

Stats Review

EC420 MSU Fall 2019

Justin Kirkpatrick

Last updated September 04, 2019

This lecture



Goal: Refresh or catch up on statistical concepts relevant to econometrics.

1. Your questions from last week
2. Random variables
3. Probability density functions (PDF/PMF, CDF)
4. Sample v. Population
5. Mean, variance (in sample, in population)
6. Correlation, independence, mean independence, iid
7. Types of distributions (normal, uniform, student-t, χ^2 , bernoulli, poisson)
8. Statistical inference
9. Useful functions and properties

Questions from last week

3 / 90

Questions from last week

- How much time do you anticipate spending on problem sets?
 - I would anticipate each problem set will take about 4-6 hours. Quicker if you're comfortable with coding.
- Is there a curve for exams?
 - Yes.
- Stata is useful in the economics field. Is it useful for marketing-related research (Stata vs. R vs. Python)?
 - Stata is quite commonly used in marketing, in my experience. It has a strong set of "canned" routines for estimation of most commonly used models. Matlab is usually used when there are no Stata-coded estimators.

4 / 90

- Are the exams multiple choice?
 - There will be a small number of multiple choice questions, but around 80% of the credit will not be multiple choice.
- If the class count decreases, will this class be cancelled?
 - I've been assured that it will not be cancelled - the University is working towards offering more of a variety in class times as part of a finish-in-four effort. I can't promise anything, but it is unlikely to be an issue. We have 11 enrolled in the class right now.

5 / 90

Questions from D2L

I answered most of the questions on D2L already, but those posted last night I haven't answered yet.

- They will be covered in this lecture (today or on Monday).

Other notes

I'm not numbering lectures as I post them, but I will use consistent titles (using those in the syllabus).

I'll get a gradebook set up on D2L soon

6 / 90

Stats Review!

7 / 90

Random Variables

8 / 90

Wooldridge defines a random variable as

"...one that takes on numerical values and has an outcome that is determined by an experiment."

When referring to a random variable (RV), we use an upper-case e.g. X

RV's have *realizations* ("...determined by an experiment")

- Like flipping a coin
- Or rolling a die

We label these using lower-case: x

And when we have multiple realizations of a RV, we can label them:

$$\{x_1, x_2, x_3\} = \{-1.74, 0.27, 1.01\}$$

9 / 90

And when we have a map of how this RV behaves, we have a **distribution** e.g.

$$X \sim N(0, 1)$$

- One important element of a distribution is the *support*, which is the possible values that the RV can take.

10 / 90

Discrete vs. Continuous

A random variable is discrete if it can take on only a finite or countably infinite number of values.

- Bernoulli (coin flip) can take on only two values. It is discrete.
- Poisson (count) can take on the values $\{0, 1, 2, \dots\}$, which is countably infinite. It is discrete.
- Normal can take on any value $\in \{-\infty, +\infty\}^*$

A fun fact about continuous variables

The probability of a continuous RV taking any one value is 0.

- $\Pr(x = 0.12482) = 0$

* \in = "in"

Probability density function

The *probability density function* (pdf) of X summarizes the information concerning the possible outcomes of X and the corresponding probabilities

$$f_X(x) = \begin{cases} p_j & \text{if } x = x_j \\ 0 & \text{otherwise} \end{cases}$$

The *pdf* tells us the probability of realizing any value within a very small window around the value. If we know the distribution, including the parameters, then we can plug in any value.

In a Normal $X \sim N(\mu, \sigma^2)$, $f_X(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$

A *pdf* must **always sum (discrete) or integrate (continuous) to one**

The *cumulative density function*, or *cdf*:

The *cdf* tells us the probability of a RV realization being less than some value:

$$\Pr(X \leq t) = F_X(t) = \int_{-\infty}^t f_X(s) ds$$

The *cdf* has a few useful properties:

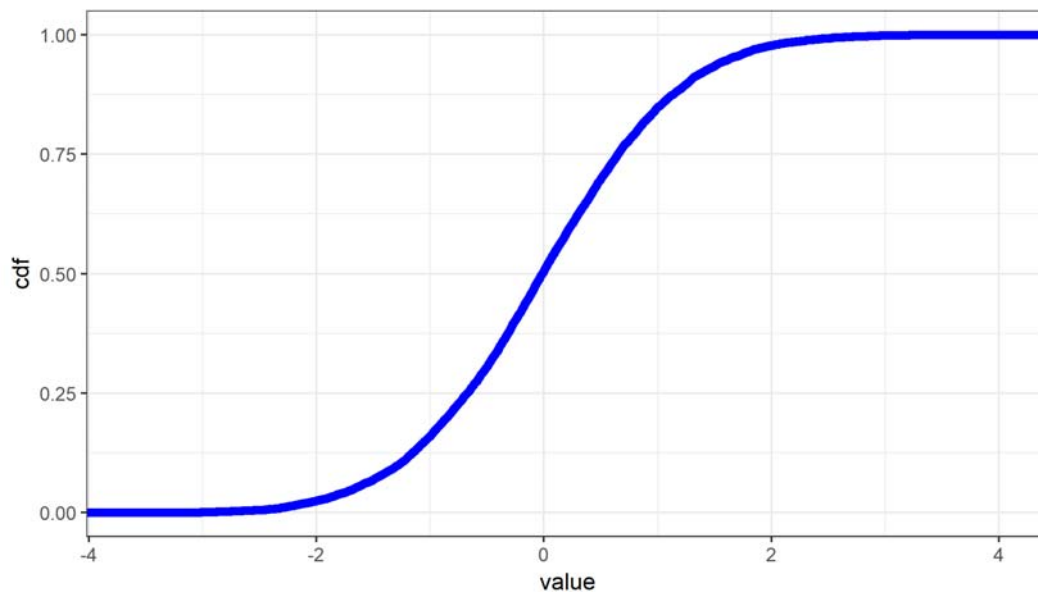
The probability of a RV X being **greater than** some value t is $1 - F_X(t)$

The derivative of the *cdf* at any value t is the *pdf* at t

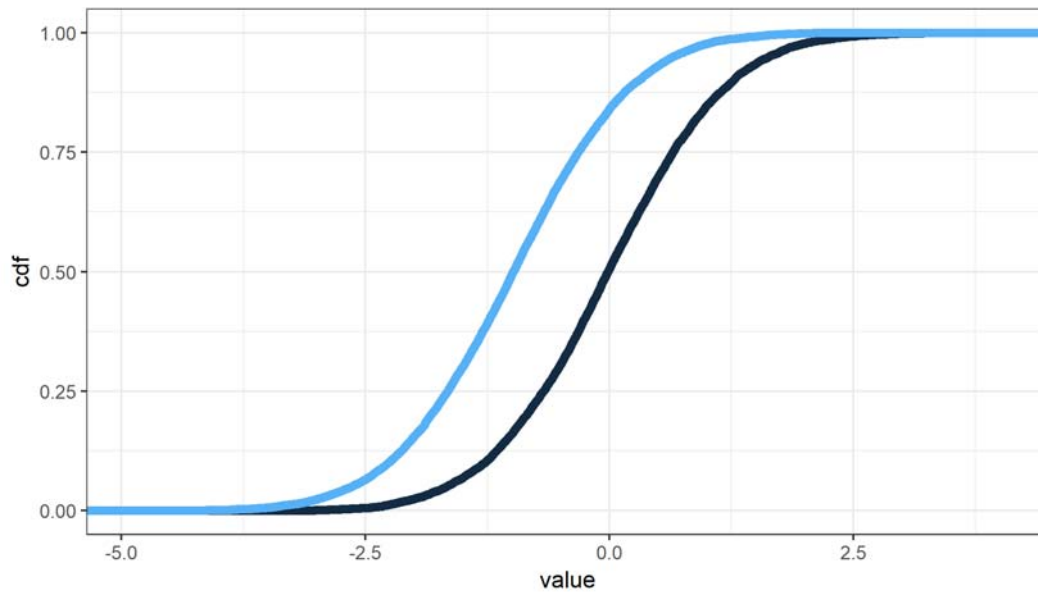
The *cdf* is always $\in [0, 1]$

13 / 90

An example



14 / 90



15 / 90

Sample v. Population

Population

The "population" is what we'd like to learn about.

- There is some feature of the population we are interested in
- But we cannot collect complete data on the population:
 - It may be too expensive
 - It just isn't possible
- In stats class, some examples might be:
 - Income
 - Age
 - Vote

16 / 90

Population

- In econometrics, we're often going to be concerned about a population *effect*:
 - Effect of a drug on a health outcome
 - Effect of a work training program

17 / 90

Sample

The sample is what we observe.

It is always taken from the population. We use N to denote it's size.

And under certain conditions* the sample tells us about the population

Population statistics will **always get a greek letter**. Sample statistics don't.

- | | |
|---------------------------------------|--------------------------------|
| • μ is the population mean | • \bar{x} is the sample mean |
| • σ is the population variance | • s^2 is the sample variance |

* We will learn a lot about these conditions this semester.

18 / 90

Stratified sampling

A **random sample** is a set of size N drawn randomly from the population.

Remember, a random sample may not be all that random if some people are more likely to be selected than others.

- For example, randomly dialing phone numbers. Does *everyone* have a phone? Are you more likely to get a person who owns a cellphone?

19 / 90

Stratified sampling

A **stratified sample** is a set of random samples drawn from conditions on the population.

- If we know that random phone number dialing over-samples people with cellphones, we could randomly draw from *people we know only have landlines* and then, separately, draw from *people we know have cellphones*.
- This requires knowing who has landline and who has cell-only.
- How many from each? Need to know approximate population proportions. In the US, 41% have landline and cell, 5.4% have landline without cell, 50.5% are cell-only, and 3% are phoneless (Blumberg and Luke, 2017)

20 / 90

Sample mean

It should surprise nobody in this room that the *sample mean* is calculated as:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{N} = \frac{1}{N} \sum_{n=1}^N x_n$$

Note that N is the sample size, n is the index, and x_n are the realizations of X

Note that we have to take *one more step* to say that $\bar{x} = \mu_X$ (the population mean of X). While \bar{x} would be the best guess (loss function minimizer), it is not necessarily μ .

\bar{x} is a *descriptive statistic* (it describes the sample x_1, \dots, x_n). We will also see that it is an *estimate* of the population statistics μ . Let's hold off on this for now.

21 / 90

Property of the mean (and summation operator)

Remember that multiplying the sample by a constant means you can take the constant out:

$$\frac{1}{N} \sum_{n=1}^N c x_n = c \bar{x}$$

But you cannot do this with the product of two RVs

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

You can move the addition or subtraction of a constant in or out:

$$\frac{1}{N} \sum_{n=1}^N (x_n - c) = \bar{x} - \frac{1}{N} N c = \bar{x} - c$$

22 / 90

The *sample average*, \bar{x} , can be treated like a constant since it is the realization from a sample.

This is important as we will be working with deviations from the mean a lot:

$$\frac{1}{N} \sum_{n=1}^N x_n - \bar{x}$$

ClassID	Score
1	4
2	5
3	3
4	5
5	5
6	2
Avg	4

One particularly important property is revealed here: the sum of all deviations from the *sample mean* is zero.

$$(4 - 4) + (5 - 4) + (3 - 4) + (5 - 4) + (5 - 4) + (2 - 4) = 0$$

$$\frac{1}{N} \sum_{n=1}^N x_n - \bar{x} = \bar{x} - \bar{x} = 0$$

23 / 90

Expected value

The expected value is a very similar concept to the mean. **In most large- N cases, it is conceptually the same.**

A sample mean exists for any set of realizations of a RV. The RV itself has an expected value.

The *expected value* of a RV is the "measure of central tendency"

- A fancy way of saying "tells us about the middle"
- Sometimes called the "first moment"

It is represented by the E operator and is read as "the expectation of X ". In a discrete RV, the Expectation is defined by:

$$E(X) = x_1 f(x_1) + x_2 f(x_2) + \dots + x_n f(x_n)$$

Remember, f is the pdf

24 / 90

For continuous RVs:

$$E(X) = \int_{-\infty}^{+\infty} x_s f(x_s) ds$$

This is the integral over all possible values of x , weighted by the pdf at that value.

Note that we don't need a sample to calculate this - we just need to know the support (the values X can take) and the pdf, $f_X(x)$.

25 / 90

Expectation operator

$E(X)$ has some useful properties (similar to the mean), where c is a constant (not a RV):

- $E(cX) = cE(X)$
- $E(cX + b) = b + cE(X)$

That is, you can add and multiply constants in and out of the expectation

26 / 90

Expectation operator

But in general:

- $E(XY) \neq E(X)E(Y)$ in general
 - This **will** be true in certain cases, which will be **very** useful to us!

In an upcoming lecture, we will think about conditional expectations:

- $E(X|Y = y)$ which is the expectation of X *conditional on* Y taking the value y

We will work a lot more with the expected value in this class

27 / 90

Expected value continued...

The *expected value* will also be the best guess of a realization of a RV

- "best" as in "minimizes the sum of differences between realizations and the guess"
- If I have to pay \$1 for every unit off I am from a realization of a RV, I am best-off by guessing the mean
- This applies for any monotonically increasing loss function
 - Loss function is any way of calculating a penalty
 - For example, it could be the square of the difference
 - Or, the absolute value of the difference
 - "Increasing" means its result increases with the input
 - This excludes *The Price Is Right* pricing

28 / 90

Variance

The *variance* of a RV is a "measure of diffusion"

- A fancy way of saying "how spread out are the likely values"
- Sometimes called the "second moment"
- Not all RVs have a finite variance; some are ∞
 - All samples do - you can always calculate the variance from a finite number of draws
 - ...but not all RVs have a finite variance

Most important, two RVs or two samples of two RVs can have the same mean but different variance

- The mean is still the "best guess"
- The variance tells us how far off your "best guess" will be on average
- Higher variance = more spread = further off

29 / 90

Mean, variance

Variance, as a general operation, is calculated by summing the squared differences between the realizations of \mathbf{X} and the mean of \mathbf{X} .

It is written as $Var(\mathbf{X})$ or σ_X^2

If we know the population mean:

$$Var(\mathbf{X}) = \sigma_X^2 = \frac{1}{N} \sum_{n=1}^N (x_i - \mu)^2$$

But we rarely know μ . But we do have a good guess: \bar{x} !

Sample variance

So, we can get a sample variance by first calculating \bar{x} , taking it as a constant, and calculating:

$$Var(\mathbf{X}) = \frac{1}{N-1} \sum_{n=1}^N (x_i - \bar{x})^2$$

30 / 90

We can write the variance using expectations as well:

(Note that I'm using the population variance here)

$$\begin{aligned}
 \frac{1}{N} \sum_{n=1}^N (x_i - \mu)^2 &= \frac{1}{N} \sum_{n=1}^N [x_i^2 - 2x_i\mu + \mu^2] \text{ (Expand polynomial)} \\
 &= \frac{1}{N} \sum_{n=1}^N x_i^2 - 2\mu \frac{1}{N} \sum_{n=1}^N x_i + \frac{1}{N} N\mu^2 \text{ (Distribute the sum)} \\
 &= E(X^2) - 2\mu E(X) + \mu^2 \\
 &= E(X^2) - 2E(X)E(X) + [E(X)]^2 \\
 &= E(X^2) - [E(X)]^2
 \end{aligned}$$

31 / 90

Operations on Variance where a and b are constants:

$$Var(aX + b) = a^2 Var(X) + 0$$

- Constants have zero variance.
- Scaling X by a scales variance by a^2 .

If you recall, we can standardize X by subtracting the mean and dividing by the standard deviation:

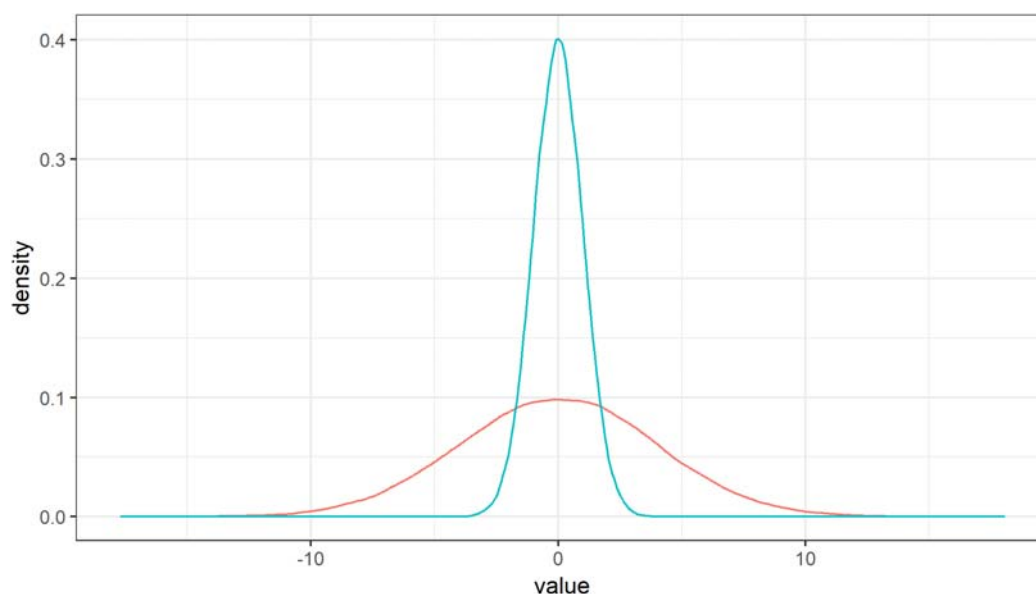
$$Z = \frac{X - \mu}{\sigma}$$

Write a as $\frac{1}{\sigma}$ and b as $-\frac{\mu}{\sigma}$.

$$\begin{aligned}
 Z &= aX + b, \\
 E(Z) &= aE(X) + b = \frac{\mu}{\sigma} - \left(\frac{\mu}{\sigma}\right) = 0, \\
 Var(Z) &= a^2 Var(X) + 0 = \frac{1}{\sigma^2} \sigma^2 = 1
 \end{aligned}$$

32 / 90

Low variance is $\sigma^2 = 1$; high variance is $\sigma^2 = 16$. Both means are zero.



33 / 90

Covariance, correlation

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

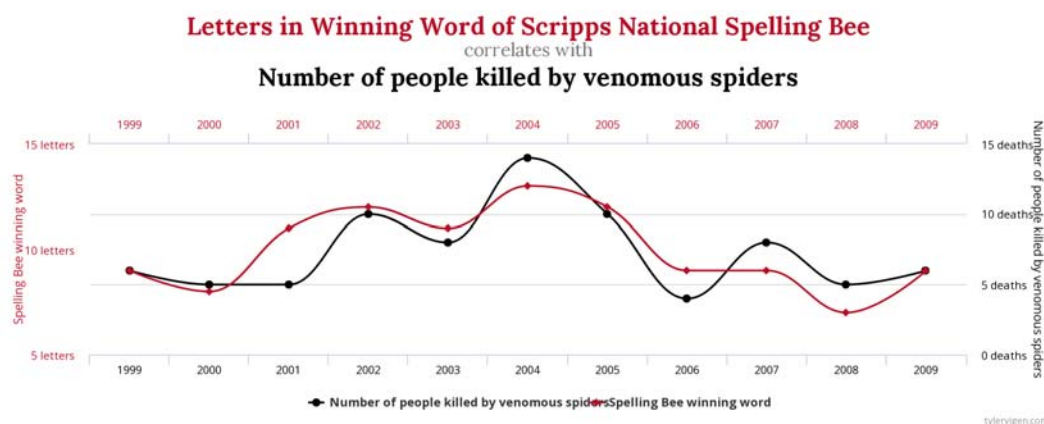
- Income and age
- Wage and education
- Corn yields and fertilizer
- Snowfall and traffic fatalities
- "Letters in Winning Word in Scripps National Spelling Bee" and "Deaths from Venemous Spiders"

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 between RV's: **covariance**.

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

34 / 90



35 / 90

Covariance, correlation

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*.

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.

36 / 90

With some algebra, we can also write covariance as:

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

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 measure of association.

37 / 90

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

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

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 ρ : $\text{Cov}(X, Y)$, $\text{Var}(X)$, $\text{Var}(Y)$, in what case would:

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

38 / 90

For constants a and b and RVs X and Y :

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X, Y)$$

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.

- 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

39 / 90

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

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

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

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

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

40 / 90

This brings 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.

41 / 90

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

Think about flipping two "fair" coins.

- If the first one lands on "heads", what do you think the second one will land on?
- The second coin would have the same probability of landing on heads, regardless of how the first coin came up
- If we were to write the pdf of the second coin, *it would not have the realization of the first in it.*
 - $\Pr(Y = \text{"heads"}) = .5$
 - See - no "X" in there at all!
 - $f_{X|Y}(x|y) = f_X(x)$

42 / 90

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:

$$X \perp Y \Rightarrow Cov(X, Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0$$

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

43 / 90

When X and Y are *not* independent ($X \not\perp Y$), 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)
- $f_{X|Y}$ changes when we learn about a realization's smoking status.
- $f_{X|Y}(x|y = 1)$ will have higher probabilities on lower values of x

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

44 / 90

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 \Longleftrightarrow).

- 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!

45 / 90

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.

46 / 90

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, $\theta = .5$

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 .

47 / 90

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 $Var(X) = \sigma^2$

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

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

48 / 90

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 standard deviation of 1.

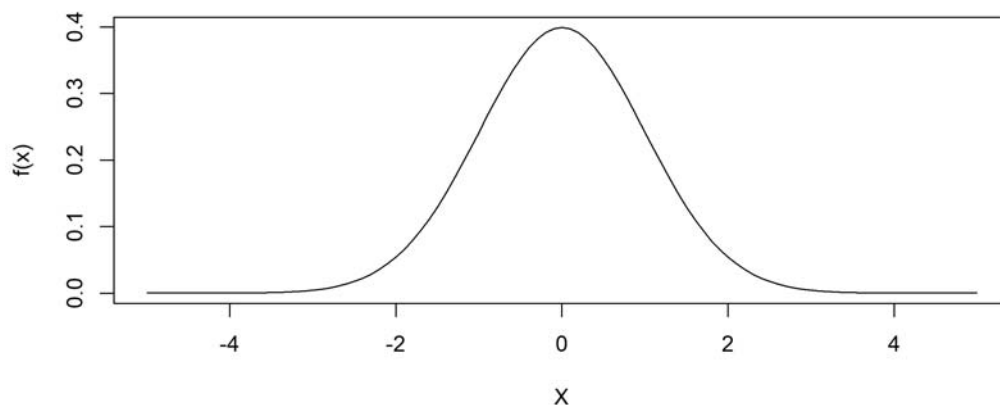
The pdf of the standard normal is $\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right]$

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 $\Phi(z)$.

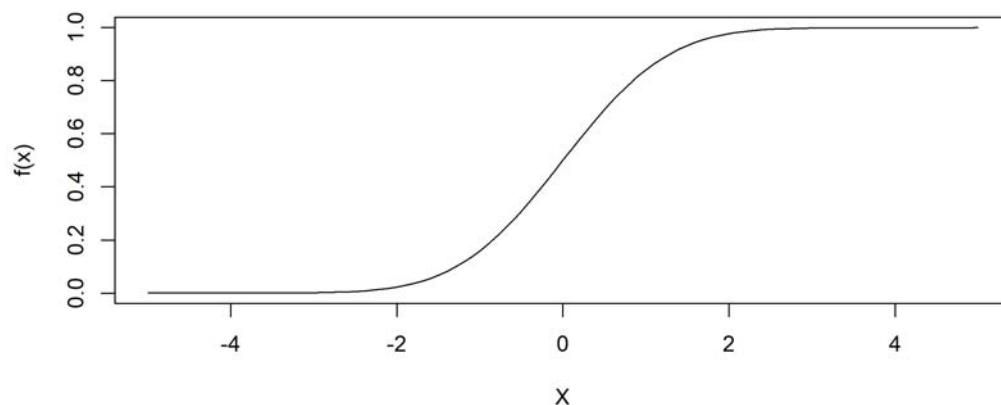
49 / 90

Standard normal pdf:



50 / 90

Standard normal cdf:



51 / 90

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 independent, identically distributed (iid) normal random variables has a normal distribution

52 / 90

The Chi-Square Distribution, χ^2

If $Z_i, i=1, \dots, n$ be iid standard normal variables. Then the sum of these variables 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.

53 / 90

The F-distribution:

If $X_1 \sim \chi_{k_1}^2$ and $X_2 \sim \chi_{k_2}^2$, 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.

54 / 90

Statistical inference

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

55 / 90

Statistical inference

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, \dots, X_n are independent random variables all drawn from a common pdf, $f(x; \theta)$, then X_1, X_2, \dots, 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, \dots, 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

56 / 90

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 .

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

57 / 90

But \mathbf{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.

58 / 90

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 $W = h(Y_1, \dots, Y_n)$, 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 earlier can be shown to be an unbiased estimate of $Var(Y) = \sigma^2$

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

59 / 90

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 $Y \sim N(\mu, \sigma_Y^2)$:

$$\begin{aligned} Var(\bar{Y}) &= Var\left(\frac{1}{N} \sum y_i\right) \\ &= \frac{1}{N^2} Var\left(\sum y_i\right) \\ &= \frac{1}{N^2} N \sigma_Y^2 \\ &= \frac{\sigma_Y^2}{N} \end{aligned}$$

So if $Y \sim N(\mu, \sigma_Y^2)$, then $\bar{Y} \sim N(\mu, \frac{\sigma_Y^2}{N})$
Note that we are assuming we know the variance

60 / 90

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

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

Which implies that:

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

± 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

61 / 90

Confidence intervals (CI)

- We know the sampling distribution of \bar{Y} .
- We know the 95% CI
- What do we do with it?

Question:

- If I draw a new sample from Y , will the 95% CI 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.

62 / 90

Wooldridge Table C-2

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

© Cengage Learning, 2013

63 / 90

Statistical inference



One problem:

We don't know σ^2 . If we did, we'd be in great shape.

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

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

64 / 90

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:

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

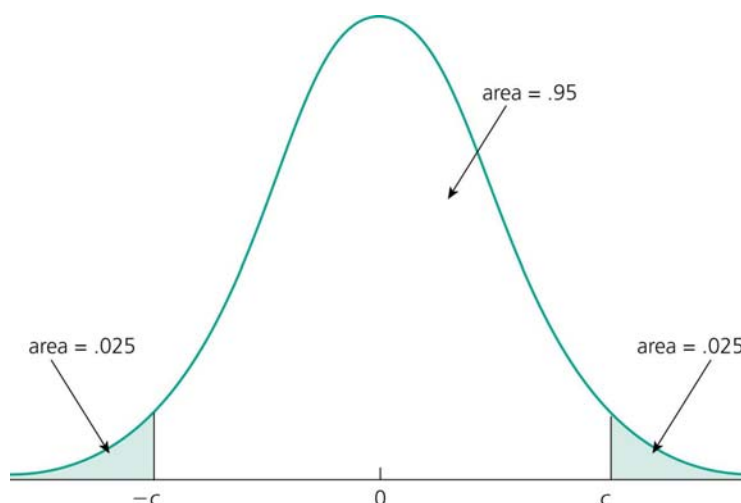
It is distributed t with $N - 1$ degrees of freedom.

- 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.

65 / 90

Critical values

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



Wooldridge Figure C-4

66 / 90

TABLE G.2 Critical Values of the t Distribution

	Significance Level				
	.10	.05	.025	.01	.005
1-Tailed:	.20	.10	.05	.02	.01
2-Tailed:	.40	.20	.10	.04	.02
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

Example: 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 $C \geq 120$ df is 1.96.
Source: This table was generated using the Stata® function invttail.

Wooldridge Figure G-2

67 / 90

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

$$N = 8$$

$$\bar{x} = 5.5$$

$$s^2 = 5.214$$

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

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

68 / 90

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

$$\pm 2.306$$

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

$$5.5 \pm 2.306 \times 0.807$$

Which is **[3.638, 7.362]**

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

69 / 90

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

Often, we are interested in these *population* questions

- Does the average person obtain more education when college is subsidized?
- This is a question with a **null hypothesis**: "is the increase in education resulting from subsidized college greater than zero?"
- 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.

70 / 90

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, $\alpha = P(\text{Reject } H_0 \mid H_0)$
- Type II error: Failing to reject the null hypothesis when it is false

71 / 90

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".

72 / 90

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.

73 / 90

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 $\mu_0 = 5$. So, we want to standardize the test statistic!

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

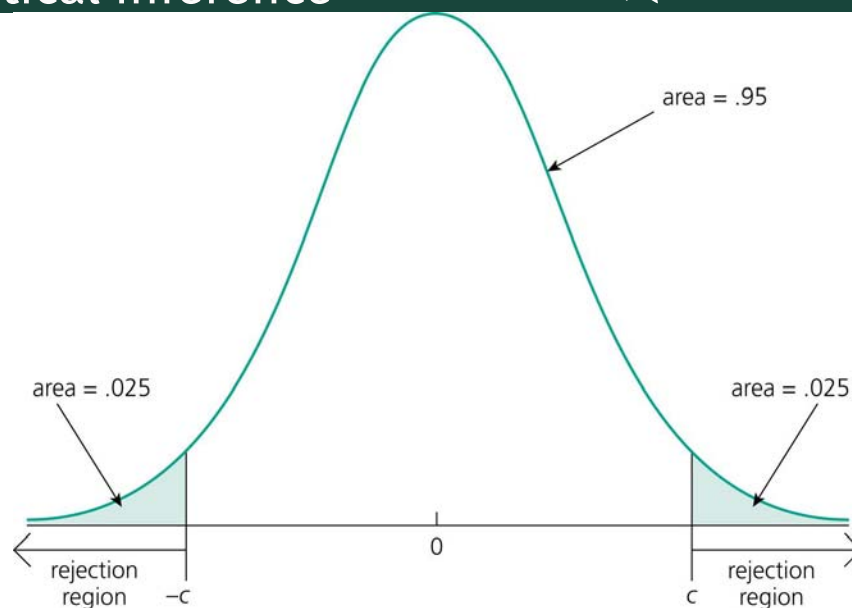
Which means we do the same with the realization: $t = \frac{\bar{y} - \mu_0}{se(\bar{y})}$, where $se(\bar{y}) = \frac{S}{\sqrt{N}}$

This lets us compare the now-standardized t to critical values of a distribution we know, t_{N-1}

74 / 90

- 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

75 / 90



Wooldridge Figure C-6

76 / 90

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

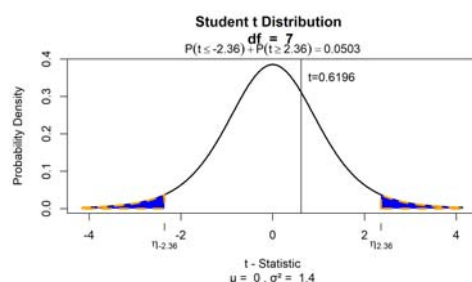
Calculate t from the sample:

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

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



77 / 90

Statistical inference

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**:

78 / 90

Some useful examples from Wooldridge

- Example C.4 - p.696: Effect of Enterprise Zones on Business Investment
- Example C.5 - p. 698: Race Discrimination 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

79 / 90

Useful functions and properties MICHIGAN STATE UNIVERSITY

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:

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

80 / 90

Linear functions

A linear function is one where the result is an affine transformation ($\beta_0 + \beta_1 x$) 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!

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

81 / 90

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:

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

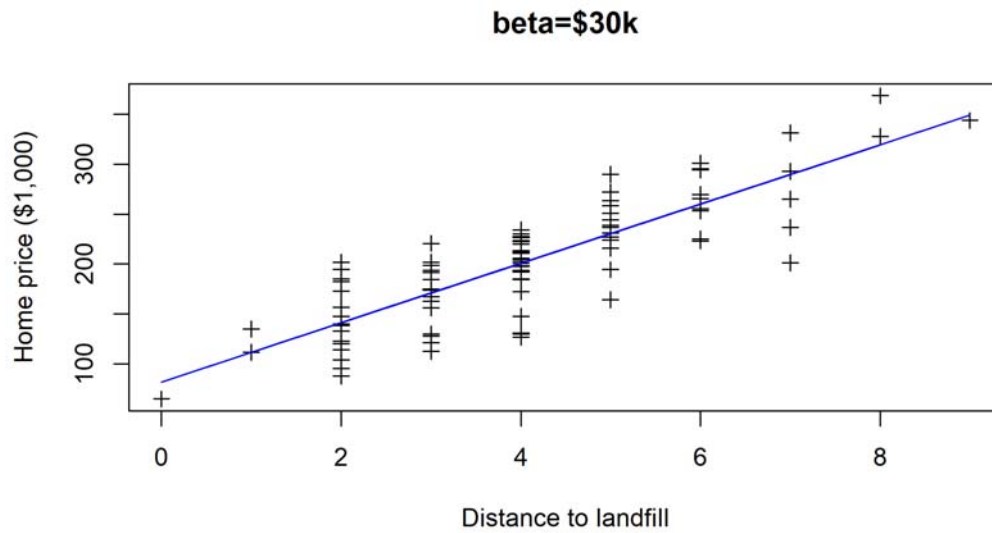
Let's call this the **marginal effect**.

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

82 / 90

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

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

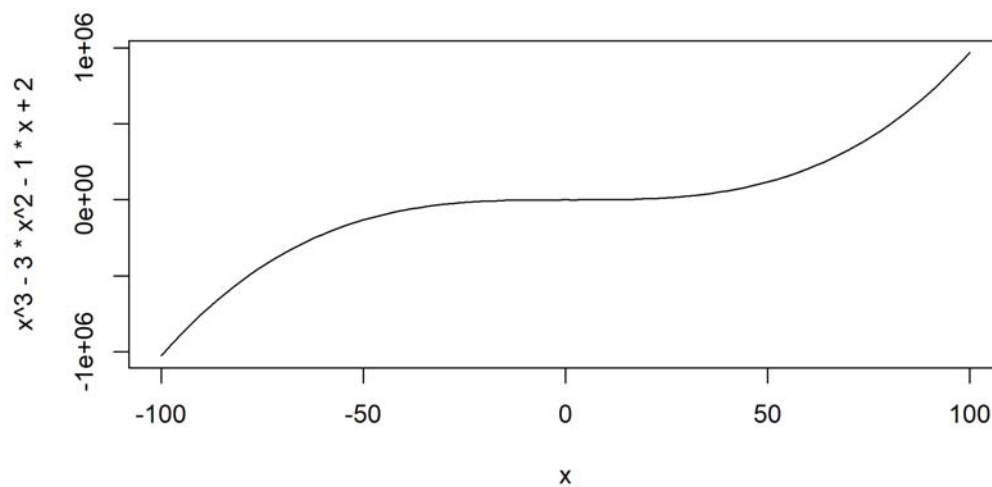


83 / 90

Useful functions and properties MICHIGAN STATE UNIVERSITY

Non-linear functions: Polynomials

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



84 / 90

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 derivative

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

- 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!

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

85 / 90

The natural log, $\ln(x)$

- | | |
|---|---|
| <ul style="list-style-type: none"> • Increasing in x • Domain is $x > 0$ • Makes really big numbers small | <ul style="list-style-type: none"> • $\ln(xy) = \ln(x) + \ln(y)$ • 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:

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

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

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

87 / 90

The natural log, $\ln(x)$

Logging both y and x is called a "log-log" or "double-log" specification

We can also log just the y 's or just the x 's.

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

- $\frac{\% \Delta y}{\Delta x} \approx (100 \times \beta_1)$ gives the percent change in y owing to a 1-unit change in x

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

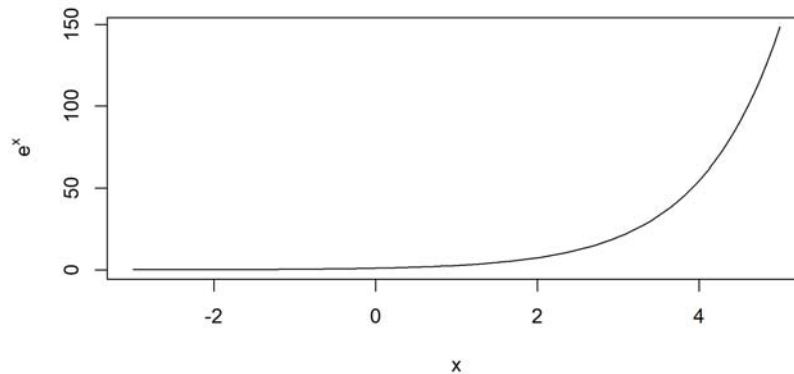
- This yields the unit change in y from a 1% increase in x

Not as useful, but not uncommon.

88 / 90

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



89 / 90

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. 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.

90 / 90