Causal inference notes

Fabian Gunzinger

2023-09-28

Table of contents

1

In this space I want to collect my notes on causal inference.

If you find the notes helpful, find any errors, or have any suggestions, please get in touch by writing to fa.gunzinger@gmail.com.

2 Statistics foundation

2.1 Sampling

- We rely on a sample to learn about a larger population.
- We thus need to make sure that the sampling procedure is free of bias, so that units in the sample are representative of those in the population.
- While representativeness cannot be achieved perfectly, it's important to ensure that non-representativeness is due to random error and not due to systematic bias.
- Random errors produce deviations that vary over repeated samples, while systematic bias persists. Such selection bias can lead to misleading and ephemeral conclusions.
- Two basic sampling procedures are simple random sampling (randomly select n units from a population of N) and stratified random sampling (randomly select n_s from each stratum S of a population of N).
- The mean outcome of the sample is denoted $\bar{x}$; that of the population, μ.
- (On stratification: why does it reduce variance? Imagine an extreme case, where the number of strata were equal to the number of different units in the sample. In this case, the variance would be zero. Number of diff units here needns be individuals, but groups of units that share all relevant characteristics)

2.2 Sampling distributions

- A sampling distribution is the distribution of a statistic (e.g. the mean) over many repeated samples. Classical statistics is much concerned with making inferences from samples about the population based on such statistics.
- When we measure an attribute of the population based on a sample using a statistic, the result will vary over repeated samples. To capture by how much it varies, we are concerned with the sampling variability.
- Key distinctions:

  - The data distribution is the distribution of the data in the sample, and its spread is measured by the standard deviation.
  - The sampling distribution is the distribution of the sample statistic, and its spread is measured by the standard error.

Figure show that:

- Data distribution has larger spread than sampling distributions (each data point is a special case of a sample with n = 1)
- The spread of sampling distributions decreases with increasing sample size

2.3 Law of large numbers and central limit theorem

- Suppose that we have a sequence of independent and identically distributed (iid) random variables $\{x_1, ..., x_n\}$ drawn from a distribution with expected value μ and finite variance σ^2, and we are interested in the mean value $\bar{x} = \frac{x_1 + ... + x_n}{n}$.
- The law or large numbers states that $\bar{x}$ converges to μ as we increase the sample size. Formally:

$$\bar{x} \to \mu \text{ as } n \to \infty.$$

- The (classical, Lindeberg-Lévy) central limit theorem describes the spread of the sampling distribution of $\bar{x}$ around μ during this convergence. In particular, it implies that for large enough n, the distribution of $\bar{x}$ will be close to a normal distribution with mean μ and variance σ^2/n. The above figures are a visual representation of this. Formally:

$$\lim_{n \to \infty} \sqrt{n}(\bar{x} - \mu) \to \mathcal{N}\left(0, \sigma^2\right).$$

- This is useful because it means that irrespective of the underlying distribution (i.e. the distribution of the values in our sequence above), we can use the normal distribution and approximations to it (such as the t-distribution) to calculate sampling distributions when we do inference. Because of this, the CLT is at the heart of the theory of hypothesis testing and confidence intervals, and thus of much of classical statistics.
- For experiments, this means that our estiamted treatment effect is normally distributed, which is what allows us to draw inferences from our experimental setting ot the population as a whole. The CLT is thus at the heart of the experimental approach.

- The CLT also explains the prevalence of the normal distribution in the natural world. Many characteristics of living things we observe and measure are the sum of the additive effects of many genetic and environmental factors, so their distribution tends to be normal. –>

2.4 Standard error

- The standard error is a measure for the variability of the sampling distribution.
- It is related to the standard deviation of the observations, σ and the sample size n in the following way:

$$se = \frac{\sigma}{\sqrt{n}}$$

- The relationship between sample size and se is sometimes called the "Square-root of n rule", since reducing the *se* by a factor of 2 requires an increase in the sample size by a factor of 4.

Derivation:

The sum of a sequence of independent random variables is:

$$T = (x_1 + x_2 + ... + x_n)$$

Which has variance

$$Var(T) = Var(x_1) + Var(x_2) + ... + Var(x_n) = n\sigma^2$$

and mean

$$\bar{x} = T/n.$$

The variance of $\bar{x}$ is then given by:

$$Var(\bar{x}) = Var\left(\frac{T}{n}\right) = \frac{1}{n^2}Var(T) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

The standard error is defined as the standard deviation of $\bar{x}$, and is thus

$$se(\bar{x}) = \sqrt{Var(\bar{x})} = \frac{\sigma}{\sqrt{n}}.$$

2.5 Bootstrap

- In practice, we often use the bootstrap to calculate standard errors of model parameters or statistics.
- Conceptually, the bootstrap works as follows:

  1) we draw an original sample and calculate our statistic
  2) we then create a blown-up version of that sample by duplicating it many times
  3) we then draw repeated samples from the large sample, recalculate our statistic, and calculate the standard deviation of these statistics to get the standard error.

- To achieve this easily, we can skip step 2) by simply sampling with replacement from the original distribution in step 3).
- The full procedure makes clear what the bootstrap results tell us, however: they tell us how lots of additional samples would behave if they were drawn from a population like our original sample.
- Hence, if the original sample is not representative of the population of interest, then bootstrap results are not informative about that population either.
- The bootstrap can also be used to improve the performance of classification or regression trees by fitting multiple trees on bootstrapped sample and then averaging their predictions. This is called “bagging”, short for “bootstrap aggregating”.
- We can use to boostrap also to calculate CIs following this algorithm:

  1) Draw a large number of bootstrap samples and calculate the statistic of interest
  2) Trim [(100-x)/2] percent of the bootstrap results on either end of the distribution
  3) The trim points are the end point of the CI.

2.6 Selection bias

Common types of selection bias in data science: - The vast search effect (using the data to answer many questions will eventually reveal something interesting by mere chance – if 20,000 people flip a coin 10 times, some will have 10 straight heads) - Nonrandom sampling - Cherry-picking data - Selecting specific time-intervals - Stopping experiments prematurely - Regression to the mean (occurs in settings where we measure outcomes repeatedly over time and where luck and skill combine to determine outcomes, since winners of one period will be less lucky next period and perform closer to the mean performer)

Ways to guard against selection bias: - have one or many holdout datasets to confirm your results.

2.7 Standard deviation vs standard error

- Standard deviation is the spread of the distribution of the values in the population of interest
- Standard error is the spread of the distribution of a sample statistic (such as the mean) based on a random sample of population values.

2.8 Degrees of freedom

In statistics, degrees of freedom generally refers to the number of values in a calculation that can vary freely.

Examples:

- Variance calculation: given that we have a mean, once we know all but one value, we also know final value, since sum of mean deviations has to be zero.
- Covariance calculation: given the two means, once we know the values for all but one x and y pair, we also know the values of the final pair. Hence, we loose one df (not clear to me why not two, given that both x and y are determined – because we treat their product as a single value? but that seems arbitrary)
- Also, why no correction when we have popultion means? See wikipedia article on variance for section on bias correction
- There is lots of confusion out there when it comes to df. For instance, you sometimes hear people say that df is the number of parameters you had to calculate on route. But this is wrong. It happens to come to the same when calculating variance, but not if you calcualte covariance (where you calculate two means beforehand, but only loose one df).

2.9 Commonly used distributions

from here

Name of the probability distribution	Probability distribution function	Mean	Variance

2.9.1 Commonly used probability distributions

The following table lists the variance for some commonly used probability distributions.

Name of the probability distribution	Probability distribution function	Mean	Variance
Binomial distribution	$\Pr(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$
Geometric distribution	$\Pr(X=k)=(1-p)^{k-1}p$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Normal distribution	$f(x\mid\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
Uniform distribution (continuous)	$f(x\mid a,b)=$ $\begin{cases}\frac{1}{b-a} & \text{for } a\le x\le b,\\ 0 & \text{for } x<a \text{ or } x>b\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential distribution	$f(x\mid\lambda)=\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Poisson distribution	$f(k\mid\lambda)=\frac{e^{-\lambda}\lambda^k}{k!}$	λ	λ

2.10 p-values – how to draw statistical conclusions ?

Limitations of reliying on pvap

- Arbitrary cutoff
- No appreciation for variation of coefficient – focus on ci instead (see Romer (2020), Imbens (2021))
- Multiple hypothesis testing (actual) – report and apply MHT-correction
- Multiple hypothesis testing (potential Gelman post)

2.11 False positive rate vs false discovery rate

- The false positive rate is $P(significantresult|notrueeffect)$
- The false discovery rate is $P(notrueeffect|significantresult)$

2.12 Confidence interval interpretations

- 95% CI for control and treatment overlap. Does this imply treatment is not significant?

2.13 Sources

- Practical statistics for data science –>

3 Regression

todo: - Use DAaidson and MacKinnon metrics book: nicely separates mechanics of ols with statistical interpretation. Do the same. - Include intro on regression from here

Linear regression is by far the most often used method to estimate relationships between outcomes and explanatory variables in applied econometrics and – probably – all of applied social science. In this article I want to to two things: define the related terminology that is often used but not defined (OLS, linear equation, etc.) and – mainly – explain a less frequently taught and understand way to think about linear regression, that in terms of its linear algebra.

There are, of course, a lot of very good resources on linear regression and OLS, and I list my favourite ones at the bottom. But none of them quite tie everything together in the way I was looking for.

There are two ways to understand linear regression: one is to think of the variables involved as dimensions and of each row as a data point – this is the way the problem is usually motivated in introductory econometrics classes. The other is the linear algebra approach – to think of each row of data as a dimension, and to think of the variable as vectors in the space formed by those dimensions.

The first approach straightforwardly links to the intuition of minimising squares, which is useful. It's the one I have learned and relied on most of my life. The second one, though, provides an alternative and very powerful way to understand what linear regression does. And, importantly, understanding the linear algebra notation simplifies much of the notation and manipulations, and opens the way to much of the literature of econometric theory, such as an understanding of the Frisch-Waugh-Lowell theorem, which was the impetus for me to dig into the linear algebra of OLS.

In this post I want to cover the following:

- Understand all the terminology related to linear regression so we fully know what we're talking about
- Understand the matrix representation of linear regression
- Understand how we can think of the least squares solution as a projection
- Understand why this is a very useful way of seeing things

3.1 Glossary

TODO: - Linear regression vs OLS

3.2 The setup

We usually start with data of the form $\{y_i, x_{i1}, \cdots, x_{ik}\}_{i=1}^{N}$, where we observe an outcome variable y_i and a set of k explanatory variables $x_i = (x_{i1}, \cdots, x_{ik})$ for each unit i in the dataset. We think that it might be reasonable to think of the outcome being linearly related to the regressors, so that, for each unit in our dataset, we can write the following linear equation:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i$$

This says that the outcome y can be thought of as a linear combination of all explanatory variables plus some error term.

TODO: What makes this a "linear" equation: - The highest power to which any regressor is raised is 1 - The coefficients are constants, not variables - Regressors are related to one another using addition and subtraction only - The resulting line (in 2-D space), plane (in 3-D space) and hyperplane (in N-D space) are linear (the term linear equation originates from the simple case where there are two regressors, one of which is a constant, in which case we get a straight line in a Cartesian plane.)

TODO: discuss all the assumptions we're making here.

TODO: Discuss the Angist & Pischke view of linear regression being good approximation even if relationship is not linear.

We thus have a system of linear equations of the form

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{1_1} + \beta_2 x_{2_1} + \ldots + \beta_k x_{k_1} + \epsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{1_2} + \beta_2 x_{2_2} + \ldots + \beta_k x_{k_2} + \epsilon_2 \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{1_n} + \beta_2 x_{2_n} + \ldots + \beta_k x_{k_n} + \epsilon_n
\end{aligned}$$

which we can rewrite in vector notation as

$$
\begin{aligned}
y_1 &= x_1'\beta + \epsilon_1 \\
y_2 &= x_2'\beta + \epsilon_2 \\
&\vdots \\
y_n &= x_n'\beta + \epsilon_n,
\end{aligned}
$$

where

$$x_i' = (x_{i1}, x_{i2}, \ldots, x_{ik})$$

is a $1 \times k$ row vector that contains all k explanatory variables for each unit i and

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$$

is a $k \times 1$ column vector that contains all k regression coefficients.

To be even more succinct, we can stack all n equations to get the matrix notation:

$$y = X\beta + \epsilon,$$

where β is defined as above,

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{pmatrix}$$

is a $n \times 1$ vector containing the n outcome variables, one for each unit in the data,

$$X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n'' \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,k} \end{pmatrix}$$

is an $n \times k$ matrix that contains all n row vectors x_i' stacked on top of each other, and

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_k \end{pmatrix}$$

a column vector containing the n error terms.

The problem

So, what do we want to do here? We have data on an outcome variable y and explanatory variables x for each unit i, and we think that it is reasonable to think that this data is generated by a process whereby y is the result of a linear combination of the xs plus some noise, which we capture in the error term. The challenge is to find the right linear combination.

3.3 Classic motivation of the solution

TODO

3.4 Linear algebra motivation of the solution

Notice how our problem here is exactly akin to the motivation for orthogonal projection discussed in **?@sec-projection**. There we had a system of linear equations of the form

$$Ax = b$$

which was overdetermined because the number of equations exceeded the number of unknowns. Our setup is equivalent. We have n equations and k unknowns (the βs), so that – in practice – there will be no solution to the system:

$$X\beta = y.$$

In other words, there is no choice of β that would linearly combine all the explanatory variables in each equation such that the result would be exactly y. We account for this by adding the error term ϵ, so that we have

$$X\beta + \epsilon = y.$$

What we do, now, is to say that we want to find that linear combination of the explanatory variables that is closest to y, so that ϵ is as small as possible, which is the same as finding the orthogonal projection of y onto X, which, traditionally, we call $\hat{y}$.

The solution is then the same as in **?@sec-projection**:[1]

$$\begin{aligned}
X'\epsilon &= 0 \\
X'(y - \hat{y}) &= 0 \\
X'(y - X\beta) &= 0 \\
X'y - X'X\beta &= 0 \\
X'X\beta &= X'y \\
\beta &= (X'X)^{-1}X'y
\end{aligned}$$

3.5 Resources

- Hayashi, Wooldridge, Verbeek, online resources

3.6 Assumptions

- Directly copied from or heavily based on https://people.duke.edu/~rnau/testing.htm. Edit and expand over time.
- In short, the assumptions are that the model is linear and additive, and that errors are iid. The latter assumption is often stated as three separate assumptions, as shown below.

[1] One thing I would always wonder about in textbook is how I knew that the condition was $X'\epsilon$ instead of $\epsilon'X$. The answer is that in cases where order doesn't matter, texts tend to choose what is more convenient for the math. We could solve $\epsilon'X$:

$$\begin{aligned}
\epsilon'X &= 0 \\
(y - \hat{y})'X &= 0 \\
(y - X\beta)'X &= 0 \\
(y' - \beta'X')X &= 0 \\
y'X - \beta'X'X &= 0 \\
\beta'X'X &= y'X \\
(\beta'X'X)' &= (y'X)' \\
X'X\beta &= X'y \\
\beta &= (X'X)^{-1}X'y
\end{aligned}$$

which gets us to the same result but in more steps.

- Note: The dependent and independent variables in a regression model do not need to be normally distributed by themselves–only the prediction errors need to be normally distributed. (In fact, independent variables do not even need to be random, as in the case of trend or dummy or treatment or pricing variables.) But if the distributions of some of the variables that are random are extremely asymmetric or long-tailed, it may be hard to fit them into a linear model whose errors will be normally distributed, and explaining the shape of their distributions may be an interesting topic all by itself. Keep in mind that the normal error assumption is usually justified by appeal to the central limit theorem, which holds in the case where many random variations are added together. If the underlying sources of randomness are not interacting additively, this argument fails to hold.

There are four principal assumptions which justify the use of linear regression models for purposes of inference or prediction:

(i) linearity and additivity of the relationship between dependent and independent variables:

  (a) The expected value of dependent variable is a straight-line function of each independent variable, holding the others fixed.

  (b) The slope of that line does not depend on the values of the other variables.

  (c) The effects of different independent variables on the expected value of the dependent variable are additive.

(ii) statistical independence of the errors (in particular, no correlation between consecutive errors in the case of time series data)

(iii) homoscedasticity (constant variance) of the errors

(a) versus time (in the case of time series data)

(b) versus the predictions

(c) versus any independent variable

(iv) normality of the error distribution.

If any of these assumptions is violated, then the forecasts, confidence intervals, and scientific insights yielded by a regression model may be (at best) inefficient or (at worst) seriously biased or misleading.

3.6.1 Linearity and additivity

Why assume linearity?

- It's the simplest non-trivial relationship and the easiest to work with
- True relationships are often at least approximately linear for ranges we care about
- Even if above not true, we can often transform variables so that relationshiop is linear

How to test linearity

- Look at scatterplot before running regression
- Look at predicted values vs residual plot after running regression (linear model leads to evenly distributed dots across a horizontal line) – can look at predicted vs observed value plot, too, in which case you expect points to fall on the 45 degree line, but this diagonal introduces additional visual noise.

What if linearity is violated?

- Add non-linear transformation like log (if data is strictly positive)
- Add regressor that is a non-liner function of a current regressor (e.g. xˆ2)
- Add new regressor that explains non-linear nature of current relationshiop

3.6.2 iid errors

Why assume iid errors?

- In many cases, it's justified by the CLT. As long as we calculate averages (sums or random variables, really), which are independent and identically distributed, the distribution will be approximately normal.
- It's convenient because it implies that the optimal coefficient estimates for linear models minimise the MSE, and because it justified using tests based on the normal family (t, F, Chi-square distribution).
- Even if the "true" error process is not normal in terms of the original units of the data, it may be possible to transform the data so that your model's prediction errors are approximately normal.
- Note that normality is required for inferenence, not for parameter estimation or prediction.

Potential issues

- Heteroskedasticity (variance larger depending on come conditions – certain covariate values)
- Interdependence (e.g. in time-series models, in the context of network effects)

How to check assumption

- Independence
    - In time-series: look at residual autocorrelation
    - In non-time series: plot independent variables vs residuals – independence suggests symmetric distribution around zero and no dependence in subsequent residuals under any ordering that is not based on the independent variables (i.e. no correlation given covariates)
- Homoskedasticity
    - Time series: plot time vs residuals
    - Non-time series: plot predicted values vs residuals and independent variables vs residuals
- Normality
    - Normal probability plot and normal quantile plot
    - Could also use formal tests (Kolmogorov-Smirnov test, Shapiro-Wilk test, Jarque-Bera test, and the Anderson-Darling test), but they are often too restrictive in practice.

How to fix violations

- Independence
    - Time-series: for mild violations (Durbin-Watson between 1.2 and 1) adding lags, for serious violations (DW $>$ 2.6) use difference model
    - Non-time series: either due to non-linearity of model or due to omitted variable.
- Homoskedasticity
    - Log transformation
    - Model seasonality with dummies (do linearly or multiplicatively (take log of dependent variable))
    - Seasonally adjust data before fitting model
- Normality

- Violations of normality often arise either because (a) the distributions of the dependent and/or independent variables are themselves significantly non-normal, and/or (b) the linearity assumption is violated. In such cases, a nonlinear transformation of variables might cure both problems.

- Another possibility is that there are two or more subsets of the data having different statistical properties, in which case separate models should be built, or else some data should merely be excluded, provided that there is some a priori criterion that can be applied to make this determination.

- In some cases, the problem with the error distribution is mainly due to one or two very large errors. Such values should be scrutinized closely: are they genuine (i.e., not the result of data entry errors), are they explainable, are similar events likely to occur again in the future, and how influential are they in your model-fitting results? If they are merely errors or if they can be explained as unique events not likely to be repeated, then you may have cause to remove them. In some cases, however, it may be that the extreme values in the data provide the most useful information about values of some of the coefficients and/or provide the most realistic guide to the magnitudes of forecast errors.

4 Miscellaneous topics

4.1 Description vs prediction vs causal inference

- Descriptive analysis describes the data. We simply turn data into meaningfull summary measures presented in tables or – better, usually! – figures. The aim here is to simply describe the data as it is, highlighting aspects that are particularly interesting or relevant to the task at hand.
- Predictive analysis predicts unobserved metric values based on observed ones. We build a model that captures the data generating process of the metric we aim to predict based on training data we observe, so that we can then predict metric values we don't observe.
- Causal inference makes statements about what would happen to outcomes if we changed the world in a particular way.

Causal inference vs prediction

- Prediction is about finding the most likely outcome based on a set of (existing) covariates. Causal inference is about finding the effect of a change in a covariate on the outcome.
- The difference is profound: when predicting, you take the features as a given and predict outcomes based on them – you're asking: "given existing features, what outcome can I expect?". When you perform causal inference, you want to know what would happen if you were to change one of the covariates – you're asking "if I were to change one covariate in a certain way, what outcome could I expect?".
- Causal inference is about manipulating covariates – to paraphrase Donald Rubin: there is no causality without manipulation.
- Technically, what this really comes down to is that in prediction, you don't care about selection bias, whereas in causal inference that's the main thing you care about.
- This also means that the role of goodness of fit is very different: for prediction, it's obviously very important – if your model explains only a very small part of the variation in the outcome, it won't be very good at predicting outcomes. In causal inference, goodness of fit doesn't matter because your aim is not to predict, but to know how the outcome changes if you change a covariate. So, you can have very low goodness of fit (lots of things outside the model predicting outcomes), but if you can precisely estimate your treatment effect, that's very valuable (you learn that regardless of all the many other factors that

determine the outcome, changing a covariate in a certain way tends to change outcomes in a certain way.)

5 Ethics

Some useful principles to consider, based on reading chapter 9 in (**kohavi2020trustworthy?**)

- Only test changes that Company policy would allow you to roll out to 100 percent of users i.e. for Meta, deliberately showing them negative content wouldn't qualify. Personally I'm not sure I agree. I think in the Meta example, it would be useful to understand the effect of exposure to negative content. I'm not gonna think about this deeply now, but I think adapting the rule to not showing levels of content users couldn't organically be exposed to on the platform might be more useful. In the Meta example, this would allow for studying the effect of negative content in a systematic way without making the experience worse for treatment users than it actually is for some users (and could well be for treatment users, too). The reason for my willingness to entertain to go further is that there is a potentially large benefit to understanding harm. Yes, there might be some cost in the very short term (and you'd obviously want ot bound that cost somehow, as I proposed above, in order to guard against slippery slopes), but if the insights gained allow you to prevent large harm indefinitely later, then that's worth considering. Two other thoughts: motivation clearly matters here. And: the possibility of a slippery slope is not generally an argument to not do something – often, as here – there are quite natural and objective ways to draw a line on how far one would be willing to go.
- Aim for equipoise: this is a situation where, ex-ante, there is no grounds to favour one variant over another. This is the normal case. (The term is borrowed from clinical trial, where clinical equipoise is the assumption in an RCT that no drug is ex-ante better than another).
- Some worthwhile experiments violate equipoise: increasing latency, disabling feature, showing more adds, all aim to help us collect data which can be useful to make tradeoffs later on, which, ultimately, can benefit users.
- Beware of behavioural experiments and deception.
- Presumptive consent: ask a small subset of users whether they would be okay participating and if they do, assume that the sentiment would generalise.
- Different from clinical trials, subjects in online trials usually have the opportunity to switch service (for sth like FB, this might be difficult).

Imbens, Guido W. 2021. “Statistical Significance, p-Values, and the Reporting of Uncertainty.” *Journal of Economic Perspectives* 35 (3): 157–74.

Romer, David. 2020. “In Praise of Confidence Intervals.” In *AEA Papers and Proceedings*, 110:55–60.