Leave-off point.

Your questions

MICHIGAN STATE UNIVERSITY

I'll add your questions here

Random variables

- "The map is not the territory"
- Described by their distribution
  - PDF/PMF (Probability Density/Mass Function)
  - CDF (Cumulative Density Function)
  - Expected Value (Mean)
  - Variance
  - E and Var are derived from PDF/CDF

Population vs. Sample

- Population statistic is what we're interested in
  - μ is the population mean
  - $E[X] = \mu_X$
  - σ^2 is the population variance
  - $Var[X] = \sigma^2$
- We **do not get to observe the population or any of its statistics**
- But we have realizations of the random variable (RV)
  - RV X has realizations $\{x_1, x_2, x_3, \cdots\}$
  - And with those realizations, we can calculate *sample mean* and *sample variance*
  - *Sample variance* has $\frac{1}{N-1}$ correction

We use greek letters like μ for (unobservable) population statistics

- And they have non-greek sample analogs like $\bar{x}$

Covariance, correlation

MICHIGAN STATE UNIVERSITY

We may be interested in how **two random variables** behave together.

- Income and age
- Wage and education
- Corn yields and fertilizer
- Snowfall and traffic fatalities
- "Letters in Winning Word in Scripps National Spelling Bee" and "Deaths from Venemous Spiders"

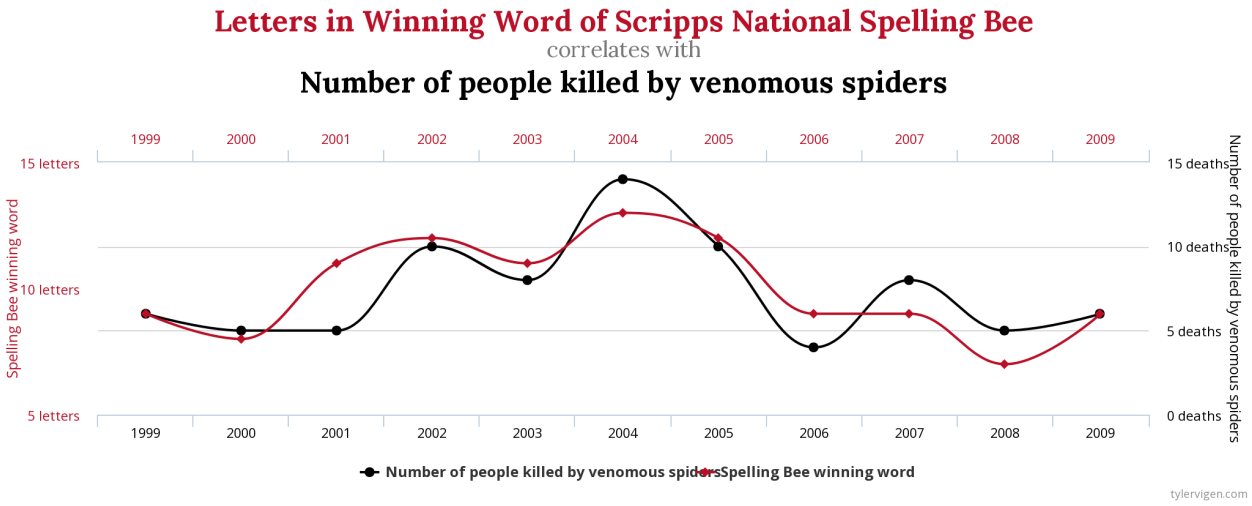
Each of these are *pairs* of RV's.

Just as we have measures of central tendency and dispersion, we will have a measure of this association betwen RV's: **covariance**.

And we can also express the relationship between the PDF's of the RV's.

Yes, it's real

MICHIGAN STATE UNIVERSITY



Let's define **Covariance** between X and Y

$$Cov(X,Y) = \sum_{n=1}^{N}(x_n - \mu_X)(y_n - \mu_Y)$$

Note that we are *summing the pairwise deviations from the mean.*

Note that this is a population concept (which will have a sample analog)

Covariance will be:

-**higher** if x is above the mean when y is also above the mean

-**lower** if x is below the mean when y is also below the mean

-**zero** if x is randomly above/below the mean when y is above/below the mean

Covariance is a measure of how closely two RV's track each other.

With some algebra, we can also write covariance as:

$$Cov(X,Y) = \sum_{n=1}^{N}(x_n - \mu_X)(y_n - \mu_Y) = E(XY) - \mu_X\mu_Y$$

Just as $E(X)$ was important to the *measure of central tendency* and $E(X^2)$ was important to the *measure of dispersion* (variance), $E(XY)$ is important to the covariance, a *measure of association*.

We can also scale the covariance so that it is between -1 and 1:

$$\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \rho \in [-1,1]$$

We will use ρ for the *correlation coefficient*, though you'll frequently see ρ used for other purposes as well.

With regards to the elements of $\rho : Cov(X,Y), Var(X), Var(Y)$, in what case would:

- $\rho = 1$?
- $\rho = -1$?
- $\rho = 0$

When we have two RV's added to each other

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

When two RVs are added to each other, their covariance must be included when calculating the variance of the sum.

- Intuitively, imagine a perfectly correlated X and Y - picture how big the variance would be if, when X is large, Y were also large.
- It would be much larger than the case where X and Y were perfectly negatively correlated.

See Wooldridge v6 Appendix B-4d for more information

For constants a and b and RVs X and Y:

$$Cov(aX, bY) = abCov(X, Y)$$

and therefore:

$$Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y)$$

When a RV is multiplied by a constant, the variance is multiplied by the *square* of the constant.

When two RVs are added to each other, their covariance must be included when calculating the variance of the sum.

Now that we've looked at covariance using the $Cov(X, Y)$ notation, what about in probability distribution function?

The *joint PDF* tells us the probability of seeing a pair of values for X and Y:

$$f_{X,Y}(x, y) = Pr(X = x, Y = y)$$

Note the subscript on f has both X and Y in it, and the joint pdf takes *two* values as inputs.

So in the joint pdf, you'll see both x and y.

- It would be strange to have X and Y correlated and not have a pdf that includes x and y

This bring us to **independence** of random variables, which will be extra-important in this class.

Two RVs are independent if and only if their joint pdf is equal to the product of the marginal pdfs

$$X \perp Y \iff f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

- $\perp$ is read as "independent"
- Sometimes you'll see $\perp\!\!\!\perp$
- The symbol $\iff$ means "if and only if"
  - This will be common in future courses
  - It means that when one side holds true, the other side will always hold true.
- $f_X(x)$ is the pdf of X, just as we used it before.
  - This will be called the *marginal pdf of X*.

Independence is *defined* in this way, but it has many implications.

Intuitively, it means that **knowing a realized value of X tells us nothing at all about the realized value of Y**.

Think about flipping a coin X and rolling a die Y

- If X lands on "heads", what do you think the die roll will be?
- The die roll would have the same probability of each number $\{1, \cdots, 6\}$, regardless of how coin X landed!
- If we were to write the pdf of the die roll Y, *it would not have the realization of the first in it.*

$$f_Y(y) = \frac{1}{6} \quad \text{for any} \quad y \in \{1, \cdots, 6\}$$

  - See - no "X" in there at all! $Y \perp X$
  - $f_{Y|X}(y|x) = f_Y(y)$ is another way of saying this

If X and Y are independent:

$$f_X = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Because $E(X)$ and $E(Y)$ are defined by their pdf's, $X \perp Y$ implies that the expectations can be separated:

$$E(XY) = E(X)E(Y)$$

And this means that $X \perp Y$ implies that:

$$X \perp Y \Rightarrow \overbrace{Cov(X,Y) = E(XY) - E(X)E(Y)}^{\text{defined before}} = \overbrace{E(X)E(Y)}^{\text{because } X \perp Y} - E(X)E(Y)$$
$$= 0$$

If X has no information about Y, then we would expect $Cov(X,Y)$ to be zero!

When X and Y are *not* independent ($\not\perp$), we can use the *conditional pdf* for X and Y. The conditional pdf is written as: $f_{X|Y}(x|y)$

- The symbol '|' is read as "conditional"
- In a discrete RV, $f_{X|Y}(x|y) = Pr(X = x|Y = y)$

Example:

Let X be life expectancy and let Y be smoking status ($Y = 1$ if a smoker).

- $f_X(x)$ is the marginal pdf of life expectancy (ignoring smoking status)
- $f_Y(y)$ is the marginal pdf of smoking status (ignoring any data we may have on life expectancy)
- $f_{X|Y}$ has a different result when we learn about a realization's smoking status.
- $f_{X|Y}(x|y = 1)$ will have higher probabilities on lower values of x here

We are pretty sure that *knowing about a random draw's smoking status* tells us something about the distribution of their life expectancy.

Of course, there is an Expectation equivalent: the *conditional expectation*:

$$E(X|Y) = \int_{-\infty}^{+\infty} x_s f_{X|Y}(x_s|Y = y)ds$$

That is, the weighted sum/integral of the *conditional pdf* of X

If X and Y are independent, then:

$$E(X|Y) = E(X)$$

This property does not always imply independence (it is not $\iff$).

- We call this property *mean independence*, meaning the *mean* of X is not affected by realizations of Y.
- This property will be crucial in the next unit.

Remember, that mean independence does not imply independence!

iid

You will see "iid" frequently - it means "independent and identically distributed". It refers to collections of random draws from random variables that are:

- *independent* - the realization of one draw does not give any information about the realization of any other draws
- *identically distributed* - each draw is from a RV that has the same pdf

When we have iid random variables, we can say something about the distribution they all share.

Since they all have the same mean (same distribution), then we can say something about the mean.

Types of distributions

Bernoulli

The *Bernoulli* distribution is a very simple distribution, best known as the "coin flip" distribution.

It is for a binary 0/1 outcome variable and has only one "parameter", θ.

For a coin flip, θ = .5

The full distribution is:

$$f_X(x) = \begin{cases} \theta & \text{if } x = 1 \\ 1 - \theta & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

This is a distribution because it sums to 1 and it covers every possible value in the domain of X.

Normal

Also known as *gaussian*, the normal is very useful. It is written as $N(\mu, \sigma^2)$

It has two parameters:

- μ is the "location" parameter
- σ^2 is the "scale" parameter
- $E(X) = \mu$ and $Var(X) = \sigma^2$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

The normal has support for all real numbers. That is, *every real number* has a non-zero probability of being realized.

Standardization of the Normal

Standardizing the normal to Z:

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

The $\sim$ is read as "is distributed"

Subtracting the (population) mean and dividing by (population) standard deviation results in a normal RV with mean 0 and variance of 1.

The pdf of the standard normal is $\phi(z) = \frac{1}{\sqrt{2\pi}} exp[-\frac{z^2}{2}]$

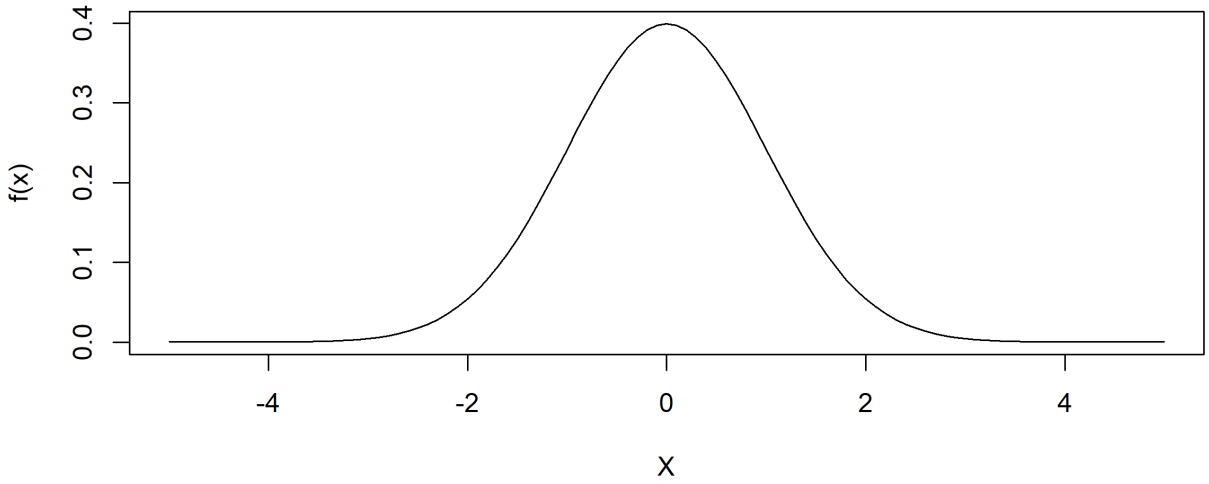
The standard normal is very useful, so it appears in a lot of texts. Thus, it gets its own notation:

- The standard normal pdf is written as $\phi(z)$.
- The standard normal cdf is written as $\Phi(z)$.

Types of distributions

MICHIGAN STATE UNIVERSITY

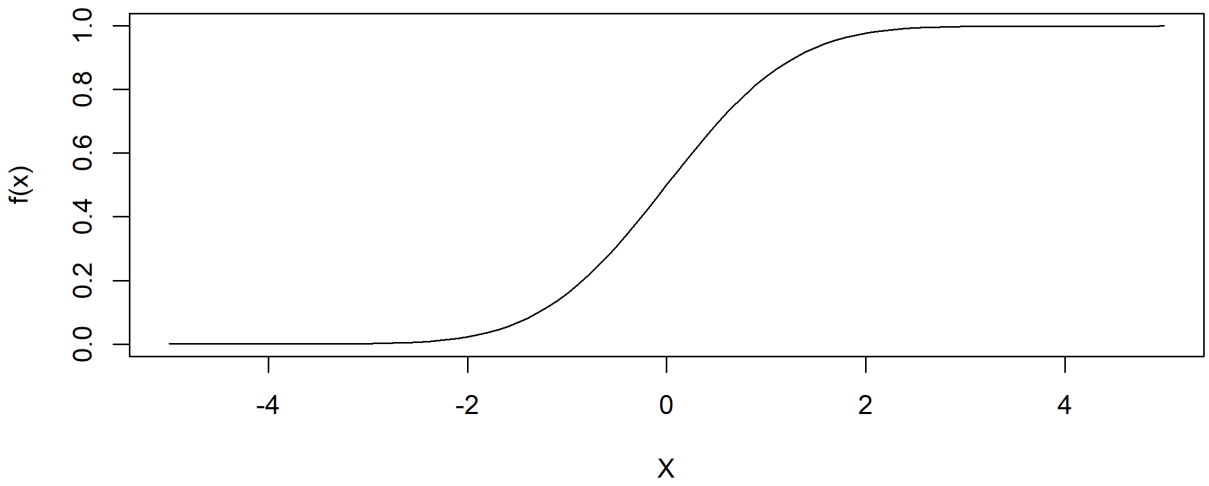
Standard normal pdf:



Types of distributions

MICHIGAN STATE UNIVERSITY

Standard normal cdf:



Properties of the Normal Distribution

- If $X \sim N(\mu, \sigma^2)$, then $Y = aX + b \sim N(a\mu + b, a^2\sigma^2)$
  - This is not true for other distributions.
- If X and Y are jointly normally distributed, then they are independent if and only if $Cov(X, Y) = 0$

Any linear combination of indpendent, identically distributed (iid) normal random variables has a normal distribution:

- Given $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$

$$\Rightarrow X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y)$$

The Chi-Square Distribution, χ^2

If Z_i, i=1,...,n is iid standard normal $N(0, 1)$, then the sum of these varaibles squared:

$X = \sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$

is distributed Chi-squared with n degrees of freedom

The t-distribution:

If $Z \sim N(0, 1)$ and $X \sim \chi_n^2$ and Z and X are independent, then:

$$T = \frac{Z}{\sqrt{\frac{X}{n}}} \sim t_n$$

T is distributed t with n degrees of freedom.

The F-distribution:

If $X_1 \sim \chi^2_{k_1}$ and $X_2 \sim \chi^2_{k_2}$, and they are independent, then:

$$F = \frac{\frac{X_1}{k_1}}{\frac{X_2}{k_2}} \sim F(k_1, k_2)$$

F is distributed F with k_1, k_2 degrees of freedom.

The F-distribution is useful when you are testing the ratio of two Chi-squared distributions.

Statistical inference

Many thanks to Prof. Herriges for the basis for this section

Our goal is to learn something about a *population* given the availability of a *sample* from that population.

- We will spend a lot of time making the connection between a sample and the population
- We have already done this when we introduced the sample mean and compared it to the (unknown) population mean

If $X_1, X_2, \ldots, X_n$ are independent random variables all drawn from a common pdf, $f(x;\theta)$, then $X_1, X_2, \ldots, X_n$ is said to be a *random sample* from $f(x;\theta)$.

- Here, the θ *parameterizes* the distribution.
- θ might be $\theta = \{\mu, \sigma^2\}$

The realization would be denoted by $x_1, x_2, \ldots, x_n$.

If we know (or will assume) a specific distribution for X but do not know the specific values of the parameters of that distribution, then we **estimate** the parameter(s) from the available sample

Estimator

The *estimator* of θ is a rule that assigns to each possible outcome of the sample a value of θ

- It is specified before any sampling. It does not depend on the sample.

For example, we may be interested in the *population mean* μ. The sample average is an *estimator* of μ.

$$\bar{Y} = \frac{1}{N}\sum_{n=1}^{N} Y_n$$

But Y is a random variable - it changes every time we take a sample. We can calculate an *estimate*:

$$\bar{y} = \frac{1}{N}\sum_{n=1}^{N} y_n$$

The *estimate* pertains to the realized sample. An *estimator* is a mapping (e.g. a "plan") for learning about the unknown parameters of the distribution.

The *estimator* is a function of random variables, so it is also a random variable. Let's generalize and refer here to the estimator as $W = h(Y_1, \ldots, Y_n)$, and θ is the population parameter.

- The distribution of the estimator is known as the **sampling distribution**
- An estimator is **unbiased** if $E(W) = \theta$
- The bias is given by $Bias(W) = E(W) - \theta$

Now, relate these to our earlier discussion on mean and variance:

- $\bar{Y}$ is an unbiased estimator of the population mean $E(Y) = \mu_Y$
- The *sample variance*, s^2 from ealier can be shown to be an unbiased estimate of $Var(Y) = \sigma^2$

$$s^2 = \frac{1}{N-1}\sum_{n=1}^{N}(Y_i - \bar{Y})^2$$

If $\bar{Y}$ is an *estimator*, a function of multiple RV's, then $\bar{Y}$ **is a Random Variable itself**

And it has a mean and variance.

- The mean of $\bar{Y}$ is the mean of the distribution of Y
- The variance of $\bar{Y}$ is related to N and σ_Y

If we assume that $Y \sim N(\mu, \sigma_Y^2)$:

$$\begin{aligned} Var(\bar{Y}) &= Var(\frac{1}{N}\sum y_i) \quad (\text{by def of average}) \\ &= \frac{1}{N^2} Var(\sum y_i) \quad (\text{By operation of variance}) \\ &= \frac{1}{N^2} N\sigma_Y^2 \quad (\text{Sum of iid } y_i) \\ &= \frac{\sigma_Y^2}{N} \quad (\text{Cancel } N) \end{aligned}$$

That gives us a *sampling distribution* for $\bar{Y}$

So if $Y \sim N(\mu, \sigma_Y^2)$, then the sampling distribution of $\bar{Y} \sim N(\mu, \frac{\sigma_Y^2}{N})$

This is a very important link because it tells us:

- $E[\bar{Y}] = \mu$, the population statistic of interest
- $Var(\bar{Y})$, so know how much dispersion there is around μ. -\textbf{IF} we know σ_Y^2.
  - Remember, $Var(Y) = \sigma_Y^2$ and $Var(\bar{Y}) = \frac{\sigma_Y^2}{N}$ are different

Central Limit Theorem (CLT)

The CLT is what tells us that $\bar{Y}$ is **normally** distributed.
In fact, it says that *any* normalized sum of iid variables is distributed normally

- "Normalized" here means "divided by N'
- **Even** when those iid variables are from a non-normal distribution
  - This is **amazing**.

Let's look at an example

Let's take a random variable, X, that is definitely not normally distributed.

We'll draw a sample of size $N = 5$ and calculate the average, $\bar{x}$.

Then, we'll do the same thing again, drawing $N = 5$ new realizations and calculating the average. And again. And again. $R = 100$ times.

Despite the fact that X is not normally distributed, if we plot all 100 values of $\bar{x}_r$ calculated from the 100 draws of $N = 5$, **it is going to look like a normal distribution!

Even more notable, it will have a variance equal to the population variance of X divided by N.

So what do we do with this?

Even if we just have *one* $\bar{y}$ from *one* draw of size N, we we can learn alot about the population mean, μ.

Law of Large Numbers

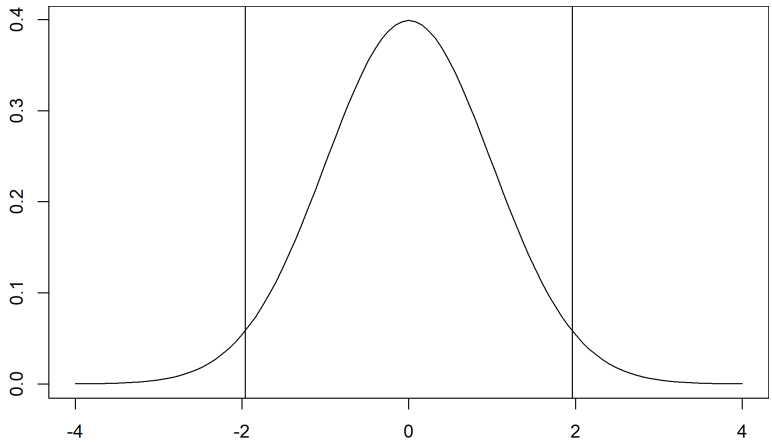
$$\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i \rightarrow E[Y] = \mu_Y \quad \text{when} \quad n \rightarrow \infty$$

Thus, we know that $E[\bar{Y}] = \mu_Y$

Now, that we have a distribution for our estimate, we can standardize it:

$$Z = \frac{\bar{Y} - \mu}{\sqrt{\sigma_Y^2/N}} \sim N(0,1)$$

Which implies that the now-standardized distribution of $\bar{Y}$, Z, looks like:



The vertical lines are where $F_(z) = \{.025, .975\}$. The area under the curve between the lines is .95

Thus, we can get our 95% CI

$$\begin{aligned} .95 &= Pr(-1.96 < \frac{\bar{Y} - \mu}{\frac{\sigma_Y}{\sqrt{N}}} < 1.96) \\ &= Pr(\bar{Y} - 1.96\frac{\sigma_Y}{\sqrt{N}} < \mu < \bar{Y} + 1.96\frac{\sigma_Y}{\sqrt{N}} \end{aligned}$$

± 1.96 is the *critical value* for a Normal with a 95% confidence interval.

- You can find this in the Z-tables in any statistics text, including Wooldridge

Confidence intervals (CI)

- We know the sampling distribution of $\bar{Y}$.
- We know the 95% CI

Question:

- If I draw a new sample from Y, will the 95% CI on $\bar{Y}$ change?
- Will the population parameter we're interested in change?

So which is correct:

A. "there is a 95 percent probability that the true value of μ falls in the estimated confidence interval."

B. "for 95% of all random samples, the constructed CI will contain μ."

B is correct.

Wooldridge Table C-2

TABLE C.2 Simulated Confidence Intervals from a Normal(μ,1) Distribution with $\mu = 2$

Replication	$\bar{y}$	95% Interval	Contains μ?
1	1.98	(1.36,2.60)	Yes
2	1.43	(0.81,2.05)	Yes
3	1.65	(1.03,2.27)	Yes
4	1.88	(1.26,2.50)	Yes
5	2.34	(1.72,2.96)	Yes
6	2.58	(1.96,3.20)	Yes
7	1.58	(.96,2.20)	Yes
8	2.23	(1.61,2.85)	Yes
9	1.96	(1.34,2.58)	Yes
10	2.11	(1.49,2.73)	Yes
11	2.15	(1.53,2.77)	Yes
12	1.93	(1.31,2.55)	Yes
13	2.02	(1.40,2.64)	Yes
14	2.10	(1.48,2.72)	Yes
15	2.18	(1.56,2.80)	Yes
16	2.10	(1.48,2.72)	Yes
17	1.94	(1.32,2.56)	Yes
18	2.21	(1.59,2.83)	Yes
19	1.16	(.54,1.78)	No
20	1.75	(1.13,2.37)	Yes

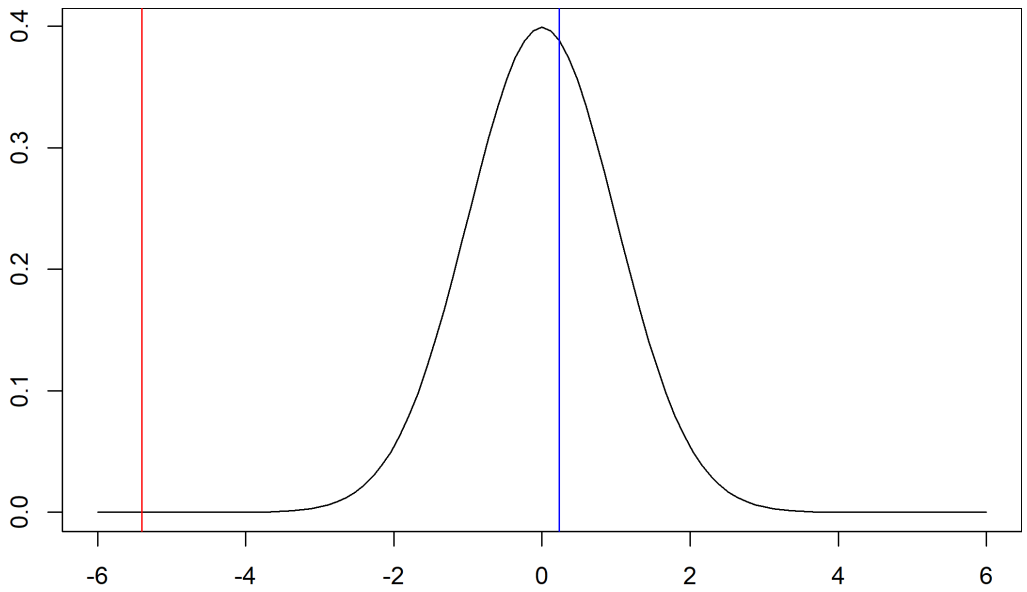
The resulting standardized distribution tells us what we would expect to see **if μ is the true population mean**.

Remember, we acted as if we knew what μ was.

Although we know that $E[\bar{Y}] = \mu$, our realization of the sample mean, $\bar{y}$, is not exactly equal to μ.

If we plug in any guess for the real μ, we can take our sample estimate of $\bar{y}$, subtract our hypothesized μ, divide by σ/N and plot it on the distribution.

Let's call our hypothesized population mean μ_0. It is the *null hypothesis*.



- If the null hypothesis is correct, the black curve is the pdf ("what we'd expect to see")
- If our statistic $\frac{\bar{y}-\mu_0}{\sigma/N}$ is the red line, do we think we have the distribution right?
- What about for the blue?

Another problem:

We can hypothesize about μ_0, but what about σ^2? We don't know σ^2. If we did, we'd be in great shape.

But we have the **sample variance** estimator from before:

$$s^2 = \frac{1}{N-1}\sum_{n=1}^{N}(y_i - \bar{y})^2$$

But because we have an estimate of σ_Y^2, our standardized statistic is no longer $\sim N(\mu_Y, \frac{\sigma_Y^2}{N})$. Now it is:

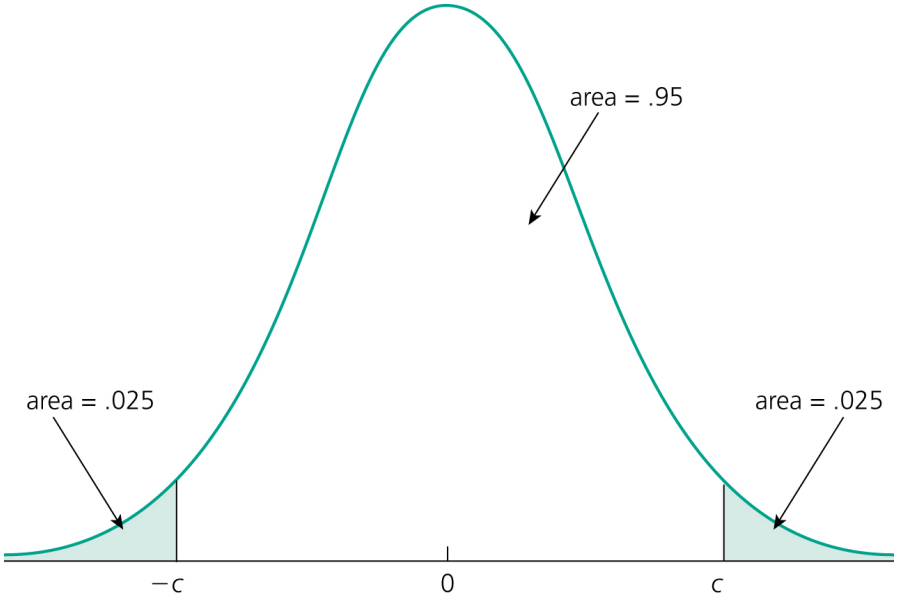
$$\frac{\bar{Y} - \mu_Y}{\sqrt{\frac{s^2}{N}}} \sim t_{N-1}$$

It is distributed t with $N-1$ degrees of freedom.

- The t-distribution is also given in your book (Table G-2)
- So we can look up the 95% critical values
- Note that the critical values will change as $N-1$ changes
- And that they get closer to the values for a Standard Normal as N gets large.

Critical values

By standardizing, we are relating the distribution of our **estimate** to a known distribution.



Wooldridge Figure C-4

TABLE G.2 Critical Values of the *t* Distribution

		Significance Level				
1-Tailed:		**.10**	**.05**	**.025**	**.01**	**.005**
2-Tailed:		**.20**	**.10**	**.05**	**.02**	**.01**
Degrees of Freedom	1	3.078	6.314	12.706	31.821	63.657
	2	1.886	2.920	4.303	6.965	9.925
	3	1.638	2.353	3.182	4.541	5.841
	4	1.533	2.132	2.776	3.747	4.604
	5	1.476	2.015	2.571	3.365	4.032
	6	1.440	1.943	2.447	3.143	3.707
	7	1.415	1.895	2.365	2.998	3.499
	8	1.397	1.860	2.306	2.896	3.355
	9	1.383	1.833	2.262	2.821	3.250
	10	1.372	1.812	2.228	2.764	3.169
	11	1.363	1.796	2.201	2.718	3.106
	12	1.356	1.782	2.179	2.681	3.055
	13	1.350	1.771	2.160	2.650	3.012
	14	1.345	1.761	2.145	2.624	2.977
	15	1.341	1.753	2.131	2.602	2.947
	16	1.337	1.746	2.120	2.583	2.921
	17	1.333	1.740	2.110	2.567	2.898
	18	1.330	1.734	2.101	2.552	2.878
	19	1.328	1.729	2.093	2.539	2.861
	20	1.325	1.725	2.086	2.528	2.845
	21	1.323	1.721	2.080	2.518	2.831
	22	1.321	1.717	2.074	2.508	2.819
	23	1.319	1.714	2.069	2.500	2.807
	24	1.318	1.711	2.064	2.492	2.797
	25	1.316	1.708	2.060	2.485	2.787
	26	1.315	1.706	2.056	2.479	2.779
	27	1.314	1.703	2.052	2.473	2.771
	28	1.313	1.701	2.048	2.467	2.763
	29	1.311	1.699	2.045	2.462	2.756
	30	1.310	1.697	2.042	2.457	2.750
	40	1.303	1.684	2.021	2.423	2.704
	60	1.296	1.671	2.000	2.390	2.660
	90	1.291	1.662	1.987	2.368	2.632
	120	1.289	1.658	1.980	2.358	2.617
	∞	1.282	1.645	1.960	2.326	2.576

Examples: The 1% critical value for a one-tailed test with 25 *df* is 2.485. The 5% critical value for a two-tailed test with large (> 120) *df* is 1.96.

Source: This table was generated using the Stata® function invttail.

Wooldridge Figure G-2

An example: years to complete a PhD

Years
9.0
6.5
4.0
7.5
5.5
4.5
1.5
5.5

Calculate $\bar{x}$ and s^2

$N = 8$

$\bar{x} = 5.5$

$s^2 = 5.214$

$\sqrt{\frac{s^2}{N}} = 0.807$

So once we standardize our $\bar{x}$, it is distributed t_{8-1}

The critical values, which we'll call t_{crit} for t_{8-1} are:

± 2.365

So the 95% confidence interval is $\bar{x} \pm t_{crit}\sqrt{\frac{s^2}{N}}$:

$$5.5 \pm 2.365 \times 0.807$$

Which is $[3.638, 7.362]$

For 95% of random samples, this confidence interval will include the true parameter.

What if we want to ask the question "does the average Ph.D. take 5 years?"

Often, we are interested in these *population* questions

- E.g.: does the average person obtain more education when college is subsidized?
- Similarly, we can ask "Does a MSU Ph.D. take five years on average?".

We call the **null hypothesis** H_0.

It is what we can test and either **reject** or **fail to reject**

- We do not ever **accept** or **confirm** a null hypothesis
- H_0 is always a point estimate (=)

We call the alternative the **alternative hypothesis**, H_1.

- H_1 is always an inequality ($>$ or $<$) or $\neq$
  - When H_1 takes the form $\neq$, it is a 2-tailed test.

Hypothesis testing

There are two types of mistakes we can make in testing a hypothesis:

- Type I error: Rejecting the null hypothesis, H_0, when it is true
  - The *significance level of a test*, α, is the *probability of a Type I error*.
  - Mathematically, α = Pr(Reject H_0 | H_0 is true)
- Type II error: Failing to reject the null hypothesis when it is false

Let's call our test statistic T

- So far, in all of our examples, our test statistic has been the population mean
  - But not always!
- The realized value (mapped from the sample to the estimate) we'll call t
- Given the test statistic, we can define a *rejection rule* which will tell us the values of t for which H_0 is rejected.
  - Think of H_0 as the "guess"
  - The rejection regions are the values which rule out that guess

For example: if our test statistic, T is "average miles driven per day" and our H_0 is something reasonable, then the rejection region will tell us the *realized* values t that would make H_0 unlikely.

- If our H_0 is "5 miles per day" and we *realize* a value $t = 100$, then we would be pretty convinced that the true value is not "5 miles per day".

The rejection region depends on the alternative hypothesis:

- $H_1 : \mu \neq \mu_0$ (where μ_0 is the hypothesized value)
- $H_1 : \mu > \mu_0$
- $H_1 : \mu < \mu_0$

In a two sided case ($H_1 : \mu \neq \mu_0$), we reject the null hypothesis if the test statistic (e.g. sample average) differs too much from the hypothesized value in either direction.

- This is most common, especially when we are testing for "has zero effect".

In a one-sided case ($H_1 : \mu > \mu_0$ or $H_1 : \mu < \mu_0$), we reject the null hypothesis if the test statistic is far above (below) the hypothesized value.

- On a right-tailed test where $H_1 : \mu > \mu_0$, a test statistic that is very *low* does not reject the null hypothesis.

Simply comparing a realization of a test statistic t to a hypothesized value, $H_0 : \mu_0 = 5$ (for example) doesn't tell us everything we need to know.

- What if $t = 5.1$?
- What if $t = 1000$?

We need to know how varied T is in the first place.

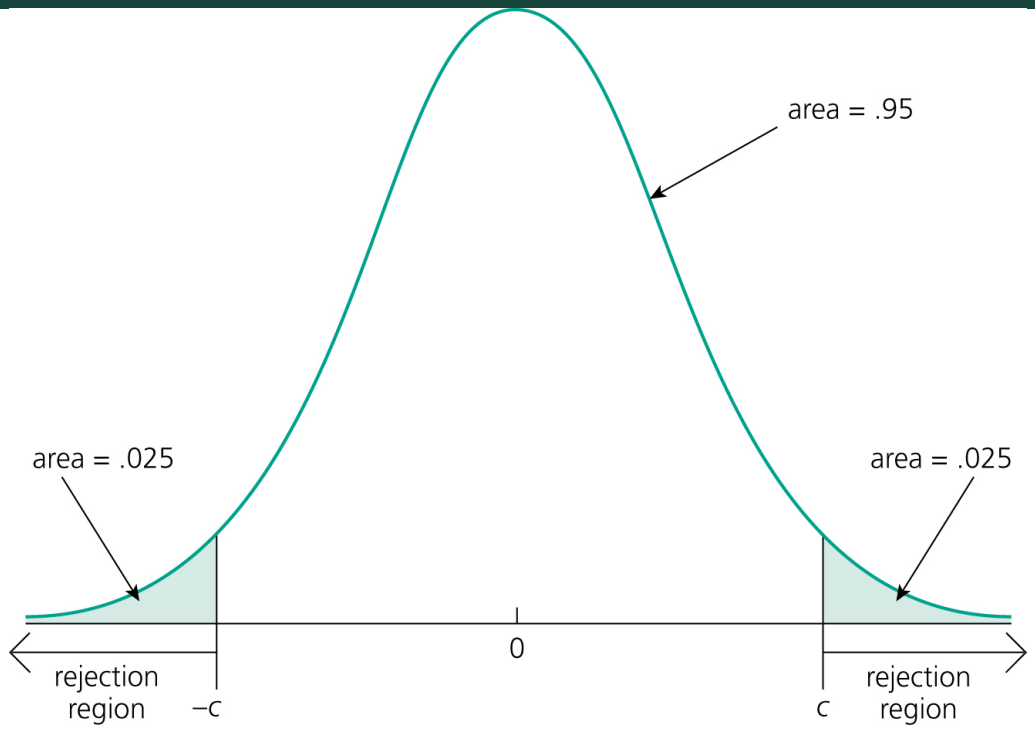
- If it's highly dispersed (high variance), then a realization of $t = 1000$ might be perfectly reasonable under the null hypothesis that $\mu_0 = 5$.

Which is why we want to standardize the test statistic!

$$T = \frac{\bar{Y} - \mu_0}{\frac{S}{\sqrt{N}}} \sim t_{N-1}$$

Which means we do the same with the realization: $t = \frac{\bar{y} - \mu_0}{se(\bar{y})}$, where $se(\bar{y}) = \frac{S}{\sqrt{N}}$

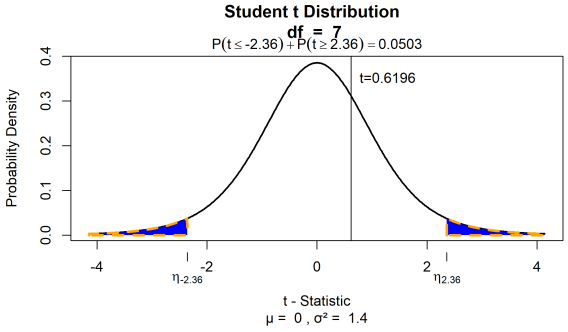
- The size of the rejection region will depend upon how confident we want to be regarding our rejection of the null
- ...or to put it another way, how small we want the probability of a Type I error (significance level) to be
- If we want a significance level to be $100 \times \alpha$, then the critical level is $c_{\frac{\alpha}{2}}$
- This splits the rejection region evenly between $\bar{y}$ being too big and $\bar{y}$ being too small



Wooldridge Figure C-6

Let's go back to our example "Years to completion of a Ph.D."...

Statistic	Value	Note
N	8	Number of observations
$\bar{x}$	5.5	Sample mean
s^2	5.214	Sample variance
$\frac{s}{\sqrt{N}}$	0.807	Std. error of the mean



Calculate t from the sample:

$$t = \frac{\bar{y} - \mu_0}{se(\bar{y})} = \frac{5.5 - 5}{0.807} = 0.6196$$

Now, does this fall into the rejection region? Is it such an extreme value that it provides evidence against $H_0 : \mu_0 = 5$?

First, we have to define the rejection region:

- For $\alpha = .95$, $\pm c_{\frac{\alpha}{2}} = \pm 2.36$

P-values

- An alternative approach computes the corresponding *p-value* for a test statistic
  - The p-value is the probability of obtaining a result equal to or 'more extreme" than what was actually observed **when the null hypothesis is true**
  - "Under the null"

$$\text{p-value} = P(|T| > |t||H_0)$$

If that probability is small, would it provide evidence *against* the null, or *for* the null?

In our example:

Some useful examples from Wooldridge

- Example C.4 - p.696: Effect of Enterprise Zones on Business Investment
- Example C.5 - p. 698: Race Discriination in Hiring
- Example C.8 - p. 700: Effect of Job Training Grants on Worker Productivity
- Example C.9 - p. 702: Effect of Freeway Width on Commute Time

We will cover:

- Linear functions
- Non-linear functions
  - Polynomials (e.g. $x^3 + 2x^2 + 10x$)
  - Natural log
  - Exponential
- Non-linear functions in equations

This will be geared towards the use of these functions in a regression of the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \epsilon$$

Linear functions

A linear function is one where the result is an affine transformation ($ax + b$) of the inputs.

$$y = \beta_0 + \beta_1 x$$

Where β_0 is the *intercept* and β_1 is the slope (which forms a straight line).

Linear functions tell us something about *changes*. Specifically, the relationship between Δx and Δy, where Δ is the change ($\Delta x = x^1 - x^0$)

- Here, x^1 is the "after change x" and x^0 is the "before change"
  - They are not exponents!

$$\begin{aligned} y^1 &= \beta_0 + \beta_1 x^1 \\ y^0 &= \beta_0 + \beta_1 x^0 \\ y^1 - y^0 &= \beta_1 (x^1 - x^0) \\ \Delta y &= \beta_1 \Delta x \end{aligned}$$

The previous result can be written as:

$$\frac{\Delta y}{\Delta x} = \beta_1$$

Which is read as "the change in y resulting from a change in x"

Note that this is the same as taking the derivative:

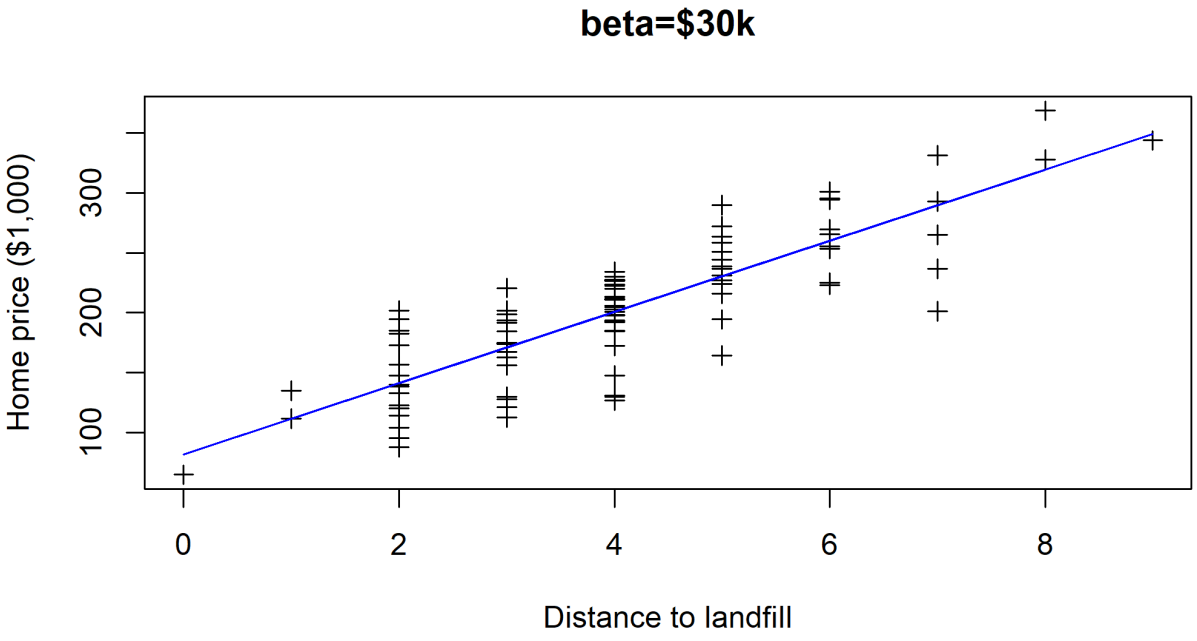
$$\frac{\partial y}{\partial x} = \beta_1$$

It is the slope of a line (rise over run)

In a linear equation relating x to y, it is the **marginal effect** of x.

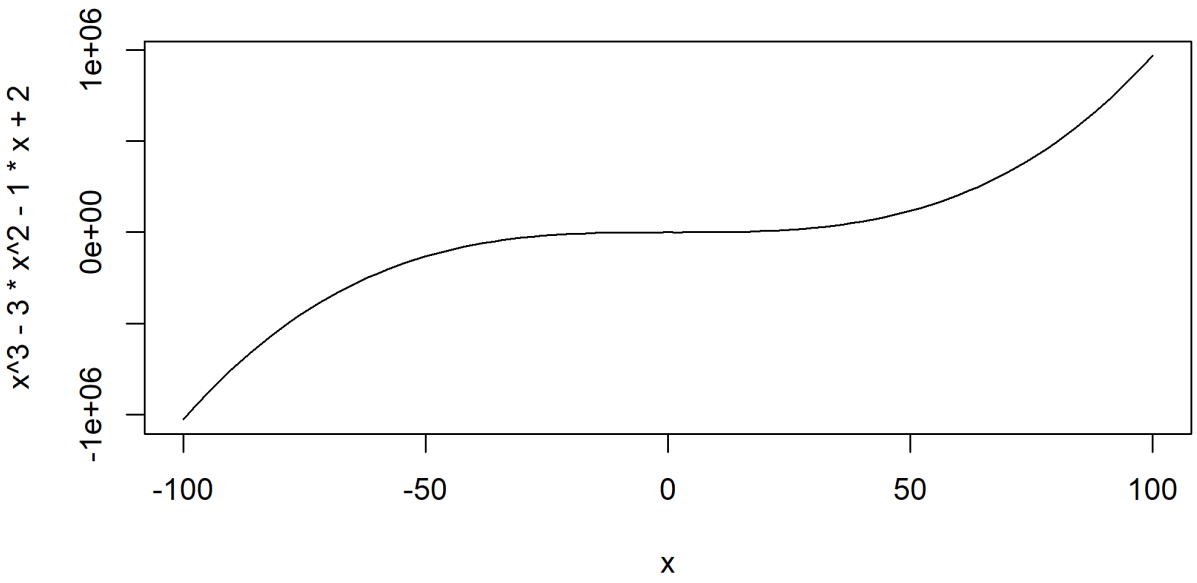
$$\text{home price} = \beta_0 + \beta_1 \times \text{distance to landfill} + \epsilon$$

$$\text{home price} = 82 + 30 \times \text{distance to landfill} + \epsilon$$



Non-linear functions: Polynomials

Polynomials take the form $ax^3 + bx^2 + cx + d$



In a linear function, the change in x (or derivative w.r.t x) is expressed only in terms of β.

In a non-linear function, the change in x depends on the value of x.

- x remains in the dervative

$$\frac{d}{dx}(ax^3 + bx^2 + cx + d) = 3ax^2 + 2bx + c$$

- The slope of the line clearly changes depending on the value of x
- The *marginal effect* is not constant across x

So a regression equation may take the form of:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$$

We can still refer to this as a linear model as it is linear in the parameters!

$$\frac{dy}{dx} \approx \beta_1 + 2\beta_2 x \quad \text{When dx is small}$$

The natural log, $ln(x)$

- Increasing in x
- Domain is $x > 0$
- Makes really big numbers small
- $ln(xy) = ln(x) + ln(y)$
- Range is $(-\infty, +\infty)$
- In economics, "log" almost always means "natural log"

The natural log, $ln(x)$

The natural log is particularly useful because of the following:

$$ln(1+x) \approx x \quad \text{when} \quad x \approx 0$$

$$\Delta ln(x) = ln(x^1) - ln(x^0) = ln\left(\frac{x^1}{x^0}\right) = ln\left(\frac{x^0 + \Delta x}{x^0}\right) =$$

$$ln\left(1 + \frac{\Delta x}{x^0}\right) \approx \frac{\Delta x}{x^0}$$

- This is the percent change in x: $\frac{\Delta x}{x}$
- $100 \times \Delta(ln(x)) = 100 \times [ln(x^1) - ln(x^0)] \approx \%\Delta x$

The natural log, $ln(x)$

If we have the relationship:

$$ln(y) = \beta_0 + \beta_1 ln(x)$$

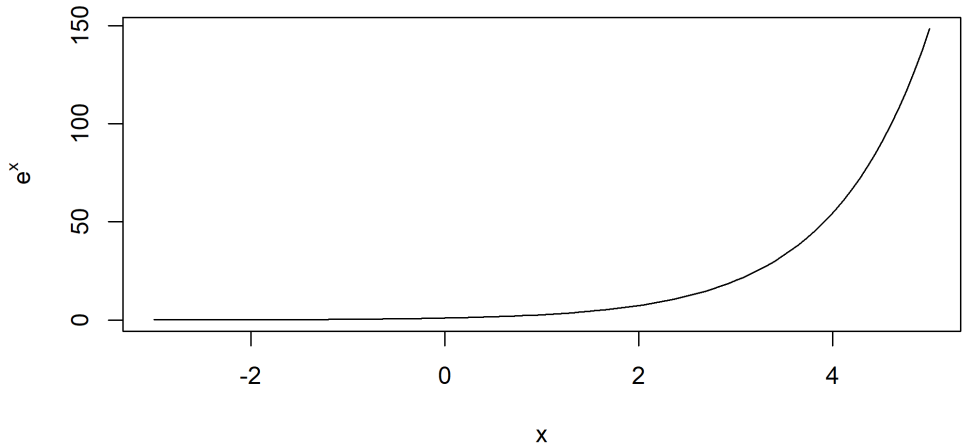
Then $\beta_1 = \frac{\Delta ln(y)}{\Delta ln(x)} = \frac{\%\Delta y}{\%\Delta x}$

And of course, the "percent change in y per 1 percent change in x" has a special name in economics:

It is the elasticity. $\epsilon = \beta_1$.

Exponential function: e^x

- Increasing in x
- $\frac{d}{dx} e^x = e^x$
- Domain is $(-\infty, +\infty)$
- Range is $(0, \infty)$
- $ln(e^x) = x$
- $e^x \times e^y = e^{x+y}$
- $\frac{e^x}{e^y} = e^{x-y}$
- Explodes quickly with large x



Whew!

We made it!

Now is the time to ask questions. Now is the time to say where you got lost.

Now is the time to visit office hours. This week office hours are by appointment (but TA's office hours are drop-in). Some, maybe even a majority, can be perplexing - this is normal and expected.

Come and talk to me during office hours and we can work through them until you're comfortable.