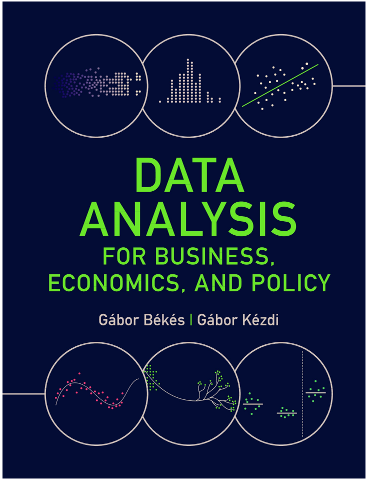
05 Generalizing from data

Gabor Bekes

Data Analysis 1: Exploration

2019

Slideshow for the Békés-Kézdi Data Analysis textbook



- Cambridge University Press, 2021 January
- Available in paperback, hardcover and e-book
- Slideshow be used and modified for educational purposes only
- **gabors-data-analysis.com**
  - Download all data and code
  - Additional material, links to references

Motivation

- *How likely is it that we shall experience losses on our investment portfolio? To answer this, you have collected and analyzed past financial information. To predict the frequency of a loss of certain magnitude for the coming calendar year, you will need to make an inference and think hard about what can be different in the future.*

Generalization

- Sometimes we analyze a dataset with the goal of learning about patterns in that dataset alone.
- In such cases there is no need to generalize our findings to other datasets.
- Example: We search for a good deal among offers of hotels, all we care about are the observations in our dataset.

- Often we analyze a dataset in order to learn about patterns that may be true in other situations.
- We are interested in finding it the relationship between
  - Our dataset
  - The situation we care about

Generalization

- Generalize the results from a single dataset to other situations.
- The act of generalization is called *inference*: we infer something from our data about a more general phenomenon because we want to use that knowledge in some other situation.

- Aspect 1: statistical inference
- Aspect 2: external validity

Statistical inference

- Uses statistical methods to make inference.
- Well-developed and powerful toolbox that helps generalizing to situations similar to our data.
- Similar to ours = general pattern represented by our dataset.

- The general pattern is an abstract thing that may or may not exist.
- If we can assume that the general pattern exists, the tools of statistical inference can be very helpful.

General patterns 1: Population and representative sample

- The cleanest example of representative data is a representative sample of a well-defined *population*.
- A sample is representative of a population if the distribution of all variables is very similar in the sample and the population.
- Random sampling is the best way to achieve a representative sample.

General patterns 2: No population but general pattern

The concept of representation is less straightforward in other setups.

- Using data with observations from the past to uncover a pattern that may be true for the future.
- Generalizing patterns observed among some products to other, similar products.

There isn't necessarily a "population" from which a random sample was drawn on purpose. Instead, we should think of our data as one that represents a general pattern.

- There is a general pattern, each year is a random realization.
- There is a general pattern, each product is a random version, all represented by the same general pattern.

External validity

- Assessing whether our data represents the same general pattern that would be relevant for the situation we truly care about.
- Externally valid case: the situation we care about and the data we have represent the same general pattern
- With external validity, our data can tell what to expect.
- No external validity: whatever we learn from our data, may turn out to be not relevant at all.

The process of inference

The process of inference

1. Consider a statistic we may care about, such as the mean.
2. Compute its *estimated value* from a dataset
3. Infer the value in the population / in the general pattern, that our data represents.

It is good practice to divide the inference problem into two.

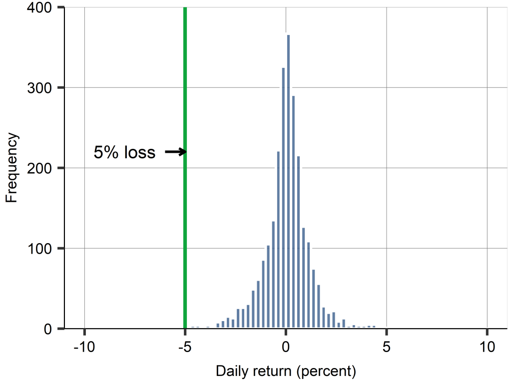
1. Use statistical inference to learn about the population, or general pattern, that our data represents.
2. Assess external validity: define the population, or general pattern we are interested in and assess how it compares to the population, or general pattern, that our data represents.

Case Study - Stock market returns: Inference

- Task: Assess the likelihood of experiencing a loss of certain magnitude on an investment portfolio from one day to the next day
- Predict the frequency of a loss of certain magnitude for the coming calendar year

- The investment portfolio is the S&P 500, a US stock market index
- Data: day-to-day returns on the S&P 500, defined as percentage changes in the closing price of the index between two consecutive days
- 11 years: 25 August 2006 to 26 August 2016. It includes 2,519 days.

Case Study - Histogram of daily returns



Note: *S&P 500 market index. Day to day (gaps ignored) changes, in percentage. From August 25 2006 to August 26 2016.*

Case Study - Stock market returns: Inference

- To define "loss", we take a day-to-day loss exceeding 5 percent.
- "loss" is a binary variable, taking 1 when the day-to-day loss exceeds 5 percent and zero otherwise.
- The statistic in the data is the proportion of days with such losses.
- It is 0.5 percent in this dataset
  - the S&P500 portfolio lost more than 5 percent of its value on 0.5 percent of the days between August 25 2006 and August 26 2016.

- Inference problem: How can we generalize this finding? What can we infer from this 0.5 percent chance for the next calendar year?

Repeated samples

- Repeated samples - the conceptual background to statistical inference
- Our data - one example of many datasets that could have been observed.
- Each datasets can be viewed as samples drawn from the population (general pattern)

- Easier concept: When our data is sample from a well-defined population - many other samples could have turned out instead of what we have.
  - Example: Mexican firms - random sample - population of firms
- Harder concept: no clear definition of population. We think of a general pattern we care about.
  - The data of returns on an investment portfolio may be thought of as a particular realization of the history of returns that could have turned out differently.

Repeated samples

- The goal of statistical inference is learning the value of a statistic in the population, or general pattern, represented by our data.
- The statistic has a distribution: its value may differ from sample to sample.
- The distribution of the statistic of interest is called its sampling distribution
  - Example: Fraction of firms with non-zero exports could be different
  - Example: The fraction of days with a 5 percent or larger loss on an investment portfolio may turn out different if we could "re-run history".

Repeated samples

- Standard deviation in this distribution: spread across repeated samples
- The standard error (SE) of the statistic = the standard deviation of the sampling distribution
- Any particular estimate is likely to be an erroneous estimate of the true value. The magnitude of that typical error is one SE.

Repeated samples

The sampling distribution of a statistic is the distribution of this statistic across repeated samples.
The sampling distribution has three important properties

1. Unbiasedness: The average of the values in repeated samples is equal to its true value (=the value in the entire population / general pattern).
2. Asymptotic normality: The sampling distribution is approximately normal. With large sample size, it is very very close.
3. Root-n convergence: The standard error (the standard deviation of the sampling distribution) is smaller the larger the samples, with a proportionality factor of the square root of the sample size.

Case Study - Stock market returns: A simulation

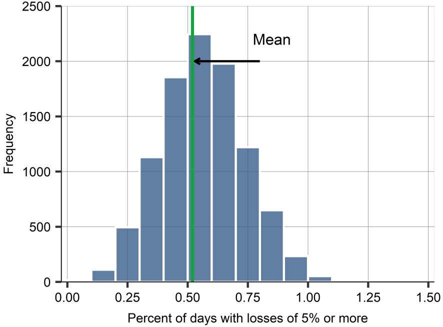
- We can not rerun history many many times...
- Simulation exercise - to better understand how repeated samples work
- Suppose the 11-year dataset is *the* population - the fraction of days with 5%+ losses is 0.5% in the entire 11 years' data. That's the true value.
- Assume we have only three years (900 days) of daily returns in our dataset.
- Task: estimate the true value of the fraction in the 11-year period from the data we have using a simulation exercise.
  1. many datasets with three years' worth of observations may be created from the 11 years' worth of data,
  2. compute the fraction of days with 5%+ losses in datasets
  3. learn about the true value

Case Study - Stock market returns: A simulation

- Do simple random sampling: days are considered one after the other and are selected or not selected in an independent random fashion.
  - This sampling destroys the time series nature
  - This is OK because daily returns are (almost) independent across days in the original dataset
- We do this 10,000 times....

Case Study - Stock market returns: A simulation

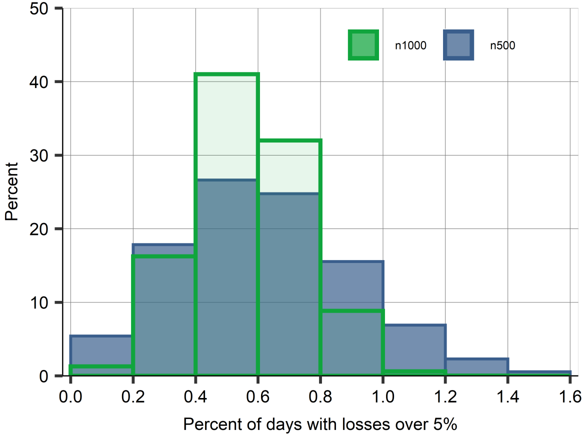
- percent of days with losses of 5% of more.
- histogram created from the 10,000 random samples, each w/ 900 obs, drawn from entire dataset
- distribution that has some spread: almost none of the days experienced such losses — 1.2 percent of the days did



Histogram of the proportion of days with losses of 5 percent or more, across repeated samples of size n=900. 10,000 random samples. Source: `sandp-stocks` data. S&P 500 market index.

Case Study - Stock market returns: Sampling distributions

- Proportion of days with losses of 5 percent or more
- Repeated samples in two simulation exercises, with n=450 and n=900. (10,000 random samples)
- Kernel density (goes to minus / can cut it at 0)
- Role of sample size: smaller sample: skewed; higher standard deviation



The standard error and the confidence interval

- Confidence interval (CI) - measure of statistical inference.
  - Recall: Statistical inference - we analyze a dataset to infer the true value of a statistic: its value in the population, or general pattern, represented by our data.
- The CI defines a range where we can expect the true value in the population, or the general pattern.
- CI gives a range for the true value with a probability
- Probability tells how likely it is that the true value is in that range
- Probability - data analysts need to picks it, such as 95%

The standard error and the confidence interval

- The "95 percent CI" gives the range of values where we think that true value falls with a 95 percent likelihood.
- Viewed from the perspective of a single sample, the chance (probability) that the truth is within the CI measured around the value estimated from that single sample is 95 percent.

- Also: we think that with 5 percent likelihood, the true value will fall outside the confidence interval.

The standard error and the confidence interval

- Confidence interval - symmetric range around the estimated value of the statistic in our dataset.
  - Get estimated value.
  - Define probability
  - Calculate CI with the use of SE
- 95 percent CI is the $\pm 2SE$ interval around the estimate from the data.
  - 90% CI is the $\pm 1.6SE$ interval, the 99 % CI is the $\pm 2.6SE$

Calculating the standard error

An important consequence of evidence from the repeated sample exercise:

- In reality, we don't get to observe the sampling distribution. Instead, we observe a single dataset
- That dataset is one of the many potential samples that could have been drawn from the population, or general pattern
- Good news: We can get a very good idea of how the sampling distribution would look like - good estimate of the standard error - even from a single sample.
- Getting SE – Option 1: Use a formula
- Getting SE – Option 2: Simulate by a new method

Calculating the standard error

Consider the statistic of the sample mean.

- Assume the values of x are independent across observations in the dataset.
- $\bar{x}$ is the estimate of the true mean value of x in the general pattern/population
- Sampling distribution is approximately normal, with the true value as its mean.

The standard error formula for the estimated $\bar{x}$ is

$$SE(\bar{x}) = \frac{1}{\sqrt{n}} Std[x] \tag{1}$$

where $Std[x]$ is the standard deviation of the variable x in the data and n is the number of observations in the data.

The standard error formula

- The standard error is larger...
  - the larger the standard deviation of the variable.
  - the smaller the sample and
- For intuition, consider $SE(\bar{x})$ vs $Std[x]$.
- Think back to the repeated samples simulation exercise:
  - $SE(\bar{x})$ = the standard error of $\bar{x}$ is the standard deviation of the various $\bar{x}$ estimates across repeated samples.
  - The larger the standard deviation of x itself, the more variation we can expect in $\bar{x}$ across repeated samples.

Case Study - Stock market returns: The standard error formula

- Let's consider our example of 11-years' of data on daily returns on the S&P 500 portfolio.
- The size of the sample is n = 2,519 so that $\sqrt{(1/n)} = 0.02$.
- The standard deviation of the fraction of 5+% losses is 0.07.
- So the $SE = 0.07 * 0.02 = 0.0014$ (0.14 percent).
- CI: the 95 percent CI is [0.22, 0.78].

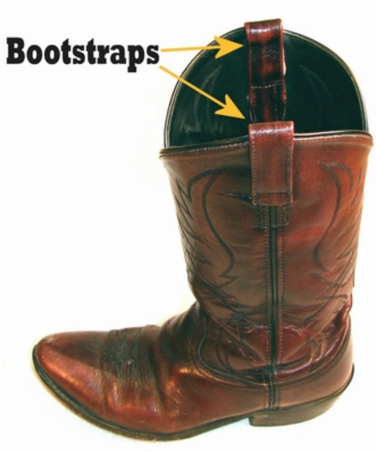
- This means that in the general pattern represented by the 11-year history of returns in our data, we can be 95 percent confident that daily losses of more than 5 percent occur with a 0.2 to 0.8 percent chance.

Take a quick stop to summarize the idea of CI

- We are interested in generalizing from our data. Statistical inference.
- Consider a statistic such as the sample mean $\bar{x}$
- Take a 95% confidence interval - where we can expect to see the true value
- CI=statistic $+/-2SE$.
- We have a formula for the SE calculated from our the data only using the standard deviation and sample size.
- Using the CI, we can now do statistical inference, generalize for the population / general pattern we care about.

The bootstrap

- Bootstrap is a method to create synthetic samples that are similar but different
- An method that is very useful in general.
- It is essential for many advanced statistics application such as machine learning
- Will not be part of exam material



The bootstrap: the motivation

- Simulation study: we took many samples of the same size from the original dataset to construct the sampling distribution of the statistic we were after. Artificial case.
- In practice we examine the entire dataset we have, and we would like to uncover the sampling distribution of a statistic in samples similar to that dataset.
- We want the sampling distribution of samples of the same size as the original dataset
- So we need to create many samples that are similar to ours AND are of the same size

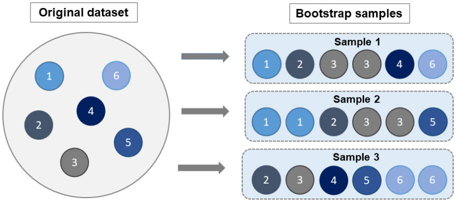
- Note: Bootstrap can be used for calculating the SE. Advantage to formula: needs fewer assumption, can be used for complicated statistics

The bootstrap

- The bootstrap method takes the original dataset and draws many repeated samples of the size of that dataset.
- The trick is that the samples are drawn *with replacement*.
- The observations are drawn randomly one by one from the original dataset; once an observation is drawn it is "replaced" to the pool so that it can be drawn again, with the same probability as any other observation.
- The drawing stops when it reaches the size of the original dataset.
- The result is a sample of the same size as the original dataset, yielding a single *bootstrap sample*.

The bootstrap

- A bootstrap sample is always the same size the original
- it includes some of the original observations multiple times,
- it does not include some of other original observations.
- We typically create 500 - 10,000 samples
- Computationally intensive but feasible, relatively fast.



The bootstrap

- We have a dataset (the sample), can compute a statistic (e.g. mean)
- Create many bootstrap samples, and get a mean value for each sample
- Bootstrap estimate of SE = standard deviation of statistic thru bootstrap samples.

Ch05_figures/bootstrap_balls2.png

The bootstrap

- The bootstrap method creates many repeated samples that are different from each other, but each has the same size as the original dataset.
- The distribution of a statistic across these repeated bootstrap samples is a good approximation to the sampling distribution we are after
  - ... what the distribution would look like across datasets similar to the original dataset.

- Bootstrap gives a good approximation of the standard error, too.
- The bootstrap estimate (or the estimate from the bootstrap method) of the standard error is simply the standard deviation of the statistic across the bootstrap samples.

Case Study - Stock market returns: Bootstrap standard error

- We estimate the standard error by bootstrap.
- Let's consider our example of 11-years' of data on daily returns on the S&P 500 portfolio.
- Do the process ———————>
-
-

The process

1. Take the original dataset and draw a bootstrap sample.
2. Calculate the proportions of days with 5%+ loss in that sample.
3. Save that value.
4. Then go back to the original dataset and take another bootstrap sample.
5. Calculate the proportion of days with 5%+ loss and save that value, too.
6. And so on, repeated many times.

Case Study - Stock market returns: The Bootstrap standard error

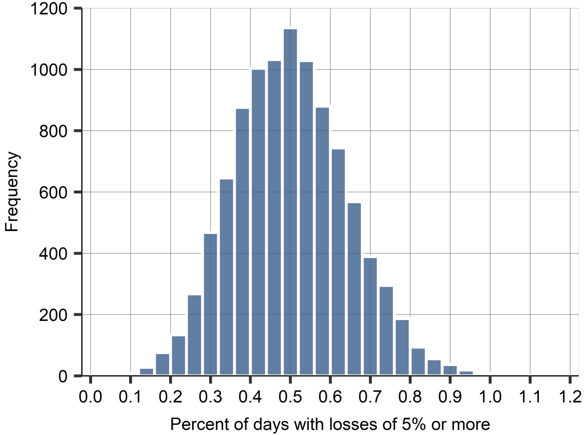
- We estimate the standard error by bootstrap.
- Let's consider our example of 11-years' of data on daily returns on the S&P 500 portfolio.
- Do the process ————————>
- End up with a new a dataset: one observations / bootstrap sample. Only variable is the estimated proportion in a sample
- **The SE is simply the standard deviation of those estimated values in this new dataset.**

The process

1. Take the original dataset and draw a bootstrap sample.
2. Calculate the proportions of days with 5%+ loss in that sample.
3. Save that value.
4. Then go back to the original dataset and take another bootstrap sample.
5. Calculate the proportion of days with 5%+ loss and save that value, too.
6. And so on, repeated many times.

Case Study - Stock market returns: The Bootstrap standard error

- 10,000 bootstrap samples with 2,519 observations
- The proportion of days with 5+ percent loss.
- Varied 0.1 percent to 1.2 percent. Mean=Median= 0.5
- Standard deviation across the bootstrap samples = 0.14
- CI: the 95 percent CI is [0.22, 0.78].



Case Study - Stock market returns: The Bootstrap standard error

- This means that in the general pattern represented by the 11-year history of returns in our data, we can be 95 percent confident that daily losses of more than 5 percent occur with a 0.22 to 0.78 percent chance.
- SE formula and bootstrap gave the **same** exact answer
- Under some conditions, this is what we expect
  - Large enough sample size
  - Observations independent
  - ... (other we overlook now)

External validity

- We discussed statistical inference: CI - uncertainty about the true value of the statistic in the population / general pattern that our data represents.
- What is the population, or general pattern, we care about?
- How close is our data to this?
- External validity is the concept that captures the similarity of our data to the population/general pattern we care about.
- High external validity: if our data is close to the population or the general pattern we care about.

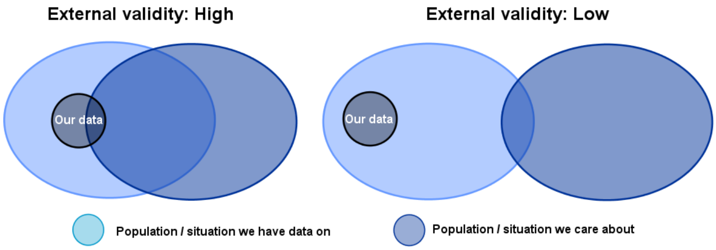
External validity

- With high external validity, the confidence interval captures all uncertainty about our estimate.
- With low external validity, it does not. Our estimate can be off.
- Unfortunately, we do not know how off. Speculative answer only.
- Thinking + intuition + theory.
- Extra uncertainty described in qualitative terms only.

- External validity is as important as statistical inference. However, it is not a statistical question.

External validity - Kidobni?

- We make inference on the population or general pattern based on the data we have at hand.
- External validity is about the population/general pattern we care about.
- Overlap indicates high vs low external validity.



External validity

- The most important challenges to external validity may be collected in three groups:
- Time: we have data on the past, but we care about the future
- Space: our data is on one country, but interested how a pattern would hold elsewhere in the world
- Sub-groups: our data is on 25-30 year old people. Would a pattern hold on younger / older people?
- Continue in DA2

External validity

- Daily 5%+ loss probability - 95 percent CI [0.2, 0.8] in our sample. This captures uncertainty for samples like ours.
- If the future one year will be like the past 11 years in terms of the general pattern that determines returns on our investment portfolio.
- However, external validity may not be high - not sure what the future holds.
- Our data: 2006-2016 dataset includes the financial crisis and great recession of 2008-2009. It does not include the dotcom boom and bust of 2000-2001. We have no way to know which crisis is representative to future crises to come.
- Hence, the real CI is likely to be substantially wider.

External validity in Big Data

- Big data: very large N
- Statistical inference not really important - CI becomes very narrow
- External validity remains as important

- 1.) Large sample DOES NOT mean representative sample
- 2.) Big data as result of actions - nature of things may change as people alter behavior, outside conditions change

Generalization - Summary

- Generalization is a key task - finding beyond the actual dataset.
- This process is made up of discussing statistical inference and external validity.
- Statistical inference generalizes from our dataset to the population using a variety of statistical tools.
- External validity is the concept of discussing beyond the population for a general pattern we care about; an important but typically somewhat speculative process.