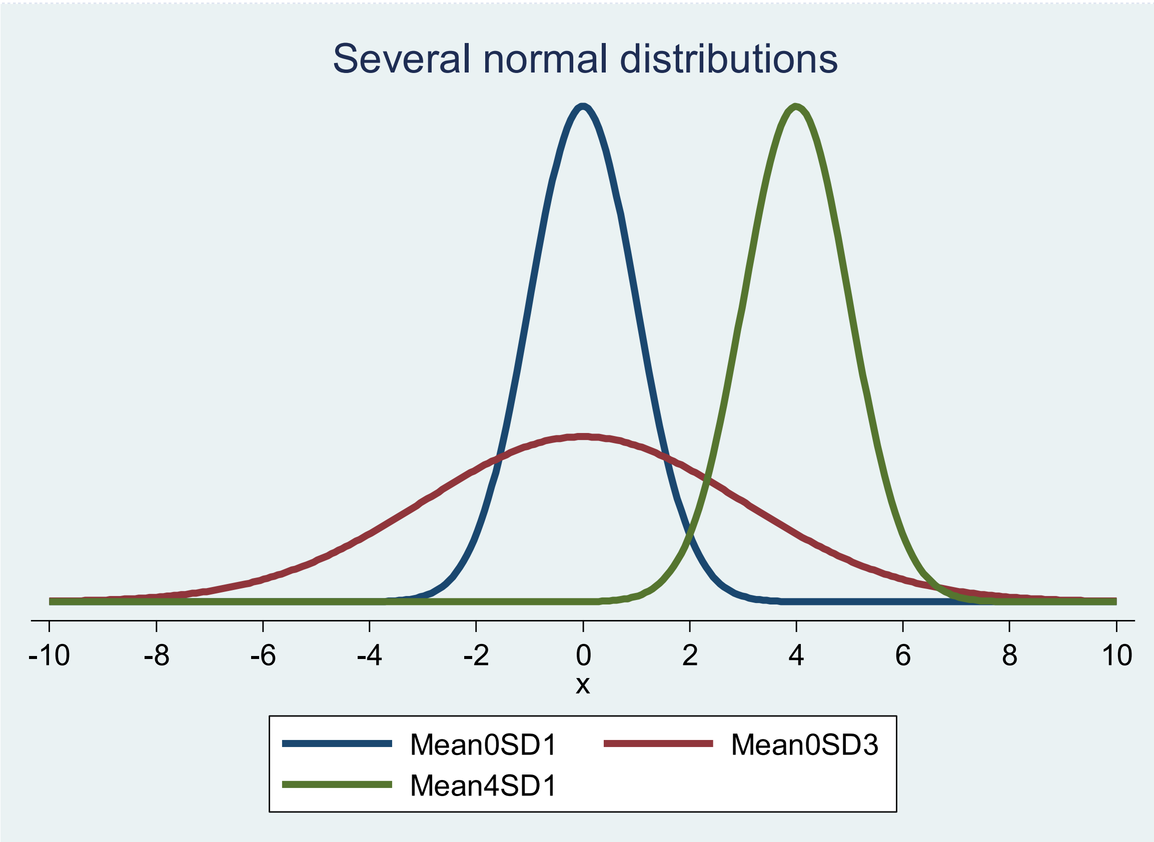
The most important distribution you are going to meet



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



We only want to deal with one Normal Distribution

The Standard Normal Distribution

- μ (the mean) and σ (the standard deviation) 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 with a mean $\mu = 0$ and standard deviation $\sigma = 1$ (and variance $\sigma^2 = 1$).
- Denoted **N(0,1)** & a formula for the distribution

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

The Standard Normal Distribution

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

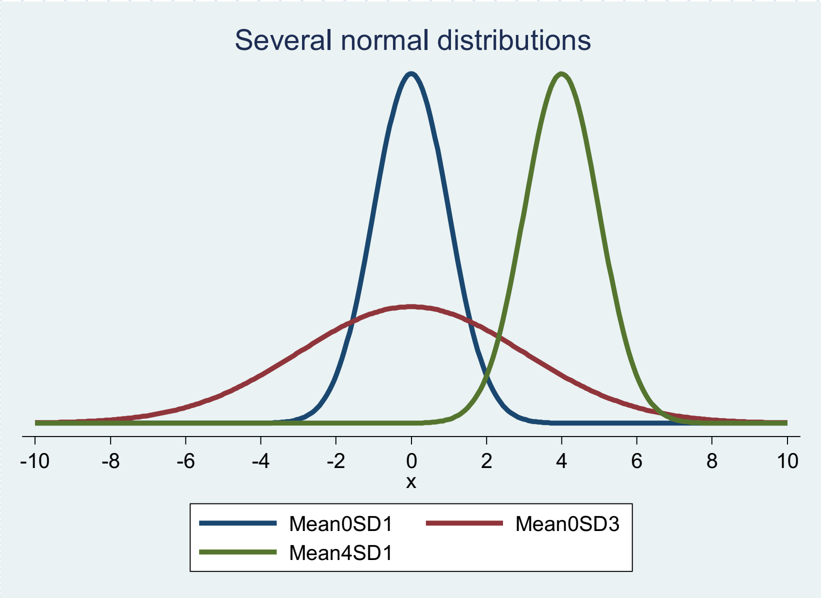
  $Z = (X - \mu)/\sigma$ and we have a *standard* normal random variable

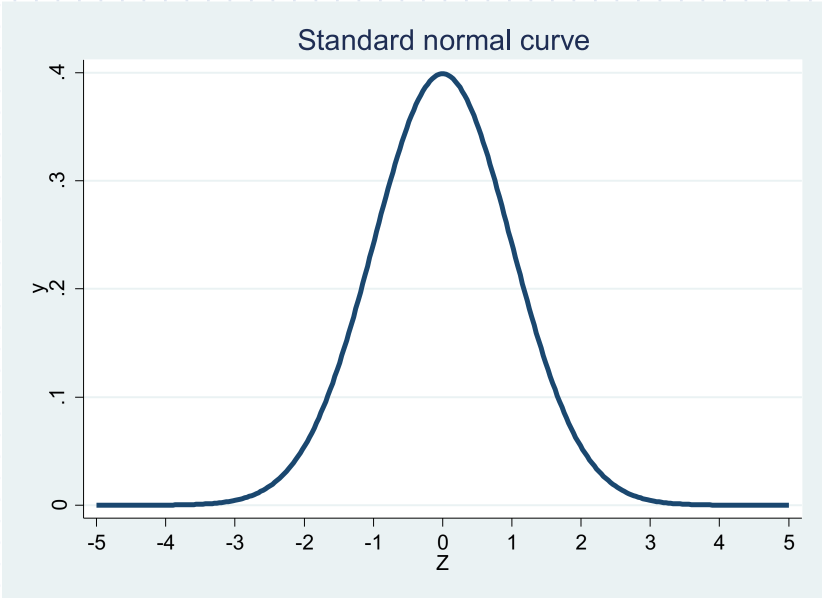
- That is, any normally distributed random variable, subtract it's mean & divide by its standard deviation, it becomes a standard normal random variable with mean=0 and standard deviation=1

- These are often called Z-scores

The Standard Normal Distribution

- If $X \sim N(\mu, \sigma)$ any mean & any standard deviation and then
- $Z = (X - \mu) / \sigma \sim N(0, 1)$





Finding probabilities of Normal distributions

- Recall that we 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

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

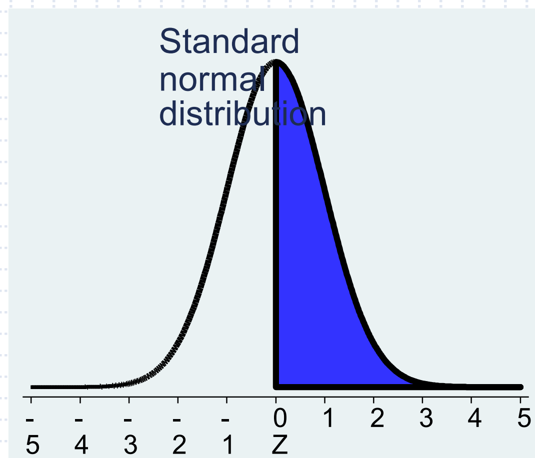
Finding probabilities of Normal distributions

- 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

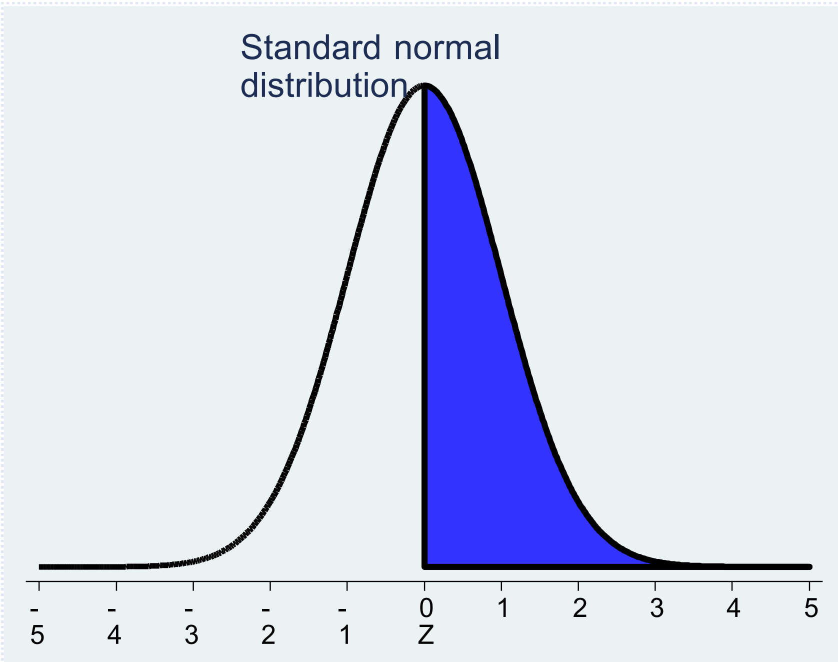
- The area under curves is calculated by taking the integral

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

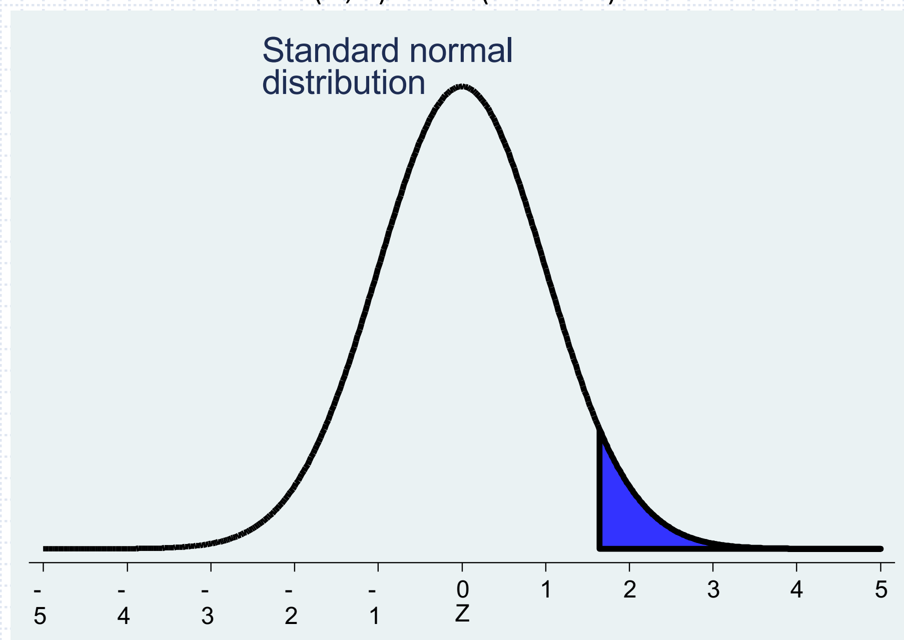


Finding probabilities of Normal distributions

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

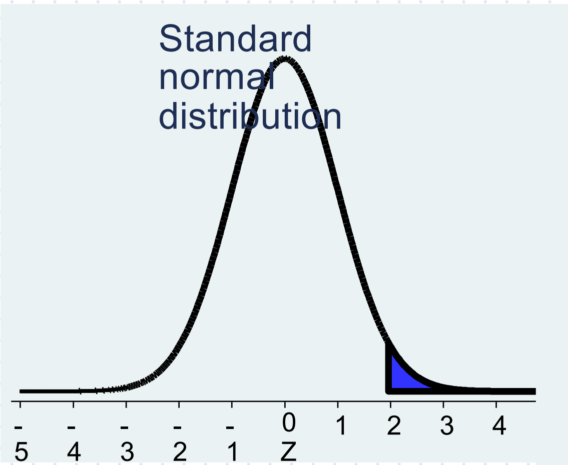


For Z ~ N(0,1) P(Z≥1.65) = 0.049



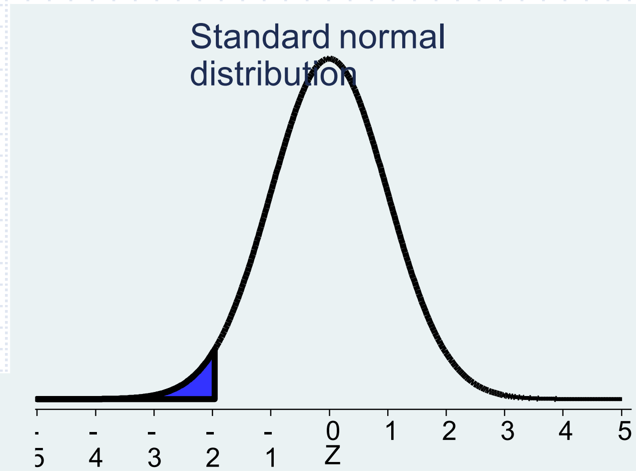
Probabilities of the Normal distribution

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



Z is symmetric

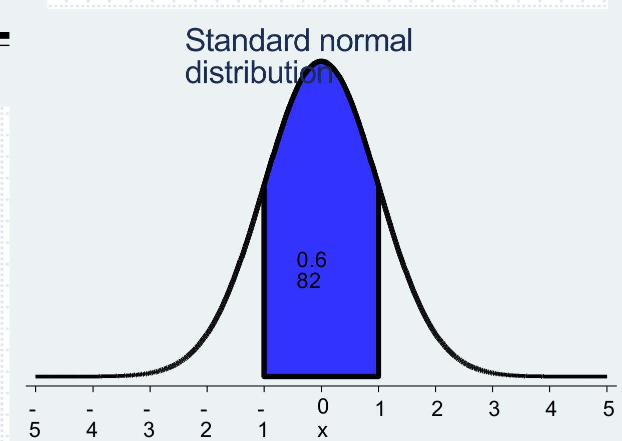
For Z ~ N(0,1) $P(Z < -1.96) = 0.025$



$P(\mu - 1\sigma \leq Z \leq \mu + 1\sigma)$ Remember

$\mu=0$ and $\sigma=1$, 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 probabilities

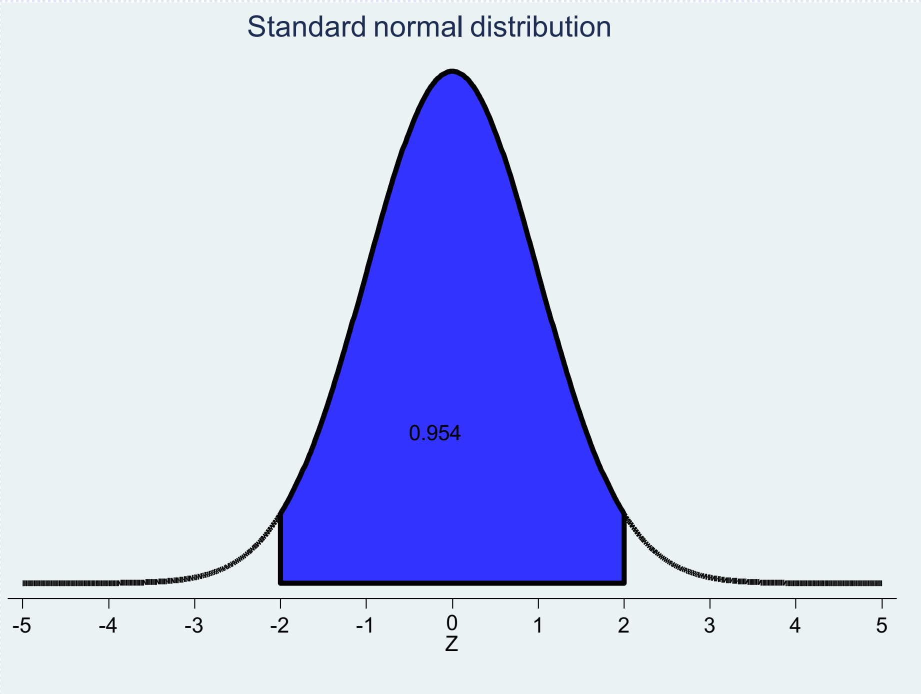
What we are often interested in:

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

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 or 5% of the curve is outside on either end



Normal distribution in STATA

- Stata will calculate standard normal probabilities for you

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

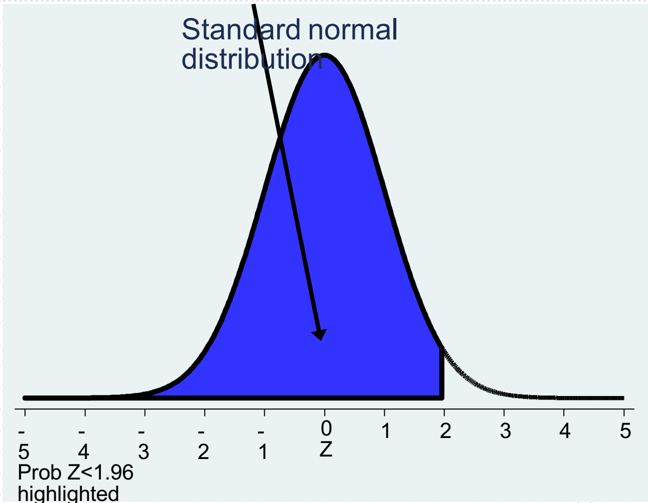
```
*** this gives you P(Z<1.96)
display normal(1.96)
.9750021
```

- 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: Probability example

- X is a random variable representing the distribution of systolic blood pressure in 18-74 year old U.S. males so ~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: Probability 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 $P(Z \geq z)=0.025$?
   ```
   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

Summary of 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