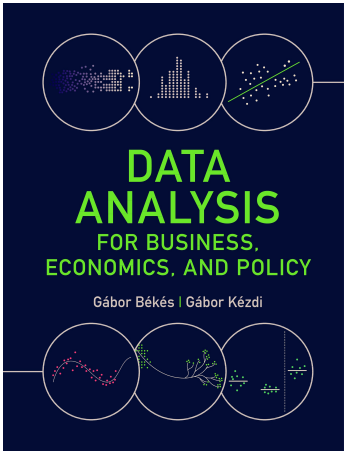
04 Comparison and correlation

Gábor Békés

Data Analysis 1: Exploration

2023

Slideshow for the Békés-Kézdi Data Analysis textbook



- Cambridge University Press, 2021
- gabors-data-analysis.com
  - Download all data and code: gabors-data-analysis.com/data-and-code/
- This slideshow is for Chapter 04

Motivation

Are larger companies better managed? Answering this question may help in benchmarking management practices in a specific company, assessing the value of a company, or estimating the potential benefits of a merger between two companies.

To answer this question you downloaded data from the World Management Survey. How should you use the data to measure firm size and the quality of management? How should you assess whether larger companies are better managed?

The y and the x

- Much of data analysis is built on comparing values of a y variable by values of an x variable, or more x variables.
- Uncover the patterns of association: whether and how observations with particular values of one variable (x) tend have particular values of the other variable (y).
- The role of y is different from the role of x.
  - it's the values of y we are interested in
  - compare observations that are different in their x values.
- It is our decision to pick y

The y and the x

- This asymmetry comes from the goal of our analysis.

- Goal 1: predicting the value of a y variable with the help of other variables - many x variables, such as x_1, x_2, ...
- The prediction itself takes place when we know the values of those other variables but not the y variable.

- Goal 2: learn about the effect of a causal variable x on an outcome variable y.
- What the value of y would be if we could change x

Conditioning, conditional distributions

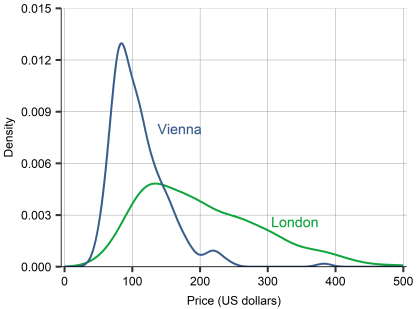
Comparison and conditioning

- Comparison –> conditioning
- We compare y, by values of x –> we condition y on x.
  - x (by the values of which we make comparisons) –> conditioning variable.
  - y –> outcome variable.

- Compare salaries of workers (y) with low and high level of education (x) –>
  - salary is the outcome
  - education is the conditioning variable.

Comparisons and conditional distributions

- The conditional distribution of a variable is the distribution of the outcome variable given the conditioning variable.
- Straightforward concept if the conditioning variable is qualitative (simple if binary)
- Comparing distributions
- Conditional distribution of prices (y, conditional on x - the city=Vienna and city=London)



Conditional statistic

- Conditional mean = mean of a variable for each value of the conditioning variable.
- The conditional expectation of variable y for different values of variable x is

$$E[y|x]$$

- This is a function: for a value of x, the conditional expectation gives number that is the expected value (mean, average) of variable y for observations that have that x value
- It gives different values based on the conditioning variable x

Case Study - Management quality and firm size

- Management quality and firm size: describing patterns of association
- Whether, and to what extent, larger firms are better managed.
- Answering this question can help understand why some firms are better managed than others.
  - Size itself may be a cause ....
  - Whether firm size is an important determinant of better management can inform policy questions such ...
- Data from the World Management Survey to investigate our question.

Case Study - Management quality and firm size

- Interviews by CEO/senior managers, based on that a score is given
- Management quality is measured as management score.
- Each score is an assessment by the survey interviewers of management practices in a particular domain
  - tracking and reviewing performance or
  - time horizon and breadth of targets, etc
- Measured on a scale of 1 (worst practice) to 5 (best practice).
- Normalized - standardized score

Case Study - Management quality and firm size

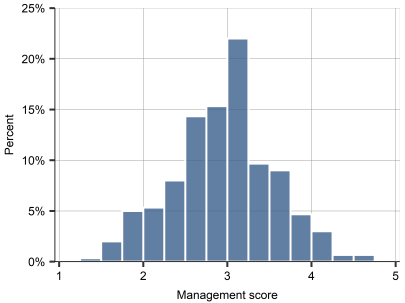
- Take 18 individual measures and average
- Our measure of the quality of management is the simple average of these 18 scores = “the" management score.
- By construction, the range of the management score is between 1 and 5 because.

Case Study - Management quality and firm size

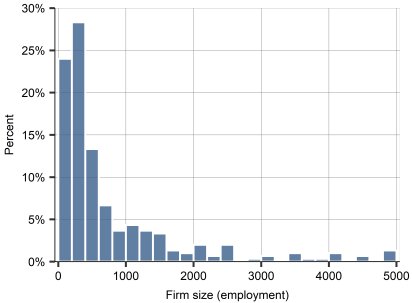
- Data from the World Management Survey to investigate our question.
- Survey introduced in class 01
- In this case study we analyze a cross-section of Mexican firms from the 2013 wave of the survey.
- Only firms with 100 – 5000 employees, N=300
- The y = measure of the quality of management. The x = measure of firm size.
- Firm size = number of employees

Case Study - Management quality and firm size

(a) Management score



(b) Firm size (number of employees)



Note: *Source: Management quality is an average score of 18 variables. Firm size is number of employees.* `wms-management-survey` *data. Mexican sample, n=300.*

Case Study - Management quality and firm size

- Management score: The mean is 2.9, the median is 3, and the standard deviation is 0.7.
- Firm size: The range of employment is 100 to 5000. The mean is 760 and the median is 350, skewness with a long right tail. Some large firms, but not extreme, kept as is.

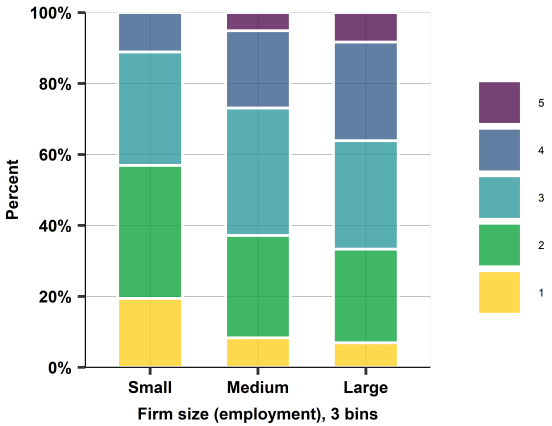
Case Study - Management quality and firm size

- Conditional probabilities
- Three bins of firm size. By number of employees: small (100–199, N=72), medium (200–999, N=156), large (1000, N=72)
- Take a single measure: Lean management score, with values 1,2,3,4,5.
- Thus, for each score variable we have 15 conditional probabilities: the probability of each of the 5 values of y by each of the three values of x – e.g., $P(y = 1|x = small)$.

Case Study - Management quality and firm size

- Lean management score 1–5
- Firm size: small, medium, large
- Conditional probability:
  - share of score=1 conditional on being a small firm is about 20%.
  - share of score=5 conditional on being a large firm is about 10%.
- Shows a pattern of association

Note: *Source: Management quality is an average score of 18 variables. Firm size is number of employees.* `wms-management-survey` *data. Mexican sample, n=300.*



Case Study - Management quality and firm size

- Conditional probabilities - can calculate average
- Three bins of employment: small (100–199, N=72), medium (200–999, N=156), large (1000, N=72)
- Mean management score is 2.68 for small firms, 2.94 for medium sized ones, and it is 3.18 for large.
- First simple evidence: larger firms have better management.

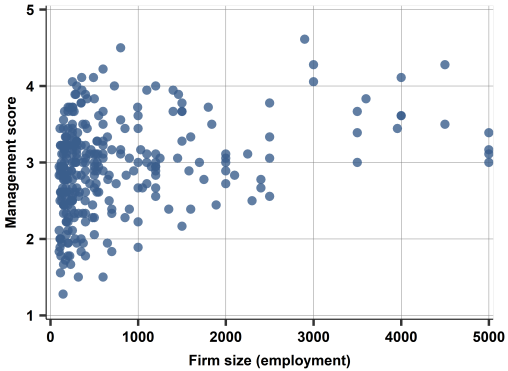
Conditional and joint distributions of two quantitative variables

- Two variables, many values
- The joint distribution of two variables shows the probabilities (frequencies) of each value combination of the two variables.
- A scatterplot is a two-dimensional graph with the values of each of the two variables measured on its two axes, and dots entered for each observation in the dataset with the combination of the values of the two variables.
- Works when dataset relatively small.
- For larger samples, we can bin values, and use "bin scatter"
- Bin scatter shows conditional means for bins we created

Case Study - Management quality and firm size

- Conditional mean and joint distribution
- How our management quality variable (y: the management score) is related to our firm size variable (x: employment)
- Scatterplot
- Bin-scatter

Case Study - Management quality and firm size



- Scatterplot
- Both x and y axis qualitative
- Each dot is an observation
- Full information on distributions

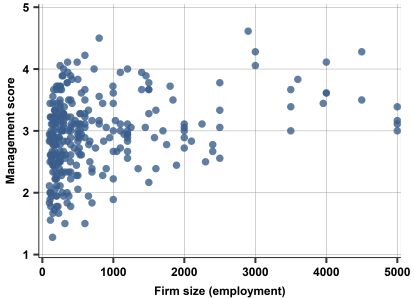
Note: *Source: Management quality is an average score of 18 variables. Firm size is number of employees.* `wms-management-survey` *data. Mexican sample, n=300.*

Case Study - Management quality and firm size

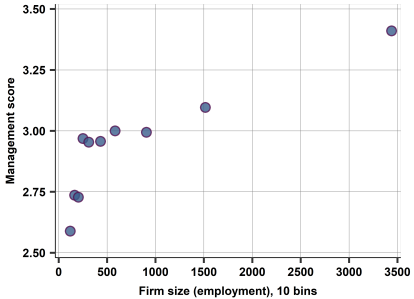
- Bin-scatter mean y conditional on three x bins and ten x bins.
- Bin-scatter: cut the distribution into 10 parts, with equal number of firms
- Show average management score as a point corresponding to the midpoint in the employment bin (e.g., 120 for the 100–120 bin).
- Dots NOT equally spread out - more frequent where more observations!

Case Study - Management quality and firm size

(a) Scatterplot



(b) 10 Bin-scatter



Note: *Source: Management quality is an average score of 18 variables. Firm size is number of employees.* `wms-management-survey` *data. Mexican sample, n=300.*

Case Study - Management quality and firm size

- Some positive association is shown, but not easy to read
- Bin-scatter - positive overall, but most for small vs medium.
- Difference in mean management quality tends to be smaller when comparing bins of larger size, suggesting a positive but nonlinear, concave pattern of association
  - (a positive concave function increases at a decreasing rate)

Boxplot and violinplot

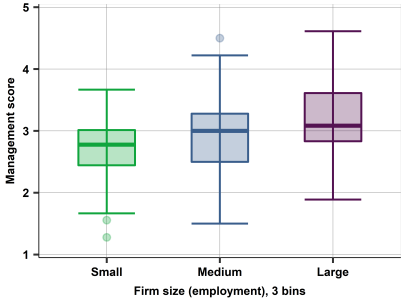
Boxplot: shows means and some features of the distribution.

- Median value of the variable, placed within a box.
- The upper side of the box is the third quartile (the 75^{th} percentile) and the lower side is the first quartile (the 25^{th} percentile).
- additional info: 1.5 times the inter-quartile range added to the third quartile and subtracted from the first quartile.
- may show individual values outside

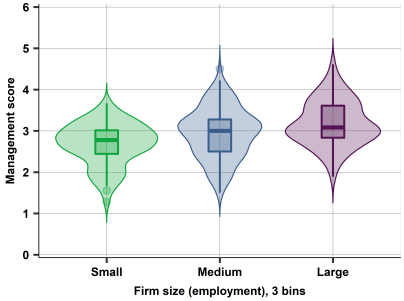
Violin plot - similar but shows density plot (kernel density as seen before)

Case Study - Management quality and firm size

(a) Box plot



(b) Violin plot



Note: *Employee retention rates: The probability of staying with the firm, in the two experimental groups. Mean denoted in boxplot. Source:* `wms-management` *dataset*

Case Study - Management quality and firm size

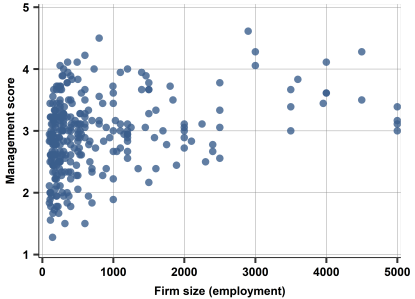
- Can look at differences along distributions to see role of potential skewness, extreme values
- Boxplot and violin plot by three bins for the latter
- show conditional distribution (conditional on x being in the actual bin)
- Both the box plots and the violin plots reveal that the median management score is higher in larger firms, reflecting the same positive association as the bin scatters and the scatterplot.
- That positive pattern is true when we compare almost any statistic of the management score: median, upper and lower quartiles, minimum and maximum.
- These figures also show that the spread of management score is somewhat smaller in smaller firms, but that difference in spread appears small.

Case Study - Management quality and firm size - detour

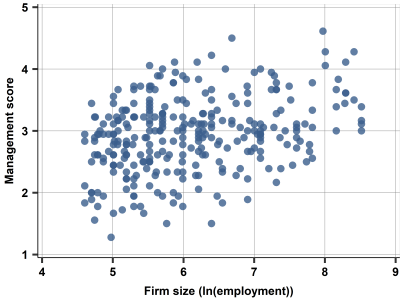
- To make the patterns more visible we apply a transformation
- The idea here is that the distribution of employment is very skewed, closer to lognormal than normal, so we take the natural logarithm of employment.
- Stretching the employment differences between firms at lower levels of employment and compressing those differences at higher levels.
- This scatterplot leads to a more spread out picture, reflecting the more symmetric distribution of the x variable.
- Here the positive association between mean management score and (log) employment is more visible.

Case Study - Management quality and firm size

(a) Scatterplot



(b) Scatterplot with Log units



Note: *Source: Management quality is an average score of 18 variables. Firm size is number of emoployees.* `wms-management-survey` *data. Mexican sample, n=300.*

Statistical dependence, correlation

Dependence and independence

- Dependence of two variables - y and x means that the conditional distributions of y - conditional on x - are not the same (x is the conditioning variable).
- Independence of y and x means the opposite: the distribution of y on x is the same, regardless of the value of x.
- Dependence of y and x, may take many forms.
  - When the value of x is different, y may be more or less spread out
  - when the value of x is different the mean of y is different.

Mean dependence

- Mean-dependence: conditional expectation $E[y|x]$ varies with the value of x.
- Mean-dependence is the extent to which conditional expectations (means) differ.
- Two variables are positively mean-dependent if the average of one variable tends to be larger when the value of the other variable is larger, too.
- Covariance and Correlation Coefficient are measures of mean dependence.

Covariance

The formula for the covariance between two variables x and y both observed in a dataset with n observations is:

$$Cov[x, y] = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{n} \tag{1}$$

- for each observation $i = 1 \dots n$
- The product within the sum in the numerator multiplies the deviation of x from its mean $(x_i - \bar{x})$ with the deviation of y from its mean $(y_i - \bar{y})$
- The entire formula is the average of these products across all observations.

Covariance

$$Cov[x, y] = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{n}$$

- If a positive deviation of x from its mean goes with a positive deviation of y from its mean the product is positive. Thus, the average of this product across all observations is positive.
- The more often a positive $x_i - \bar{x}$ goes together with a positive $y_i - \bar{y}$ the larger positive is the covariance.
- Or, the larger are the positive deviations that go together the larger the covariance.

The correlation coefficient

$$Corr[x, y] = \frac{Cov[x, y]}{Std[x]Std[y]} \tag{2}$$

$$-1 \leq Corr[x, y] \leq 1 \tag{3}$$

- The correlation coefficient is the standardized version of the covariance.
- The covariance may be any positive or negative number, while the correlation coefficient is bound to be between negative one and positive one.

Dependence, mean-dependence, correlation

- If two variables are independent, they are also mean-independent and thus the conditional expectations are all the same: $E[y|x] = E[y]$ of any value of x.
- Is this true the other way around?

Dependence, mean-dependence, correlation

- If two variables are independent, they are also mean-independent and thus the conditional expectations are all the same: $E[y|x] = E[y]$ of any value of x.
- But the reverse is not true.
- Can have zero correlation but mean dependence (e.g., a symmetrical U-shaped conditional expectation has an average of zero),
- Can have zero correlation and zero mean dependence without complete independence (e.g., the spread of y may be different for different values of x).

Dependence, mean-dependence, correlation

- Covariance or the correlation coefficient allow for all kinds of variables, including binary variables and ordered qualitative variables as well as quantitative variables.
- The covariance and the correlation coefficient will always be zero if the two variables are mean-independent, positive if positively mean-dependent, and negative if negatively mean-dependent.

- However, they are more appropriate measures for quantitative variables. That's because the differences $y_i - \bar{y}$ and $x_i - \bar{x}$ make more sense when y and x are quantitative variables.

Case Study - Management quality and firm size

- The covariance between firm size and the management score is 177.
- The standard deviation of firm size is 977, the standard deviation of management score is 0.6.
- Positive mean-dependence: firm size tends to be higher at firms with better management.
- the correlation coefficient is 0.30 $(177/(977 * 0.6))$.
- This suggests a positive and moderately strong association.
- Management quality–firm size correlation varies considerably across industries?

Case Study - Management quality and firm size

Table: Measures of management quality and their correlation with size by industry

Industry	Management–firm size correlation	Observations
Auto	0.50	26
Chemicals	0.05	69
Electronics	0.33	24
Food, drinks, tobacco	0.05	34
Materials, metals	0.32	50
Textile, apparel	0.29	43
Wood, furniture, paper	0.28	29
Other	0.44	25
All	0.30	300

Note: *Employee retention rates: The probability of staying with the firm, in the two experimental groups. Source: working_from_home dataset*

Measuring a latent concept with many observed variables

- Often a concept is hard, even impossible, to measure.
- Latent variables - while we can think of them as a variable there is no single observed variable to measure them.
- Quality of management at a firm - it is a concept that may be measured with a collection of variables, not a single one of them
- IQ - measured by a series of quiz-like questions.
- The problem here is how to combine multiple observed variables

Condensing information

If a dataset has more than one variable aimed to measure the same latent variable how should we combine them? Alternatives:

- Use one observed variable only
- Take the average (or sum) of all observed variables
- Use principal component analysis (PCA) to combine all observed variables

Condensing information 1: Using a single variable

- Using one measured variable and exclude the rest has the advantage of easy interpretation.
- It has the disadvantage of discarding potentially useful information contained in the other measured variables.
- Can be often a sensible start

Condensing information 2: Using a sum

- Taking the average of all measured variables makes use of all information.
- If all measured using the *same scale* this approach, simple and a natural interpretation

- When variables measured in different scales, simple average is difficult to interpret and meaningless

Condensing information 2: Using a sum

- Taking the average of all measured variables makes use of all information.
- If all measured using the *same scale* this approach, simple and a natural interpretation

- When variables measured in different scales, simple average is difficult to interpret and meaningless
- Need bring it to common scale - standardization: subtracting the mean and dividing with the standard deviation (class 03)
- The result is a series of variables with zero mean and standard deviation of one
- This standardized measure is called a "z-score" or "score"

Condensing information 3 using different weights

- Principal component analysis (PCA) is a method to give potentially different weights to the observable variables for creating a weighted average.
- The weights are constructed in such a way that observed variables that are better measures receive higher weights.
- PCA finds out which observed variable is a better measure by examining how they would be correlated with the weighted average.
- Bit of black box method

We suggest using the z-score

- Use z-score - simple average of multiple observed variables after making sure that they are measured on the same scale
- Simple, easy to understand
- Transparent
- Typically marginally different to PCA
- Pay attention
  - Look at correlation signs, you may check it first (PCA is better here)
  - Sensitive to extreme values

Case Study - Management quality and firm size

- The latent concept here is the overall quality of management.
- The observable variables are the 18 "score" variables.
- Each of the 18 scores were measured on a scale of 1 (worst practice) to 5 (best practice).
- The overall management score variable was the simple average of these.
- Looking at correlation of each of the sixteen observed variables with the average score [0.45 to 0.73]. Close, not same.
- PCA could be slightly better. In practice average score and PCA very similar.

Comparison and variation in x

- Variation in the conditioning variable is necessary to make comparisons.
- If no variation in the conditioning variable
  - all observations have the same values
  - impossible to make comparisons
- Example: Uncover the effect of price changes on sales –> need many observations with different price values.
- Generalization: The more variation is there in the conditioning variable the better are the chances for comparison.

Comparison and variation in x

- source of variation in the conditioning variable
- why values of the conditioning variable may differ across observations.
- Option 1: experimental data
- Option 2: observational data

Comparison in Experimental data

- We have an intervention or treatment.
- Value of the conditioning variable differs across observations because the person running the experiment made them different. Also: treatment variable
- There is controlled variation - a rule deciding treatment
- Experiment - comparing one or more outcome variables across the various values of a treatment variable

- Example: drug trial
  - Medical experiment - some patients receive the drug while others receive a placebo
  - Outcome is recovery from the illness or not
  - Control (treatment) variable is gets the drug or not

Comparison with observational data

- Most data used in business, economics and policy analysis are observational.
- In observational data, no variable is fully controlled.
- Typical variables in such data are the results of the decisions
- The source of variation in these variables may have multiple sources
- People's choices, decisions, interactions, expectations, etc.
- Compare the value of the outcome variable for different values of the conditioning variable.
- Much harder interpretation

Source of variation important for causal analysis

Experimental data

- Easy - if conditioning variable is experimentally controlled -
- Made sure that differences in the outcome variable are due to that variable only
- Example. Randomly give aspirin vs placebo.
- Any difference in stroke likelihood is due to treatment

Observational data

- Hard - many other things may be different when the value of the conditioning variable differs
- Example: observe people aspirin taking routine
- Any difference in stroke likelihood could be for other reasons
- E.g. Aspirin takers may have chosen to take Aspirin because they experienced a stroke already

AI and patterns

- AI is great to give you a first review of patterns – similar to a few lines of code, or panda profiler in Python
- judgment of correlation (weak, strong) is often wrong.
- you need to know what pattern to pursue

See separate pdf for example

Summary

- Be explicit about what y and x are in your data and how they are related to the question of your analysis. $E[y|x]$ is mean y conditional on x.
- For qualitative variables, correlation can be shown by summarizing conditional probabilities (frequencies).
- For quantitative variables, scatterplots offer a visual insight to the pattern of the relationship.
- The correlation coefficient captures a simple measure of mean dependence.

Functional form: ln transformation

- ComparsiFrequent nonlinear patterns better approximated with y or x transformed by taking relative differences:
- In cross-sectional data usually there is no natural base for comparison.
- Taking the natural logarithm of a variable is often a good solution in such cases.
- When transformed by taking the natural logarithm, differences in variable values we *approximate relative differences*.
  - Log differences works because differences in natural logs approximate percentage differences!

Logarithmic transformation - interpretation

- $ln(x)$ = the natural logarithm of x
  - Sometimes we just say log x and mean $ln(x)$. Could also mean log of base 10. Here we use $ln(x)$
- x needs to be a positive number
  - $ln(0)$ or $ln(\textit{negative number})$ do not exist
- Log transformation allows for comparison in relative terms – percentages!

  Claim:

$$\ln(x + \Delta x) - \ln(x) \approx \frac{\Delta x}{x}$$

- The difference between the natural log of two numbers is approximately the relative difference between the two for small differences.

Logarithmic Functions of y and/or x

- $ln(x)$ = the natural logarithm of x
  - Sometimes we just say log x and mean $ln(x)$. Could also mean log of base 10. Here we use $ln(x)$
- x needs to be a positive number
  - $ln(0)$ or $ln(negative\ number)$ do not exist
- Log transformation allows for comparison in relative terms (percentage), because:

$$ln(x + y) - ln(x) = ln(1 + \frac{\Delta x}{x}) \approx \frac{\Delta x}{x} \tag{4}$$

- Numerically:
  - $\ln(1.01) = 0.0099 \approx 0.01$
  - $\ln(1.1) = 0.095 \approx 0.1$

Logarithmic Functions of y and/or x

- Alternatively, from the relationship in calculus:

$$\frac{dln(x)}{dx} = \frac{1}{x}$$

  - So that, for small Δx,

$$\frac{ln(x + \Delta x) - ln(x)}{\Delta x} \approx \frac{1}{x}$$

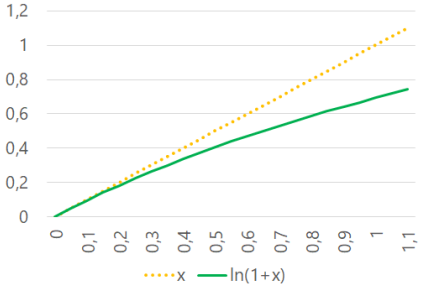
  - And thus

$$ln(x + \Delta x) - ln(x) \approx \frac{\Delta x}{x}$$

- i.e., the difference between the natural log of two numbers is approximately the relative difference between the two
  - For small differences

Log Approximation of Small Relative Diffs

- Log differences approximate small relative differences
- When x is small
  - 0.3 or smaller
  - the log approximation is close
- But for larger x, there is a difference,
  - And may have to calculate percentage change by hand



Ln(x) vs percentage

- Log differences approximate relative differences (percent)
- log difference - we mean $ln(x)$
  - x needs to be a positive number
- Log differences approximate small relative differences - say below 0.3
  - A difference of 0.1 log units corresponds to a 10% difference
  - For larger positive differences, the log difference is smaller
    - A log difference of +1.0 corresponds to a +170% difference
  - For larger negative differences, the log difference is larger in absolute value
    - A log difference of -1.0 corresponds to a -63% difference